Journal of Modern Applied Statistical Methods

Volume 10 | Issue 2 Article 35

11-2-2011

Vol. 10, No. 2 (Full Issue)

JMASM Editors

Journal of Modern Applied Statistical Methods
November 2011, Vol. 10, No. 2, 403-750

1538 – 9472/11/$95.00

Journal Of Modern Applied Statistical Methods

Journal of Modern Applied Statistical Methods
November 2011, Vol. 10, No. 2, 403-750

1538 – 9472/11/$95.00

Journal Of Modern Applied Statistical Methods

Invited Articles

Regular Articles

**deceased*

Journal of Modern Applied Statistical Methods
November 2011, Vol. 10, No. 2, 403-750

1538 – 9472/11/$95.00

Journal of Modern Applied Statistical Methods
November 2011, Vol. 10, No. 2, 403-750

1538 – 9472/11/$95.00

Journal of Modern Applied Statistical Methods
November 2011, Vol. 10, No. 2, 403-750

1538 – 9472/11/$95.00

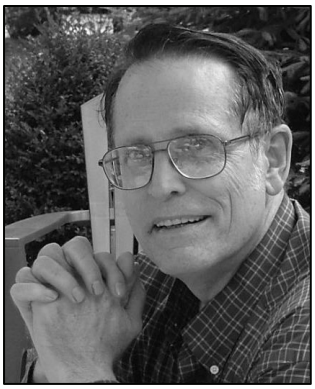
JMASM is an independent print and electronic journal (http://www.jmasm.com/), publishing (1) new statistical tests or procedures, or the comparison of existing statistical tests or procedures, using computer-intensive Monte Carlo, bootstrap, jackknife, or resampling methods, (2) the study of nonparametric, robust, permutation, exact, and approximate randomization methods, and (3) applications of computer programming, preferably in Fortran (all other programming environments are welcome), related to statistical algorithms, pseudo-random number generators, simulation techniques, and self-contained executable code to carry out new or interesting statistical methods.

Journal of Modern Applied Statistical Methods
November 2011, Vol. 10, No. 2, 403-417

1538 – 9472/11/$95.00

Invited Article

Robustness, Power and Interpretability of Pairwise Tests of Discriminant Functions in MANOVA




Philip H. Ramsey
Queens College of the City University of New York, Flushing, NY

Patricia P. Ramsey
Fordham University, New York, NY

Priscila Hachimine
Nancy Andiloro
Graduate Center of the City University of New York, New York, NY

Limiting follow-up hypotheses to be tested can reduce problems relating to the control of Type I and Type II errors in multivariate analysis of variance (MANOVA). Such limitations can also improve the interpretability of results. The importance of sample size, shape of population distribution, within-group correlations and heterogeneity of variances are demonstrated. The protected greatest characteristic root (GCR) procedure is shown to work well for small, group size, N (≤ 10). The unprotected GCR is shown to work well for larger N.

Key words: Any-pair power, discriminant functions, MANOVA, pair-wise test.

Introduction

Testing for the significance of differences in means of k groups on p variables can be accomplished with multivariate analysis of variance (MANOVA). The full, null hypothesis is

$$H_0: \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 \ldots = \boldsymbol{\mu}_k,$$

where $\boldsymbol{\mu}_i$ ($i = 1, \ldots, k$) is the vector of population means for group i on the p variables. The hypothesis degrees of freedom is $df_h = k - 1$. In the general case, the parameter, $s = \min(p, df_h)$. In MANOVA a variety of test statistics for the null hypothesis are possible. Taking p x p matrices, **H** and **E**, of the sum-of-products for hypotheses and error respectively as

$$\mathbf{H} = \sum_{i=1}^{k} \boldsymbol{n}_i (\bar{\mathbf{X}}_i - \bar{\mathbf{X}})(\bar{\mathbf{X}}_i - \bar{\mathbf{X}})', \quad (1)$$

Philip Ramsey is a Professor of Psychology. His research interests include psychometrics, applied statistics, and multiple comparisons. Email: Philip.Ramsey@qc.cuny.edu. Patricia Ramsey passed away in 2011. She was a Professor in the Graduate School of Business at Fordham University. Priscila Hachimine is a graduate student in psychology at CUNY. Email: phachimine@gmail.com. Nancy Andiloro is a graduate student in educational psychology at CUNY. Email: nancy6183@gmail.com.

and

$$\mathbf{E} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (\mathbf{X}_{ij} - \bar{\mathbf{X}}_{\mathbf{j}})(\mathbf{X}_{ij} - \bar{\mathbf{X}}_{\mathbf{j}})', \tag{2}$$

where $\mathbf{X}_{ij}$ is the j^{th} of n_i observation vectors in group i, $\bar{\mathbf{X}}_{\mathbf{i}}$ is the mean vector for the i^{th} group and $\bar{\mathbf{X}}$ is the grand mean vector. The s, nonzero eigenvalues of $\mathbf{HE}^{-1}$ can be designated as λ_1, …, λ_s in order from largest to smallest. Equivalently, the s, nonzero eigenvalues (also called characteristic roots) of $\mathbf{H(H+E)^{-1}}$ can be designated as θ_1, … θ_w in order from largest to smallest. Each corresponding member of the respective sets of eigenvalues can be related by $\theta = \lambda/(1 + \lambda)$.

Multivariate Test Procedures

The four, most common MANOVA test statistics are:

1. The Pillai-Bartlett trace, $V = \sum_{i=1}^{s} \theta_i$;
2. Wilks' likelihood ratio, $W = \prod_{i=1}^{s} (1 - \theta_i)$;
3. The Hotelling-Lawley trace, $T = \sum_{i=1}^{s} \lambda_i$; and
4. Roy's greatest characteristic root (GCR), $R = \theta_1$.

Computer packages such as SPSS and SAS typically provide approximate and sometimes exact p values for each of these four test statistics.

Routines for Testing

In each of the following routines s is defined as shown above and $df_E = \Sigma(N_i - 1)$. One method of evaluating V for a group of k means is with an F test (Pillai, 1955; Seber, 1984, p. 564) defined by

$$F = \frac{cV}{b(s - V)},$$

where $c = df_E - p + s$, and $b = \max(p, k - 1)$. To test at level α requires critical value, $CV = F_{1-\alpha}$ (sb, sc). This method is designated here as VPB.

Two, more accurate F tests for V are available (Muller, 1998). Method 1 is

$$F = \frac{df_2}{df_1} \frac{V}{d - V}, \tag{4}$$

where $df_1 = p(k - 1)$,

$$df_2 = \frac{[p(k-1)+2]df_E}{df_E(k+p)+(k+1)(k-2)}(df_E + k - 1 - p),$$

and

$$d = \frac{p(k-1)+df_2}{df_2 + k - 1}.$$

To test at level α requires $CV = F_{1-\alpha}(df_1, df_2)$. This method is designated here as VM1.

For Method 2 (Muller, 1988) the F test is

$$F = \frac{df_2}{df_1} \frac{V}{s - V}, \tag{5}$$

where

$$K = \frac{1}{s(df_E + k - 1)} \left[\frac{s(df_E + s - p)(df_E + k + 1)(df_E + k - 2)}{df_E(df_E + k - 1 - p)} \right] - 2$$

$df_1 = p(k - 1)K$, $c = df_E - p + s$, and $df_2 = scK$. To test at level α requires $CV = F_{1-\alpha}(df_1, df_2)$. This method is designated here as VM2.

One method of evaluating W for a group of k mean vectors is with an F test (Rao, 1951; Seber, 1984, p. 41) defined by

$$F = \frac{1 - U}{U} \frac{df_2}{df_1}, \tag{6}$$

where

$$t = \sqrt{\frac{p^2(k-1)^2 - 4}{p^2 + (k-1)^2 - 5}},$$

$$f = \frac{\mathrm{df_E} - (p - k + 2)}{2}$$

$$g = \frac{\mathrm{p(k\text{-}1)\text{-}2}}{2},$$

$$\mathrm{df_1} = p(k - 1),$$

$$\mathrm{df_2} = \mathrm{ft} - \mathrm{g},$$

and

$$U = W^{1/t}.$$

To test at level α requires CV = $F_{1-\alpha}$ (df_1, df_2). This method is designated here as WLR. It can be shown that (6) provides an exact *F* test for $p = 1, 2$ or $k = 2, 3$ (Seber, 1984, pp. 40-41).

One method of evaluating *T* for a group of *k* mean vectors is with an *F* test (McKeon, 1974; Seber, 1984, p. 39) defined by

$$F = \frac{\mathrm{T}}{\mathrm{c}} \qquad (7)$$

where

$$B = \frac{(\mathrm{df_E} + \mathrm{k\text{-}p\text{-}2})(\mathrm{df_E}\text{-}1)}{(\mathrm{df_E}\text{-}\mathrm{p}\text{-}3)(\mathrm{df_E}\text{-}\mathrm{p})},$$

$$b = 4 + \frac{\mathrm{a+2}}{\mathrm{B\text{-}1}},$$

and

$$c = \frac{\mathrm{a(b\text{-}2)}}{\mathrm{b}(\mathrm{df_E}\text{-}\mathrm{p}\text{-}1)}.$$

To test at level α requires CV = $F_{1-\alpha}$ (a, b). This method is designated here as THL.

Routines for computing *p* values for Roy's *R* are either quite complex or rather crude. The versions used by statistical packages are not very accurate. For example, SAS prints a footnote on output warning that the corresponding *F* ratio for *R* is an upper bound. Consequently, the *p* value is a lower bound. Therefore, a *p* value of .04 would only tell the user that the exact *p* value is no less than .04. It would be more helpful to know that the exact *p* value was no greater than some value. Tables of critical values for *R* are available (Harris, 2001, pp. 518-531; Sever, 1984, pp. 593-598). Routines described by Harris (2001) were used to determine *p* values and critical values in the present study; the method is designated here as GCR.

Pairwise testing on a discriminant function can be performed as described by Harris (2001, p. 222). The *F* test for the difference between a given pair of means on the discriminant function is compared to a critical value, F_{CRIT}. The value of F_{CRIT} is found from $df_E(\theta_{CRIT})/(1 - \theta_{CRIT})$ where θ_{CRIT} is the critical value for *R*.

Noncentrality

In the non-null case, the *p* x *p* matrix Φ can be defined as

$$\Phi = \sum_{i=1}^{k} n_i(\boldsymbol{\mu}_i - \boldsymbol{\mu})(\boldsymbol{\mu}_i - \boldsymbol{\mu})', \qquad (8)$$

where $\boldsymbol{\mu}$ is the grand mean vector of the population.

Take the *p* x *p* matrix Γ as

$$\Gamma = \Phi\Sigma^{-1},$$

where Σ is the population covariance matrix. The *p* eigenvalues of Γ are $\gamma_1, \ldots, \gamma_p$. The noncentrality parameter, δ^2, is

$$\delta^2 = \sum_{i=1}^{p} \gamma_i. \qquad (10)$$

Populations vary along a continuum from a concentrated structure where γ_1 is the only nonzero eigenvalue of Γ to a diffuse structure where *s* eigenvalues of Γ are nonzero. When the usual MANOVA assumptions are satisfied the most powerful tests of the four listed above for evaluating a concentrated structure would be *R*. For the diffuse structure the most powerful of the four tests would be *V* (Olson, 1974).

Robustness

Investigations of various testing procedures have shown marked differences in robustness (Olson, 1974). All test procedures in MANOVA have reduced control of Type I and Type II errors in the presence assumption

failure. The most extreme problems occur for *R* and the least for *V*.

Follow-Up Tests

Roy's *R* has been found to be more useful than *V*, *W*, or *T* for finding specific differences between groups (Bird & Hadzi-Pavlovic, 1983). In order to improve the robustness and interpretability of significant group differences, Bird and Hadzi-Pavlovic, (1983) proposed limiting the testing of group contrasts in two ways. First, they proposed the examination of group differences on single dependent variables, sums of dependent variables, differences between dependent variables, or combinations of these. That is, complex weightings of dependent variables used to form discriminant functions were avoided.

The second restriction was a limitation of the contrasts on group means to be tested. A moderate restriction on contrasts allows only one subset of means to be compared to another subset. With $k = 4$ there would be only 25 possible contrasts (6 pairwise, 3 pairs versus another pair, 12 pairs versus a single & 4 triples versus a single). With $p = 2$ dependent variables there would be four variables for testing (2 dependent variables, one sum, & one difference). That would allow only 100 contrasts to be tested. For $p = 6$ the total number of contrasts to be tested would be 9,100.

A strong restriction on the permissible contrasts for $k = 4$ would allow the 25 contrasts about the 4 groups to be applied only to each dependent variable. With $p = 2$, there would be only 50 tests performed. With $p = 6$ there would be 150. Bird and Hadzi-Pavlovic, (1983) reported considerable improvement in Type I error control under assumption failure with both moderate and strong restrictions. A univariate, Bonferroni- Scheffé (B-S) approach was also considered by testing contrasts on each dependent variable using the Scheffé (1953) procedure at level α/p. They also suggest the possibility of a so-called protected *R* test in which *R* is applied to testing contrasts only after a significant overall test such as *V*.

In an attempt to increase power, Sheehan-Holt (1998) considered a partially restricted condition. Sheehan-Holt placed no restriction on the variable thus allowing the testing of group contrasts on any discriminant function. For $k = 4$, the 25 contrasts would be tested on the first discriminant function. If the first discriminant function were limited to pairwise testing there would be only six tests of group differences on the discriminant function for $k = 4$.

A Monte Carlo Study

The present restriction on group contrasts to be tested is limited to pairwise testing. For $k = 4$, the six contrasts constitute fewer group contrasts than any considered by Bird and Hadzi-Pavlovic, (1983) or Sheehan-Holt (1998). However, the present investigation applies those group contrasts to all significant discriminant functions.

Seven procedures were used to test the full null hypothesis: VPB, VM1, VM2, THL, WLR, GCR, and the Bonferroni-Scheffé (B-S). The first five procedures follow a significant overall test with pairwise testing based on *R*. These five methods are examples of a protected *R* test. The GCR procedure also applies pairwise testing as an unprotected *R* test.

Conditions investigated included $k = 4$, common group sizes *N* of 10, 15 and 20, and $p = 4$. The population covariance matrix was varied to produce either uncorrelated variates ($\mathbf{\Sigma} = \mathbf{I}$) or $\mathbf{\Sigma}$ with all variables correlated by a common correlation ρ of either 0.71 or –0.2. For non-null conditions δ^2 was varied over a range of several values to produce power values in the neighborhood of 0.50.

Covariance Heterogeneity

Following Bird and Hadzi-Pavlovic (1983) and Olson (1974), heterogeneity was introduced by multiplying all variates in Group 1 by a constant chosen to produce a value of the coefficient of variation, *C*, (Box, 1954). If the variances in Group 1 are all initially set at $\sigma^2 = 1$ and a value *d* is the multiplicative value, C^2 can be calculated as

$$C^2 = \frac{1}{k\bar{\sigma}^4}\sum_{i=1}^{k}(\sigma_i^2 - \bar{\sigma}^2)^2, \qquad (11)$$

where $\sigma_i^2 = d$ for $i = 1$ and 1 for $i \neq 1$, and

$$\bar{\sigma}^2 = \frac{\sum_{i=1}^{k} \sigma_i^2}{k}.$$

Bird and Hadzi-Pavlovic (1983) used $C = 0.4$ as moderate covariance heterogeneity and $C = 0.8$ as substantial covariance heterogeneity. Thus, C^2 values would be .16 for moderate and 0.64 for substantial covariance heterogeneity. However, Olson (1974) investigated values as high as $C^2 = 2.4$. The present investigation examined values as high as $C^2 = 2.0$. Olson's (1974) results seem to suggest that error rates approach an upper limit for very high values of C^2.

Nonnormality

Previous studies have given little consideration to failure of the normality assumption. Some degree of kurtosis has been investigated showing relative little effect. However, the degree of kurtosis is not clear. For example, the fourth moment calibration was not reported.

Micceri (1989) reported many distributions that were clearly nonnormal. However, the data sets reported by Micceri are not as extreme as those used in many studies evaluating statistical robustness. Among skewed distributions, Micceri identified the most extreme distributions as being typified by the exponential distribution with standardized third and fourth moments as ($\sqrt{\beta_1} = 2.0$, $\beta_2 = 9.0$). Among symmetric, platykurtic distributions Micceri represented the shape as typical of the uniform distribution ($\sqrt{\beta_1} = 0.0$, $\beta_2 = 1.8$). Among symmetric, leptokurtic distributions Micceri identified the shape as double exponential ($\sqrt{\beta_1} = 0.0$, $\beta_2 = 6.0$).

To investigate the effects of distribution shape, four shapes were considered: the normal, uniform, exponential, and double exponential. The three nonnormal shapes represent the most extreme conditions reported by Micceri (1989). The uniform distribution was easily produced directly from the generated random numbers. The double exponential was approximated as a *t* distribution with *df* = 6. This *t* distribution has the same third and fourth moments as the double exponential distribution. The exponential distribution was approximated by Johnson's (1949) S_B method as described by Tadikamalla (1980) with $\sqrt{\beta_1} = 2.0$ and $\beta_2 = 9$.

Each simulated experiment was replicated 10,000 times. Significant differences in Type I error rates can be identified as deviating from an expected interval about the nominal rejection rates. For rejection rates between 0.0 and 1.0 the standard error (SE) depends on the value of the rate. If x is the proportion of replications exceeding a critical value, the SE is $[x(1 - x)/10000]^{1/2}$. For x = 0.5 the SE would be a maximum and have a value, SE = $\sqrt{.000025} = 0.005$ so a 50% rejection rate would be included in a 2SE interval from 0.49 to 0.51 in approximately 95% of the simulations. An x of 0.05 would have SE = $\sqrt{.00000475} = 0.002179$ and a 2SE interval from 0.045641 to 0.054358. Thus rates even as small as 5% will usually be estimated to differ from the correct value by no more than about 0.0044.

Even after Type I error rates are identified as significantly different from nominal levels and not due to chance, an additional question arises. How much deviation from the nominal level is acceptable to a given researcher? Bradley (1978) has suggested that a real error rate that differs from the intended nominal rate, α, by no more than 0.1α is negligibly non-robust. Thus, a rate of $\alpha = 0.05$ should not exceed 0.055 to be negligibly non-robust. Bradley (1978) also suggested that rates above $1.5\,\alpha$ (0.075 for $\alpha = 0.05$), should never be accepted as robust. All researchers must make their own decisions but an upper limit of 0.075 for the 0.05-level test seems a useful guideline.

Power rates require a different approach. To compare power rate for two statistical procedures requires that they have the same, or in some sense equivalent, control of Type I errors. If one procedure has true Type I error rates that never exceed the nominal level and a second procedure has true Type I error rates that never exceed one half the nominal level then both are limiting the Type I error rate to no more than the nominal level: Power rates can be expected to be higher for the first procedure but that may not always be the case.

Any uniformly, higher power rate for one of two such procedures justifies identifying it as more powerful. Higher power rates in specific conditions may guide a researcher to select a procedure based on conditions of the investigation. If power rates are uniformly higher but small then other factors such as ease of application may be considered. Einot and Gabriel (1975) used such an argument in the univariate case to support a slightly less powerful procedure. Power advantages less than 0.1 might be ignored, but advantages above 0.2 might be designated as substantial and override other considerations. Again, all researchers must make their own decisions.

McNemar's (1947) test of correlated proportions was used to test the significance of the difference between proportions as power rates in the non-null conditions. For greater efficiency the procedures were placed in order with consecutive procedures tested for power differences. The order is VPB, VM1, VM2, THL, WLR, GCR, and B-S.

Results

Type I Error Rates

Table 1 presents the Type I error rates for seven procedures with $k = 4$, equal N of 10, three values of ρ, four population distributions, and $C^2 = 1.6$. The overall maximum error rates are in bold print. Those are also the maximum error rates for the same conditions when C^2 has values 0.0, 0.8, and 1.2. Clearly, with C^2 values as high as 1.6, the error rates are well above the Bradley upper limit of 0.075. None of the procedures is robust by this criterion for that value of C^2.

The maximum error rates in Table 1 all occur for populations with an exponential distribution. This suggests that differences in skewness are more important than differences in kurtosis. Only differences in kurtosis were investigated in the previous studies (Bird & Hadzi-Pavlovic, 1983; Olson, 1974; Sheehan-Holt, 1998).

Table 2 presents summaries for N values of 10, 15 and 20 including the maximum rates for the results shown in Table 1. In every case the maximum error rate was found for the exponential population but could be for any one of the three values of ρ.

As shown in Table 2 (a) with N = 10, the $C^2 = 0.0$ condition shows all seven procedures to have a maximum Type I error rate below the nominal 0.05 level even when the maximum is taken over three values of ρ and four population distributions. When C^2 rises to 0.8, only VPB, the original testing formula for the V statistic is below the nominal level. However, VM1 and VM2 have maximum rates almost identical to the nominal level. Also, THL, WLR, and GCR satisfy the 0.075 limit to robustness. The Bonferroni-Scheffé is not robust for $C^2 \geq 0.8$.

If the $C^2 = 0.64$ definition of substantial covariance heterogeneity is accepted as suggested by Bird and Hadzi-Pavlovic (1983), the VPB combination of testing V and pairwise testing with R is robust for that condition. The same conclusion is probably justified for VM1 and VM2.

In all parts of Table 2 the Bonferroni-Scheffé, B-S, procedure has a simple, almost linear relationship between error rates and C^2. The greater the covariance heterogeneity the higher is the Type I error rate. The situation is quite different for the other six, multivariate procedures. Table 2 (b) presents results for $N = 15$. Even for $C^2 = 2.0$ the first five procedures have no more than negligible non-robustness (i.e. ≤ 0.055). GCR does exceed that limit but only for the most extreme case and is always robust (i.e. ≤ 0.075).

Table 2 (c) presents results for $N = 20$. All six multivariate procedures are conservative (i.e. rates ≤ 0.05). Even GCR is conservative and the protection of another procedure may not be needed. The greater control of Type I errors for all multivariate procedures as shown in Table 2(c) suggests that protected tests are not needed for sample sizes this large. The maximum Type I error rate for GCR is 0.0369 occurs for $C^2 = 0.8$.

Power Rates

For $N = 10$ the five protected procedures (VPB, VM1, VM2, THL, WLR) provide varying control of Type I errors for C^2 vales from 0.0 to about 0.8. The B-S procedure provides poor control in the same conditions of C^2. However, B-S represents a useful alternative provided it can be equated in Type I error control. Repeated testing of these six procedures (VPB, VM1,

Table 1: Type I Error Rates for Seven Pairwise Testing Procedures for k = 4, N = 10, $\alpha = .05$, $C^2 = 1.6$ and a True, Full-Null Hypothesis

ρ	Population	VPB	VM1	VM2	THL	WLR	GCR	B-S
0.00	Normal	.0240	.0242	.0241	.0256	.0253	.0262	.1046
	Uniform	.0327	.0333	.0333	.0347	.0340	.0348	.1193
	Exponential	.0788	.0814	.0810	.0893	.0875	.0921	**.1827**
	Double Exponential	.0206	.0208	.0207	.0225	.0219	.0229	.0786
0.71	Normal	.0279	.0284	.0284	.0300	.0296	.0311	.0844
	Uniform	.0329	.0335	.0335	.0346	.0345	.0349	.0864
	Exponential	.0792	.0814	.0814	.0904	.0886	.0927	.1142
	Double Exponential	.0236	.0238	.0237	.0253	.0248	.0256	.0745
–0.20	Normal	.0254	.0261	.0261	.0281	.0272	.0283	.1086
	Uniform	.0295	.0297	.0297	.0309	.0304	.0313	.1199
	Exponential	**.0823**	**.0855**	**.0852**	**.0943**	**.0914**	**.0977**	.1664
	Double Exponential	.0198	.0203	.0203	.0220	.0215	.0228	.0892

Notes: C^2 = measure of variance heterogeneity, ρ = correlation, VPB = *V* tested by Pillai, (1955) formula, VM1 = *V* tested by Muller (1988) Method 1, VM2 = *V* tested by Muller (1988) Method 2, THL = *T* tested by McKeon, (1974), WLR = *W* tested by Rao, (1951), GCR = *R* tested by Harris, (2001), B-S = Bonferroni-Scheffé. Pairwise testing of first six procedures done by ρ (see Harris, 2001, p. 222); Maximum value for each column in **bold**.

VM2, THL, WLR, B-S) showed that each would limit the Type I error rate to a maximum .05 in the conditions of Table 2(a) provided they were applied at the nominal rates of 0.0115, 0.0093, 0.0095, 0.0016, 0.0036 and 0.0024, respectively.

Any-pair power is defined as the probability of detecting one or more true differences between pairs of population means. Table 3 presents the any-pair power rates for the six procedures applied to the first discriminant function for data from four population distributions, $k = 4$, $N = 10$ and a diffuse noncentrality structure.

The most powerful procedure in all conditions is VM1 testing *V* with Muller's Method 1. McNemar's test showed each procedure to be significantly different from the one to the right provided the difference was at least 0.0006 or more. However, many differences are quite small. The power advantage of VM1 over the other protected *R* procedures can be seen in Table 3 to be modest. The power advantage of VM1 over VPB and

Table 2: Maximum Over Three ρ values, and Four Populations for Type I Error Rates for Seven Pairwise Testing Procedures for $k = 4$, $\alpha = .05$, C^2 = measure of variance heterogeneity, and a True, Full-Null Hypothesis

C^2	VPB	VM1	VM2	THL	WLR	GCR	B-S
(a) $N = 10$							
0.0	.0175	.0186	.0185	.0232	.0210	.0301	.0277
0.8	.0473	.0508	.0505	.0639	.0588	.0735	.1008
1.2	.0669	.0706	.0702	.0850	.0799	.0916	.1573
1.6	**.0823**	**.0855**	**.0852**	**.0943**	**.0914**	**.0977**	**.1827**
(b) $N = 15$							
0.0	.0172	.0179	.0178	.0208	.0196	.0265	.0241
0.8	.0421	.0425	.0425	.0467	.0452	.0507	.0936
1.2	.0509	.0511	.0510	**.0542**	.0532	**.0565**	.1427
1.6	.0524	.0525	.0525	.0533	.0531	.0537	.1736
2.0	**.0534**	**.0534**	**.0534**	.0534	**.0534**	.0535	**.2053**
(c) $N = 20$							
0.0	.0201	.0208	.0207	.0228	.0216	.0292	.0237
0.8	**.0328**	**.0334**	**.0333**	**.0349**	**.0342**	**.0369**	.0930
1.2	.0325	.0327	.0327	.0332	.0331	.0336	.1285
1.6	.0322	.0322	.0322	.0323	.0323	.0323	.1533
2.0	.0297	.0297	.0297	.0297	.0297	.0297	**.1882**

Notes: VPB = *V* tested by Pillai, (1955) formula, VM1 = *V* tested by Muller (1988) Method 1, VM2 = *V* tested by Muller (1988) Method 2, THL = *T* tested by McKeon, (1974), WLR = *W* tested by Rao, (1951), GCR = *R* tested by Harris, (2001), B-S = Bonferroni-Scheffé; Maximum value for each column in **bold**.

VM2 is always less than 0.01. The power advantage of VN1 over WLR is always less than 0.06. The greatest power advantage of VM1 over any protected *R* procedure is over THL but is always less than 0.15.

The power advantage of VM1 over B-S can be quite large. For normal populations the maximum is 0.4744 (= 0.6712 – 0.1968). For the other distributions the maximum power advantages are 0.4750 (= 0.6559 – 0.1809) for uniform distributions, 0.4453 (= 0.7514 – 0.3061) for exponential distributions, and 0.4652 (= 0.7062 – 0.2410) for double exponential distributions.

The maximum power advantages of VM1 over B-S for diffuse noncentrality structures and $C^2 = 0$ (i.e. homogeneous covariances) are shown in Table 4(a) for each of the four population distributions and three values of ρ. The power advantages vary from 0.4453 to 0.8896.

The same conditions reported in Table 3 were investigated for a diffuse noncentrality structure but $C^2 = 1.6$. The maximum power advantages of VM1 over B-S for a diffuse

Table 3: Any-Pair Power of Five Procedures on the First Discriminant Function and B-S for N = 10, $\alpha = .05$, Four Distributions, A Diffuse Non-centrality Structure and Four Non-centrality Values and $C^2 = 0.0$

Population	δ^2	VPB	VM1	VM2	THL	WLR	B-S
Normal	30.0	.6679	**.6712**	.6694	.5478	.6366	.1968
	24.3	.5233	**.5303**	.5260	.3909	.4797	.1277
	19.2	.3760	**.3829**	.3775	.2425	.3275	.0733
	14.7	.2537	**.2610**	.2558	.1436	.2072	.0453
Uniform	30.0	.6526	**.6559**	.6542	.5275	.6172	.1809
	24.3	.4983	**.5038**	.5008	.3678	.4558	.1150
	19.2	.3536	**.3603**	.3560	.2271	.3073	.0672
	14.7	.2277	**.2354**	.2308	.1256	.1886	.0388
Exponential	30.0	.7479	**.7514**	.7486	.6434	.7214	.3061
	24.3	.6000	**.6048**	.6026	.4697	.5588	.1907
	19.2	.4580	**.4637**	.4602	.3143	.4054	.1169
	14.7	.3117	**.3196**	.3151	.1820	.2612	.0575
Double Exponential	30.0	.7028	**.7062**	.7044	.5970	.6767	.2410
	24.3	.5592	**.5650**	.5615	.4249	.5141	.1561
	19.2	.4072	**.4145**	.4093	.2704	.3594	.0874
	14.7	.2728	**.2788**	.2755	.1627	.2314	.0503

Notes: VPB = *V* tested by Pillai, (1955) formula, VM1 = *V* tested by Muller (1988) Method 1, VM2 = *V* tested by Muller (1988) Method 2, THL = *T* tested by McKeon, (1974), WLR = *W* tested by Rao, (1951), GCR = *R* tested by Harris, (2001), B-S = Bonferroni-Scheffé; Maximum value for each row in **bold**.

noncentrality structures are shown in Table 4(b) for each of the four population distributions and three values of ρ. The power advantages vary for 0.2238 to 0.7288.

The same conditions reported in Table 3 and Table 4(a) were investigated for a concentrated noncentrality structure where group differences existed only along a single dimension. The maximum power advantages of VM1 over B-S for a concentrated noncentrality structures are shown in Table 4(c) for each of the four population distributions and three values of ρ. The power advantages vary from –0.1454 to 0.5335. Of course, the negative advantage means that B-S has a power advantage over VM1 as high as 0.1454. This occurs only for ρ = 0.71 but for all four population distributions.

The same conditions reported in Table 4(b) were investigated for a concentrated noncentrality structure where group differences existed only along a single dimension. The maximum power advantages of VM1 over B-S for concentrated noncentrality structures are shown in Table 4(d) for each of the four population distributions and three values of ρ. The power advantages vary for –0.4019 to 0.4827. Again the negative advantage means

Table 4: Any-Pair Power Advantage of VM1 Over B-S for $k = 4$, $N = 10$, a = .05, and $C^2 = 0.0$ or 1.6

	ρ		
Population	0.0	0.71	–0.2
(a) Diffuse Noncentrality Structure with $C^2 = 0$			
Normal	.4744	.8748	.6418
Uniform	.4750	.8854	.6501
Exponential	.4453	.8696	.5997
Double Exponential	.4652	.8708	.6245
(b) Diffuse Noncentrality Structure with $C^2 = 1.6$			
Normal	.2975	.7288	.6217
Uniform	.3809	.7311	.6258
Exponential	.2238	.6543	.5781
Double Exponential.	.2920	.7259	.5974
(c) Concentrated Noncentrality Structure with $C^2 = 0$			
Normal.	.5133	-.1454	.8724
Uniform	.5335	-.1327	.8873
Exponential	.4579	-.1043	.8505
Double Exponential	.4780	-.1155	.8561
(d) Concentrated Noncentrality Structure with $C^2 = 1.6$			
Normal	.0487	-.3668	.4170
Uniform	.0484	-.4019	.4827
Exponential	.3826	-.1549	.3070
Double Exponential	.0553	-.3369	.4756

that B-S has a power advantage over VM1 as high as 0.4019. This occurs only for ρ = 0.71 and for all four population distributions.

As shown in Table 2(b), all six multivariate procedures, VPB, VM1, VM2, THL, WLR, and GCR, showed good control of Type I errors for $N = 15$. In the most extreme conditions each of these procedures has a Type I error rate slightly above the nominal level. Even GCR, with no additional multivariate test, had a maximum rate of only 0.0565. Although that exceeds Bradley's negligible nonrobustness limit of 0.055, it might be adequate for some researchers. The rates at which each of the seven procedures must be performed to limit the actual Type I error rate to the nominal 0.05 level are 0.044, 0.044, 0.044, 0.044, 0.044, 0.044, 0.0005 respectively for VPB, VM1, VM2, THL, WLR, GCR, and B-S.

Table 5 presents the power advantages of GCR over B-S for $N = 15$ just as did Table 4 for the power advantage of VM1 over B-S. In Table 5, the greater power for B-S over GCR for ρ = 0.71 with concentrated noncentrality structures occurs only for the heterogeneous covariance condition.

The power advantage of GCR over B-S for ρ = 0.0 in Table 5(d) is less than 0.1 for all populations and becomes slightly negative for exponential distributions.

Table 5: Any-Pair Power Advantage of GCR Over B-S for $k = 4$, $N = 15$, $\alpha = .05$, and $C^2 = 0.0$ or 2.0

	ρ		
Population	0.0	0.71	–0.2
(a) Diffuse Noncentrality Structure with $C^2 = 0.0$			
Normal	.6516	.8984	.7528
Uniform	.6511	.8975	.7744
Exponential	.6030	.9049	.7346
Double Exponential	.6243	.9041	.7354
(b) Diffuse Noncentrality Structure with $C^2 = 2.0$			
Normal	.4737	.8081	.5719
Uniform	.5399	.8219	.6010
Exponential	.3207	.7342	.4180
Double Exponential.	.4110	.8002	.5380
(c) Concentrated Noncentrality Structure with $C^2 = 0.0$			
Normal.	.7970	.3290	.9241
Uniform	.8205	.3448	.9288
Exponential	.8159	.3556	.9187
Double Exponential	.7827	.3264	.9284
(d) Concentrated Noncentrality Structure with $C^2 = 2.0$			
Normal	.0607	-.3958	.5498
Uniform	.0618	-.4434	.5415
Exponential	-.0304	-.1979	.3274
Double Exponential	.0584	-.3700	.5469

Table 6 presents the power advantages of GCR over B-S for $N = 20$. The conservative Type I error rejection rate GCR implies that the procedure must be applied at a lenient rate of 0.099 to limit the rate to 0.05. In contrast B-S must be applied at a rate of 0.0008. The power advantages of GCR over B-S in Table6 are similar to those of Table 5.

Conclusion

The present investigation extends the previous work of Bird and Hadzi-Pavlovic (1983) and Sheehan-Holt (1998) on follow-up tests for MANOVA to pairwise testing on the discriminant functions. As shown in Tables 1 and 2, Type I error rates can be quite high depending upon ρ (the correlation between dependent variables), the population distribution, sample size N, and especially the covariance heterogeneity, C^2.

For samples of size, $N = 10$, and only moderate covariance heterogeneity (i.e. $C^2 = 0.8$), Three protected tests, VPR, VM1, and VM2, provide good control of Type I errors even for realistic nonnormality. Even for slightly higher covariance heterogeneity (i.e. $C^2 = 1.2$), these three protected R procedures are below Bradley's (1978) 1.5 α limit for robustness.

Power comparisons in the present investigation used adjusted alpha levels so that

Table 6: Any-Pair Power Advantage of GCR Over B-S for $k = 4$, $N = 20$, $\alpha = 0.05$ and $C^2 = .0$, 0.8 or 2.0

	ρ		
Population	0.0	0.71	–0.2
(a) Diffuse Noncentrality Structure with $C^2 = 0.0$			
Normal	.7159	.9364	.8048
Uniform	.7226	.9400	.8217
Exponential	.6883	.9343	.7683
Double Exponential	.7072	.9396	.7856
(b) Diffuse Noncentrality Structure with $C^2 = 2.0$			
Normal	.5582	.8397	.6291
Uniform	.5946	.8569	.6584
Exponential	.4314	.7616	.5010
Double Exponential.	.5059	.8214	.6094
(c) Concentrated Noncentrality Structure with $C^2 = 0.0$			
Normal.	.8386	.4161	.9649
Uniform	.8349	.4313	.9674
Exponential	.8159	.4446	.9517
Double Exponential	.8258	.4099	.9590
(d) Concentrated Noncentrality Structure with $C^2 = 0.8$			
Normal	.3355	-.2174	.7411
Uniform	.333	-.2292	.7420
Exponential	.2579	-.0916	.6048
Double Exponential	.3461	-.1822	.7541

power could be compared when all methods provided the same control of Type I errors. Table 3 shows a clear advantage in power over all procedures for homogeneous covariance and diffuses noncentrality condition for VM1. However, the power advantage over VPB and VM2 is only modest. The power advantage of VM1 over the Bonferroni-Scheffé (B-S) is shown in Tables 3 and 4 to be as high as 0.8854 but can be as low as –0.1454. On balance, the protected multivariate approach of VM1 is clearly superior to the univariate approach of B-S.

As shown in Table 2(b), a minimum sample size of about 15 is sufficient for GCR to provide adequate control of Type I errors even without the addition of the alternative protection of an additional multivariate test. Table 5 shows the power advantage of GCR over B-S to range from 0.9049 to –0.3958. As was true for Table 4 results, the power advantage of B-S is almost exclusively in conditions where $\rho = 0.71$. A univariate-based follow-up is most powerful when dependent variables are highly, positively correlated.

Table 6 provides power advantages for GCR over B-S for $N = 20$. These rates range from 0.94 to –0.2174 and are similar to those in Table 5. Although B-S can be powerful even when applied at a reduced alpha level to control

Type I errors, it would still not be practical in those conditions. Continually applying a test at different alpha levels is tedious and requires a large table of appropriate alpha levels.

Discriminant functions are more difficult to interpret than are simple combinations of dependent variables. However, MANOVA may profitably be considered not just as combined dependent variables but rather a blending of several ANOVAs and factor analysis. A discriminant function can be considered an approximation to a latent variable. The correlation between each dependent variable and the discriminant function could be used to identify the latent variable just as is done in factor analysis using factor loadings.

If a new statistical package is being developed, it might be desirable to replace the traditional VPB with VM1. However, the existing VPB reported by many statistical packages such as SAS and SPSS should provide adequate results in a protected *R* test for small *N*.

Numerous additional conditions could be considered. Various patterns of correlations might have an effect. More powerful methods of pairwise testing then the Scheffé could be considered if one is willing to consider only pairwise testing. The higher rejection rates of such powerful pairwise tests are also likely to produce even higher Type I error rates. More extreme nonnormality than is considered there can be investigated.

Example

Baumann, Seifert-Kessell, and Jones (1992) report comparing three strategies for teaching reading comprehension to fourth-graders. One strategy was Think-Aloud (TA). A second strategy was Direct Reading Activity (DRA). The third was Direct Reading and Thinking Activity (DRTA). The two dependent variables were Error Detection Task (Y_1) and Degrees of Reading Power (Y_2). There were 21 students in each of the three groups. The means and standard deviations were:

TA

Y_1	Y_2
M = 7.7727	M = 43.4545
SD = 3.9271	SD = 7.8603

DRA

Y_1	Y_2
M = 6.6818	M = 42.0455
SD = 2.7669	SD = 6.6151

DRTA

Y_1	Y_2
M = 6.2273	M = 46.6364
SD = 2.0915	SD = 7.6441

Analysis in SAS produces:

Eigenvalues

	λ	θ
Root 1	.165844	.142252
Root 2	.019988	.019596

Eigenvectors

	Y_1	Y_2
Root 1	-.038037	.017307
Root 2	.027758	.008466

s = 2, m = –0.5, n = 30

Statistic	Value	P-Value
Wilks' Lambda	0.84093942	0.0286
Pillai's Trace	0.16184815	0.0284
Hotelling-Lawley Trace	0.18583147	0.0290
Roy's Greatest Root	0.16584380	0.0321

Dividing each eigenvector element by the square root of the sum of squared values for the eigenvector, convert each subjects' dependent variable scores to a score on the first discriminant function.

$$DF1 = 0.414159Y_2 - 0.910204Y_1$$

Group	1	2	3			
N	21	21	21	$MS_E = 9.0893$		
Mean	10.9223	11.3317	13.6468	Value	SS	F
Contrast 1	-1	0	1	2.7245	77.9405	8.58*
Contrast 2	-1	1	0	0.4094	1.7599	0.19
Contrast 3	0	-1	1	2.3151	56.2767	6.19

s	n	m	$\theta_{.95}$	$df_E(\theta_{.95})/(1-\theta_{.95})$	CV
2	30	–0.5	0.1287	30(0.1287)/(0.8713) =	6.73

Group 3 (DRTA) is significantly higher than Group 1 (TA) on the first discriminant function at $\alpha = 0.05$. The average, within-group correlation between Y_1 and DF1 is –0.50. The average, within-group correlation between Y_2 and DF1 is 0.54. The two, dependent variables have about the same size relationship to DF1, however, Y_1 is inversely related whereas Y_2 is directly related to DF1. Y_1 was measuring the number of errors to be detected so it is negatively related to Y_2, reading power. DF1 is a composite measure of error detection and reading power.

The three groups failed to differ significantly on either dependent variable even at $\alpha = 0.10$. A significant B-S would require group differences on at least one dependent variable to be significant at the 0.025 level.

References

Baumann, J. F., Seifert-Kessell, N., & Jones, L. A. (1992). Effect of think-aloud instruction on elementary students' comprehension monitoring abilities. *Journal of Reading*, *24*, 143-172.

Bird, K. D., & Hadzi-Pavlovic, D. (1983). Simultaneous test procedures and the choice of a test statistic in MANOVA. *Psychological Bulletin*, *93*, 167-178.

Box, G. E. P. (1954). Some theorems on quadratic forms applied in the study of analysis of variance problems: Effects of inequality of variance in the one-way classification. *Annals of Mathematical Statistics*, *25*,290-302.

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology*, *31*, 141-152.

Einot, I., & Gabriel, K. R. (1975). A study of the powers of several methods of multiple comparisons. *Journal of the American Statistical Association*, *70*, 574-583.

Harris, R. J. (1985) Extending the GCR tables: n < 1 and n > 1000. *Multivariate Behavioral Research*, *20*, 475-481.

Harris, R. J. (2001). *A primer of multivariate statistics*. Mahwah, NJ: Lawrence Erlbaum.

McNemar, Q. (1947). Note on the sampling error of the differences between correlated proportions or percentages. *Psychometrika*, *12*, 153-157.

McKeon, J. (1974). F approximations to the distribution of Hotelling's T^2. *Biometrika*, *61*, 381-383.

Micceri, T. (1989).The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, *105*,156-166.

Muller, K. E. (1998). A new F approximation of the Pillai-Bartlett trace under H_0, *Journal of Computational and Graphical Statistics*, *7*, 131-137.

Pillai, K. C. S. (1955). Some new test criteria in multivariate analysis. *Annals of Mathematical Statistics*, *26*, 117-121.

Rao, C. R. (1951). An asymptotic expansion of the distribution of Wilks' criterion. *Bulletin of the Institute of International Statistics*, *33*, 177-180.

Roy, S. N. (1966). Sensitivity comparisons among tests of the general linear hypotheses, *Journal of the American Statistical Association*, *61*, 415-435.

Olson, C. L. (1974). Comparative robustness of six tests in multivariate analysis of variance. *Journal of the American Statistical Association*, *69*, 894-908.

Seber, G. A. F. (1984). *Multivariate observations*. New York, NY: Wiley.

Scheffé, H. (1953). A method for judging all contrasts in the analysis of variance (Corr: V56 p229) *Biometrika*, *40*, 87-104.

Sheehan-Holt, J. K. (1998). MANOVA simultaneous test procedures: The power and robustness of restricted multivariate contrasts. *Educational and Psychological Measurement*, *58*, 861-881.

Tadikamalla, P. R. (1980). On simulating non-normal distributions. *Psychometrika*, *45*, 273-279.

Journal of Modern Applied Statistical Methods
November 2011, Vol. 10, No. 2, 418-423

1538 – 9472/11/$95.00

Regular Articles

Type I Error Rates of the Two-Sample Pseudo-Median Procedure

Nor Aishah Ahad
Universiti Utara Malaysia,
Kedah, Malaysia

Abdul Rahman Othman
Universiti Sains Malaysia,
Penang, Malaysia

Sharipah Soaad Syed Yahaya
Universiti Utara Malaysia,
Kedah, Malaysia

The performance of the pseudo-median based procedure is examined in terms of controlling Type I error for a two independent groups test. The procedure is a modification of the one-sample Wilcoxon statistic using the pseudo-median of differences between group values as the central measure of location. The proposed procedure was shown to have good control of Type I error rates under the study conditions regardless of distribution type.

Key words: Mann-Whitney-Wilcoxon, pseudo-median, *t*-test, type I error.

Introduction

Testing the equality of central tendency parameters between two independent samples by controlling Type I error is a common statistical problem. If an underlying distribution is normally distributed with equal population variances, the most suitable test statistic to use is the Student's *t*-test. Student's *t*, however, is sensitive to non-normal data and heterogeneity of variances. Under these situations, Welch's approximate test (Welch, 1938) usually offers the best practical solution, but this statistic does not adequately control Type I error probabilities under non-normal distributions.

To surmount the problem of non-normality, researchers typically seek nonparametric test alternatives, such as the Mann-Whitney-Wilcoxon, which is believed to be effective against violations of normality. Although ranking methods are often useful when samples are obtained from heavy-tailed distributions, they are influenced by unequal variances similar to parametric tests (Pratt, 1964; Zimmerman & Zumbo, 1992). Further, nonparametric methods are more appropriate for non-normal symmetric data. Many attempts have been made to deal with asymmetric distributions. In this study, a method to handle the problem of asymmetric data, as well as heterogeneity of variances, is suggested. The method is known as the pseudo-median procedure, where the pseudo-median of differences between group values are employed as the central measure of location with the one-sample nonparametric Wilcoxon procedure in a two group setting. The pseudo-median of a distribution F is defined to be the median of the distribution $(Z_1 + Z_2)/2$, where Z_1 and Z_2 are all possible differences between two observations from each group. Z_1 and Z_2 are independent and have the same distribution as F (Hoyland, 1965; Hollander & Wolfe, 1999).

The pseudo-median is a location parameter. The estimation of this parameter is accomplished using the Hodges-Lehmann estimator. According to Hollander and Wolfe (1999), the Hodges-Lehmann estimator $\left(\hat{\theta}\right)$ is a consistent estimator of the pseudo-median, which in general may differ from the median. However, when F is symmetric, the median and pseudo-median coincide. The pseudo-median is selected as the central measure of location because it is convenient and the asymptotic properties of the pseudo-median are the same as

Nor Aishah Ahad is an Academician in the School of Quantitative Sciences. Email: aishah@uum.edu.my. Abdul Rahman Othman is a Professor in the School of Distance Education. Email: oarahman@usm.my. Sharipah Soaad Syed Yahaya is an Associate Professor in the School of Quantitative Sciences. Email: sharipah@uum.edu.my.

median. In this study, the performance of the pseudo-medians procedure in terms of Type I error was measured via Monte Carlo simulation. Because the sampling distribution of this pseudo-median procedure is intractable, the bootstrap method was used to arrive at the significant values.

Methodology

This study addresses both symmetric and asymmetric distribution and the methods applied to the two types of distributions are very different. Let $X_1 = \left(X_{11}, X_{12}, ..., X_{1n_1}\right)$ and $X_2 = \left(X_{21}, X_{22}, ..., X_{2n_2}\right)$ be samples from distributions F_1 and F_2 respectively. The pseudo-median is defined as:

$$\begin{aligned}\hat{d} &= \text{Median}\left(\frac{D_{ij} + D_{i'j'}}{2}\right) \\ &= \text{Median}\left(\frac{\left(X_{1i} - X_{2j}\right) + \left(X_{1i'} - X_{2j'}\right)}{2}\right)\end{aligned} \quad (1)$$

where $i \neq i'$ and $j \neq j'$. When F_1 and F_2 are symmetric, d can be defined as the difference between the centers of symmetry. Hence, the hypothesis is given as:

$$\begin{aligned} &H_0 : d = 0 \\ &\quad versus \\ &H_1 : d \neq 0. \end{aligned} \quad (2)$$

Let $D_{ij} = X_{1i} - X_{2j}$, $i = 1, 2, ..., n_1$, $j = 1, 2, ..., n_2$ and $N = n_1 n_2$. The statistic is a one-sample Wilcoxon statistic based on the ND_{ij}'s. Let R_{ij} denote the rank of $\left|D_{ij}\right|$. The indicator function and the statistic are expressed as:

$$e_{ij} = \begin{cases} 0, & D_{ij} < 0 \\ 0.5, & D_{ij} = 0 \\ 1, & D_{ij} > 0 \end{cases} \quad (3)$$

and

$$W = \sum_{i=}^{n_1} \sum_{j=1}^{n_2} R_{ij} e_{ij}. \quad (4)$$

The modification of the Wilcoxon procedure is performed by adding the pseudo-median value to the second sample to form a new sample, $X_2 + \hat{d}$. The aligned difference, based on the location-aligned samples, becomes:

$$\hat{D}_{ij} = X_{1i} - \left(X_{2j} + \hat{d}\right) = D_{ij} - \hat{d}. \quad (5)$$

Let $\hat{R}_{ij}$ denote the rank of $\left|\hat{D}_{ij}\right|$. The indicator function and the aligned statistic are expressed as:

$$\hat{e}_{ij} = \begin{cases} 0, & \hat{D}_{ij} < 0 \\ 0.5, & \hat{D}_{ij} = 0 \\ 1, & \hat{D}_{ij} > 0 \end{cases} \quad (6)$$

and

$$\hat{W} = \sum_{i=}^{n_1} \sum_{j=1}^{n_2} \hat{R}_{ij} \hat{e}_{ij}. \quad (7)$$

Because the second sample was realigned with the estimate *d*, it is necessary to find the pseudo sampling distribution for the estimate *W*. Use of a bootstrap procedure is proposed in order to construct the hypothesis test. Separately bootstrap n_i observations from X_1 group and n_j observations from $X_2 + \hat{d}$ group to obtain bootstrap samples, X_1^* and X_2^*. The bootstrap difference becomes $D_{ij}^* = X_{1i}^* - X_{2j}^*$ where R_{ij}^* denotes the rank of $\left|D_{ij}^*\right|$. The indicator function and the bootstrap statistic can be defined as:

$$e_{ij}^* = \begin{cases} 0, & D_{ij}^* < 0 \\ 0.5, & D_{ij}^* = 0 \\ 1, & D_{ij}^* > 0 \end{cases} \quad (8)$$

and

$$W^* = \sum_{i=}^{n_1} \sum_{j=1}^{n_2} R_{ij}^* e_{ij}^*. \quad (9)$$

The steps to obtain the *p* value using the bootstrap method for symmetric distribution are as follows:

1. Calculate W from X_1 and X_2.
2. Calculate $\hat{d}$ from X_1 and X_2.
3. Add $\hat{d}$ to X_2 to form a new sample, $X_2 + \hat{d}$.
4. Calculate $\hat{W}$ from X_1 and the new sample in step 3.
5. Generate bootstrap samples by randomly sampling with replacement n_i observations from the X_1 group, and n_j observations from the new sample in step 3 yielding X_1^* and X_2^*.
6. Calculate W^* from the bootstrap samples, X_1^* and X_2^*.
7. Find $\left(W^* - \hat{W}\right)$.
8. Repeat Steps 5 - 7 B times.
9. Compare the value of $\left(W^* - \hat{W}\right)$ with $\left(W - E(W \mid H_0)\right)$.

   Let $U = \left(W^* - \hat{W}\right) > \left(W - E(W \mid H_0)\right)$ and $L = \left(W^* - \hat{W}\right) < \left(W - E(W \mid H_0)\right)$.

10. Calculate the *p* value as $\frac{2}{B} \times \min(\#L, \#U)$.

For asymmetric distributions, the difference between the centers of symmetry between the two groups cannot be assumed to be zero; therefore, to ensure the setting for the null condition, a constant a must be determined and added to the members of the second sample. The value of a is obtained via simulation. For example, let X_1 and X_2 be two skewed distributions where the standard deviations need not be the same. Let $Y_1 = (Y_{11}, Y_{12})$ and $Y_2 = (Y_{21}, Y_{22})$ represent the new generated samples of size two, which have the same distribution with X_1 and X_2, respectively. Compute a_i as follows:

$$a_i = median\left[\frac{(Y_{11} - Y_{21}) + (Y_{12} - Y_{22})}{2}\right] \quad (10)$$

Repeat the process of generating new samples of size two 9,999 times and repeat the computation of a_i to obtain $a_1, a_2, ..., a_{10,000}$. Therefore, the median of $a_1, a_2, ..., a_{10,000}$ is the value of a.

For asymmetric distributions, the steps to obtain the *p* value using a bootstrap method are the same except for one small alteration in step 1. In this step, a constant a is introduced to the members of the second sample (X_2) to form a new sample, $X_2 new$. Steps 2-10 proceed as noted, with the one difference that X_2 has become $X_2 new$.

To study the robustness of this procedure, four variables were manipulated to create conditions known to highlight the strengths and weaknesses of the test for the equality of location parameters. The variables are (1) types of distributions, (2) degree of variance inequality, (3) balanced/unbalanced sample sizes, and (4) pairings of unequal group variance and sample sizes. In this study, empirical Type 1 error rates were collected and later compared under various study conditions.

The number of groups and sample sizes were fixed. This study covered only the two groups case with total sample size of $N = 40$. This value was later divided into two groups forming the balanced and unbalanced design. For the balanced design, the value is equally divided into $n_1 = n_2 = 20$, and for the unbalanced design the groups were divided into $n_1 = 15$ and $n_2 = 25$. To investigate the distribution types, this study focused on (1) heavy tailed symmetric non-normal distribution, and (2) heavy tailed asymmetric distribution.

The normal distribution was used as the basis for comparison. The symmetric non-normal distribution was generated from a *g*-and-*h* distribution (Hoaglin, 1985); specifically, $g = 0$ and $h = 0.225$ with skewness $(\gamma_1) = 0$ and kurtosis $(\gamma_2) = 154.84$ was chosen for investigation. The Chi-square with three degrees of freedom ($\gamma_1 = 1.63$ and $\gamma_2 = 4$) was selected to represent the asymmetric distribution.

The pseudo-random normal variates were generated using the SAS generator RANDGEN function (SAS Institute, 1999); this involved the (RANDGEN(Y, 'NORMAL')) function to generate normal variates with means equals to zero and standard deviation equals to one. To generate data from the *g*-and-*h* distribution, standard unit normal variables (Z_{ij}) were converted to the *h* random variates via

$$Y_{ij} = Z_{ij} \exp\left(\frac{hZ_{ij}^2}{2}\right). \qquad (11)$$

For the Chi-square distribution, data were generated using the (RANDGEN(Y, 'CHISQUARE', 3)) function.

Apart from the types of distribution, two other manipulated variables were the degrees of variance inequality and pairings of variances and group sizes. The nature of pairings of variances and sample sizes affect Type I error rates (Keselman, et al., 1998; Keselman, Othman, Wilcox & Fradette, 2004; Othman, et al., 2004). The variances were manipulated in the following manner: In the case of equal variances, both group variances were set at 1; for the unequal case, the variances were set at 1 and 36.

For positive pairings, the group with the largest number of observations was paired with the group having largest variance, and the group with the smallest number of observations was paired with the group having smallest variance. For the negative pairings, the group with largest number of observations was paired with the group having the smallest variance, and the group with smallest number of observations was paired with the group having the largest variance. This condition was included in the investigation because the direction (positive/negative) of the pairings has been shown to exert some effect on the results. Positive pairings typically produce conservative results and negative pairings tend to produce liberal results (Keselman, Wilcox, Othman & Fradette, 2002; Cribbie & Keselman, 2003; Othman, et al., 2004; Syed Yahaya, Othman & Keselman, 2004, 2006). Therefore, both positive and negative pairings were evaluated.

The operating characteristics of the procedures investigated in this study could be described as extreme because they substantially depart from homogeneity and normality. These conditions were used because it is reasonable to assume that, if a procedure works under the most extreme conditions, it will probably also work under most conditions likely to be encountered by researchers.

The simulation program was written in SAS/IML (SAS Institute, 1999). For each condition examined, 5,000 data sets were generated and within each data set, 599 bootstrap samples were obtained. The level of significance was set at $\alpha = 0.05$.

Results

To evaluate whether the test is robust (insensitive to assumption violations) under each particular condition, the Bradley criterion of robustness (Bradley, 1978) was employed. According to this criterion, for the five percent nominal level used in this study, a test is considered robust if its empirical Type I error rate is within [0.025, 0.075]. Correspondingly, a test is considered to be non-robust if, for a particular condition, its Type I error rate is not contained in this criterion. This criterion was chosen because it provides a reasonable standard for judging robustness. The empirical Type I error rates for the pseudo-median procedure (PM), *t*-test and Mann-Whitney-Wilcoxon (MWW) across all distributions are displayed in Table 1.

With respect to the procedures, results show that all Type I error rates for the pseudo-median procedure are robust under Bradley's liberal criterion and are very close to the nominal level (0.05) regardless of distribution or conditions. The disparity between Type I error rates from balanced and unbalanced designs is minuscule and the rates are consistent across the

investigated conditions. The *t*-test also produces robust Type I error rates for all distributions and conditions, however, for the Chi square distribution, the Type I error rates inflate to a level above 0.065 when the variances are unequal and worsen under negative pairing. For the Mann-Whitney-Wilcoxon test, half of the Type I error rates are above the robustness criterion under unequal variances, especially negative pairing. The Type I error rates for MWW under the Chi-square distribution are too liberal and not robust except under the homogeneous variance condition.

In terms of distributional shapes, the Chi-square distribution produced better empirical Type I error rates compared to the *g*-and-*h* distribution in most conditions for the pseudo-median procedure. Higher values of Type I error rates from Chi-square distribution are apparent for the *t*-test and Mann-Whitney-Wilcoxon.

With respect to variance equality and inequality, results show a contradiction between symmetric and asymmetric distributions for both the pseudo-median and the *t*-test. For the $g = 0$, $h = 0.225$ distributions, homogeneous variances produced greater Type I error rates compared to heterogeneous variances. For the Chi-square distribution, homogeneous variances produced smaller Type I error rates compared to heterogeneous variances. However, no specific pattern could be identified for the Mann-Whitney-Wilcoxon test.

With respect to the pairings of group sizes and variances, results show that the *g*-and-*h* distribution produced liberal (> 0.05) Type I error rates for the pseudo-median procedure and conservative (< 0.05) results for the *t*-test. The Chi-square distribution for the pseudo-median procedure produced conservative Type I error rates for the positive pairing, and liberal results for the negative pairing. The *t*-test produced liberal results for both pairings

Table1: Empirical Type I Error Rates of Pseudo-Medians Procedure, *t*-test and Mann-Whitney- Wilcoxon*

Method	Distribution	Group Sizes			
		(20, 20)		(15, 25)	
		Variance (1:1)	Variance (1:36)	Variance (1:36) +ve pairing	Variance (36:1) -ve pairing
PM	Normal g=0, h=0.225 χ^2_3	0.0552 0.0588 0.0454	0.049 0.0544 0.0504	0.0486 0.0518 0.0476	0.0492 0.0532 0.055
t-test	Normal g=0, h=0.225 χ^2_3	0.054 0.0522 0.052	0.052 0.0458 0.0696	0.0492 0.0448 0.0654	0.0514 0.044 0.0736
MWW	Normal g=0, h=0.225 χ^2_3	0.0516 0.0516 0.052	**0.0912** **0.0854** **0.2428**	0.0458 0.0436 **0.1812**	**0.1142** **0.108** **0.2398**

*Bolded entries indicate Type I error rates of the test exceeding the 0.075 criterion.

Conclusion

The purpose of this study was to investigate how well the pseudo-medians procedure responded to the violations of assumptions compared to the traditional *t*-test and Mann-Whitney-Wilcoxon method. The procedure was tested the heavy-tailed distributions, namely the $g = 0$ and $h = 0.225$ and the Chi-square with three degrees of freedom. Results show that the Type I error rates for the pseudo-median procedure and the *t*-test are robust under Bradley's criterion of robustness and close to the nominal value. The nature of the sample sizes - balanced or unbalanced - did not show much difference in the procedure's ability to control Type I error rates.

The pseudo-median procedure performed better than *t*-test, especially for a skewed distribution with unbalanced design and heterogeneous variances. This procedure also outperforms the popular Mann-Whitney-Wilcoxon method in most conditions. The pseudo-median procedure was observed to have good control of Type I error rates, regardless of distributions under the study conditions. The pseudo-median procedure can thus be recommended as an alternative for testing the differences between two groups, particularly when assumptions of normality and variance homogeneity are not met.

References

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology*, *31*, 321-339.

Cribbie, R. A., & Keselman, H. J. (2003). The effects of nonnormality on parametric, nonparametric, and model comparison approaches to pairwise comparisons. *Educational and Psychological Measurement*, *63*(*4*), 615-635.

Hoaglin, D. C. (1985). Summarizing shape numerically: The g- and h- distributions. In *Exploring data tables, trends, and shapes*, D. C. Hoaglin, F. Mosteller & J. W. Tukey (Eds.), 461-513. New York: NY: Wiley.

Hollander, M., & Wolfe, D. A. (1999). *Nonparametric statistical methods* (2nd Ed.). New York: NY: Wiley.

Hoyland, A. (1965). Robustness of the Hodges-Lehmann estimates for shift. *The Annals of Mathematical Statistics*, *36*, 174-197.

Keselman, H. J., et al. (1998). Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA, and ANCOVA analyses. *Review of Educational Research*, *68*(*3*), 350-386.

Keselman, H. J., et al. (2004). The new and improved two-sample *t*-test. *American Psychological Society*, *15*(*1*), 57-51.

Keselman, H. J., et al. (2002). Trimming, transforming statistics, and bootstrapping: Circumventing the biasing effects of heteroscedasticity and nonnormality. *Journal of Modern Applied Statistical Methods*, *1*(*2*), 288-309.

Othman, A. R., et al. (2004). Comparing measures of the "typical" score across treatment groups. *British Journal of Mathematical and Statistical Psychology*, *57*(*2*), 215-234.

Pratt, J. W. (1964). Robustness of some procedures for two-sample location problem. *Journal of the American Statistical Association*, *59*, 665-680.

SAS Institute, Inc. (1999). *SAS/IML User's Guide version 8*. Cary, NC:SAS Institute.

Syed Yahaya, S. S., Othman, A. R., & Keselman, H. J. (2004). Testing the equality of location parameters for skewed distributions using S_1 with high breakdown robust scale estimators. In *Theory and Applications of Recent Robust Methods, Series: Statistics for Industry and Technology*, M. Hubert, G. Pison, A. Struyf & S. Van Aelst (Eds.), 319-328. Birkhauser: Basel, Switzerland.

Syed Yahaya, S. S., Othman, A. R., & Keselman, H. J. (2006). Comparing the "typical score" across independent groups based on different criteria for trimming. *Methodoloski Zvezki-Advances in Methodology and Statistics*, *3*, 49-62

Welch, B. L. (1938). The significance of the difference between two means when the population variances are unequal. *Biometrika*, *29*, 350-362.

Zimmerman, D. W., & Zumbo, B. D. (1992). Parametric alternatives to the Student *t* test under violation of normality and homogeneity of variance. *Perceptual Motor Skills*, *74*, 835-844.

Journal of Modern Applied Statistical Methods
November 2011, Vol. 10, No. 2, 424-435

1538 – 9472/11/$95.00

Modified Ratio and Product Estimators for Population Mean in Systematic Sampling

Housila P. Singh Rajesh Tailor Narendra Kumar Jatwa
Vikram University,
Ujjain, M. P., India

The estimation of population mean in systematic sampling is explored. Properties of a ratio and product estimator that have been suggested in systematic sampling are investigated, along with the properties of double sampling. Following Swain (1964), the cost aspect is also discussed.

Key words: Population mean, exponential estimator, systematic sampling, bias, mean squared error.

Introduction

Systematic sampling is one of the simplest sampling procedures adopted in practice and is operationally more convenient than simple random sampling. Apart from the simplicity of its concept and execution, systematic sampling is likely to be more precise than simple random sampling and even more precise than stratified sampling under certain specific conditions. In sample surveys it is common to use of auxiliary information to increase the precision of estimates of population parameters. The ratio method of estimation is a good example in this context; the ratio method of estimation is consistent, biased and gives more reliable estimates than those based on simple averages (Cochran, 1963).

If an auxiliary variate x positively (high) correlated with the study variate y is obtained for each unit in the sample and the population mean $\overline{X}$ of the auxiliary variate x is known, the classical ratio estimator for the population mean $\overline{Y}$ of the study variate y is defined by

$$\overline{y}_R = \overline{y}\frac{\overline{X}}{\overline{x}} \tag{1.1}$$

where $\overline{y}$ and $\overline{x}$ are the sample means of the study variate y and the auxiliary variate x respectively, that is, the simple averages of y and x based on the sample.

If the auxiliary variate x is negatively (high) correlated with the study variate then the classical product estimator for population mean $\overline{Y}$ of the study variate y is defined by

$$\overline{y}_P = \overline{y}\frac{\overline{x}}{\overline{X}}, \tag{1.2}$$

which was first developed by Robson (1957) and later rediscovered by Murthy (1964).

Bahl and Tuteja (1991) suggested modified ratio and product estimators for estimating the population mean $\overline{Y}$ respectively as

$$\overline{y}_{\mathrm{Re}} = \overline{y}\exp\left(\frac{\overline{X}-\overline{x}}{\overline{X}+\overline{x}}\right) \tag{1.3}$$

and

$$\overline{y}_{Pe} = \overline{y}\exp\left(\frac{\overline{x}-\overline{X}}{\overline{x}+\overline{X}}\right). \tag{1.4}$$

Housila P. Singh is a Professor in the School of Studies in Statistics. E-mail him at: hpsujn@rediffmail.com. Rajesh Tailor is a Reader in the School of Studies in Statistics. His research interests are in the field of sampling techniques. Email him at: tailorraj@gmail.com. Narenda Kumar Jatwa is a research scholar in the School of Studies in Statistics. Email him at: jatwanarendra@gmail.com.

Under simple random sampling without replacement (SRSWOR), the variances of $\bar{y}_R$, $\bar{y}_P$, $\bar{y}_{\mathrm{Re}}$ and $\bar{y}_{Pe}$ to the first degree of approximation are given, respectively, by

$$\mathrm{Var}\left(\bar{y}_{\mathrm{R}}\right)_{\mathrm{random}} = \left(\frac{1}{\mathrm{n}}-\frac{1}{\mathrm{N}}\right)\left[\mathrm{S}_{\mathrm{y}}^2+\mathrm{R}^2\mathrm{S}_{\mathrm{x}}^2-2\mathrm{R}\rho_{\mathrm{xy}}\mathrm{S}_{\mathrm{x}}\mathrm{S}_{\mathrm{y}}\right] \tag{1.5}$$

$$\mathrm{Var}\left(\bar{y}_{\mathrm{P}}\right)_{\mathrm{random}} = \left(\frac{1}{\mathrm{n}}-\frac{1}{\mathrm{N}}\right)\left[\mathrm{S}_{\mathrm{y}}^2+\mathrm{R}^2\mathrm{S}_{\mathrm{x}}^2+2\mathrm{R}\rho_{\mathrm{xy}}\mathrm{S}_{\mathrm{x}}\mathrm{S}_{\mathrm{y}}\right] \tag{1.6}$$

$$\mathrm{Var}\left(\bar{y}_{\mathrm{Re}}\right)_{\mathrm{random}} = \left(\frac{1}{\mathrm{n}}-\frac{1}{\mathrm{N}}\right)\left[\mathrm{S}_{\mathrm{y}}^2+(1/4)\mathrm{R}^2\mathrm{S}_{\mathrm{x}}^2-\rho_{\mathrm{xy}}\mathrm{S}_{\mathrm{x}}\mathrm{S}_{\mathrm{y}}\right] \tag{1.7}$$

and

$$\mathrm{Var}\left(\bar{y}_{\mathrm{Pe}}\right)_{\mathrm{random}} = \left(\frac{1}{\mathrm{n}}-\frac{1}{\mathrm{N}}\right)\left[\mathrm{S}_{\mathrm{y}}^2+(1/4)\mathrm{R}^2\mathrm{S}_{\mathrm{x}}^2+\rho_{\mathrm{xy}}\mathrm{S}_{\mathrm{x}}\mathrm{S}_{\mathrm{y}}\right] \tag{1.8}$$

where

$$S_y^2=\frac{1}{(N-1)}\sum_{i=1}^{N}\left(y_i-\bar{Y}\right)^2$$

and

$$S_x^2=\frac{1}{(N-1)}\sum_{i=1}^{N}\left(x_i-\bar{X}\right)^2$$

are population mean squares of the study variate y and the auxiliary variate x respectively, ρ_{xy} is the correlation coefficient between x and y and $R=\frac{\bar{Y}}{\bar{X}}$.

Under the SRSWOR sampling scheme

$$Var(\bar{y})=\left(\frac{1}{n}-\frac{1}{N}\right)S_y^2. \tag{1.9}$$

Hasel (1942) and Griffth (1945-46) found systematic sampling to be efficient and convenient in sampling from certain natural populations like forest areas for estimating the volume of timber. In the case of estimating the volume of timber the leaf area or the girth of the tree may be taken as the auxiliary variable (Swain, 1964).

The properties of the ratio estimator $\bar{y}_R$ under systematic sampling have been discussed by Swain (1964) and Shukla (1971) presented the properties of product estimator $\bar{y}_P$. This article discusses the properties of the modified ratio and product estimators $\bar{y}_{\mathrm{Re}}$ and $\bar{y}_{Pe}$ in systematic sampling in the cases of single and double sampling and comparisons are made.

Modified Estimators in Systematic Sampling: Single Sampling

Suppose N units in the population are numbered from 1 to N in some order. To select a sample of n units, if a unit at random is taken from the first k units and every k^{th} subsequent unit, then $N=nk$. This sampling method is similar to that of selecting a cluster at random out of k clusters (each cluster containing n units), made such that i^{th} cluster contains serially numbered units $i,i+k,i+2k,...,i+(n-1)k$. After sampling of n units, observe both the study variate y and auxiliary variate x. Let y_{ij} *and* x_{ij} denote the observations regarding the variate y and variate x respectively on the unit bearing the serial number $i+(j-1)k$ in the population $(i=1,2,...,k;\ j=1,2,...,n)$. If the i^{th} sampling unit is taken at random from the first k units, then $\bar{y}_{sy}$ and $\bar{x}_{sy}$ are defined as:

$$\bar{y}_{sy}=\bar{y}_{i.}=\frac{1}{n}\sum_{j=1}^{n}y_{ij},$$

and

$$\bar{x}_{sy}=\bar{x}_{i.}=\frac{1}{n}\sum_{j=1}^{n}x_{ij}.$$

Suggested Estimator

Assuming the population mean $\overline{X}$ of the auxiliary variate x is known, Swain (1964) suggested the ratio estimator of population mean $\overline{Y}$ of the study variate y based on the systematic samples as

$$\overline{y}_{Rsy} = \overline{y}_{sy}\frac{\overline{X}}{\overline{x}_{sy}} \tag{2.1}$$

and Shukla (1971) proposed the product estimator based on systematic samples as

$$\overline{y}_{Psy} = \overline{y}_{sy}\frac{\overline{x}_{sy}}{\overline{X}}. \tag{2.2}$$

The variances of $\overline{y}_{sy}$ and $\overline{x}_{sy}$ are given approximately by

$$Var\left(\overline{y}_{sy}\right) = \left(\frac{N-1}{N}\right)\frac{S_y^2}{n}\left\{1+\rho_y(n-1)\right\} \tag{2.3}$$

where S_y^2 is the population mean square for the variate y and ρ_y is the intra-class correlation between the units of a cluster corresponding to the y variate and is given by

$$\rho_y = \frac{E\left(y_{ij}-\overline{Y}\right)\left(y_{ij'}-\overline{Y}\right)}{E\left(y_{ij}-\overline{Y}\right)^2}$$

$$= \left[\left\{\frac{1}{kn(n-1)}\sum_{i=1}^{k}\sum_{j\neq j'=1}^{n}\left(y_{ij}-\overline{Y}\right)\left(y_{ij'}-\overline{Y}\right)\right\}\left\{\frac{kn}{(kn-1)S_y^2}\right\}\right] \tag{2.4}$$

and

$$Var\left(\overline{x}_{sy}\right) = \left(\frac{N-1}{N}\right)\frac{S_x^2}{n}\left\{1+\rho_x(n-1)\right\} \tag{2.5}$$

where S_x^2 and ρ_x bear the same meanings as for the study variate $y's$.

For large N, the variances of $\overline{y}_{Rsy}$ and $\overline{y}_{Psy}$ to the first degree of approximation are respectively given by

$$\begin{aligned}\mathrm{Var}\left(\overline{y}_{Rsy}\right) &= \frac{1}{n}\left[S_y^2 + R^2S_x^2 - 2R\rho_{xy}S_xS_y\right] \\ &\quad + \frac{1}{n}\left[\rho_y(n-1)S_y^2 + R^2\rho_x(n-1)S_x^2\right. \\ &\quad \left. -2R\rho_{xy}S_xS_y\left\{\sqrt{\left\{1+\rho_y(n-1)\right\}\left\{1+\rho_x(n-1)\right\}}-1\right\}\right],\end{aligned} \tag{2.6}$$

and

$$\begin{aligned}\mathrm{Var}\left(\overline{y}_{Psy}\right) &= \frac{1}{n}\left[S_y^2 + R^2S_x^2 + 2R\rho_{xy}S_xS_y\right] \\ &\quad + \frac{1}{n}\left[\rho_y(n-1)S_y^2 + R^2\rho_x(n-1)S_x^2\right. \\ &\quad \left. +2R\rho_{xy}S_xS_y\left\{\sqrt{\left\{1+\rho_y(n-1)\right\}\left\{1+\rho_x(n-1)\right\}}-1\right\}\right],\end{aligned} \tag{2.7}$$

Assuming the intraclass correlation to be the same for both the variates y and x, for example, $\rho_y = \rho_x = \rho$, then the $Var\left(\overline{y}_{Rsy}\right)$ and $Var\left(\overline{y}_{Psy}\right)$ respectively reduce to

$$\begin{aligned}\mathrm{Var}\left(\overline{y}_{Rsy}\right) &= \frac{1}{n}\left[S_y^2 + R^2S_x^2 - 2R\rho_{xy}S_xS_y\right] \\ &\quad + \frac{\rho(n-1)}{n}\left[S_y^2 + R^2S_x^2 - 2R\rho_{xy}S_xS_y\right] \\ &= \mathrm{Var}\left(\overline{y}_R\right)_{random}\left\{1+\rho(n-1)\right\}\end{aligned} \tag{2.8}$$

and

$$\begin{aligned}\mathrm{Var}\left(\overline{y}_{Psy}\right) &= \frac{1}{n}\left[S_y^2 + R^2S_x^2 + 2R\rho_{xy}S_xS_y\right] \\ &\quad + \frac{\rho(n-1)}{n}\left[S_y^2 + R^2S_x^2 + 2R\rho_{xy}S_xS_y\right] \\ &= \mathrm{Var}\left(\overline{y}_P\right)_{random}\left\{1+\rho(n-1)\right\}\end{aligned} \tag{2.9}$$

Following Bahl and Tuteja (1991), the following modified ratio and product estimators for population mean $\overline{Y}$ are defined respectively as

$$\bar{y}_{\mathrm{Re}sy} = \bar{y}_{sy}\exp\left(\frac{\bar{X}-\bar{x}_{sy}}{\bar{X}+\bar{x}_{sy}}\right) \qquad (2.10)$$

and

$$\bar{y}_{\mathrm{Pesy}} = \bar{y}_{sy}\exp\left(\frac{\bar{x}_{sy}-\bar{X}}{\bar{x}_{sy}+\bar{X}}\right). \qquad (2.11)$$

To obtain the biases and variances of the estimators $\bar{y}_{\mathrm{Re}sy}$ and $\bar{y}_{Pesy}$, $\bar{y}_{sy} = \bar{Y}(1+e_0)$, $\bar{x}_{sy} = \bar{X}(1+e_1)$, is written such that $E(e_0) = E(e_1) = 0$ and

$$\left.\begin{aligned}
&\mathrm{E}\left(\mathrm{e}_0^2\right) = \\
&\quad = \frac{\mathrm{Var}\left(\bar{\mathrm{y}}_{sy}\right)}{\bar{\mathrm{Y}}^2} \\
&\quad = \left(\frac{\mathrm{N}-1}{\mathrm{N}}\right)\left(\frac{\mathrm{C}_y^2}{\mathrm{n}}\right)\left\{1+(\mathrm{n}-1)\rho_y\right\}, \\
&\mathrm{E}\left(\mathrm{e}_1^2\right) = \\
&\quad = \frac{\mathrm{Var}\left(\bar{\mathrm{x}}_{sy}\right)}{\bar{\mathrm{X}}^2} \\
&\quad = \left(\frac{\mathrm{N}-1}{\mathrm{N}}\right)\left(\frac{\mathrm{C}_x^2}{\mathrm{n}}\right)\left\{1+(\mathrm{n}-1)\rho_x\right\}, \\
&\mathrm{E}\left(\mathrm{e}_0\mathrm{e}_1\right) = \\
&\quad = \frac{\mathrm{Cov}\left(\bar{\mathrm{x}}_{sy},\bar{\mathrm{y}}_{sy}\right)}{\bar{\mathrm{X}}\bar{\mathrm{Y}}} \\
&\quad = \left(\frac{\mathrm{N}-1}{\mathrm{N}}\right)\left(\frac{\rho_{xy}\mathrm{C}_x\mathrm{C}_y}{\mathrm{n}}\right)\sqrt{\left\{1+(\mathrm{n}-1)\rho_y\right\}\left\{1+(\mathrm{n}-1)\rho_x\right\}}
\end{aligned}\right\} \qquad (2.12)$$

where $C_y = S_y/\bar{Y}$ and $C_x = S_x/\bar{X}$ are the population coefficients of variation of y and x respectively.

$$\begin{aligned}
\bar{\mathrm{y}}_{\mathrm{Re}sy} &= \bar{\mathrm{Y}}(1+\mathrm{e}_0)\exp\left\{-\frac{\mathrm{e}_1}{(2+\mathrm{e}_1)}\right\} \\
&= \bar{\mathrm{Y}}(1+\mathrm{e}_0)\exp\left\{-\frac{\mathrm{e}_1}{2}\left(1+\frac{\mathrm{e}_1}{2}\right)^{-1}\right\} \\
&= \bar{\mathrm{Y}}(1+\mathrm{e}_0)\left[1-\frac{\mathrm{e}_1}{2}\left(1+\frac{\mathrm{e}_1}{2}\right)^{-1}+\frac{\mathrm{e}_1^2}{8}\left(1+\frac{\mathrm{e}_1}{2}\right)^{-2}-\ldots\right] \\
&= \bar{\mathrm{Y}}(1+\mathrm{e}_0)\left[\begin{aligned}&1-\frac{\mathrm{e}_1}{2}\left(1-\frac{\mathrm{e}_1}{2}+\frac{\mathrm{e}_1^2}{8}-\ldots\right)\\&+\frac{\mathrm{e}_1^2}{8}\left(1-\mathrm{e}_1+\frac{3}{8}\mathrm{e}_1^2-\ldots\right)-\ldots\end{aligned}\right] \\
&= \bar{\mathrm{Y}}(1+\mathrm{e}_0)\left[1-\frac{\mathrm{e}_1}{2}+\frac{3}{8}\mathrm{e}_1^2-\ldots\right] \\
&= \bar{\mathrm{Y}}\left[1+\mathrm{e}_0-\frac{\mathrm{e}_1}{2}-\frac{\mathrm{e}_0\mathrm{e}_1}{2}+\frac{3}{8}\mathrm{e}_1^2-\ldots\right]
\end{aligned}$$

or

$$\left(\bar{y}_{\mathrm{Re}sy}-\bar{Y}\right) \cong \bar{Y}\left[e_0-\frac{e_1}{2}+\frac{3}{8}e_1^2-\frac{e_0e_1}{2}\right]. \qquad (2.13)$$

Taking the expectations of both sides in (2.13) and using the results given by (2.12) the bias of the ratio estimator $\bar{y}_{\mathrm{Re}sy}$ to the first degree of approximation is obtained as

$$\begin{aligned}
&\mathrm{B}\left(\bar{\mathrm{y}}_{\mathrm{Re}sy}\right) = \\
&= \left(\frac{\mathrm{N}-1}{\mathrm{nN}}\right)\bar{\mathrm{Y}}\left[\begin{aligned}&\frac{3}{8}\mathrm{C}_x^2\left\{1+\rho_x(\mathrm{n}-1)\right\}\\&-\frac{1}{2}\rho_{xy}\mathrm{C}_x\mathrm{C}_y\sqrt{\left\{1+\rho_y(\mathrm{n}-1)\right\}\left\{1+\rho_x(\mathrm{n}-1)\right\}}\end{aligned}\right] \\
&= \left(\frac{\mathrm{N}-1}{\mathrm{nN}}\right)\bar{\mathrm{Y}}\left(\frac{\mathrm{C}_x^2}{8}\right)\left[\begin{aligned}&3\left\{1+\rho_x(\mathrm{n}-1)\right\}\\&-4\mathrm{c}\sqrt{\left\{1+\rho_y(\mathrm{n}-1)\right\}\left\{1+\rho_x(\mathrm{n}-1)\right\}}\end{aligned}\right] \\
&= \left(\frac{\mathrm{N}-1}{\mathrm{nN}}\right)\bar{\mathrm{Y}}\left(\frac{\mathrm{C}_x^2}{8}\right)\left\{1+\rho_x(\mathrm{n}-1)\right\}\left[3-4\mathrm{c}\sqrt{\frac{\left\{1+\rho_y(\mathrm{n}-1)\right\}}{\left\{1+\rho_x(\mathrm{n}-1)\right\}}}\right]
\end{aligned} \qquad (2.14)$$

where $c = \rho_{xy} C_y / C_x$.

Squaring both sides of (2.13) and neglecting terms of $e's$ having power greater than two results in

$$\left(\bar{y}_{\mathrm{Re}\,sy} - \bar{Y}\right)^2 = \bar{Y}^2 \left(e_0^2 + \frac{e_1^2}{4} - e_0 e_1 \right) \tag{2.15}$$

Taking the expectations of both sides in (2.15) and using the results given by (2.12) provides the variance of the modified ratio estimator $\bar{y}_{\mathrm{Re}\,sy}$ as

$$\mathrm{Var}\left(\bar{y}_{\mathrm{Re}\,sy}\right) = \left(\frac{N-1}{nN}\right)\bar{Y}^2 \begin{bmatrix} \left\{1+\rho_y(n-1)\right\}C_y^2 \\ +\left\{1+\rho_x(n-1)\right\}\frac{C_x^2}{4} \\ -\rho_{xy}C_xC_y\sqrt{\left\{1+\rho_y(n-1)\right\}\left\{1+\rho_x(n-1)\right\}} \end{bmatrix}$$

For large N , the above expression reduces to

$$\mathrm{Var}\left(\bar{y}_{\mathrm{Re}\,sy}\right) = \frac{S_y^2}{n} \begin{bmatrix} \left\{1+\rho_y(n-1)\right\}S_y^2 \\ +\left\{1+\rho_x(n-1)\right\}R^2\frac{S_x^2}{4} \\ -R\rho_{xy}S_xS_y\sqrt{\left\{1+\rho_y(n-1)\right\}\left\{1+\rho_x(n-1)\right\}} \end{bmatrix} \tag{2.16}$$

and in the case where $\rho_y = \rho_x = \rho$, it reduces to

$$\mathrm{Var}\left(\bar{y}_{\mathrm{Re}\,sy}\right) = \frac{1}{n}\left[S_y^2 + \frac{1}{4}R^2S_x^2 - R\rho_{xy}S_xS_y \right]\left\{1+\rho(n-1)\right\}. \tag{2.17}$$

From (1.7) and (2.17):

$$Var\left(\bar{y}_{\mathrm{Re}\,sy}\right) = Var\left(\bar{y}_{\mathrm{Re}}\right)_{random}\left\{1+\rho(n-1)\right\}. \tag{2.18}$$

The efficiency of the modified ratio method of estimation using systematic samples with respect to modified ratio method of estimation using sample random sampling is

$$\frac{Var\left(\bar{y}_{\mathrm{Re}}\right)_{random}}{Var\left(\bar{y}_{\mathrm{Re}\,sy}\right)} = \frac{1}{\left\{1+\rho(n-1)\right\}}. \tag{2.19}$$

As expected, the ratio method of estimation with systematic samples will be more efficient if $\rho < 0$. The minimum value that ρ can take is $-\left(\frac{1}{n-1}\right)$, when the reduction in variance is 100%.

Further expressing (2.11) in terms of $e's$:

$$\begin{aligned} \bar{y}_{\mathrm{Pesy}} &= \bar{Y}(1+e_0)\exp\left\{\frac{e_1}{2}\left(1+\frac{e_1}{2}\right)^{-1}\right\} \\ &= \bar{Y}(1+e_0)\left[1+\frac{e_1}{2}\left(1+\frac{e_1}{2}\right)^{-1}+\frac{e_1^2}{8}\left(1+\frac{e_1}{2}\right)^{-2}+\ldots\right] \\ &= \bar{Y}(1+e_0)\left[1+\frac{e_1}{2}\left(1-\frac{e_1}{2}+\ldots\right)+\frac{e_1^2}{8}(1-e_1+\ldots)-\ldots\right] \\ &= \bar{Y}(1+e_0)\left[1+\frac{e_1}{2}-\frac{1}{8}e_1^2-\ldots\right] \\ &= \bar{Y}\left[1+e_0+\frac{e_1}{2}+\frac{e_0e_1}{2}-\frac{1}{8}e_1^2+\ldots\right] \end{aligned}$$

or

$$\left(\bar{y}_{Pesy} - \bar{Y}\right) \cong \bar{Y}\left[e_0 + \frac{e_1}{2} + \frac{e_0e_1}{2} - \frac{1}{8}e_1^2 \right]. \tag{2.20}$$

Taking the expectations of both sides of (2.20) and using the results given by (2.12) provides the bias of the product estimator $\bar{y}_{Pesy}$ to the first degree of approximation as

$$B\left(\bar{y}_{Pesy}\right) =$$

$$= \left(\frac{N-1}{nN}\right)\frac{\bar{Y}}{8}\begin{bmatrix} 4\rho_{xy}C_xC_y\sqrt{\left\{1+\rho_y(n-1)\right\}\left\{1+\rho_x(n-1)\right\}} \\ -C_x^2\left\{1+\rho_x(n-1)\right\} \end{bmatrix}$$

$$= \left(\frac{N-1}{nN}\right)\left(\frac{\bar{Y}}{8}\right)C_x^2\left\{1+\rho_x(n-1)\right\}\left[4c\sqrt{\frac{\left\{1+\rho_y(n-1)\right\}}{\left\{1+\rho_x(n-1)\right\}}}-1\right] \tag{2.21}$$

Squaring both sides of (2.19) and neglecting terms of $e's$ having power greater than two results in:

$$\left(\bar{y}_{Pesy}-\bar{Y}\right)^2 = \bar{Y}^2\left(e_0^2+\frac{e_1^2}{4}+e_0e_1\right). \tag{2.22}$$

Taking the expectations of both sides in (2.22) and using the results given by (2.12) provides the variance of the modified product estimator $\bar{y}_{Pesy}$ as:

$$\mathrm{Var}\left(\bar{y}_{Pesy}\right) = \left(\frac{N-1}{nN}\right)\bar{Y}^2\begin{bmatrix} \left\{1+\rho_y(n-1)\right\}C_y^2 \\ +\left\{1+\rho_x(n-1)\right\}\frac{C_x^2}{4} \\ +\rho_{xy}C_xC_y\sqrt{\left\{1+\rho_y(n-1)\right\}\left\{1+\rho_x(n-1)\right\}} \end{bmatrix}.$$

For large N, this expression reduces to

$$\mathrm{Var}\left(\bar{y}_{Pesy}\right) = \frac{S_y^2}{n}\begin{bmatrix} \left\{1+\rho_y(n-1)\right\}S_y^2 \\ +\left\{1+\rho_x(n-1)\right\}R^2\frac{S_x^2}{4} \\ +R\rho_{xy}S_xS_y\sqrt{\left\{1+\rho_y(n-1)\right\}\left\{1+\rho_x(n-1)\right\}} \end{bmatrix}. \tag{2.23}$$

In the casewhere $\rho_y = \rho_x = \rho$, the expression (2.23) reduces to

$$\mathrm{Var}\left(\bar{y}_{Pesy}\right) = \frac{1}{n}\bar{Y}^2[C_y^2+\left[\frac{1}{4}\right]C_x^2+\rho_{xy}]\left\{1+\rho(n-1)\right\}S$$

$$\mathrm{Var}\left(\bar{y}_{Pesy}\right) = \frac{1}{n}\left[S_y^2+\frac{1}{4}R^2S_x^2+R\rho_{xy}S_xS_y\right]\left\{1+\rho(n-1)\right\} \tag{2.24}$$

From (1.8) and (2.24):

$$Var\left(\bar{y}_{Pesy}\right) = Var\left(\bar{y}_{Pe}\right)_{random}\left\{1+\rho(n-1)\right\}. \tag{2.25}$$

The efficiency of the modified product method of estimation using systematic samples with respect to modified product method of estimation using sample random sampling is

$$\frac{Var\left(\bar{y}_{Pe}\right)_{random}}{Var\left(\bar{y}_{Pesy}\right)} = \frac{1}{\left\{1+\rho(n-1)\right\}} \tag{2.26}$$

which is greater than unity if:

$$\begin{aligned} &\mathrm{Var}\left(\bar{y}_{Pe}\right)_{random} > \mathrm{Var}\left(\bar{y}_{Pesy}\right), \\ &\frac{1}{1+\rho(n-1)} > 1, \\ &\text{i.e., if } 1 > 1+\rho(n-1), \\ &\text{i.e., if } 0 > \rho(n-1), \\ &\text{i.e., if } \rho < 0. \end{aligned} \tag{2.27}$$

Thus, the modified product method of estimation using systematic samples will be more efficient than the modified product method of estimation with simple random samples if $\rho < 0$. The minimum value that ρ can take is $-\left(\frac{1}{n-1}\right)$ and, in this case, $Var\left(\bar{y}_{Pesy}\right) = 0$, that is, the reduction in variance of $\bar{y}_{Pesy}$ is 100%.

Comparison of Modified Ratio $\bar{y}_{\mathrm{Re}sy}$ and Product $\bar{y}_{Pesy}$ Estimators with Usual Unbiased Estimator $\bar{y}_{sy}$, Ratio Estimator $\bar{y}_{Rsy}$ and Product Estimator $\bar{y}_{Psy}$

For large N, the variance of $\bar{y}_{sy}$ in (2.3) reduces to

$$Var(\bar{y}_{sy}) = \frac{S_y^2}{n}\{1+\rho_y(n-1)\}. \quad (2.28)$$

From (2.16) and (2.28)

$$\begin{aligned}
&\mathrm{Var}(\bar{y}_{\mathrm{Re}sy}) - \mathrm{Var}(\bar{y}_{sy}) \\
&= \frac{1}{4n}\left[\begin{array}{l}\{1+\rho_x(n-1)\}R^2S_x^2 \\ -4R\rho_{xy}S_xS_y\sqrt{\{1+\rho_y(n-1)\}\{1+\rho_x(n-1)\}}\end{array}\right] \\
&= \frac{R^2S_x^2\{1+\rho_x(n-1)\}}{4n}\left[1-4\rho_{xy}\frac{S_y}{S_xR}\sqrt{\frac{\{1+\rho_y(n-1)\}}{\{1+\rho_x(n-1)\}}}\right] \\
&= \frac{R^2S_x^2\{1+\rho_x(n-1)\}}{4n}\left[1-4\frac{\beta}{R}\sqrt{\frac{\{1+\rho_y(n-1)\}}{\{1+\rho_x(n-1)\}}}\right] \\
&= \frac{R^2S_x^2\{1+\rho_x(n-1)\}}{4n}\left[1-4c\sqrt{\frac{\{1+\rho_y(n-1)\}}{\{1+\rho_x(n-1)\}}}\right]
\end{aligned}$$

which is negative if:

$$1-4c\sqrt{\frac{\{1+\rho_y(n-1)\}}{\{1+\rho_x(n-1)\}}} < 0,$$

$$\text{i.e., if } \frac{1}{4} < c\sqrt{\frac{\{1+\rho_y(n-1)\}}{\{1+\rho_x(n-1)\}}},$$

$$\text{i.e., if } c\sqrt{\frac{\{1+\rho_y(n-1)\}}{\{1+\rho_x(n-1)\}}} > \frac{1}{4},$$

$$\text{i.e., if } c > \frac{1}{4}\sqrt{\frac{\{1+\rho_x(n-1)\}}{\{1+\rho_y(n-1)\}}}, \quad (2.29)$$

where $\beta = \rho_{xy}\frac{S_y}{S_x}$ is the population regression coefficient of y on x and $c = \frac{\beta}{R}$.

From (2.6) and (2.16)

$$\begin{aligned}
&\mathrm{Var}(\bar{y}_{\mathrm{Re}sy}) - \mathrm{Var}(\bar{y}_{Rsy}) \\
&= \frac{S_y^2}{n}\left[\begin{array}{l}-\frac{3}{4}R^2S_x^2\{1+\rho_x(n-1)\} \\ +R\rho_{xy}S_xS_y\sqrt{\{1+\rho_y(n-1)\}\{1+\rho_x(n-1)\}}\end{array}\right] \\
&= \left(\frac{R^2S_x^2S_y^2}{n}\right)\{1+\rho_x(n-1)\}\left[\frac{\beta}{R}\sqrt{\frac{\{1+\rho_y(n-1)\}}{\{1+\rho_x(n-1)\}}}-\frac{3}{4}\right] \\
&= \left(\frac{R^2S_x^2S_y^2}{n}\right)\{1+\rho_x(n-1)\}\left[c\sqrt{\frac{\{1+\rho_y(n-1)\}}{\{1+\rho_x(n-1)\}}}-\frac{3}{4}\right]
\end{aligned}$$

which is negative if:

$$c\sqrt{\frac{\{1+\rho_y(n-1)\}}{\{1+\rho_x(n-1)\}}} - \frac{3}{4} < 0$$

$$\text{i.e., if } c < \frac{3}{4}\sqrt{\frac{\{1+\rho_x(n-1)\}}{\{1+\rho_y(n-1)\}}}. \quad (2.30)$$

Thus, from (2.29) and (2.30) it follows that the modified ratio estimator $\bar{y}_{\mathrm{Re}sy}$ is more efficient than usual unbiased estimator $\bar{y}_{sy}$ and Swain's (1964) estimator $\bar{y}_{Rsy}$ if:

$$\frac{1}{4}\sqrt{\frac{\{1+\rho_x(n-1)\}}{\{1+\rho_y(n-1)\}}} < c < \frac{3}{4}\sqrt{\frac{\{1+\rho_x(n-1)\}}{\{1+\rho_y(n-1)\}}}$$

i.e., if

$$\frac{1}{4}\frac{C_x}{C_y}\sqrt{\frac{\{1+\rho_x(n-1)\}}{\{1+\rho_y(n-1)\}}} < \rho_{xy} < \frac{3}{4}\frac{C_x}{C_y}\sqrt{\frac{\{1+\rho_x(n-1)\}}{\{1+\rho_y(n-1)\}}} \quad (2.31)$$

when the intraclass correlation coefficients for both the variates are same (i.e. $\rho_y = \rho_x = \rho$), then condition (2.31) reduces to:

$$\frac{1}{4}\frac{C_x}{C_y} < \rho_{xy} < \frac{3}{4}\frac{C_x}{C_y}. \qquad (2.32)$$

From (2.7), (2.23) and (2.28)

$$\mathrm{Var}\left(\bar{y}_{\mathrm{Pesy}}\right) - \mathrm{Var}\left(\bar{y}_{\mathrm{sy}}\right) = \left(\frac{R^2 S_x^2 S_y^2}{n}\right)\left\{1+\rho_x(n-1)\right\}\left[\frac{1}{4} + c\sqrt{\frac{\left\{1+\rho_y(n-1)\right\}}{\left\{1+\rho_x(n-1)\right\}}}\right] \qquad (2.33)$$

and

$$\mathrm{Var}\left(\bar{y}_{\mathrm{Pesy}}\right) - \mathrm{Var}\left(\bar{y}_{\mathrm{Psy}}\right) = \left(\frac{R^2 S_x^2 S_y^2}{n}\right)\left\{1+\rho_x(n-1)\right\}\left[-\frac{3}{4} - c\sqrt{\frac{\left\{1+\rho_y(n-1)\right\}}{\left\{1+\rho_x(n-1)\right\}}}\right] \qquad (2.34)$$

It follows from (2.33) and (2.34) respectively that the proposed modified product estimator $\bar{y}_{Pesy}$ is more efficient than

(i) usual unbiased estimator $\bar{y}_{sy}$ if

$$Var\left(\bar{y}_{Pesy}\right) < Var\left(\bar{y}_{sy}\right),$$

$$\text{i.e., if } \frac{1}{4} + c\sqrt{\frac{\left\{1+\rho_y(n-1)\right\}}{\left\{1+\rho_x(n-1)\right\}}} < 0,$$

$$\text{i.e., if } c < -\frac{1}{4}\sqrt{\frac{\left\{1+\rho_x(n-1)\right\}}{\left\{1+\rho_y(n-1)\right\}}}. \qquad (2.35)$$

(ii) Shukla's (1971) product estimator $\bar{y}_{Psy}$ if

$$Var\left(\bar{y}_{Pesy}\right) < Var\left(\bar{y}_{Psy}\right),$$

$$\text{i.e., if } -\frac{3}{4} - c\sqrt{\frac{\left\{1+\rho_y(n-1)\right\}}{\left\{1+\rho_x(n-1)\right\}}} < 0,$$

$$\text{i.e., if } -\frac{3}{4} < c\sqrt{\frac{\left\{1+\rho_y(n-1)\right\}}{\left\{1+\rho_x(n-1)\right\}}},$$

$$\text{i.e., if } c > -\frac{3}{4}\sqrt{\frac{\left\{1+\rho_x(n-1)\right\}}{\left\{1+\rho_y(n-1)\right\}}}. \qquad (2.36)$$

Thus, from (2.35) and (2.36) it follows that the proposed modified product estimator $\bar{y}_{Pesy}$ is more efficient than usual unbiased estimator $\bar{y}_{sy}$ and Shukla's (1971) estimator $\bar{y}_{Psy}$ if

$$-\frac{3}{4}\sqrt{\frac{\left\{1+\rho_x(n-1)\right\}}{\left\{1+\rho_y(n-1)\right\}}} < c < -\frac{1}{4}\sqrt{\frac{\left\{1+\rho_x(n-1)\right\}}{\left\{1+\rho_y(n-1)\right\}}} \qquad (2.37)$$

i.e., if

$$-\frac{3}{4}\frac{C_x}{C_y}\sqrt{\frac{\left\{1+\rho_x(n-1)\right\}}{\left\{1+\rho_y(n-1)\right\}}} < \rho_{xy} < -\frac{1}{4}\frac{C_x}{C_y}\sqrt{\frac{\left\{1+\rho_x(n-1)\right\}}{\left\{1+\rho_y(n-1)\right\}}} \qquad (2.38)$$

In the case where $\rho_y = \rho_x = \rho$, the condition (2.38) reduces to:

$$-\frac{3}{4}\frac{C_x}{C_y} < \rho_{xy} < -\frac{1}{4}\frac{C_x}{C_y}. \qquad (2.39)$$

Modified Estimators in Systematic Sampling: Two-Phase (or Double) Sampling

If the population mean $\bar{X}$ of the auxiliary variable x is not known before start of the survey, then it may be more efficient to conduct the sampling in two-phase (or double) sampling by taking a large preliminary sample to estimate the population mean $\bar{X}$. This method is a powerful and cost effective (economical) procedure and, therefore, has role to play in survey sampling (Hidiroglou & Sarndal, 1998; Hidiroglou, 2001).

In the present situation the population is divided into k clusters of n units each according to the previous rule and λ clusters (λ being less than k) and selected to observe only the auxiliary variate x, while another cluster is selected to observe both y and x variates (Swain, 1964). If $\bar{x}$ is the mean of the $x's$ from the selected λ clusters, then

$$\bar{x}' = \frac{1}{\lambda n}\sum_{i=1}^{\lambda}\sum_{j=1}^{n} x_{ij}, \qquad (3.1)$$

such that $E(\bar{x}') = \bar{X}$, that is, $\bar{x}'$ is an unbiased estimator of the population mean $\bar{X}$.Swain (1964) suggested the double sampling ratio estimator with systematic samples as

$$\bar{y}_{Rsy}^{(d)} = \bar{y}_{sy}\left(\frac{\bar{x}'}{\bar{x}_{sy}}\right). \qquad (3.2)$$

The double sampling version of product estimator $\bar{y}_{Psy}$ in (2.29) is defined by

$$\bar{y}_{Psy}^{(d)} = \bar{y}_{sy}\left(\frac{\bar{x}_{sy}}{\bar{x}'}\right). \qquad (3.3)$$

Case I

For large N, $\rho_y = \rho_x = \rho$ and, if the first set of λ clusters and the second cluster are chosen randomly and independently, the variances of the double sampling ratio $\left(\bar{y}_{Rsy}^{(d)}\right)$ and product $\left(\bar{y}_{Psy}^{(d)}\right)$ based on systematic samples to the first degree of approximation are respectively given by

$$\mathrm{Var}\left(\bar{y}_{\mathrm{Rsy}}^{(\mathrm{d})}\right) = \frac{1}{\mathrm{n}}\begin{bmatrix}\mathrm{S_y^2 + R^2S_x^2}\\ -2\mathrm{R}\rho_{\mathrm{xy}}\mathrm{S_yS_x}\end{bmatrix}\left\{1+\rho(\mathrm{n}-1)\right\} + \frac{\mathrm{R^2S_x^2}}{\lambda \mathrm{n}}\left\{1+\rho(\mathrm{n}-1)\right\} \qquad (3.4)$$

and

$$\mathrm{Var}\left(\bar{y}_{\mathrm{Psy}}^{(\mathrm{d})}\right) = \frac{1}{\mathrm{n}}\begin{bmatrix}\mathrm{S_y^2 + R^2S_x^2}\\ +2\mathrm{R}\rho_{\mathrm{xy}}\mathrm{S_yS_x}\end{bmatrix}\left\{1+\rho(\mathrm{n}-1)\right\} + \frac{\mathrm{R^2S_x^2}}{\lambda \mathrm{n}}\left\{1+\rho(\mathrm{n}-1)\right\}. \qquad (3.5)$$

Case II

For large N, $\rho_y = \rho_x = \rho$ and if the second cluster is chosen randomly from the first set of selected clusters, the variances of the double sampling ratio and product estimators based on systematic sampling are respectively given by

$$\mathrm{Var}\left(\bar{y}_{\mathrm{Rsy}}^{(\mathrm{d})}\right) = \frac{1}{\mathrm{n}}\begin{bmatrix}\mathrm{S_y^2 + R^2S_x^2}\\ -2\mathrm{R}\rho_{\mathrm{xy}}\mathrm{S_yS_x}\end{bmatrix}\left\{1+\rho(\mathrm{n}-1)\right\} + \frac{1}{\lambda \mathrm{n}}\left[\begin{Bmatrix}2\mathrm{R}\rho_{\mathrm{xy}}\mathrm{S_yS_x}\\ -\mathrm{R^2S_x^2}\end{Bmatrix}\begin{Bmatrix}1+\\ \rho(\mathrm{n}+1)\end{Bmatrix}\right] \qquad (3.6)$$

and

$$\mathrm{Var}\left(\bar{y}_{\mathrm{Psy}}^{(\mathrm{d})}\right) = \frac{1}{\mathrm{n}}\begin{bmatrix}\mathrm{S_y^2 + R^2S_x^2}\\ +2\mathrm{R}\rho_{\mathrm{xy}}\mathrm{S_yS_x}\end{bmatrix}\left\{1+\rho(\mathrm{n}-1)\right\} + \frac{1}{\lambda \mathrm{n}}\left[\begin{Bmatrix}-2\mathrm{R}\rho_{\mathrm{xy}}\mathrm{S_yS_x}\\ -\mathrm{R^2S_x^2}\end{Bmatrix}\begin{Bmatrix}1\\ +\rho(\mathrm{n}-1)\end{Bmatrix}\right] \qquad (3.7)$$

Following Bahl and Tuteja (1991) and Singh and Vishwakarma (2007) a modified double sampling ratio estimator based on systematic sampling is proposed a

$$\bar{y}_{\mathrm{Re}\,sy}^{(d)} = \bar{y}_{sy}\exp\left(\frac{\bar{x}'_{sy} - \bar{x}_{sy}}{\bar{x}'_{sy} + \bar{x}_{sy}}\right), \qquad (3.8)$$

and the modified double sampling product estimator based on systematic sampling

$$\bar{y}_{Pesy}^{(d)} = \bar{y}_{sy} \exp\left(\frac{\bar{x}_{sy} - \bar{x}'_{sy}}{\bar{x}_{sy} + \bar{x}'_{sy}}\right). \qquad (3.9)$$

Case I

For large N, $\rho_y = \rho_x = \rho$ and if the first set of λ clusters and the second cluster is chosen randomly and independently, the variances of the modified double sampling ratio and product estimators based on systematic samples to the first degree of approximation are respectively given by

$$\mathrm{Var}\left(\bar{y}_{\mathrm{Re}sy}^{(d)}\right) = \frac{1}{n}\begin{bmatrix} S_y^2 + (1/4)R^2S_x^2 \\ -R\rho_{xy}S_yS_x \end{bmatrix}\{1+\rho(n-1)\} + \frac{R^2S_x^2}{4\lambda n}\{1+\rho(n-1)\} \qquad (3.10)$$

and

$$\mathrm{Var}\left(\bar{y}_{\mathrm{Pesy}}^{(d)}\right) = \frac{1}{n}\begin{bmatrix} S_y^2 + (1/4)R^2S_x^2 \\ +R\rho_{xy}S_yS_x \end{bmatrix}\{1+\rho(n-1)\} + \frac{R^2S_x^2}{4\lambda n}\{1+\rho(n-1)\} \qquad (3.11)$$

Case II

If the second cluster is selected randomly from the first set of selected clusters, then the variances of the double sampling ratio $\bar{y}_{\mathrm{Re}sy}^{(d)}$ and product $\bar{y}_{Pesy}^{(d)}$ estimators to the first degree of approximation are respectively given by

$$\mathrm{Var}\left(\bar{y}_{\mathrm{Re}sy}^{(d)}\right) = \frac{1}{n}\left[S_y^2 + (1/4)R^2S_x^2 - R\rho_{xy}S_yS_x\right]\{1+\rho(n-1)\} + \frac{1}{\lambda n}\left[\{R\rho_{xy}S_yS_x - (1/4)R^2S_x^2\}\{1+\rho(n+1)\}\right] \qquad (3.12)$$

and

$$\mathrm{Var}\left(\bar{y}_{\mathrm{Pesy}}^{(d)}\right) = \frac{1}{n}\left[S_y^2 + (1/4)R^2S_x^2 - R\rho_{xy}S_yS_x\right]\{1+\rho(n-1)\} + \frac{1}{\lambda n}\left[\{-R\rho_{xy}S_yS_x - (1/4)R^2S_x^2\}\{1+\rho(n-1)\}\right] \qquad (3.13)$$

For large N and $\rho_y = \rho$, the variance of usual unbiased estimator $\bar{y}_{sy}$ is given by

$$Var(\bar{y}_{sy}) = \frac{S_y^2}{n}\{1+\rho(n-1)\}. \qquad (3.14)$$

Efficiency Comparisons

From (3.4), (3.5), (3.10), (3.11) and (3.14) in Case I it can be shown that the proposed estimator $\bar{y}_{\mathrm{Re}sy}^{(d)}$ is more efficient than

(a) the usual unbiased estimator $\bar{y}_{sy}$ if

$$\rho_{xy} > \frac{C_x}{4C_y}\left(1+\frac{1}{\lambda}\right) \qquad (3.15)$$

(b) Swain's (1964) estimator $\bar{y}_{Rsy}^{(d)}$ if

$$\rho_{xy} < \frac{3C_x}{4C_y}\left(1+\frac{1}{\lambda}\right) \qquad (3.16)$$

and that $\bar{y}_{Pesy}^{(d)}$ is better than

(a) the usual unbiased estimator $\bar{y}_{sy}$ if

$$\rho_{xy} < -\frac{C_x}{4C_y}\left(1+\frac{1}{\lambda}\right) \qquad (3.17)$$

(b) the product estimator $\bar{y}_{Psy}^{(d)}$ if

$$\rho_{xy} > -\frac{3C_x}{4C_y}\left(1+\frac{1}{\lambda}\right). \qquad (3.18)$$

Combining {(3.15) and (3.16)} and {(3.17) and (3.18)} shows that the proposed estimator $\bar{y}_{\mathrm{Re}\,sy}$ is more efficient than $\bar{y}_{sy}$ and $\bar{y}_{Rsy}^{(d)}$ if

$$\frac{C_x}{4C_y}\left(1+\frac{1}{\lambda}\right) < \rho_{xy} < \frac{3C_x}{4C_y}\left(1+\frac{1}{\lambda}\right) \tag{3.19}$$

and the proposed modified product estimator $\bar{y}_{Pesy}^{(d)}$ is better than $\bar{y}_{sy}$ and $\bar{y}_{Psy}^{(d)}$ if

$$-\frac{3C_x}{4C_y}\left(1+\frac{1}{\lambda}\right) < \rho_{xy} < -\frac{C_x}{4C_y}\left(1+\frac{1}{\lambda}\right). \tag{3.20}$$

From (3.6), (3.7), (3.12), (3.13) and (3.14) in case II it can be established that the proposed estimator $\bar{y}_{\mathrm{Re}\,sy}^{(d)}$ is better than

(a) the usual unbiased estimator $\bar{y}_{sy}$ if

$$\rho_{xy} > \frac{1}{4}\frac{C_x}{C_y} \tag{3.21}$$

(b) the ratio estimator $\bar{y}_{Rsy}^{(d)}$ if

$$\rho_{xy} < \frac{3}{4}\frac{C_x}{C_y} \tag{3.22}$$

and $\bar{y}_{Pesy}^{(d)}$ is more efficient than

(a) the usual unbiased estimator $\bar{y}_{sy}$ if

$$\rho_{xy} < -\frac{1}{4}\frac{C_x}{C_y} \tag{3.23}$$

(b) the product estimator $\bar{y}_{Psy}^{(d)}$ if

$$\rho_{xy} > -\frac{3}{4}\frac{C_x}{C_y}. \tag{3.24}$$

Combining (3.21) and (3.22) shows that the proposed estimator $\bar{y}_{\mathrm{Re}\,sy}^{(d)}$ is more efficient than $\bar{y}_{sy}$ and $\bar{y}_{Rsy}^{(d)}$ if

$$\frac{1}{4}\frac{C_x}{C_y} < \rho_{xy} < \frac{3}{4}\frac{C_x}{C_y},$$

a condition which is usually met in practice. Further from (3.23) and (3.24) it follows that the proposed estimator $\bar{y}_{Pesy}^{(d)}$ is better than $\bar{y}_{sy}$ and $\bar{y}_{Psy}^{(d)}$ if:

$$-\frac{3}{4}\frac{C_x}{C_y} < \rho_{xy} < -\frac{1}{4}\frac{C_x}{C_y}.$$

Cost Aspect

Following Swain (1964), let the cost function be of the form

$$C^* = c_0 n + c_1 \lambda n = (c_0 + c_1\lambda)n \tag{3.25}$$

where:

C^* = total cost,

c_0 = cost for observing a pair of (y,x) on a sampling unit, and

c_1 = cost for observing x on any unit of λ clusters.

From (3.10), (3.11), (3.12) and (3.13), note that all the four variance formulae are of the form:

$$V = \frac{V_1}{n}\{1+\rho(n-1)\} + \frac{V_2}{\lambda n}\{1+\rho(n-1)\}. \tag{3.26}$$

The optimum values of n and λ can be obtained by minimizing the variance function for a given cost. The value of λ which minimizes the variance function can be obtained by the equation

$$\frac{\partial V}{\partial \lambda} = 0,$$

where

$$V = V_1\left[\frac{(c_0+c_1\lambda)}{C}\left\{1+\rho\left(\frac{C}{c_0+c_1\lambda}-1\right)\right\}\right] + V_2\left[\frac{(c_0+c_1\lambda)}{C\lambda}\left\{1+\rho\left(\frac{C}{c_0+c_1\lambda}-1\right)\right\}\right]. \quad (3.27)$$

Differentiating (3.27) with respect to λ and equating to zero results in

$$\frac{\partial V}{\partial \lambda} = 0$$

$$= V_1\frac{c_1}{C}(1-\rho) - \frac{V_2}{\lambda^2}\left\{\rho+(1-\rho)\frac{c_0}{C}\right\}$$

$$\Rightarrow V_1\frac{c_1}{C}(1-\rho) = \frac{V_2}{\lambda^2}\left\{\rho+(1-\rho)\frac{c_0}{C}\right\}$$

$$\Rightarrow \lambda^2 = \frac{V_2}{V_1}\left[\frac{c_0}{c_1}+\frac{\rho}{(1-\rho)}\frac{C}{c_1}\right]$$

which gives

$$\lambda_{opt} = \sqrt{\frac{V_2}{V_1}\left[\frac{c_0}{c_1}+\frac{\rho}{(1-\rho)}\frac{C}{c_1}\right]}. \quad (3.28)$$

Substituting (3.28) in (3.25) results in

$$n_{opt} = \frac{C}{(c_0+c_1\lambda_{opt})} = \frac{C}{c_0+c_1\sqrt{\frac{V_2}{V_1}\left[\frac{c_0}{c_1}+\frac{\rho}{(1-\rho)}\frac{C}{c_1}\right]}}, \quad (3.29)$$

and substitution of (2.28) and (3.29) in (3.26) yields the minimum variance

$$V_{opt} = \frac{V_1}{n_{opt}}\left\{1+\rho\left(n_{opt}-1\right)\right\} + \frac{V_2}{\lambda_{opt}n_{opt}}\left\{1+\rho\left(n_{opt}-1\right)\right\}. \quad (3.30)$$

References

Cochran, W. G. (1946). Relative accuracy of systematic and stratified random samples for a certain class of populations. *Annals of Mathematical Statistics*, *17*, 164-177.

Robson, D. S. (1957).Application of multivariate polykays to the theory of unbiased ratio-type estimation. *Journal of the American Statistical Association*, *59*, 1225-1226.

Murthy, M. N. (1964). Product method of estimation. *Sankhya*, *A*(*26*), 69-74.

Bahl, S., & Tuteja, R. K. (1991). Ratio and product type exponential estimators. Information and optimization. *Science*, *12*, 159-163.

Hasel, H. A. (1942). Estimation of volume of timber stands by strip sampling. *Annals of Mathematical Statistics*, *13*, 179-206.

Griffth, A. L. (1945-46). The efficiency of enumerations. *Indian Forest Leaflets*, *83-93*. Forest-Research Institute: Dehra Dun.

Swain, A. K. P. C. (1964). The use of systematic sampling ratio estimate. *Journal of the Indian Statistical Association*, *2*,160-164.

Shukla, N. D. (1971). Systematic sampling and product method of estimation.In *Proceedings of all India Seminar on Demography and Statistics*. B.H.U. Varanasi: India.

Singh, H. P., & Vishwakarma, G. K. (2007). Modified exponential ratio and product estimators for finite population mean in double sampling. *Austrian Journal of Statistics*, *36*, 217-225.

Journal of Modern Applied Statistical Methods
November 2011, Vol. 10, No. 2, 436-446

1538 – 9472/11/$95.00

Comparison of Several Tests for Combining Several Independent Tests

Madhusudan Bhandary
Columbus State University
Columbus, GA

Xuan Zhang
Pracs Institute, Ltd.
Fargo, ND

Several tests for combining p-values from independent tests have been considered to address a particular common testing problem. A simulation study shows that Fisher's (1932) Inverse Chi-square test is optimal based on a power comparison of several different tests.

Key words: Omnibus test, omnibus hypothesis, p-value, Kolmogorov-Smirnov test, Tippett's test, Wilkinson's test, Inverse Chi-square test, Inverse normal test, Logit test.

Introduction

Tests for statistical significance of combined results were possibly the first statistical procedures developed for quantitative research synthesis. Combined test procedures were developed to combine the results of significance tests from different research studies.

Combining data from similar studies, as opposed to data derived from a single study, is important in Statistics. This study is a review of so-called omnibus statistical methods for testing the statistical significance of combined results. The procedures are called omnibus or non-parametric because they do not depend on the form of the underlying data, but only on the exact significance levels commonly called p-values. A key point is that observed p-values derived from continuous test statistics have a uniform distribution under the null hypothesis regardless of the test statistics or distribution from which they arise. The non-parametric nature of combined significance tests gives great flexibility in applications. Such tests can be used to combine any independent tests of hypotheses, even though the individual tests examine somewhat different hypotheses. For example, combined significance tests may be used to summarize the results of 10 studies each of which examined the effect of a treatment on a different outcome variable. Such a procedure would test whether the treatment produced a superior outcome on any of the dimensions investigated. These procedures can also be used in research synthesis to combine the results of studies that test the same conceptual hypothesis by different methods.

Many statistical tests are available for testing the significance for combining results. This study examines the most widely used tests. Nine different tests were compared, these are: Kolmogorov-Smirnov, Tippett's, Wilkinson's (for r = 2, 3, 4, 5), Inverse Chi-square, Inverse normal and Logit test. The objective of this study was to perform a comprehensive comparison of the performance of these tests based on their powers. A simulation study was conducted and the powers of the tests were compared. It was found that Fisher's (1932) Inverse Chi-square test was optimal based on the power comparison of the different tests.

Madhusudan Bhandary is a Professor in the Department of Mathematics. Email him at: bhandary_madhusudan@columbusstate.edu.
Xuan Zhang is a statistician. Email him at: zhangxuannd@yahoo.com.

p-Value Calculation: Normal Distribution

Let $X_1, X_2, ..., X_n$ be a random sample from $N(\mu, \sigma^2)$. Let $\bar{X}$ be the sample mean and let u be the observed value of the sample mean. Let $\Phi(.)$ be the distribution function of the standard normal distribution.

Case 1

$$H_0 : \mu = \mu_0 (specified)$$
$$H_1 : \mu < \mu_0$$

$$p-value = \Pr(\bar{X} \le u / H_0)$$
$$= P(\frac{\sqrt{n}(\bar{X} - \mu_0)}{\sigma} \le \frac{\sqrt{n}(u - \mu_0)}{\sigma})$$
$$= \Phi(\frac{\sqrt{n}(u - \mu_0)}{\sigma})$$

Case 2

$$H_0 : \mu = \mu_0 (specified)$$
$$H_1 : \mu > \mu_0$$

$$p-value = \Pr(\bar{X} \ge u / H_0)$$
$$= 1 - P(\bar{X} < u / H_0)$$
$$= 1 - P(\frac{\sqrt{n}(\bar{X} - \mu_0)}{\sigma} < \frac{\sqrt{n}(u - \mu_0)}{\sigma})$$
$$= 1 - \Phi(\frac{\sqrt{n}(u - \mu_0)}{\sigma})$$

Case 3

$$H_0 : \mu = \mu_0 (specified)$$
$$H_1 : \mu \ne \mu_0$$

$$p-value = P(\left|\frac{\sqrt{n}(\bar{X} - \mu_0)}{\sigma}\right| \ge \left|\frac{\sqrt{n}(u - \mu_0)}{\sigma}\right| / H_0)$$
$$= 2\left[1 - \Phi(\left|\frac{\sqrt{n}(u - \mu_0)}{\sigma}\right|)\right]$$

p-Value Calculation: Exponential Distribution

Let $X_1, X_2, ..., X_n$ be a random sample from EXP(μ). Let $\bar{X}$ be their sample mean.

$$H_0 : \mu = \mu_0 (specified)$$
$$H_1 : \mu \ne \mu_0$$

The likelihood ratio test is given by:

$$\Lambda = \frac{\underset{H_0}{Sup}\, L(\mu)}{\underset{\Omega}{Sup}\, L(\mu)} = (\frac{\bar{X}}{\mu_0})^n e^{-\frac{n\bar{X}}{\mu_0} + n}$$

where, Ω is the parameter space. For large n,

$$p-value = P(\Lambda \le observed\,\Lambda / H_0)$$
$$= P(-2\ln\Lambda \le -2\ln(observed\,\Lambda) / H_0)$$
$$= P(\chi_1^2 \ge -2\ln(observed\,\Lambda))$$

Omnibus Hypotheses and Omnibus Tests

Suppose n independent investigators have set about testing the validity of some null hypothesis:

H_0 : The population mean is $\mu_0 (specified)$

versus

H_1 : The population mean $\mu < \mu_0$

Each investigator will select a random sample from the population under focus, collect the relevant data, apply the appropriate test, and then report the *p*-value. The sample size could vary from investigator to investigator. The information provided by the investigators can be summarized as follows:

Investigator	Sample Size	*p*-Value
1	n_1	p_1
2	n_2	p_2
…	…	…
n	n_n	p_n

The objective is to determine if the null hypothesis is universally true. If the null hypothesis is true overall then, theoretically, $p_1, p_2, ...p_n$ should be a random sample of size n from a uniform distribution over (0, 1). In order to test the merit of the hypothesis overall, a test statistic must be built that is a function of the data $p_1, p_2, ...p_n$. A multitude of tests have been proposed in this connection, but before presenting a plethora of tests, the above problem must be generalized.

Assume m independent investigators, each investigating a hypothesis testing problem where H_{0i} is the null hypothesis proposed by investigator i, and H_{1i} is the alternative i = 1, 2, …, m. Each investigator collects data, tests his/her hypothesis and reports a p-value. This scenario can be summarized as follows:

Investigator	Null Hypothesis	Alternative Hypothesis	p-value
1	H_{01}	H_{11}	p_1
2	H_{02}	H_{12}	p_2
…	…	…	…
m	H_{0m}	H_{1m}	p_m

Postulating that the omnibus hypothesis, H_0 : H_{0i} is true for all i, versus the alternative H_1 : at least one H_{1i} is true, the data to decide in this issue are $p_1, p_2, ...p_m$. Theoretically, each p_i has a uniform distribution over (0, 1) if H_{0i} is true. If the omnibus null hypothesis is true, $p_1, p_2, ...p_m$ are independently, identically uniformly distributed over (0, 1). Now replace both the omnibus null and alternate hypotheses with the following equivalent hypotheses:

H_0: $p_1, p_2, ...p_m$ is a random sample from a uniform distribution over (0, 1), versus

H_1: $p_1, p_2, ...p_m$ is a random sample from a distribution which is not a uniform distribution over (0, 1).

Several tests have been developed to test the validity of the above modified hypotheses.

Test 1: Kolmogorov-Smirnov Test

The Kolmogorov-Smirnov (KS) test was originally proposed in the 1930's by Kolmogorov (1933) and Smirnov (1939). The KS test is only appropriate for testing data against a continuous distribution. The KS test statistic is defined as follows:

$$D = \underset{0<p<1}{Sup} \left| \hat{F}(p) - p \right|,$$

where $\hat{F}$ is the empirical distribution function of the data $p_1, p_2, ...p_m$. The exact distribution of D under H_0 has been worked out and a table of critical values is available.

Test 2: Tippett's Test

The first test of the significance of combined results was proposed by Tippett (1931), who noted that, if $p_1, p_2, ...p_m$ are independent p-values from continuous test statistics, then each has a uniform distribution under H_0. The test procedure is as follows: Reject H_0 if $p_{(1)} < 1-(1-\alpha)^{1/m}$, where $p_{(1)}$ = minimum of $p_1, p_2, ...p_m$. The p-value of the test is $=1-(1-p_{(1)})^m$.

Test 3: Wilkinson's Test

Wilkinson (1951) provided a generalization of Tippett's procedure that uses not just the smallest but the r[th] smallest p-value, $p_{(r)}$, as a test statistic, where $p_{(1)} \le p_{(2)} \le ... \le p_{(m)}$ are the ordered p-values (order statistics) obtained from $p_1, p_2, ...p_m$. The test procedure is given as follows: Reject H_0 if $p_{(r)} < p_{r,\alpha}$, where $p_{r,\alpha}$ is a critical value for $p_{(r)}$, or use a critical number $m_{r,\alpha}$ of p-values that are smaller than a fixed level α. Wilkinson described his procedure in terms of the number of significant p-values, that is, those that are smaller than α. He provided tables of the probability of obtaining m or more significant results at the α = 0.05 and α = 0.01 levels (that is, m or more p-values less than 0.05 or 0.01) for m < 25. Nomographs extending Wilkinson's tables to m = 100 for α = 0.05 and to m = 500 for α = 0.01 are given in Sakoda, Cohen and Beall (1954). Because $p_{(r)}$ has a beta distribution with parameters r and n-r+1, tables of the incomplete beta function can be used to obtain critical values of $p_{(r)}$ directly.

Test 4: The Inverse Chi-Square Test

One of the most widely used combination procedures is from Fisher (1932). Given m independent studies and p-values $p_1, p_2, ...p_m$, Fisher's procedure uses the product $p_1 p_2 ... p_m$ to combine p-values. He used a connection between the uniform distribution and the Chi-square distribution – namely, that if u has a uniform distribution, then $-2\ln u$ has a Chi-square distribution with 2 degrees of freedom. Consequently, when H_{0i} is true, $-2\ln p_i$ has a Chi-square distribution with 2 degrees of freedom, therefore, $-2\ln(p_1 p_2 ... p_n) = \sum_{i=1}^{m} -2\ln p_i$ also has a Chi-square distribution with 2m degrees of freedom. Due to this fact, no special tables are needed for the Fisher method. The test procedure becomes, reject H_0 if $T = -2\sum_{i=1}^{m} \ln p_i \geq c$, where the critical value c is obtained from the upper tail of the chi-square distribution with 2m degrees of freedom.

Test 5: The Inverse Normal Test

Another procedure for combining p-values is the inverse normal method proposed independently by Stouffer, et al. (1949) and by Liptak (1958). This procedure involves transforming each p-value to the corresponding normal score and then averaging. More specifically, defining Z_i by $p_i = \Phi(Z_i)$, where $\Phi(x)$ is the standard normal cumulative distribution function. When H_0 is true, the statistic

$$Z = \frac{Z_1 + Z_2 + ... + Z_m}{\sqrt{m}} = \frac{\Phi^{-1}(p_1) + \Phi^{-1}(p_2) + ... + \Phi^{-1}(p_m)}{\sqrt{m}}$$

has the standard normal distribution. Thus, H_0 is rejected whenever Z exceeds the appropriate critical value of the standard normal distribution.

Test 6: The Logit Test

The method of combining m independent p-values, $p_1, p_2, ...p_m$, suggested by George (1977) and investigated by Mudholkar and George (1979) transforms each p-value into a logit, $\ln(\frac{p}{1-p})$, and then combine the logits via the statistic

$$L = \ln\frac{p_1}{1-p_1} + ... + \ln\frac{p_m}{1-p_m}.$$

The exact distribution of L is not simple, but when H_0 is true, Mudholkar and George (1979) showed that the distribution of L (except for a constant) can be closely approximated by Student's t-distribution with 5m+4 degrees of freedom. Therefore, the test procedure is reject H_0 if $L^* = |L|\sqrt{\frac{(0.3)(5m+4)}{m(5m+2)}} > c$ where the critical value c is obtained from the t-distribution with 5m+4 degrees of freedom. (Note that the term 0.3 is more accurately given by $\frac{3}{\pi^2}$.) For large values of m, $L^* \simeq \left(\frac{0.55}{\sqrt{m}}\right) L.$

Methodology

Monte Carlo Simulation

A Monte Carlo simulation study was conducted to compare the performance of the omnibus test statistics described on the basis of estimated powers when the underlying data distributions are normal and exponential. The sample sizes used were 10 and100. The omnibus hypotheses are:

$$H_0 : \mu = 5$$

versus

$$H_1 : \mu \neq 5$$

The maintenance of significance levels was checked for each of the nine tests (for Test 3, r = 2, 3, 4, 5 were used), under each sample size and population mean, and for two distributions:

normal and exponential. Empirical error rates for each case were estimated by first simulating 10,000 different samples with specified sample size and population mean (μ_0) from a population with a specified distribution.

The test of interest was performed on each sample and it was determined if the null hypothesis was rejected at the 5% significance level. The empirical error rates for that test were then computed as the proportion of times the null hypothesis was rejected at each significance level. A test was considered acceptable at the 5% significance level if the error rates were between 0.044 and 0.056. The range represents a 99% confidence interval for the stated significance level.

Results

Tables 1-4 display the estimated powers of each test statistic investigated at the 0.05significance level; Figures 1-4 show the power curves.

Conclusion

Of the nine test statistics considered, the Inverse Chi-square test gives the highest power in almost every simulation, regardless of the number of populations, sample size or parameter values. The second highest power observed was with the Inverse Normal test. The minimum *p* test almost always gave the lowest power. In general, the Inverse Chi-Square proved superior by performing consistently in simulations for a wide range of cases.

References

Fisher, R. A. (1932). *Statistical methods for research workers* (*4th Ed*.). London: Oliver & Boyd.

Kolmogorov, A. N. (1933). Sulla determinazione empirica di una legge di distribuzione. *Institute of Italian Attuari. Gorn*., *4*, 1-11.

Smirnov, N. V. (9139). On the estimation of the discrepancy between empirical curves of distribution for two independent samples. *Bulletin of the University of Moscow*, *2*(*2*), 3-14.

Tippett, L. H. C. (1931). *The methods of statistics* (*1st Ed.*). London: Williams & Norgate.

Wikinson, B. (1951). A statistical consideration in psychological research. *Psychological* Bulletin, *48*, 156-158.

Sakoda, J. M., Cohen, B. H., & Beall, G. (9154). Test of significance for a series of statistical tests. *Psychological Bulletin*, *51*, 172-175.

Stouffer, S. A., Suchman, E. A., Devinney, L. C., Star, S. A., & Williams, R. M., Jr. (1949). *The American soldier 1949: Adjustment during army life*. Princeton, NJ: Princeton University Press.

Liptak, T. (1958). On the combination of independent tests. *Magyar Tudomanyos Akademia Matematikai Kutato Intezetenek Kozlemenyei*, *3*, 1971-1977.

George, E. O. (1977). *Combining independent one-sided and two-sided statistical tests: Some theory and applications.* Unpublished doctoral dissertation, University of Rochester.

Mudholkar, G. S., & George, E. O. (1979). The logit method for combining probabilities. In *Symposium on optimizing methods in statistics*, J. Rustagi (Ed.). 345-366. New York, NY: Academic Press.

Table 1: Normal Distribution n = 10, nrep = 10,000, $\alpha = 0.05$

μ	KS	P(1)	P(2)	P(3)	P(4)	P(5)	INV-CHI	INV-NORM	LOGIT
4.0	0.9712	0.8352	0.9467	0.9705	0.9730	0.9709	0.9926	0.9901	0.9827
4.1	0.9171	0.7380	0.8915	0.9241	0.9286	0.9259	0.9735	0.9657	0.9441
4.2	0.8108	0.6101	0.7879	0.8354	0.8443	0.8328	0.9225	0.9025	0.8556
4.3	0.6635	0.4830	0.6530	0.6973	0.7105	0.6910	0.8096	0.7796	0.7009
4.4	0.4901	0.3691	0.4919	0.5300	0.5412	0.5190	0.6511	0.6060	0.5119
4.5	0.3308	0.2643	0.3514	0.3738	0.3705	0.3607	0.4632	0.4254	0.3351
4.6	0.2051	0.1830	0.2253	0.2384	0.2351	0.2226	0.2982	0.2658	0.1891
4.7	0.1275	0.1158	0.1315	0.1418	0.1395	0.1296	0.1614	0.1488	0.1000
4.8	0.0817	0.0816	0.0853	0.0845	0.0844	0.0852	0.0946	0.0871	0.0658
4.9	0.0622	0.0578	0.0568	0.0589	0.0595	0.0552	0.0601	0.0586	0.0492
5.0	0.0529	0.0501	0.0509	0.0505	0.0517	0.0525	0.0492	0.0483	0.0472
5.1	0.0610	0.0576	0.0575	0.0580	0.0587	0.0573	0.0629	0.0595	0.0533
5.2	0.0835	0.0785	0.0822	0.0839	0.0856	0.0821	0.0913	0.0848	0.0622
5.3	0.1168	0.1194	0.1399	0.1410	0.1360	0.1263	0.1667	0.1460	0.0999
5.4	0.1975	0.1750	0.2214	0.2306	0.2301	0.2168	0.2887	0.2618	0.1860
5.5	0.3328	0.2599	0.3433	0.3743	0.3767	0.3598	0.4649	0.4187	0.3308
5.6	0.4853	0.3651	0.4985	0.5387	0.5352	0.5186	0.6506	0.6089	0.5176
5.7	0.6747	0.4825	0.6502	0.7005	0.7130	0.6946	0.8148	0.7789	0.7044
5.8	0.8079	0.6088	0.7831	0.8307	0.8363	0.8272	0.9166	0.8963	0.8484
5.9	0.9159	0.7273	0.8864	0.9264	0.9261	0.9197	0.9726	0.9631	0.9404
6.0	0.9688	0.8351	0.9518	0.9733	0.9746	0.9707	0.9935	0.9918	0.9832

Table 2: Normal Distribution n = 100, nrep = 10,000, $\alpha = 0.05$

μ	KS	P(1)	P(2)	P(3)	P(4)	P(5)	INV-CHI	INV-NORM	LOGIT
4.0	1.0000	0.9947	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
4.1	1.0000	0.9713	0.9998	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
4.2	1.0000	0.8992	0.9969	0.9995	0.9998	1.0000	1.0000	1.0000	1.0000
4.3	1.0000	0.7659	0.9681	0.9936	0.9983	0.9995	1.0000	1.0000	1.0000
4.4	0.9995	0.5823	0.8658	0.9480	0.9782	0.9888	1.0000	0.9999	0.9997
4.5	0.9811	0.4042	0.6596	0.7916	0.8662	0.9098	0.9978	0.9934	0.9876
4.6	0.7915	0.2547	0.4307	0.5320	0.6075	0.6601	0.9298	0.8763	0.8212
4.7	0.4084	0.1572	0.2358	0.2899	0.3263	0.3569	0.6109	0.5204	0.4148
4.8	0.1627	0.0980	0.1179	0.1351	0.1495	0.1519	0.2440	0.2052	0.1353
4.9	0.0834	0.0633	0.0687	0.0705	0.0683	0.0699	0.0795	0.0729	0.0522
5.0	0.0765	0.0475	0.0477	0.0466	0.0508	0.0539	0.0524	0.0511	0.0498
5.1	0.0864	0.0658	0.0668	0.0684	0.0684	0.0717	0.0848	0.0765	0.0545
5.2	0.1587	0.0997	0.1178	0.1314	0.1456	0.1548	0.2423	0.2063	0.1364
5.3	0.4102	0.1619	0.2367	0.2925	0.3360	0.3651	0.6093	0.5189	0.4105
5.4	0.8004	0.2626	0.4307	0.5503	0.6245	0.6747	0.9314	0.8825	0.8274
5.5	0.9805	0.4072	0.6626	0.7906	0.8600	0.8998	0.9975	0.9931	0.9872
5.6	0.9997	0.5881	0.8672	0.9507	0.9768	0.9875	1.0000	1.0000	0.9999
5.7	1.0000	0.7548	0.9661	0.9945	0.9991	0.9996	1.0000	1.0000	1.0000
5.8	1.0000	0.8929	0.9957	0.9996	1.0000	1.0000	1.0000	1.0000	1.0000
5.9	1.0000	0.9674	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
6.0	1.0000	0.9940	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Figure 1: Normal Distribution n = 10, nrep = 10,000, α = 0.05

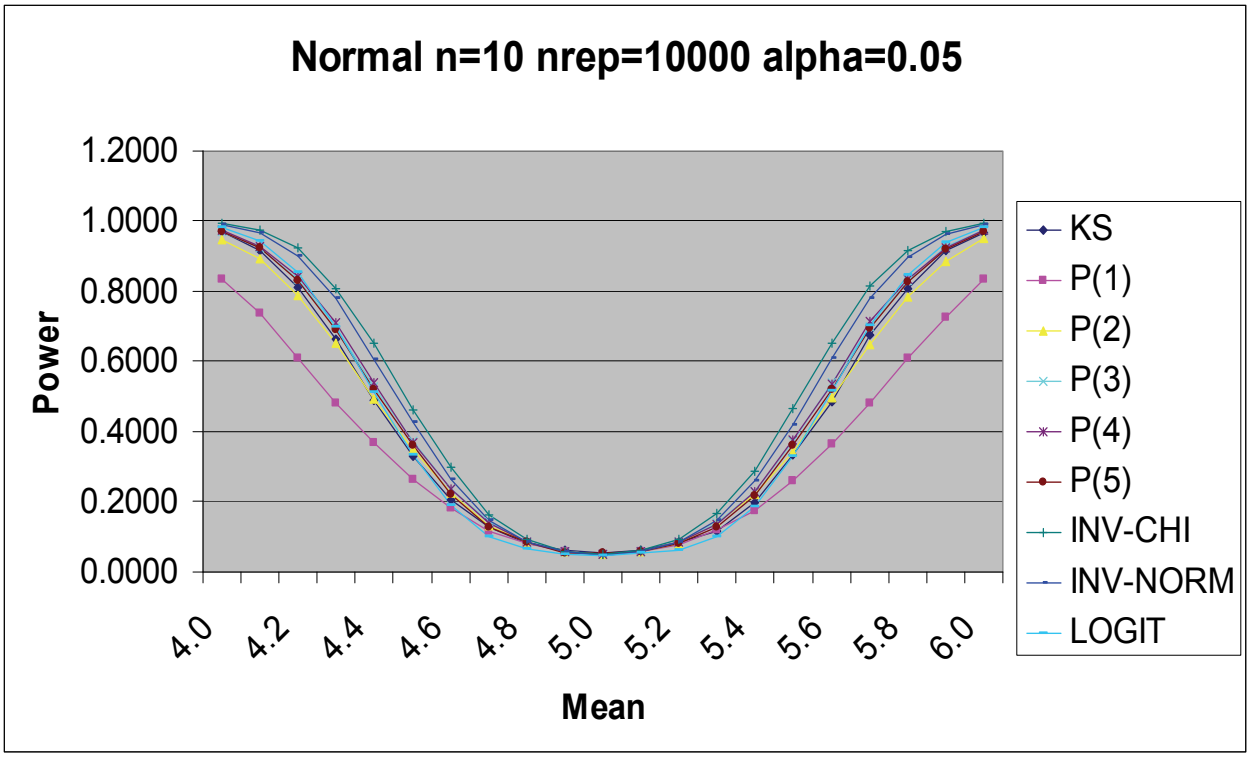


Figure 2: Normal Distribution n = 100, nrep = 10,000, α = 0.05

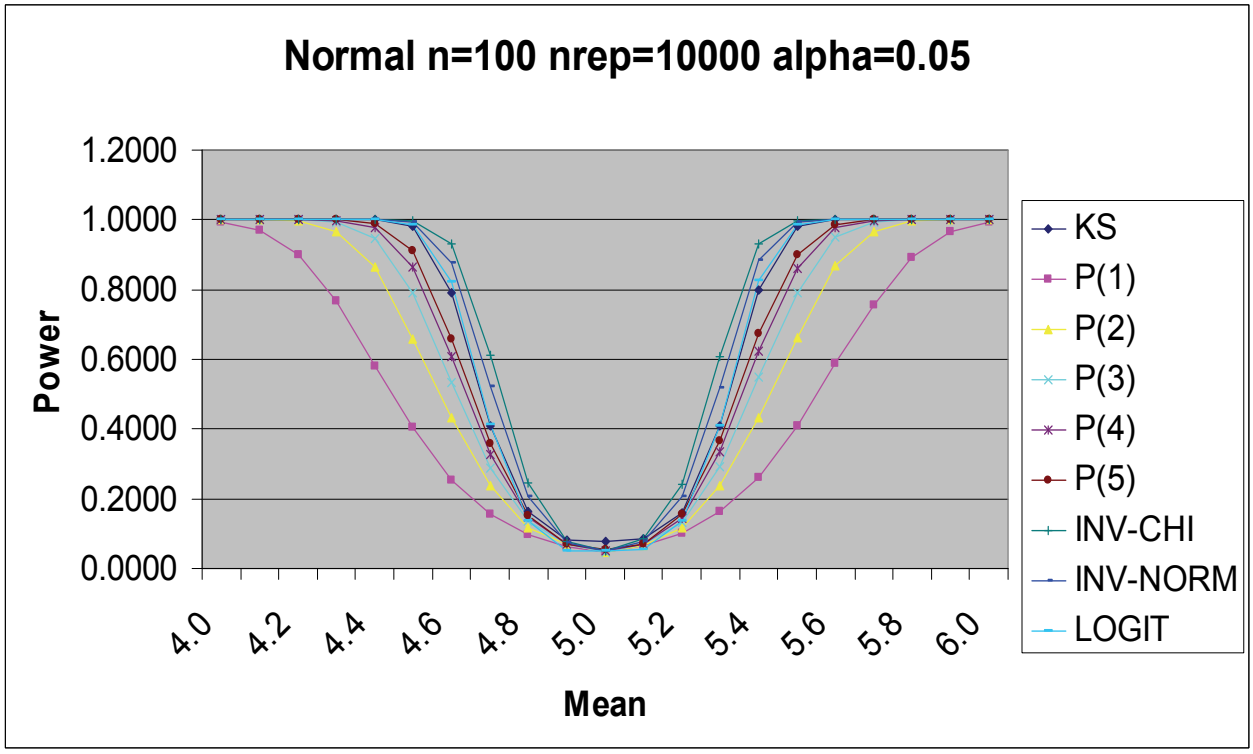


Table 3: Exponential Distribution n = 10, nrep = 10,000, $\alpha = 0.05$

μ	KS	P(1)	P(2)	P(3)	P(4)	P(5)	INV-CHI	INV-NORM	LOGIT
4.0	0.8962	0.6269	0.8249	0.8834	0.9039	0.8990	0.9560	0.9503	0.9141
4.1	0.7830	0.5042	0.7091	0.7804	0.7964	0.7944	0.8823	0.8701	0.8064
4.2	0.6375	0.3978	0.5681	0.6391	0.6609	0.6545	0.7619	0.7394	0.6505
4.3	0.4689	0.3033	0.4425	0.4964	0.5089	0.5020	0.6033	0.5731	0.4712
4.4	0.3293	0.2340	0.3229	0.3611	0.3698	0.3545	0.4451	0.4164	0.3164
4.5	0.2167	0.1786	0.2253	0.2452	0.2496	0.2384	0.3017	0.2814	0.1962
4.6	0.1366	0.1222	0.1479	0.1556	0.1548	0.1486	0.1811	0.1712	0.1133
4.7	0.0940	0.0897	0.1039	0.1053	0.1014	0.1006	0.1185	0.1103	0.0742
4.8	0.0716	0.0670	0.0686	0.0722	0.0685	0.0686	0.0743	0.0711	0.0577
4.9	0.0636	0.0518	0.0578	0.0561	0.0558	0.0517	0.0553	0.0553	0.0510
5.0	0.0572	0.0481	0.0524	0.0546	0.0526	0.0526	0.0557	0.0554	0.0511
5.1	0.0599	0.0552	0.0589	0.0564	0.0536	0.0554	0.0562	0.0536	0.0493
5.2	0.0713	0.0651	0.0669	0.0683	0.0774	0.0682	0.0749	0.0695	0.0581
5.3	0.0904	0.0884	0.0953	0.0999	0.0953	0.0926	0.1123	0.1048	0.0746
5.4	0.1275	0.1289	0.1474	0.1496	0.1422	0.1382	0.1748	0.1532	0.1086
5.5	0.1754	0.1734	0.2067	0.2141	0.2069	0.1946	0.2586	0.2286	0.1647
5.6	0.2369	0.2236	0.2801	0.2966	0.2804	0.2650	0.3588	0.3097	0.2339
5.7	0.3293	0.2902	0.3807	0.3971	0.3825	0.3596	0.4907	0.4345	0.3528
5.8	0.4445	0.3738	0.4886	0.5048	0.5007	0.4770	0.6202	0.5643	0.4774
5.9	0.5556	0.4670	0.5907	0.6278	0.6209	0.5966	0.7407	0.6935	0.6100
6.0	0.6666	0.5468	0.6933	0.7277	0.7274	0.6898	0.8314	0.7867	0.7243

Table 4: Exponential Distribution n = 100, nrep = 10,000, α = 0.05

μ	KS	P(1)	P(2)	P(3)	P(4)	P(5)	INV-CHI	INV-NORM	LOGIT
4.0	1.0000	0.8712	0.9963	0.9998	1.0000	1.0000	1.0000	1.0000	1.0000
4.1	1.0000	0.7470	0.9732	0.9968	0.9991	0.9998	1.0000	1.0000	1.0000
4.2	1.0000	0.6055	0.9049	0.9717	0.9895	0.9962	1.0000	1.0000	1.0000
4.3	0.9992	0.4563	0.7605	0.8903	0.9437	0.9691	1.0000	0.9999	0.9996
4.4	0.9808	0.3351	0.5863	0.7254	0.8072	0.8605	0.9979	0.9938	0.9866
4.5	0.8378	0.2387	0.3969	0.5073	0.5945	0.6458	0.9436	0.9056	0.8500
4.6	0.5128	0.1590	0.2455	0.3097	0.3630	0.4033	0.7017	0.6235	0.5127
4.7	0.2440	0.1077	0.1451	0.1740	0.1910	0.2106	0.3639	0.3086	0.2167
4.8	0.1166	0.0728	0.0876	0.0952	0.1012	0.1078	0.1483	0.1276	0.0835
4.9	0.0846	0.0565	0.0608	0.0578	0.0613	0.0613	0.0736	0.0725	0.0558
5.0	0.0804	0.0487	0.0495	0.0475	0.0503	0.0527	0.0519	0.0506	0.0485
5.1	0.0796	0.0534	0.0551	0.0619	0.0598	0.0624	0.0670	0.0649	0.0483
5.2	0.1134	0.0820	0.0890	0.0908	0.0971	0.1043	0.1468	0.1276	0.0820
5.3	0.2106	0.1198	0.1539	0.1738	0.1972	0.2111	0.3360	0.2789	0.1877
5.4	0.4033	0.1845	0.2568	0.3201	0.3585	0.3834	0.6240	0.5311	0.4205
5.5	0.6834	0.2675	0.4090	0.5053	0.5665	0.6185	0.8805	0.7972	0.7174
5.6	0.9024	0.3724	0.5874	0.7071	0.7770	0.8231	0.9815	0.9551	0.9275
5.7	0.9852	0.5028	0.7528	0.8647	0.9114	0.9397	0.9986	0.9956	0.9913
5.8	0.9984	0.6325	0.8851	0.9519	0.9769	0.9869	1.0000	0.9998	0.9995
5.9	0.9999	0.7602	0.9566	0.9879	0.9954	0.9985	1.0000	1.0000	1.0000
6.0	1.0000	0.8643	0.9867	0.9981	0.9994	1.0000	1.0000	1.0000	1.0000

Figure 3: Exponential Distribution n = 10, nrep = 10,000, α = 0.05

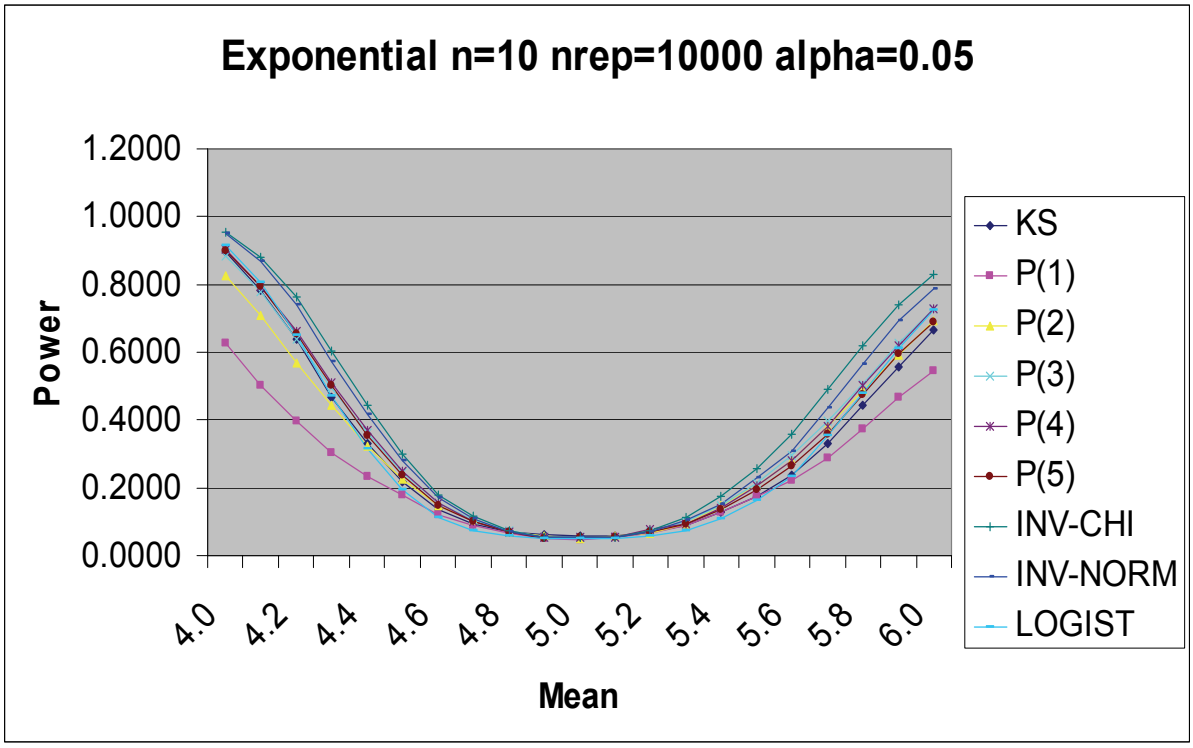
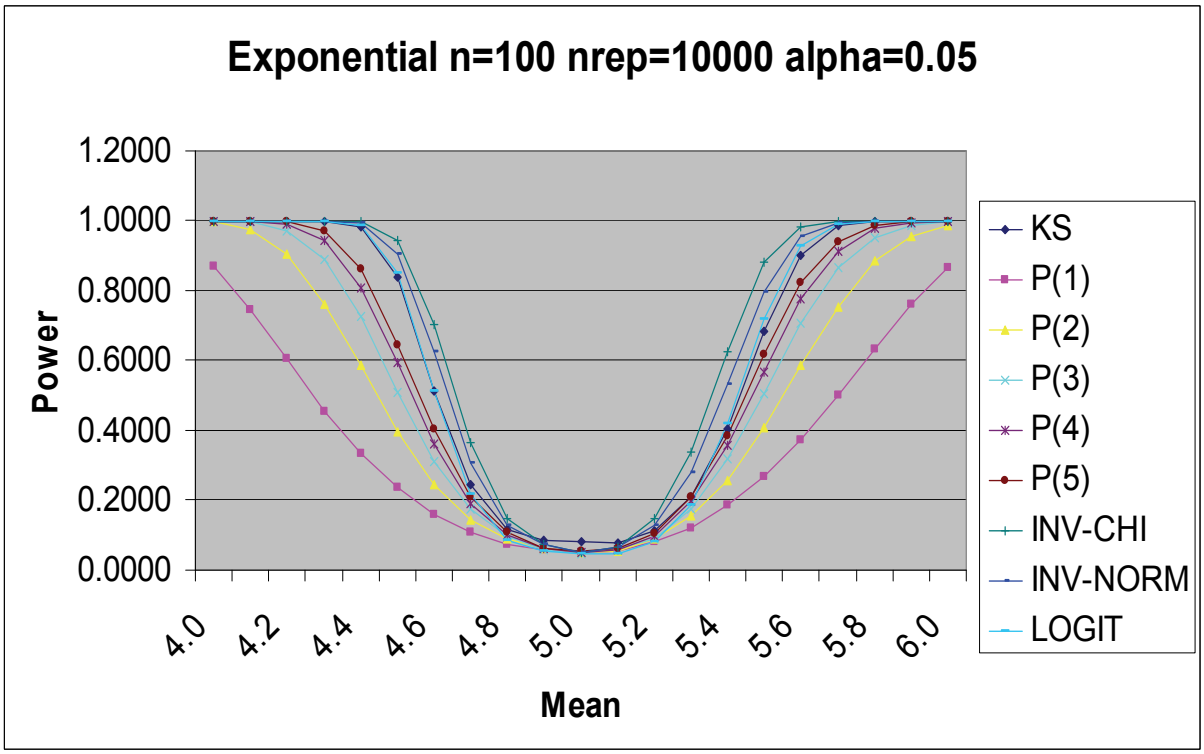


Figure 4: Exponential Distribution n = 10, nrep = 10,000, α = 0.05



Journal of Modern Applied Statistical Methods
November 2011, Vol. 10, No. 2, 447-461

1538 – 9472/11/$95.00

A Permutation Test for Compound Symmetry with Application to Gene Expression Data

Tracy L. Morris
University of Central Oklahoma, Edmond, OK

Mark E. Payton
Oklahoma State University, Stillwater, OK

Stephanie A. Santorico
University of Colorado Denver, Denver, CO

The development and application of a permutation test for compound symmetry is described. In a simulation study the permutation test appears to be a level-α test and is robust to non-normality. However, it exhibits poor power, particularly for small samples.

Key words: Compound symmetry, covariance matrix, gene expression, intraclass correlation, microarray, nonparametric inference, permutation test.

Introduction

Determining the underlying covariance or correlation structure of a data set can be challenging. The classical parametric method of testing for some hypothesized covariance structure involves using a likelihood ratio statistic that converges in distribution to a Chi-square random variable (Wilks, 1946). One common covariance structure, in which all of the variances are equal and all of the covariances are equal, is compound symmetry. One of the requirements of the likelihood ratio test (LRT) for compound symmetry is that the data be sampled from a multivariate normal population. Because the LRT is not robust to departures from normality (Huynh & Mandeville, 1979; Keselman, et al., 1980)a nonparametric test for compound symmetry would be very useful. In particular, permutation tests (PTs) have minimal to no distributional assumptions, do not require random samples and allow any combination of sample size and number of variables:

Tracy L. Morris is an Assistant Professor in the Department of Mathematics and Statistics. Email her at: tmorris2@uco.edu. Mark E. Payton is a Professor in the Department of Statistics. Email him at: mark.payton@okstate.edu. Stephanie A. Santorico is an Associate Professor in the Department of Mathematical and Statistical Sciences. Email her at: stephanie.santorico@ucdenver.edu.

Existing Tests for Compound Symmetry

Wilks (1946) was the first to develop a test for compound symmetry. This is a test of $H_0 : \boldsymbol{\Sigma} = \boldsymbol{\Sigma}_{CS}$, where

$$\boldsymbol{\Sigma}_{CS} = \sigma^2 \left[(1-\rho)\mathbf{I}_p + \rho \mathbf{1}_p \mathbf{1}_p' \right]; \qquad (1)$$

σ^2 is the common variance; ρ is the common pairwise correlation; $\mathbf{I}_p$ is the $p \times p$ identity matrix; and $\mathbf{1}_p$ is a $p \times 1$ unit vector. The classical approach to testing for compound symmetry involves the use of a LRT. Let $\mathbf{x}_i$, $i=1, \ldots, n$ be p-component vectors distributed according to $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The LRT criterion for this test is given by

$$\lambda = \frac{\left|\hat{\boldsymbol{\Sigma}}\right|^{n/2}}{\left[\left(s^2\right)^p (1-r)^{p-1} \left(1+(p-1)r\right) \right]^{n/2}},$$

where $\hat{\boldsymbol{\Sigma}}$ is the maximum likelihood estimator (MLE) of $\boldsymbol{\Sigma}$ under $H_a : \boldsymbol{\Sigma} \neq \boldsymbol{\Sigma}_{CS}$ and s and r are the MLEs of σ and ρ, respectively, under H_0.

Wilks (1946) determined the exact distribution of $\lambda^{2/n}$ for $p = 2$ and 3; however, the derivation of the exact distribution for larger values of p is too complex to be of practical use.

Therefore, the asymptotic distribution is most commonly used. Specifically, $-n\log\lambda^{2/n}$ is asymptotically distributed as a Chi-square random variable with $\frac{1}{2}p(p+1)-2$ degrees of freedom. As with other LRTs, this is a good approximation when n is considerably larger than p, but is poor when n is close to p. Therefore, the corrected LRT (CLRT) derived by Box (1950) is preferred. Box showed that $-(n-1)C\log\lambda^{2/n}$ is asymptotically distributed as a Chi-square random variable with $\frac{1}{2}p(p+1)-2$ degrees of freedom where

$$C = 1 - \frac{p(p+1)^2(2p-3)}{6(n-1)(p-1)(p^2+p-4)}.$$

The LRT for compound symmetry is actually just an extension of an earlier test of $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}$ developed by Mauchly (1940). Consequently, the LRT for compound symmetry suffers from the same limitations as Mauchly's test. Specifically, it is not a level-α test (Boik, 1975; Cornell, et al., 1992), is not robust to non-normality (Huynh & Mandeville, 1979; Keselman, et al., 1980), and requires $n > p$. The CLRT alleviates the problems with the type I error rate (except when n is close to p). It is not, however, robust to non-normality, and also requires $n > p$.

Wilks' (1946) work was subsequently extended. Lee, Krishnaiah and Chang (1976) determined that the Chi-square approximation for the distribution of the likelihood ratio statistic for compound symmetry is adequate for so-called practical purposes, and Votaw (1948) developed a test for compound symmetry in subsets of variates. Still other authors have explored similar tests for the structure of correlation rather than covariance matrices (Aitkin, 1969; Aitkin, Nelson & Reinfurt, 1968).

Tests for compound symmetry based on spatial signs and ranks have been developed more recently. Marden (1999) introduced one such rank-based test utilizing the differences between the estimated variances and covariances under the alternative hypothesis and the estimated variances and covariances under the null hypothesis. Two subsequent studies extended this work. The first used a permutation testing procedure where the usual LRT statistic was computed for the spatial ranks (Gao & Marden, 2001). In the second, a Hotelling T^2-type statistic was derived and shown to converge in distribution to a Chi-square random variable (Marden & Gao, 2002). The latter article also presents a similar test based on spatial signs. Marden & Gao performed a small simulation study ($n = 100$ & $p = 3$) for these tests and found both the rank and sign tests to be level-α tests when simulating data from spherically symmetric distributions.

Other authors have considered tests for sphericity based on spatial signs and ranks (Hallin & Paindaveine, 2006; Sirkiä, Taskinen, Oja & Tyler, 2009). These tests can also be used to test for compound symmetry by first applying an appropriate data transformation. All of these rank and sign tests are superior to the LRT for compound symmetry in that they broaden the family of distributions to which a test for compound symmetry can be applied. They are also applicable in cases in which $n \le p$. Unfortunately, these tests still have distributional assumptions: they require that data be sampled from a multivariate elliptical distribution.

Methodology

When the assumptions of parametric procedures are violated, PTs have been used as alternatives. Specifically, PTs reduce or eliminate distributional assumptions (Fisher, 1936; Good, 2005) and allow the use of nearly any test statistic; they are also valid for any combination of n and p. As with any statistical procedure, however, PTs have limitations. The greatest of which is that they can be computationally intensive even for moderate sample sizes. With continued advances in technology, PTs have become more feasible for larger sample sizes; however, there still exists a limit at which the computing time required to examine all possible permutations of the data is prohibitive. In such cases, a random sample of permutations may be selected to compute an approximate p-value (Dwass, 1957). These tests are commonly known as Monte Carlo PTs (MCPT).

Given the benefits of PTs and the limitations of LRTs for testing for the structure of a covariance matrix, it is the purpose of this research to develop a PT for compound symmetry. Before describing this test, note that covariance matrices are invariant to changes in location. Therefore, it was assumed throughout this study that the variable means are equal. If the variable means are unequal, the raw data can be easily centered by calculating $\mathbf{x}_i - \boldsymbol{\mu}$ or $\mathbf{x}_i - \bar{\mathbf{x}}$ depending on whether $\boldsymbol{\mu}$ is assumed known or unknown, respectively.

Proposed PT Test for Compound Symmetry

Let $\mathbf{x}_i$, i=1,…,n be identically distributed, p-variate vectors of observations on each of n subjects. The objective is to test $H_0 : \boldsymbol{\Sigma} = \boldsymbol{\Sigma}_{CS}$ where $\boldsymbol{\Sigma}$ is the covariance matrix of the distribution of $\mathbf{x}_i$, and $\boldsymbol{\Sigma}_{CS}$ has the compound symmetry structure given in (1). Good (2005) argues that the observations within each vector are exchangeable if either (i) the observations are independent, or (ii) they are normally distributed with equal covariances. The first of these conditions is a special case of compound symmetry, called sphericity, in which the variances are all equal and the covariances are all zero. In this case, the PT makes no distributional assumptions. The second set of conditions requires multivariate normality with equal covariances. Under the null hypothesis, the covariances are assumed equal and it appears from the simulation results presented herein that a weaker distributional assumption may be sufficient for practical purposes. Specifically, it appears that equivalent marginal distributions will suffice.

Because covariance matrices are symmetric, one possible test statistic can be computed by summing the absolute differences between the elements on or above the diagonal of the covariance matrix obtained from the observed data and the elements on or above the diagonal of the hypothesized covariance matrix estimated from the observed data. In matrix notation:

$$D = \mathbf{1}'_{\frac{1}{2}p(p+1)} \left| \mathrm{vec}\left(\boldsymbol{\Sigma}_{obs} - \hat{\boldsymbol{\Sigma}}_{CS} \right) \right|,$$

where $\boldsymbol{\Sigma}_{obs}$ is the covariance matrix obtained from the observed data;

$$\hat{\boldsymbol{\Sigma}}_{CS} = \overline{s^2} \left[(1 - \bar{r}) \mathbf{I}_p + r \mathbf{1}_p \mathbf{1}'_p \right];$$

vec($\mathbf{M}$) is a vector of the elements on or above the diagonal of $\mathbf{M}$; and $\overline{s^2}$ and $\bar{r}$ are the means of the sample variances and correlations, respectively. This test statistic is computed for each possible permutation of the data and the proportion of test statistic values greater than or equal to the one obtained from the original data is the p-value. Note that D can also be used to test for a specific common variance and/or correlation by substituting the specified value for $\overline{s^2}$ and/or $\bar{r}$, respectively, rather than estimating these values as described previously.

Results

Type I Error

One-thousand simulations were run using R version 2.10.1 (R, 2009) for various combinations of n (=5, 10, 25, 50, 100) and p (=3, 5, 10, 20). Due to the extremely large number of permutations required to carry out the PTs for any reasonable values of n and p, MCPTs were used in the simulations. For each simulation, a p-variate data set was generated and the MCPT, CLRT and sign test for sphericity (SIGN) were performed. The sign test for sphericity is available in the SpatialNP package for R (Sirkiä, Nordhausen & Oja, 2009).

One-thousand random permutations of the centered data were sampled for each MCPT. In practice, a much larger sample of permutations would be used for individual tests (usually 10,000 permutations); however, for a simulation study of this size, such a large number proved to be prohibitive. Therefore, 1,000 permutations were sampled for each MCPT based on the suggestions of Jöckel (1986) and Manly (1997). For the CLRT and SIGN test, the asymptotic Chi-square distributions were used to determine approximate 5% critical values.

Four different multivariate distributions (normal, uniform, double exponential and two-parameter exponential) were investigated. For

the multivariate normal distribution, data were generated in R using the mvrnorm function within the MASS add-on package (Venables & Ripley, 2002). For the multivariate uniform distribution data were generated using a procedure described in Falk (1999), and for the multivariate double exponential and two-parameter exponential distributions a procedure described in Vale and Maurelli (1983) was used.

The simulated type I error rates for the tests of compound symmetry are displayed in Figure 1. Simulations were run for n = 5, 10, 15, 25, 50, 100, $p = 3$, $\sigma^2 = 9$, and $\rho = 0.6$. For normally distributed data, the three tests are comparable with respect to the simulated type I error rates, with the CLRT and SIGN test appearing to be slightly conservative, particularly for small samples. The MCPT appears to be fairly robust to non-normality, especially when the underlying distribution is symmetric (normal, uniform, double exponential); however, in the case of the two-parameter exponential data, the MCPT appears to be too liberal with respect to the simulated type I error rates, especially for small samples. The CLRT appears to be too conservative for uniform data and much too liberal for double exponential and two-parameter exponential data, in the latter case achieving a simulated type I error rate as high as 0.352 for $n = 100$.

These results are consistent with those of Huynh and Mandeville (1979) who performed a simulation study of Mauchly's (1940) test of sphericity and found that for light-tailed distributions the LRTs were conservative and for heavy-tailed distributions, the type I error rates exceeded the nominal rate. The SIGN test performs very well with respect to the simulated type I error rates for double exponential data; however, the simulated type I error rates are extremely high for uniform (as high as 1.000 for $n = 50$) and two-parameter exponential data (as high as 0.604 for $n = 100$). This is undoubtedly due to the assumption of the SIGN test that the data be sampled from a multivariate elliptical distribution.

One disadvantage of the LRTs is that they do not exist when $p \geq n$; due to this, type I error rates tend to inflate as p approaches n. Figure 2 displays the simulated type I error rates for $n = 25$, $p = 3,5,10,20$, $\sigma^2 = 9$, and $\rho = 0.6$. From these results it is clear that the CLRT is not a level-α test, even for normally distributed data, when p is close to n; and the SIGN test suffers from the same problems as in Figure 1 for non-elliptical data. Consequently, the MCPT is the best choice, with respect to the simulated type I error rates of the three tests for uniform and two-parameter exponential data, even though the MCPT is too liberal in the latter case.

Power

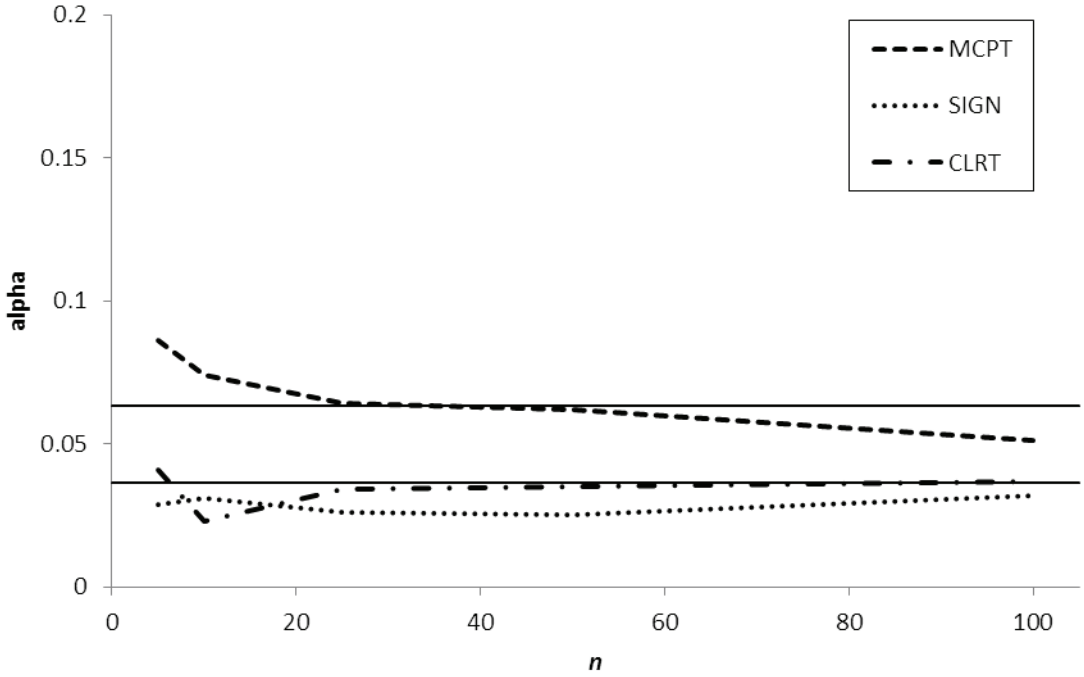
The power of the tests of compound symmetry to detect heteroscedasticity and serial correlation was studied. The MCPT, SIGN test and CLRT were all conducted for various combinations of n, p and distribution; however, because the SIGN test is not a level-α test for uniform and two-parameter exponential data and the CLRT is not a level-α test for double exponential and two-parameter exponential data the power results for these cases are largely excluded in the following discussion, but are presented in Figures 3 and 4 for completeness.

Figure 3 shows the simulated power of the test of compound symmetry versus heteroscedasticity. Specifically, multivariate data were generated from distributions with covariance matrices having diagonal elements given by 1, 1+d/(p-1), 1+2d/(p-1), …, 1+d and zero off diagonal elements, where d represents the difference between the first and last (or smallest and largest) diagonal elements. Figure 3 displays the power results for n = 5, 10, 25, 50, $p = 3$, $d = 4$ and $\rho = 0$.

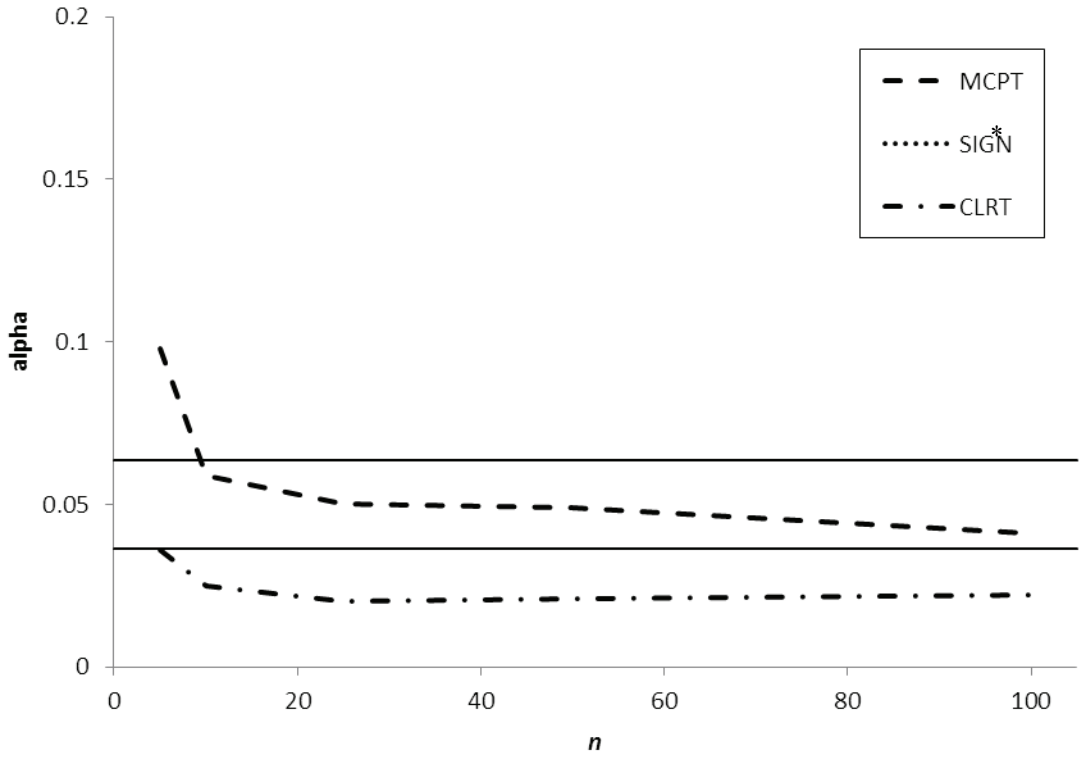
For normally distributed data the power of the CLRT is greater than that for the MCPT and SIGN test in most cases, but the MCPT performs fairly well, achieving a power of 0.983 when $n = 50$. The true benefit of the MCPT is observed in the non-normal cases. For uniformly distributed data; the simulated power of the MCPT is greater than or equal to that of the CLRT except for $n = 25$ (0.941 for the MCPT and 0.943 for the CLRT). For double exponential data the simulated powers of the MCPT and SIGN test are very close with the MCPT slightly more powerful for small samples ($n = 5,10,25$) and the SIGN test slightly more powerful for large samples ($n = 50$). For two-

Figure 1: Simulated Type I Error Rates for the Test of Compound Symmetry ($p = 3, \sigma^2 = 9, \rho = 0.6$)

a. Normal



b. Uniform

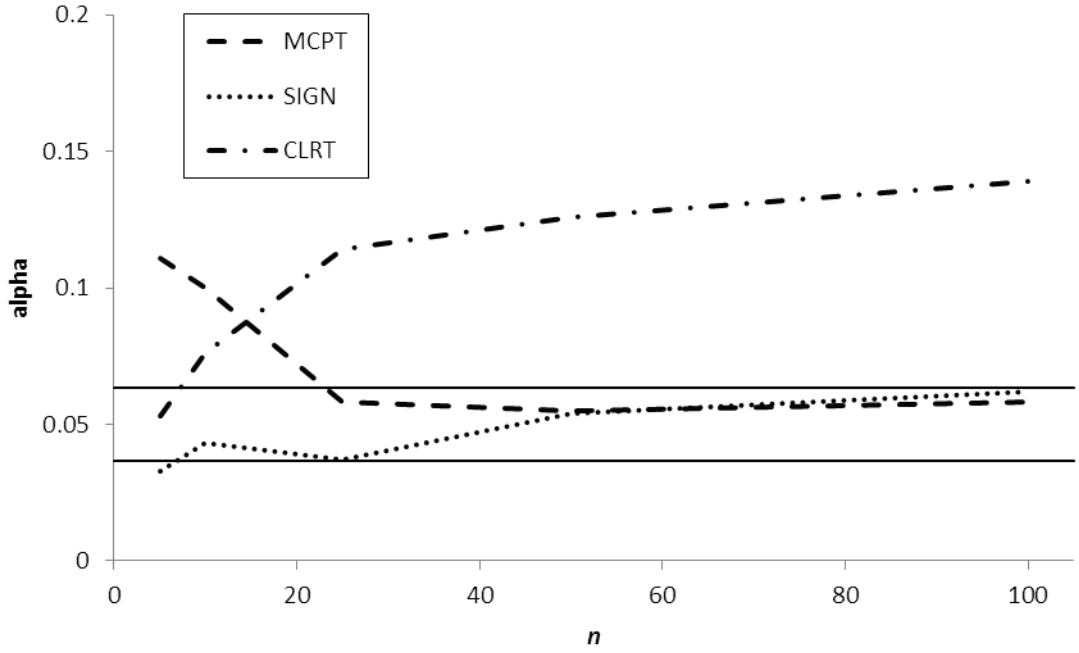


* The type I error rates for this test are greater than 0.2 for all simulated values of *n*.

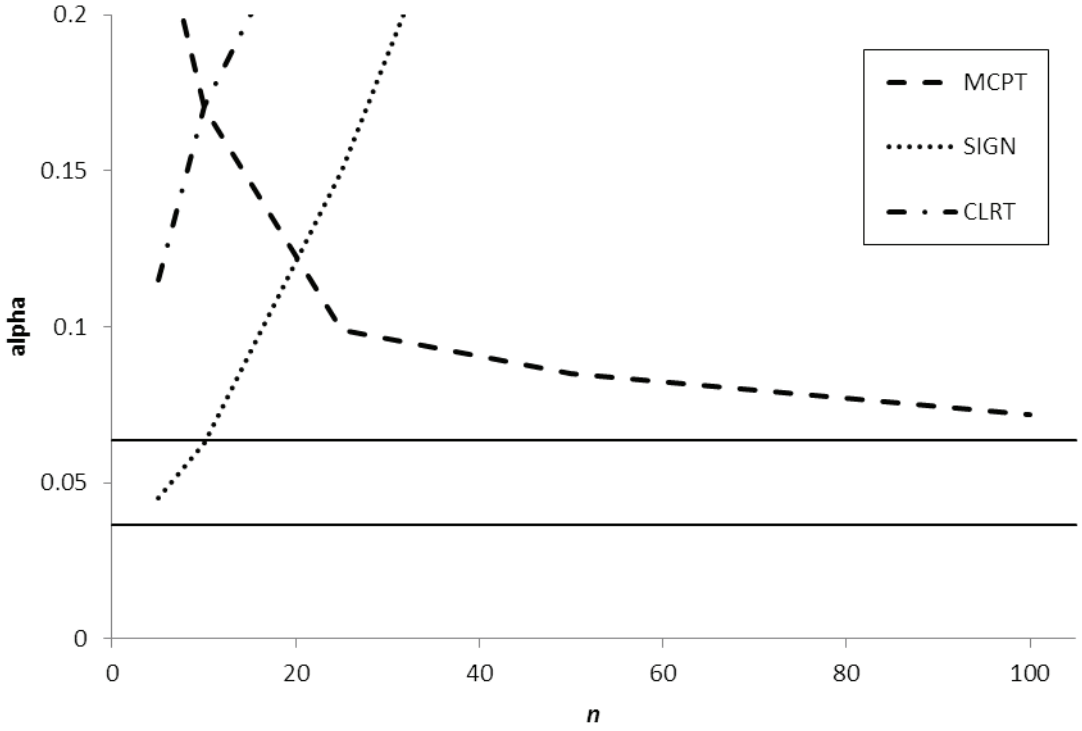
Note: The horizontal lines correspond to $0.05 \pm 1.96\sqrt{(0.05)(0.95)/1000}$

Figure 1 (continued): Simulated Type I Error Rates for the Test of Compound Symmetry ($p = 3, \sigma^2 = 9, \rho = 0.6$)

c. Double Exponential



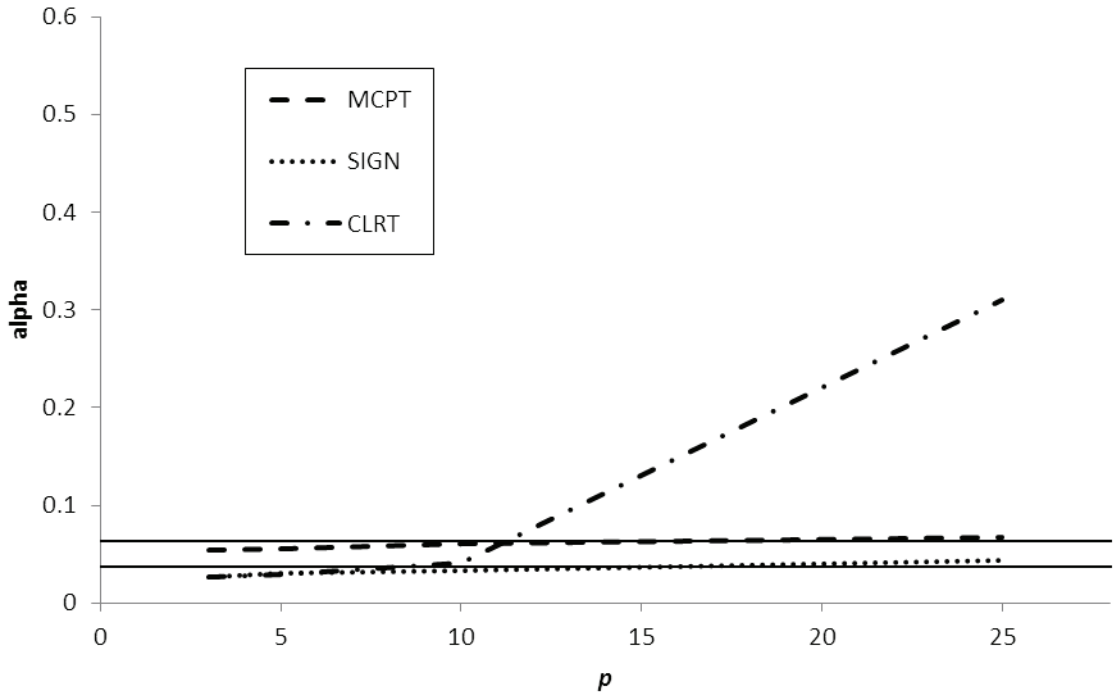
d. Two-Parameter Exponential



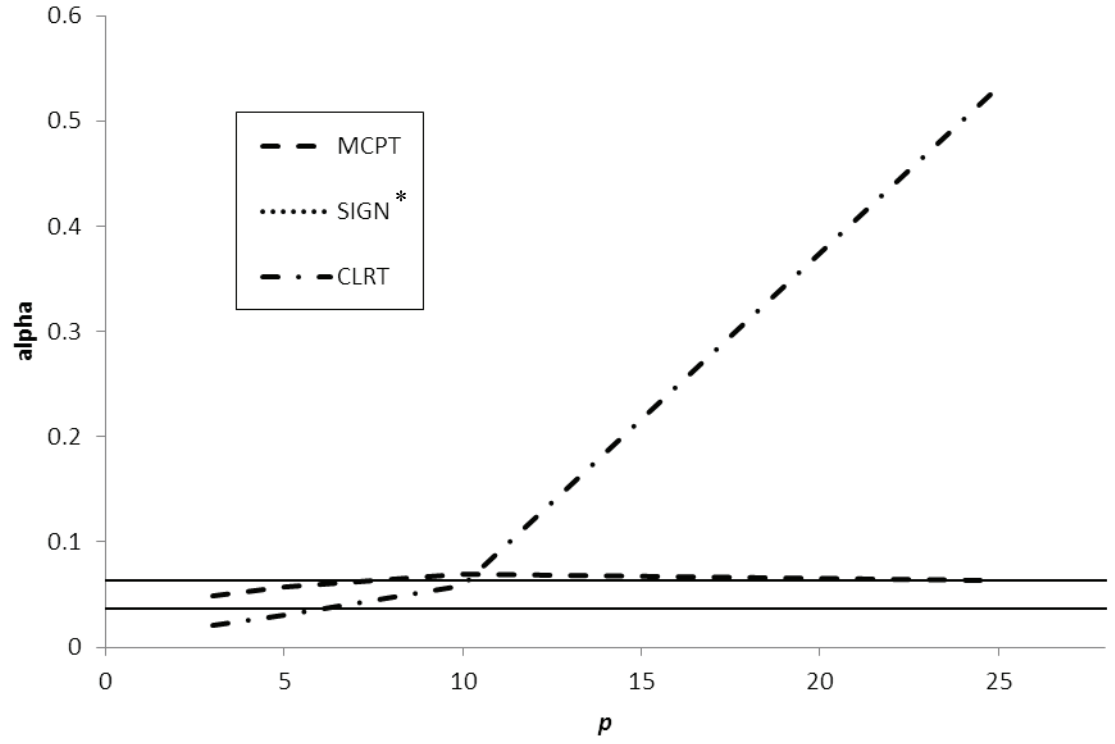
Note: The horizontal lines correspond to $0.05 \pm 1.96\sqrt{(0.05)(0.95)/1000}$

Figure 2: Simulated Type I Error Rates for the Test of Compound Symmetry ($n = 25, \sigma^2 = 9, \rho = 0.6$)

a. Normal



b. Uniform

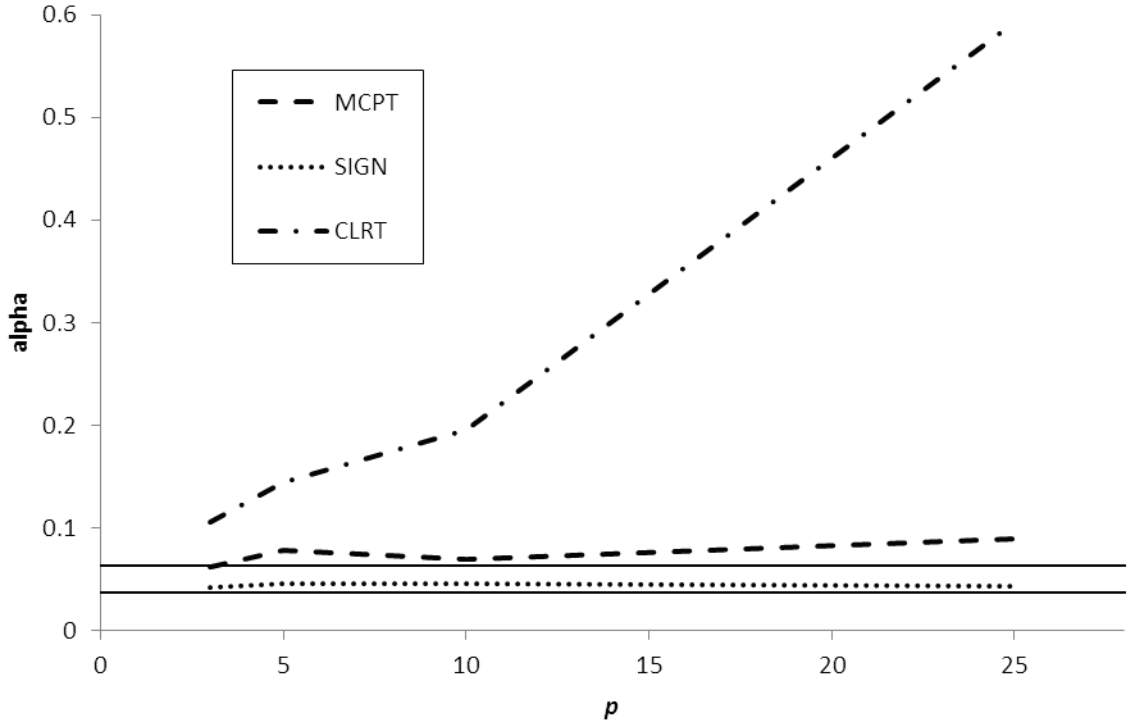


* The type I error rates for this test are greater than 0.2 for all simulated values of n.

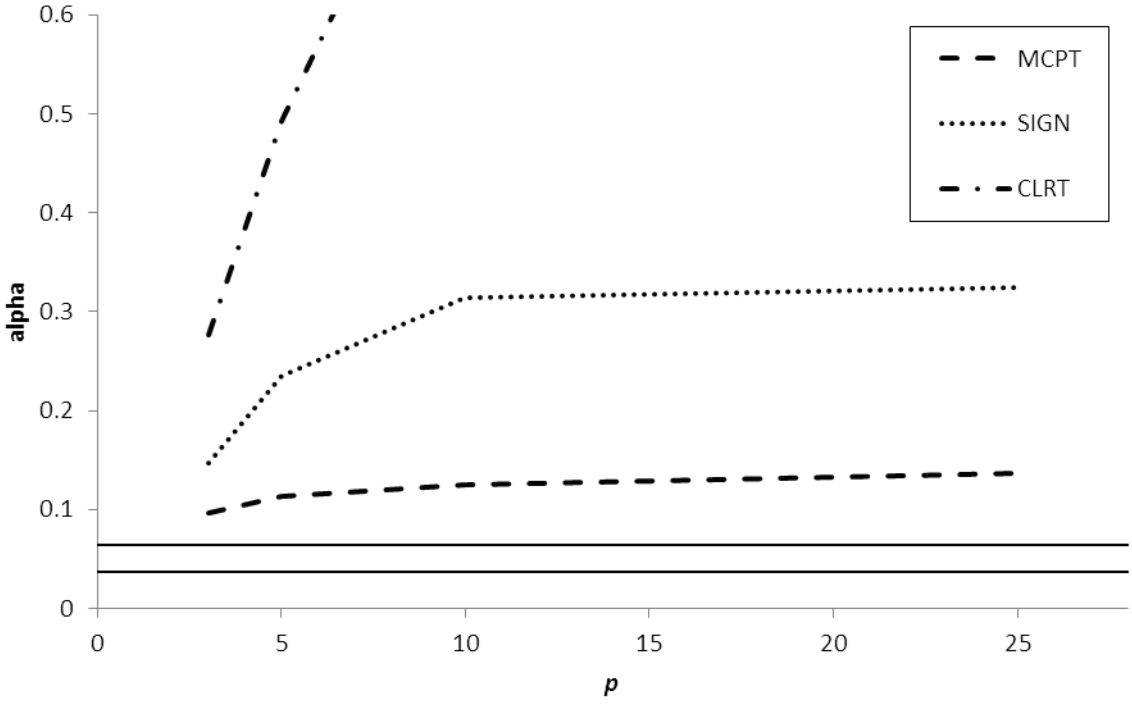
Note: The horizontal lines correspond to $0.05 \pm 1.96\sqrt{(0.05)(0.95)/1000}$

Figure 2 (continued): Simulated Type I Error Rates for the Test of Compound Symmetry ($n = 25, \sigma^2 = 9, \rho = 0.6$)

c. Double Exponential



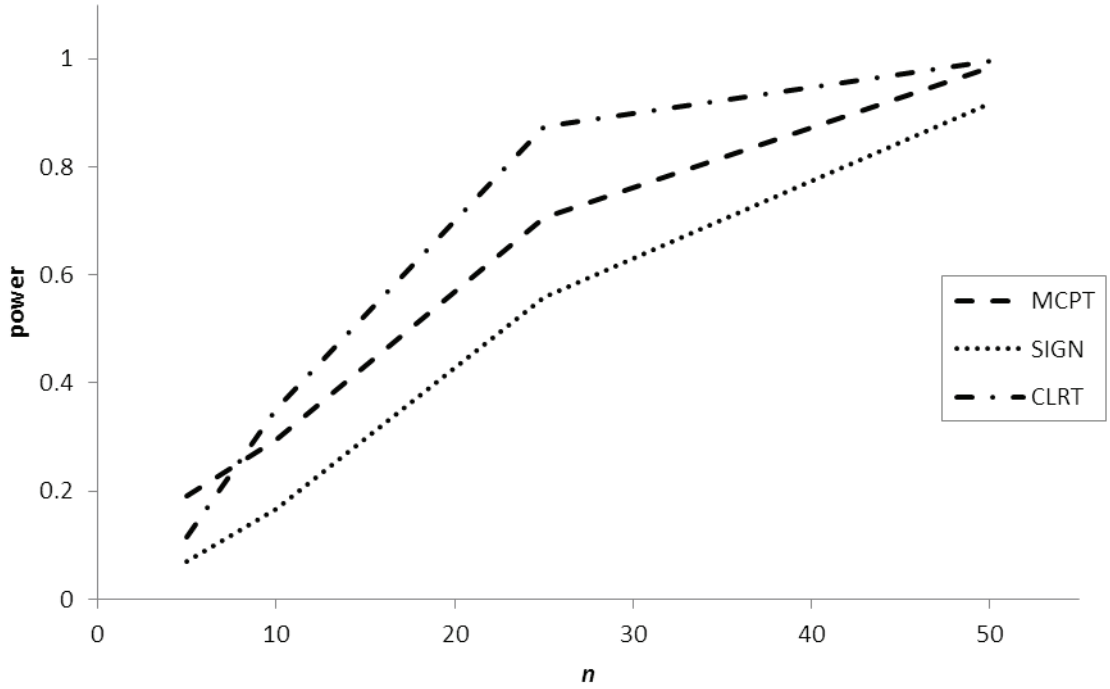
d. Two-Parameter Exponential



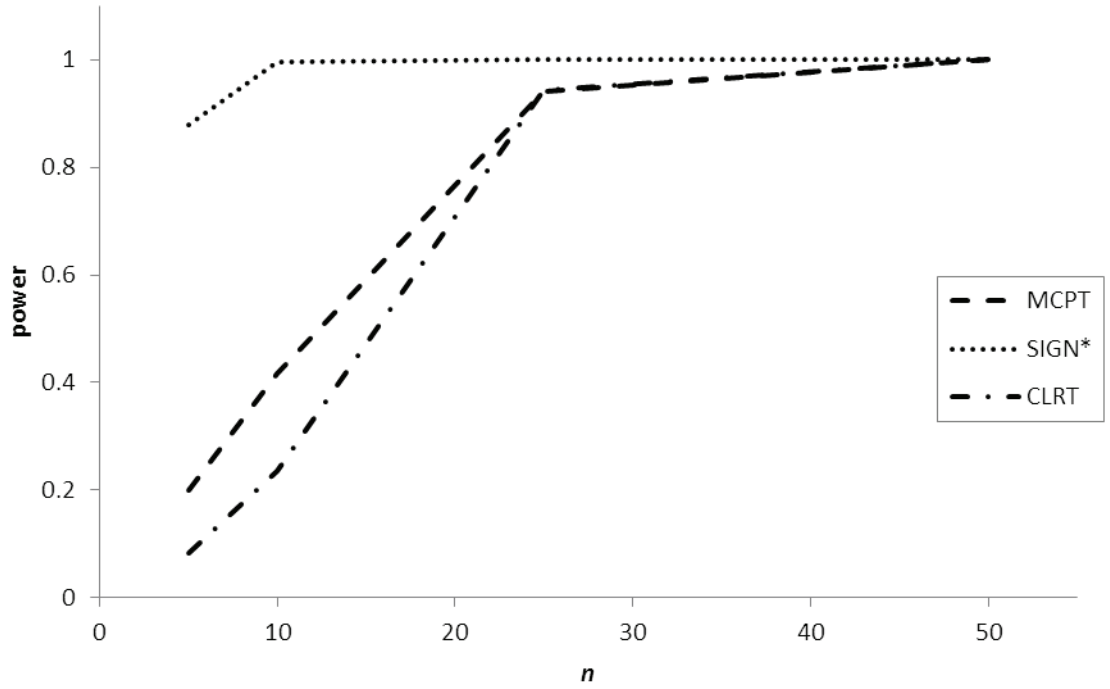
Note: The horizontal lines correspond to $0.05 \pm 1.96\sqrt{(0.05)(0.95)/1000}$

Figure 3: Simulated Power for the Test of Compound Symmetry vs. Heteroscedasticity (p=3, ρ = 0, d = 4)*

a. Normal



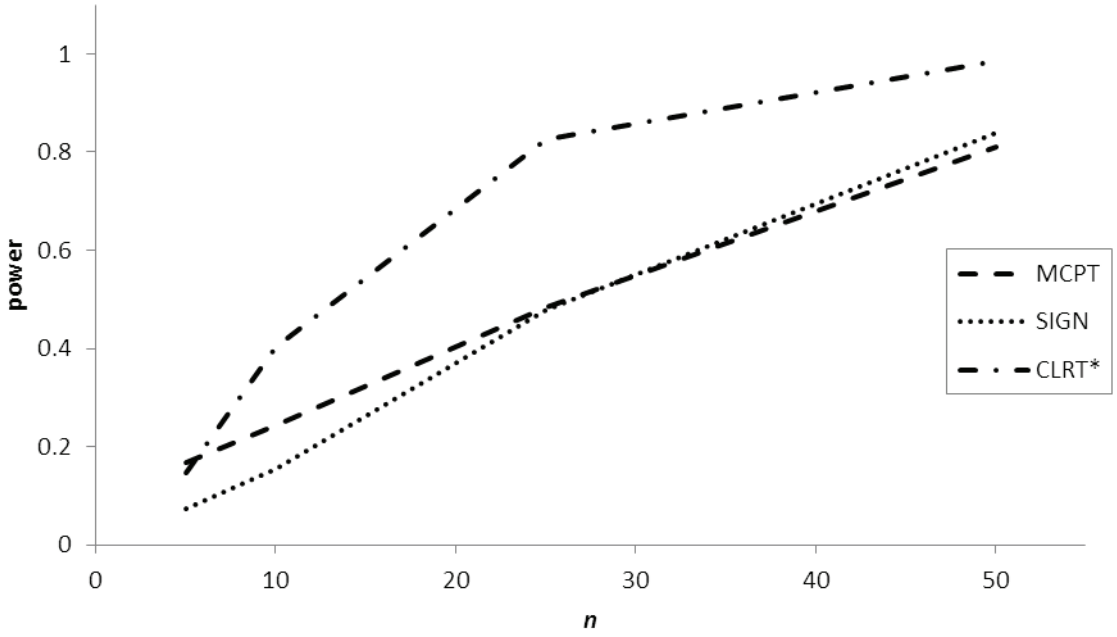
b. Uniform



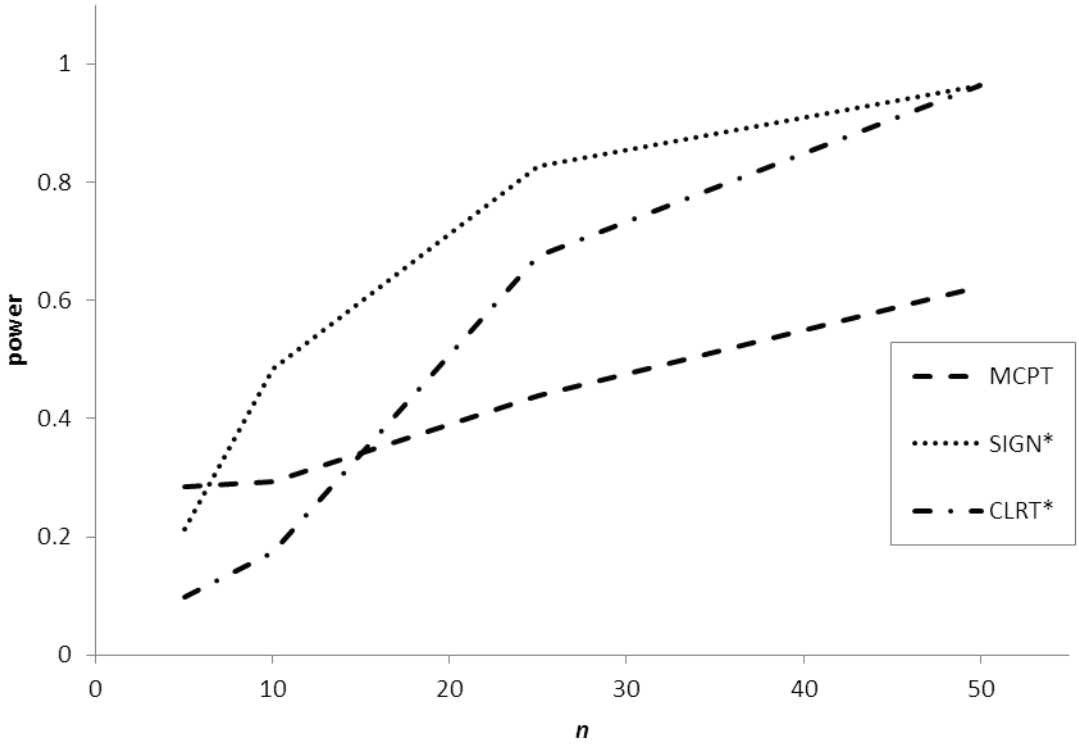
*These are not level-α tests.

Figure 3 (continued): Simulated Power for the Test of Compound Symmetry vs. Heteroscedasticity (p=3, $\rho = 0$, $d = 4$)*

c. Double Exponential



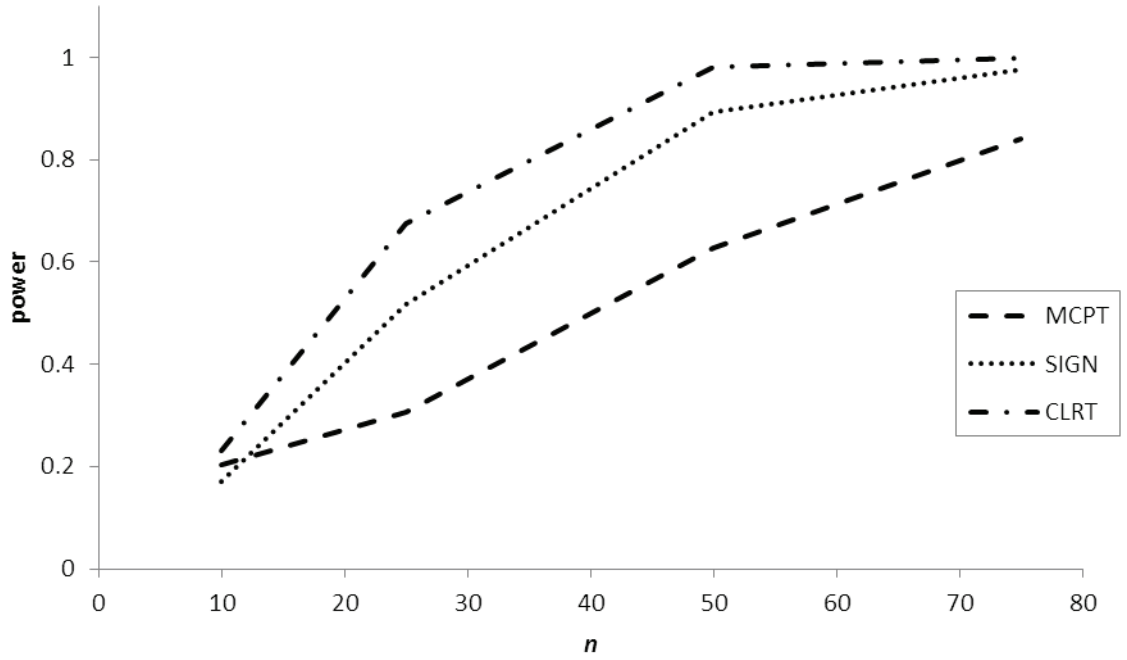
d. Two-Parameter Exponential



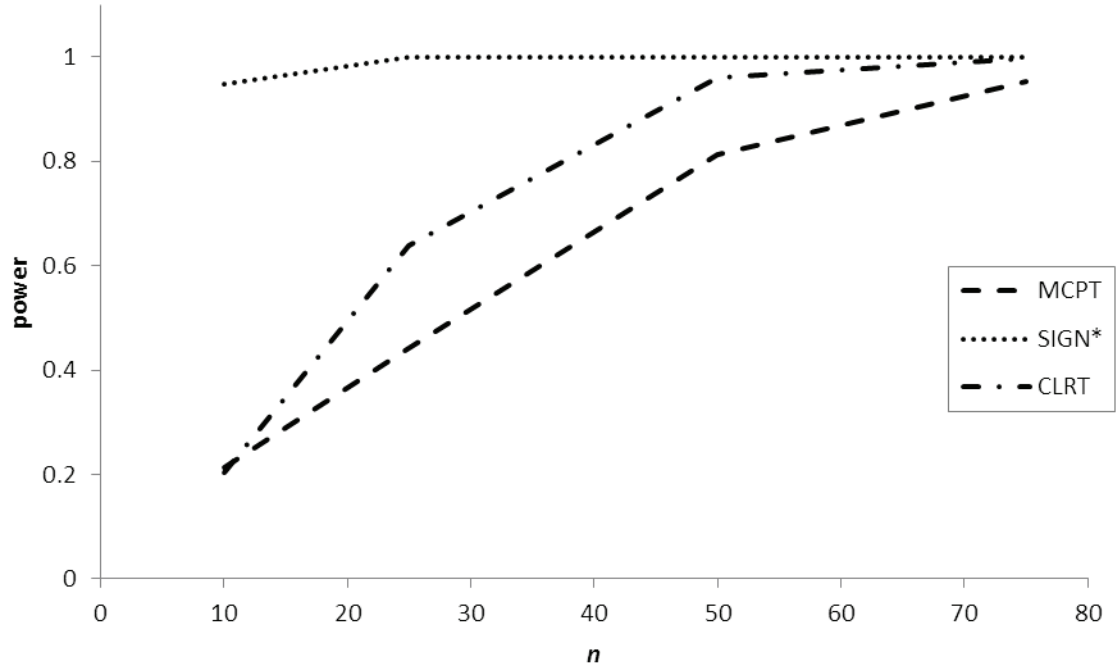
*These are not level-α tests.

Figure 4: Simulated Power for Test of Compound Symmetry vs. Serial Correlation ($p = 5, \sigma^2 = 1, \rho = 0.6$)*

a. Normal



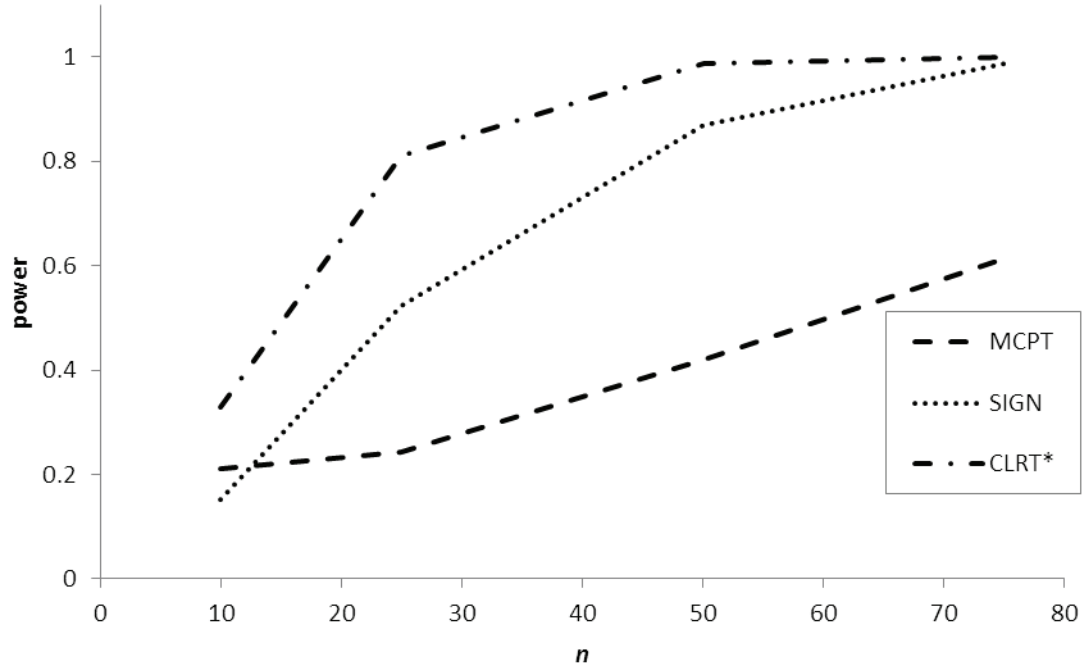
b. Uniform



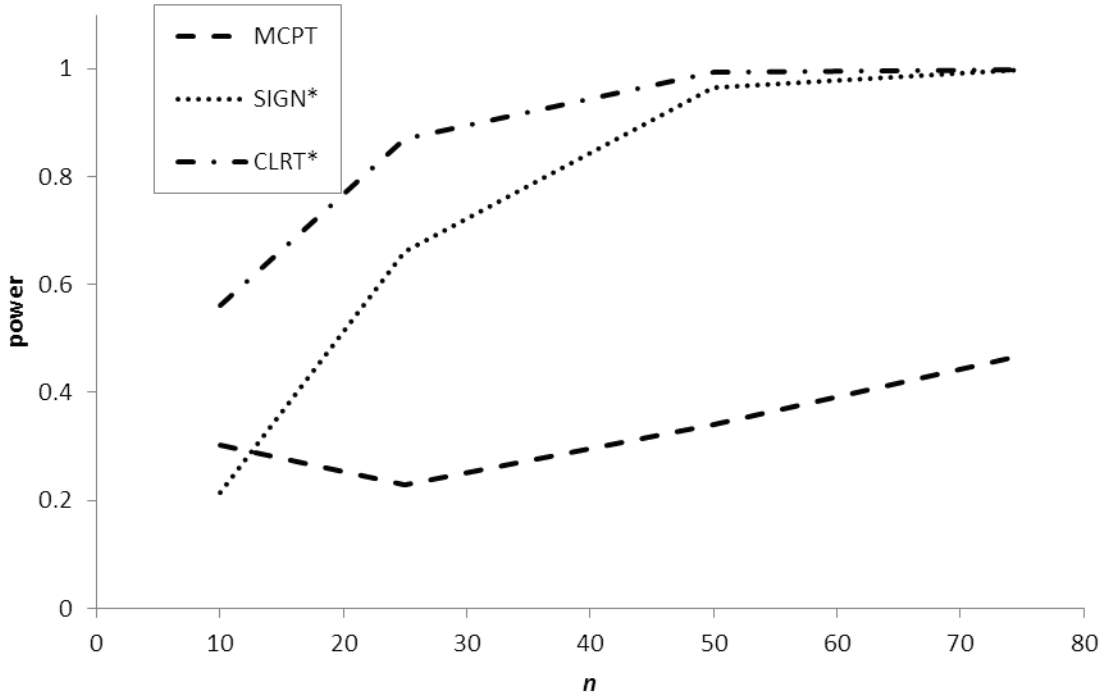
*These are not level-α tests.

Figure 4 (continued): Simulated Power for Test of Compound Symmetry vs. Serial Correlation ($p = 5, \sigma^2 = 1, \rho = 0.6$)*

c. Double Exponential



d. Two-Parameter Exponential



*These are not level-α tests.

parameter exponential data, even though the MCPT is slightly liberal, it is the best choice of the three tests given that the CLRT and SIGN test have simulated type I error rates that are much too high; however the MCPT in this case is not very powerful, only achieving a simulated power of 0.624 for $n = 50$.

Figure 4 displays the simulated power of the test of compound symmetry versus the serial correlation structure given by

$$\boldsymbol{\Sigma}_{SC} = \frac{\sigma^2}{1-\rho^2}\begin{bmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{p-1} \\ \rho & 1 & \rho & \cdots & \rho^{p-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{p-1} & \rho^{p-2} & \rho^{p-3} & \cdots & 1 \end{bmatrix}$$

where $\sigma^2/(1-\rho^2)$ is the common variance of the p variables and ρ is the correlation between successive observations of the variables. Figure 4 displays the power results for n = 10, 25, 50, 75, $p = 5$, $\sigma^2 = 1$, and $\rho = 0.6$.

Figure 4 is very similar to Figure 3 for the CLRT and SIGN test, but the MCPT appears to be less powerful at detecting serial correlation than heteroscedasticity. However, it is difficult to make direct comparisons between these two situations because the degree to which the simulated alternatives depart from compound symmetry cannot be quantified.

Application

Consider a data set reported in Monks, et al. (2004). In this study, 15 Centre d'Etude du Polymorphisme Humain (CEPH) families were selected and the expression for 23,499 genes was measured in lymphoblastoid cell lines; of these, 762 genes were found to be expressed and heritable. Three of the genes (NM_001081, NM_002125, and V00522) are known to have a linkage to the same location on chromosome 6; consequently, interest lies in determining whether there is a compound symmetry covariance structure with respect to these three genes. Among the 15 families included in the CEPH study there were 47 grandparents. These grandparents were the oldest generation included in the study; therefore, it is assumed that no genetic material is shared among them. Only the 47 grandparents were included in the analysis.

It is common in genetic studies to standardize gene expression data; therefore, the covariance and correlation matrices are equivalent. The sample covariance matrix among these three genes is estimated to be

$$\hat{\boldsymbol{\Sigma}} = \begin{bmatrix} 1 & 0.823 & 0.896 \\ 0.823 & 1 & 0.824 \\ 0.896 & 0.824 & 1 \end{bmatrix},$$

and the hypothesis to be tested is $H_0 : \boldsymbol{\Sigma} = \boldsymbol{\Sigma}_{CS}$ vs. $H_a : \boldsymbol{\Sigma} \neq \boldsymbol{\Sigma}_{CS}$. In all, $(3!)^{47} \approx 3.74 \times 10^{36}$ permutations of the raw data are possible. Consequently, a random sample of 10,000 permutations was selected for the MCPT. The p-values for the three tests are 0.9904 for the MCPT, 0.3042 for the CLRT and 0.0664 for the SIGN test. In each case, the null hypothesis would not be rejected at the 0.05 level, but the three p-values are very different. According to the Shapiro-Wilk test of multivariate normality, there is evidence that these data are not from a multivariate normal population ($p = 0.00016$), violating the assumptions of the CLRT. Given that the structure of $\hat{\boldsymbol{\Sigma}}$ does not deviate much from compound symmetry, it may also be speculated that the data may not have a multivariate elliptical distribution which could explain the unusually low p-value for the SIGN test.

Conclusion

With somewhat recent advances in technology permutation tests are becoming more feasible and – consequently – more common; this article proposed such a test for the compound symmetry covariance structure. Our simulation study indicates that the MCPT is robust to non-normality (more so when the data are symmetrically distributed), an issue with the CLRT, but is generally not as powerful as the CLRT when the data are normally distributed. The MCPT is also an improvement over the SIGN test in that the MCPT appears to be robust to non-elliptical distributions (again, more so when the data are symmetrically distributed).

One additional – and probably more common situation – that was not considered herein is the case of data sets in which the variables are not all equally distributed. Because the PT requires either independent observations or normally distributed observations with equal covariances for exchangeability, it is suspected that the PT would not perform well in this case, at least for extreme differences in distribution.

This article presented only the PT for the compound symmetry structure. According to Good (2005) this particular test requires multivariate normality and equal covariances for the exchangeability of the data. Evidence presented shows that this test is robust to departures from normality, but the situation of unequal covariances has not been addressed. A data transformation such that a PT for the structure of any covariance matrix can be achieved by applying the PT for compound symmetry to the transformed data is currently under development.

Another issue with the CLRT is that it does not exist for cases in which $p \geq n$. Although the PT exists in these cases, evidence exists to show that it is not a level-α test. Consequently, alternative test statistics are being considered that will alleviate this problem.

Acknowledgements

This research was supported in part by the Eugene and Doris Miller fellowship of Oklahoma State University and the Office of Research and Grants at the University of Central Oklahoma. The authors would also like to thank Dr. Mauricio Subieta and the High Performance Computer Center at Oklahoma State University for assistance with running the simulations for this study.

References

Aitkin, M. A. (1969). Some tests for correlation matrices. *Biometrika*, *56*, 443-446.

Aitkin, M. A., Nelson, W. C., & Reinfurt, K. H. (1968). Tests for correlation matrices. *Biometrika*, *55*, 327-334.

Boik, R. J. (1975). Interactions in the analysis of variance: A procedure for interpretation and a Monte Carlo comparison of univariate and multivariate methods for repeated measures designs. *Dissertation Abstracts International*, *36*, 2908B. (UMI No. 7527837).

Box, G. E. P. (1950). Problems in the analysis of growth and wear curves. *Biometrics*, *6*, 362-389.

Cornell, J. E., Young, D. M., Seaman, S. L., & Kirk R. E. (1992). Power comparisons of eight tests for sphericity in repeated measures designs. *Journal of Educational Statistics*, *17*, 233-249.

Dwass, M. (1957). Modified randomization tests for nonparametric hypotheses. *The Annals of Mathematical Statistics*, *28*, 181-187.

Falk, M. (1999). A simple approach to the generation of uniformly distributed random variables with prescribed correlations. *Communications in Statistics: Simulation and Computation*, *28*, 785-791.

Fisher, R. A. (1936). The coefficient of racial likeness and the future of craniometry. *The Journal of the Anthropological Institute of Great Britain and Ireland*, *66*, 57-63.

Gao, Y. & Marden, J. I. (2001). Some rank-based hypothesis tests for covariance structure and conditional independence. *Contemporary Mathematics*, *287*, 97-109.

Good, P. (2005). *Permutation, parametric and bootstrap tests of hypotheses* (*3rd Ed.*). New York, NY: Springer.

Hallin, M., & Paindaveine, D. (2006). Semiparametrically efficient rank-based inference for shape I. Optimal rank-based tests for sphericity. *The annals of statistics*, *34*, 2707-2756.

Huynh, H., & Mandeville, G.K. (1979). Validity conditions in repeated measures designs. *Psychological Bulletin*, *86*, 964-973.

Jöckel, K. H. (1986). Finite sample properties and asymptotic efficiency of Monte Carlo tests. *The Annals of Statistics*, *14*, 336-347.

Keselman, H. J., Rogan, J. C., Mendoza, J. L., & Breen, L. J. (1980). Testing the validity conditions of repeated measures *F* tests. *Psychological Bulletin*, *87*, 479-481.

Lee, J. C., Krishnaiah, P. R., & Chang, T. C. (1976). On the distribution of the likelihood ratio test statistic for compound symmetry. *South African Statistical Journal*, *10*, 49-62.

Manly, B. F. J. (1997). *Randomization and Monte Carlo methods in biology* (2nd ed.). London: Chapman & Hall.

Marden, J. I. (1999). Multivariate rank tests. In S. Ghosh (Ed.), *Multivariate analysis, design of experiments, and survey sampling*, 401-432. New York, NY: Marcel Dekker.

Marden, J., & Gao, Y. (2002). Rank-based procedures for structural hypotheses on covariance matrices. *Sankhyā: the Indian Journal of Statistics*, *64*, 653-677.

Mauchly, J. W. (1940). Significance test for sphericity of a normal *n*-variate distribution. *The Annals of Mathematical Statistics*, *11*, 204-209.

Monks, S. A., et al. (2004). Genetic inheritance of gene expression in human cell lines. *American Journal of Human Genetics*, *75*, 1094-1105.

R Development Core Team (2009). *R: A language and environment for statistical computing* (*Version 2.10.1*). Available from http://www.R-project.org.

Sirkiä, S., Nordhausen, K., & Oja, H. (2009). *SpatialNP: Multivariate nonparametric methods based on spatial signs and ranks* (*Version 1.0-1*). Available from http://CRAN.R-project.org/package=SpatialNP.

Sirkiä, S., Taskinen, S., Oja, H., & Tyler, D. E. (2009). Tests and estimates of shape based on spatial sings and ranks. *Journal of Nonparametric Statistics*, *21*, 155-176.

Vale, C. D., & Maurelli, V. A. (1983). Simulating multivariate nonnormal distributions. *Psychometrika*, *48*, 465-471.

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S* (*2nd Ed*.). New York: NY: Springer.

Votaw, D. F. (1948). Testing compound symmetry in a normal multivariate distribution. *The Annals of Mathematical Statistics*, *19*, 447-473.

Wilks, S. S. (1946). Sample criteria for testing equality of means, equality of variances, and equality of covariances in a normal multivariate distribution. *The Annals of Mathematical Statistics*, *17*, 257-281.

Journal of Modern Applied Statistical Methods
November 2011, Vol. 10, No. 2, 462-475

1538 – 9472/11/$95.00

Robust Inference for Regression with Spatially Correlated Errors

Juchi Ou Jeffrey M. Albert
Case Western Reserve University,
Cleveland, OH

A robust variance estimator for a regression model with spatially correlated errors is proposed using the estimated empirical covariogram. Simulations studies show unbiasedness and robustness for the OLS but not for the GLS estimates. The new robust variance estimation method is applied to hospital quality data.

Key words: Ordinary least squares, generalized least squares, robust variance estimation, hospital quality, semivariogram.

Introduction

In observational studies, an objective of interest is to compare the mean response of exposed and unexposed units. Commonly, the effect of an exposure or treatment on an outcome is evaluated via conventional linear regression models that assume independence of errors. For geographical data, observations and corresponding errors may be spatially correlated rather than independent. One unbiased estimator of an exposure effect in a linear regression model is the ordinary least squares estimator (OLS). This estimator is known to be the best linear unbiased estimator (BLUE) when the errors are independent with a constant variance. However, when errors are correlated, this estimator may be inefficient. Furthermore, its standard variance estimator may be biased. To improve precision for correlated data, methods that take into account the correlation structure, such as maximum likelihood (ML) estimation and generalized least squares (GLS) are of interest for evaluating an exposure effect.

Juchi Ou earned a Ph.D. in Biostatistics in 2010 from Case Western Reserve University. Email: jxo37@cwru.edu. Jeffrey M. Albert is an Associate Professor of Biostatistics in the Department of Epidemiology and Biostatistics at the School of Medicine. His research interests include longitudinal data analysis and causal inference. Email: jma13@case.edu.

A number of researchers have studied regression models with serially or spatially correlated errors. For example, Lee & Lund (2004) provided expressions for the OLS variances for autocorrelated errors and proposed confidence intervals based on their derived variance. The empirical coverage probabilities of their confidence intervals were close to the 95%target value when the sample size was large (at least 500). Athough Lee & Lund studied the variance for time series autocorrelation structures, their results require extension to regression models where errors are correlated in a space.

Basu & Reinsel (1994) compared the OLS and GLS estimators when errors follow a spatial unilateral first-order autoregressive moving average model; they found that the difference between variances of the two estimators were small unless the spatial correlation was close to 1. They investigated autocorrelation models; however, regression model errors could follow other spatial structures, such as a spatial Gaussian or spatial exponential model. Mardia and Marshall (1984) developed ML estimators for regression parameters in the spatial context assuming the errors follow a spatial Gaussian distribution.

A limitation of previous methods of inference for spatial data is that they rely on a correct specification of the covariance structure. When the covariance matrix is unknown, methods for variance estimation that are robust to covariance model misspecification are of interest. In the context of longitudinal data, a

well-known robust method to improve variance estimators for correlated data is the sandwich variance estimator (Diggle, et al., 2003). However, this estimator is not suitable for spatially correlated data that involve a single multivariate observation as opposed to multiple independent vectors. Furthermore, previous researches have given little attention to properties of estimators of the variance of effect estimates for spatially correlated errors.

This article develops estimators for mean differences along with robust variance estimators in a regression model with spatially correlated errors. A new robust (sandwich) variance estimator for exposure effects is proposed using the empirical variogram for spatially correlated errors. Although this approach may be applied to the maximum likelihood estimate, the focus here is on the methods of ordinary and generalized least squares. The appeal of the latter is that is has computational advantages over ML estimation and retains equivalent asymptotic efficiency (Charnes, et al., 1976).

The OLS and GLS estimators, along with the proposed versus standard variance estimators, are assessed via simulation studies. Simulation data were generated under either a spatial Gaussian or spatial exponential model, both of which are commonly used to analyze spatial data. As an applied example, data is analyzed to assess the effect of urban versus rural locations on the number of full-time equivalents (FTE) for registered nurses. Previous researchers investigating this question (Rosenblatt, et al., 2006; Jiang, et al., 2006) did not consider the spatial pattern of hospitals in assessing the difference in mean FTE. Therefore, the proposed methods are applied to consider the difference in mean FTE between urban and rural hospitals taking into account spatial correlations among hospitals. The data analyzed are from two databases: hospital financial reports from the Office of Statewide Health Planning and Development, and HCUP State Inpatient Databases (SID).

Methodology

Assume a linear regression model, standard (OLS and GLS) approaches for estimations of regression parameters and that the outcomes (Y(s)) and covariates (X(s)) at location s are linearly related. Also, the errors, e(s), for this linear regression model are allowed to be correlated, where s is an index for a spatial location. This model is as follows:

$$Y(s) = X(s)\beta + e(s);\ \ e(s) \sim N(0; \Sigma), \qquad (1)$$

where Σ represents the variance-covariance matrix for the error vector. The argument, (s), will be dropped for ease of notation.

For correlated errors, two common estimators of regression parameters (β) are the ordinary least squares (OLS) and the generalized least squares (GLS) estimators. The OLS estimator of regression parameters is

$$\hat{\beta}_{ols} = (X'X)^{-1}X'Y\ ; \qquad (2)$$

and the corresponding naïve variance estimator for $\hat{\beta}_{ols}$ is

$$Var(\hat{\beta}_{ols}) = \hat{\sigma}^2 (X'X)^{-1}\ , \qquad (3)$$

where $\hat{\sigma}^2$ is the sample variance of residuals. Another estimator of regression parameters is the GLS estimator,

$$\hat{\beta}_{gls} = (X'W^{-1}X)^{-1}X'W^{-1}Y\ , \qquad (4)$$

where W is the working matrix and it is equal to the estimated covariance matrix. The corresponding naïve variance estimator is

$$Var(\hat{\beta}_{gls}) = (X'W^{-1}X)^{-1}\ . \qquad (5)$$

Both the OLS and the GLS point estimators are unbiased, but the variance of the GLS estimator is smaller than that of the OLS estimator (Bloomfield & Watson, 1975) when W^{-1} is equal to the true covariance matrix. In the conventional, so-called naïve or model-based, approach, the covariance structure for the OLS variance estimator is assumed to follow the independence model whereas that for the GLS variance estimator is assumed to be proportional

to the working weight matrix W. In the context of longitudinal data, Liang & Zeger (1986) showed that the point estimator for β via generalized estimating equations (GEE) is consistent even if the correlation matrix is misspecified. However, when the assumed covariance structure is different from the true covariance model, the naïve variance estimator is inconsistent.

Robust Variance Estimator

The model-based variance estimators described above may be inadequate when the spatial covariance structure is unknown with the possibility of being misspecified. In the case of longitudinal data, where there are multiple measurements for each subject, a robust (sandwich) variance estimator is available (Diggle, et al., 2003). The robust variance estimator for the generalized least squares estimator $\hat{\beta}_{gls}$ is

$$\mathrm{Var}(\hat{\beta}_{gls})=(X'W^{-1}X)^{-1}X'W^{-1}\hat{V}W^{-1}X(X'W^{-1}X)^{-1}, \quad (6)$$

where $\hat{V}$ is a block-diagonal matrix with non-zero block $\hat{V}_0$ which may be estimated via restricted maximum likelihood estimation (REML). Letting Y_{hij} denote the j^{th} measurement on the i^{th} unit in the h^{th} group, the sample mean for the measurement j in group h is

$$\hat{\mu}_{hj}=\frac{1}{m_h}\sum_{i=1}^{m_h}Y_{hij}, h=1,...,g; i=1,...,m_h; j=1,...,n, \quad (7)$$

and the REML estimator is

$$\hat{V}_0=\left(\sum_{h=1}^{g}m_h-g\right)\sum_{h=1}^{g}\sum_{i=1}^{m_h}(Y_{hi}-\hat{\mu}_h)(Y_{hi}-\hat{\mu}_h)', \quad (8)$$

where $Y_{hi}=(Y_{hi1},...,Y_{hin})'$ and $\hat{\mu}_h=(\hat{\mu}_{h1},...,\hat{\mu}_{hn})'$. For this estimator, no assumption exists regarding the structure of means and covariance matrix.

In the case of longitudinal data where there are independent realizations of the correlated responses, sample estimates of the variance and covariance parameters are generally used to obtain the empirical estimate of V. For spatial data, there is only one (multivariate) observation and the above robust estimator would not be a good estimator. For this case, an empirical covariogram is used in place of the empirical variance-covariance matrix used for longitudinal data.

Variogram

Assume the spatial process to be second-order stationary and isotropic, where stationarity means that absolute coordinates are unimportant and isotropic means that the spatial correlations are the same in different directions (i.e., north-south versus west-east). For a spatial process Y(s): $s \in D \subset R^2$, one common tool to measure spatial correlations is the semivariogram for geostatistical data. The semivariogram $(\gamma^*(s_i, s_j) \equiv \gamma(s_i - s_j) = \gamma(h))$ is defined as a function of the distance (h) of two locations (s_i, s_j),

$$\gamma(h)=\frac{1}{2}\mathrm{Var}[Y(s_i)-Y(s_i+h)]. \quad (9)$$

If the spatial process (Y(s)) is second-order stationary, the semivariogram can be expressed in terms of the covariance function, $C(h)$, and

$$\gamma(h)=C(0)-C(h). \quad (10)$$

There are two important components for a semivariogram: the sill and the spatial range. The sill is defined as the asymptote of the variogram function, and the range is the distance at which the sill is reached.

Two commonly used variogram models are the spatial Gaussian and the spatial exponential models. Their covariance functions are as follows:

1. Gaussian model: $C_g(h)=\sigma^2\exp\{-(h/\alpha)^2\}$, and

2. Exponential model: $C_x(h)=\sigma^2\exp\{-(h/\alpha)\}$,

where α and σ^2 represent the spatial range and the sill, respectively, and h is the distance between two locations. The semivariograms for these two models are shown in Figure 1. As the distance increases, the semivariogram increases. The parameters $\theta \equiv (\alpha, \sigma^2)$ for a variogram model ($\gamma(h,\theta)$) may be estimated by iteratively reweighted least squares (IWLS) to minimize the following expression,

$$\sum |N(h)|(\hat{\gamma}(h)-\gamma(h,\theta))^2 , \qquad (11)$$

where N(h) is the number of distinct pairs of locations at distance h and $\hat{\gamma}(h)$ is an estimate of the semivariogram.

To avoid a parametric assumption regarding the spatial model, the moment-based empirical semivariogram could be used to estimate the semivariogram. The empirical (Matheron) semivariogram ($\hat{\gamma}$) for two observed measurements ($Y(s_i),Y(s_j)$) with distance h between two different locations (s_i,s_j) is

$$\hat{\gamma}(h)=\frac{1}{2|N(h)|}\sum_{N(h)}(Y(s_i)-Y(s_j))^2 , \qquad (12)$$

where |N(h)| is the number of measurement pairs with distance h. The corresponding empirical covariogram estimator for the covariance function, C(h) is as follows

$$\hat{C}(h)=\frac{1}{|N(h)|}\sum_{N(h)}(Y(s_i)-\bar{Y})(Y(s_j)-\bar{Y}) , \qquad (13)$$

where $\bar{Y}$ is the average of all Y(s). In this study, the empirical covariogram estimator is used to estimate the variance-covariance matrix.

Simulation Study

Data Generation

Using a 10x10 grid, two different covariance structures for the errors in Model 1 were studied: spatial Gaussian and spatial exponential. In general, the sill for a covariance structure varies from 0.01 to over 100. Therefore, the sill for both covariance structures was set to 9 in this study. The spatial ranges were set to 2, 5 or 10 in order to compare weak, modified and strong correlations between locations on a 10x10 grid. A binary covariate (X, with values 0 and 1) was generated from the binomial distribution with probability of X = 1 equal to 0.5 and the outcome (Y) was generated from the linear model

$$Y = 2X + \varepsilon, \qquad (14)$$

that is, the outcome was linearly related with the binary covariate with slope 2.

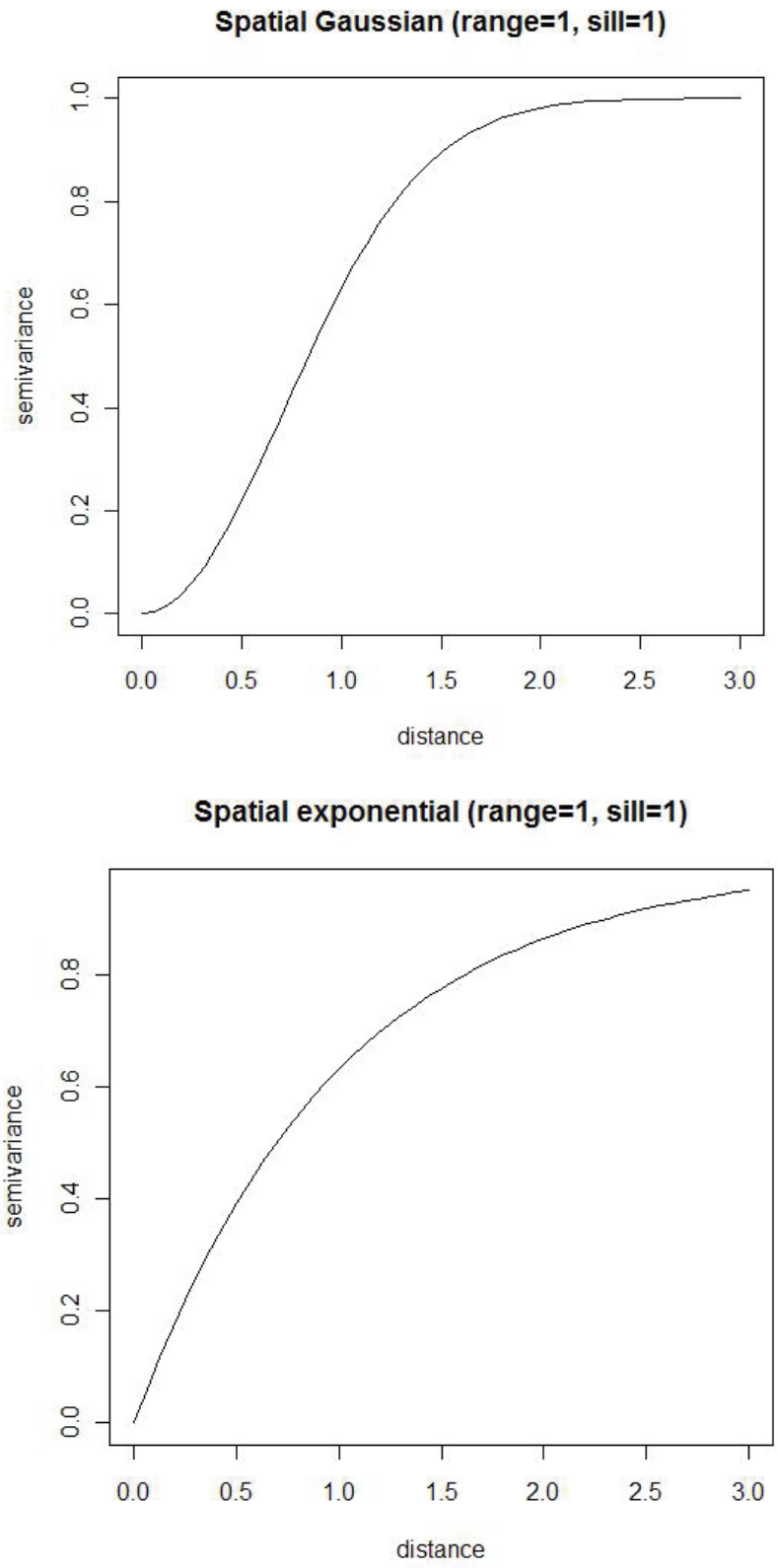
Estimator of the Exposure/Treatment

Two point estimators for the exposure/treatment effect were studied, namely, OLS (ordinary least squares) and GLS (generalized least squares) estimators. In addition, the working matrix of the GLS estimator was estimated based on either independence (OLS residuals), spatial Gaussian or spatial exponential.

Variance Estimator of the Treatment Effect

The naïve variance estimators as well as the sandwich variance estimators were evaluated. For the sandwich variance estimator, the variance-covariance matrix could be the spatial Gaussian $\left(\hat{C}_g\right)$, spatial exponential $\left(\hat{C}_x\right)$ or the spatial empirical covariance structure $\left(\hat{C}\right)$. The variance estimators for the OLS point estimator are as follows: independence, $\hat{\sigma}^2(X'X)^{-1}$; empirical, $(X'X)^{-1}X'\hat{C}X(X'X)^{-1}$; Gaussian, $(X'X)^{-1}X'\hat{C}_gX(X'X)^{-1}$; and Exponential: $(X'X)^{-1}X'\hat{C}_xX(X'X)^{-1}$.

Figure 1: Semivariogram Models



Where $\hat{C}$, $\hat{C}_g$ and $\hat{C}_x$ represent the spatial empirical covariance, the estimated spatial Gaussian covariance and the estimated spatial exponential covariance matrices. The variance estimators for the GLS point estimator are naïve, $(X'W^{-1}X)^{-1}$, and empirical, $(X'W^{-1}X)^{-1}X'W^{-1}\hat{C}W^{-1}X(X'W^{-1}X)^{-1}$, where W^{-1} would be either the spatial Gaussian or the spatial exponential covariance matrix, and $\hat{C}$ is the empirical covariance matrix.

The bias and MSE of the OLS and GLS point estimators of the regression coefficient were computed. The bias and MSE for 1,000 replications are obtained as

$$\text{Bias} = \frac{1}{1000}\sum(\hat{\beta}_i - 2), \qquad (15)$$

$$\text{MSE} = \frac{1}{1000}\sum(\hat{\beta}_i - 2)^2. \qquad (16)$$

In addition, the relative bias for each estimator ($\hat{\theta}$, that is, $\hat{\beta}$ or $\hat{V}(\hat{\beta})$) was calculated. This relative bias is defined as

$$\text{RB} = \frac{\hat{\theta}-\theta}{\theta}. \qquad (17)$$

Results

Spatial Gaussian Errors Data: OLS

The bias of the ordinary least squares estimator (OLS) and its corresponding variance estimator, in the case where the errors are spatially correlated over a 10 * 10 grid, are shown in Table 1. When the covariance matrix for errors is spatial Gaussian distributed, the bias of the OLS estimator is smaller (closer to 0.01) for all examined spatial ranges. The corresponding MSE decreases as the spatial range increases. Among the four variance estimators, the estimator using the independence covariance structure has the largest difference from the true variance for each spatial range. As the strength of spatial correlation (that is, the range) increases, the bias of the independence variance estimator increases. Both the empirical and the Gaussian variance estimators underestimate the variance. In addition, the empirical estimated variance is closer to the true value than the two estimators based on incorrect covariance models (independence and exponential) and has similar bias to the estimator using the correct covariance model (Gaussian), over varying range values.

Spatial Gaussian Errors Data: GLS

Working weight matrices for the GLS estimator based on the Gaussian and the exponential spatial covariance models were considered. The results for the Gaussian and exponential working matrices are shown in Table 2. For the Gaussian working matrix, the bias of the estimated effect is small for the each strength of the spatial correlations. The bias for the Gaussian working matrix is reduced at least 80% from the OLS estimators. The bias of the naïve estimated variance is smaller than that of the empirical estimator when the true working matrix (Gaussian model) was fit. However, as the spatial correlation increases, the relative bias of the naïve and empirical variance become more similar. When the exponential working matrix is used for the spatial Gaussian errors data, the biases of the GLS estimated effect are also small, and the bias is reduced at least 46.4% from the OLS estimators. In this case, the naïve and empirical variance estimators both have large biases which are similar in magnitude.

Spatial Exponential Errors Data: OLS

A second simulation involved the generation of spatial exponential errors. The bias and MSE for the ordinary least squares estimators (OLS) and its corresponding variance estimators are shown in Table3. The bias of the estimated effect is smaller than 0.005 for all examined spatial ranges. The independence estimator overestimates the variance of the effect for all examined spatial ranges and the spatial empirical estimator slightly underestimates the variance. The spatial empirical estimated variance is closer to the true value than the other estimated variances. The exponential variance estimator for the OLS estimator, though it uses the correct covariance model, underestimates the variance for all examined spatial ranges. The Gaussian variance estimator overestimates the variance when the spatial range is larger than 5.

Spatial Exponential Errors Data: GLS

For the spatial exponential errors data, two working weight matrices for the generalized least squares (GLS) estimator are considered: the spatial Gaussian and the spatial exponential covariance models. The results for the GLS effect estimators are shown in Table 4. For both Gaussian and exponential working matrices, the biases of estimated effects are smaller than 1% for all examined spatial ranges. When data are spatial exponential correlated across a study space (spatial range at 10), the biases of the GLS effect estimators are smaller than that of the OLS estimator. The bias reduction is 37.1% for a strongly spatial correlation. For the spatial exponential errors data, the relative bias decreases as the spatial range increases. When the spatial correlation (spatial range) increases, the MSE decreases.

For simulated data with exponential errors, the naïve (based on the correct working covariance matrix) and empirical variance estimates have positive biases for all examined spatial ranges. The bias of the naïve estimated variance is smaller than that of the empirical estimated variance. For all examined spatial correlations, the MSE of the GLS with incorrect (Gaussian) working matrix is larger than corresponding MSE of the GLS with correct (exponential) working matrix for the spatial exponential errors data.

Table 1: OLS-Bias and Variance Estimator for Spatial Gaussian Errors for 1,000 Replications

Range	OLS-Bias	MSE			Variance		
			TRUE	Indep*	Em*	Gau* (correct)	Ex*
2	0.0069	0.339	0.354	0.334	0.343	0.346	0.342
5	0.0108	0.136	0.146	0.222	0.146	0.132	0.142
10	0.0103	0.033	0.033	0.096	0.031	0.030	0.060

*indep: independent; Em: empirical; Gau: Gaussian; Ex: exponential

Table 2: GLS Bias and Variance Estimator for Spatial Gaussian Errors for 1,000 Replications

Range	Gaussian Working Matrix (Correct)					
	GLS-Bias	RB*	MSE	Variance		
				True(sim)*	Naïve	Em*
2	0.00138	-80.00%	0.0091	0.0091	0.0256	4.4886
5	0.00004	-99.60%	0.0020	0.0019	0.0019	0.9598
10	0.00089	-91.40%	0.0004	0.0004	0.0007	0.0333

Range	Exponential Working Matrix (Incorrect)					
	GLS-Bias	RB*	MSE	Variance		
				True(sim)*	Naïve	Em*
2	-0.0037	-46.40%	0.0238	0.0238	0.0928	0.0982
5	0.0005	-95.40%	0.0014	0.0014	0.0565	0.0314
10	0.0003	-97.10%	0.0008	0.0008	0.0403	0.0121

*RB: relative bias; True(sim): simulated variance; Em: empirical

Example

Background

A common cause of adult hospitalization is pneumonia. Several pneumonia inpatient management measures are provided by the Centers for Medicare & Medicaid Service. Among these quality measures, a blood culture prior to first antibiotic administration is recommended (Waterer & Wunderink, 2001; Metersky, et al., 2004). For care services in the hospitals, nurse staffing plays an important role. Kovner, et al. (2000, 2002) found that lower nurse staffing levels resulted in significantly higher rates of pneumonia. Rosenblatt, et al. (2006) and Jiang, et al. (2006) showed that the full-time equivalent (FTE) for registered nurses were significantly different between rural and urban community health centers in the US. However, although these studies assumed the hospital outcomes to be independent, they did not take into account possible spatial correlations among hospitals.

Data Source and Sample

This research is interested in examining the association between the FTEs for registered nurses and hospital location (urban versus rural). In general, one FTE represents 2,080 work hours

Table 3: OLS-Bias and Variance Estimator for Spatial Exponential Errors for 1,000 Replications

Range	OLS-Bias	MSE	Variance				
			TRUE	Indep*	Em*	Gau*	Ex* (correct)
2	0.0026	0.277	0.307	0.317	0.302	0.301	0.300
5	0.0041	0.171	0.185	0.223	0.185	0.187	0.177
10	0.0035	0.099	0.106	0.143	0.106	0.110	0.104

*indep: independent; Em: empirical; Gau: Gaussian; Ex: exponential

Table 4: GLS Bias and Variance Estimator for Spatial Exponential Errors for 1,000 Replications

Range	Gaussian Working Matrix (Incorrect)					
	Bias	RB*	MSE	Variance		
				True(sim)*	Naïve	Em*
2	-0.0037	42.30%	0.135	0.135	0.147	0.187
5	-0.0042	2.40%	0.056	0.056	0.061	0.091
10	-0.0032	-8.60%	0.029	0.029	0.030	0.050

Range	Exponential Working Matrix (Correct)					
	Bias	RB*	MSE	Variance		
				True(sim)*	Naïve	Em*
2	-0.0033	26.90%	0.130	0.130	0.146	0.187
5	-0.0031	-24.40%	0.054	0.054	0.066	0.092
10	-0.0022	-37.10%	0.027	0.027	0.034	0.050

*RB: relative bias; True(sim): simulated variance; Em: empirical

within a year to a fulltime worker. Here, the outcome of interested was FTEs for registered nurse per occupied bed. Data for this outcome, available in hospitals financial reports, was provided by the Office for Statewide Health Planning and Development (OSHPD). The binary predictor, hospital location (urban/rural), was taken from the Healthcare Cost and Utilization Project (HCUP) California State Inpatient Database (SID); this predictor was denoted as location. In addition, the report for pneumonia quality measures of inpatient management was provided by the Centers for Medicare & Medicaid Service. Data was merged from these three sources restricting the sample to hospitals in the State of California in 2004. The resulting dataset included 186 hospitals that reported: the above pneumonia quality measure, the number of registered nurse FTEs per occupied bed and hospital location.

The spatial correlation for each model variable was assessed via the test by Diblasi & Bowman (2001). The semivariograms of the response (FTE) and predictor (location) with their corresponding p-value of the spatial correlation test are shown in Figure 2. Both variables were spatially correlated across hospitals in California in 2004.

OLS Result

The effect of hospital location on the number of FTEs for registered nurses was estimated using the ordinary least squares (OLS). OLS estimates, the independence variance estimate,and three spatial variance estimates (empirical, spatial Gaussian, exponential structure) are shown in Table 5, along with standardized effect estimates (estimated effect divided by the square root of the estimated variance). The OLS estimated mean difference for FTE between urban and rural hospitals was 0.3018. The independence and spatial empirical variance estimates were close and both were less than 0.1. These two variance estimators both provided standardized effect estimates greater than 3.9. The spatial Gaussian and exponential variance estimates were larger, and their respective standardized estimates of 2.2 and 1.99, smaller than the other two estimates. Thus, all methods indicated an effect of the hospital locations on FTE with higher mean FTEs at the urban hospitals. The standardized effects based on the spatial Gaussian and spatial exponential estimated variances suggested marginal evidences; by contrast, the standardized effects based on independence and the empirical estimated indicated strong evidences of a location effect. The conclusions, based on California hospitals, are substantially the same as previous study results for United States health centers.

The semivariograms of OLS residuals are shown in Figure 3. The line in the left figure is the fitted spatial Gaussian structure with estimated spatial range and sill equal to 0.43 and 0.08. The line in the right figure is to the fitted spatial exponential structure with estimated range and sill equal to 0.50 and 0.11. Both theoretical semivariogram models (i.e., Gaussian and exponential) were close to empirical semivariogram when the distance was smaller than 2. However, these two models were far from empirical semivariogram when the distance was larger than 2.

GLS Result

For comparison, GLS estimators were considered under the same models as examined for the OLS estimators. Thus, estimated spatial Gaussian and exponential structures were used as the working weight matrices for GLS estimators. The results for the point and variance estimates are shown in Table 6. Compared to the OLS estimated effects, the two GLS estimated effects were larger. For each working weight matrix, both the naïve and the empirical variance estimates were less than 0.01. The empirical variance estimate was smaller than the naïve estimated variance for both the Gaussian and exponential working matrices. All three GLS standardized effect estimates were greater than 3.5 and one of them was as high as 3.73. All GLS standardized effect estimates indicated strong evidences of an effect of location on FTE, with a higher mean FTE at urban hospitals. Thus, the conclusion based on the GLS estimators with either a spatial Gaussian or exponential working matrices, agree with that given above for the OLS estimators.

Figure 2: Semivariograms of Response and Predictor

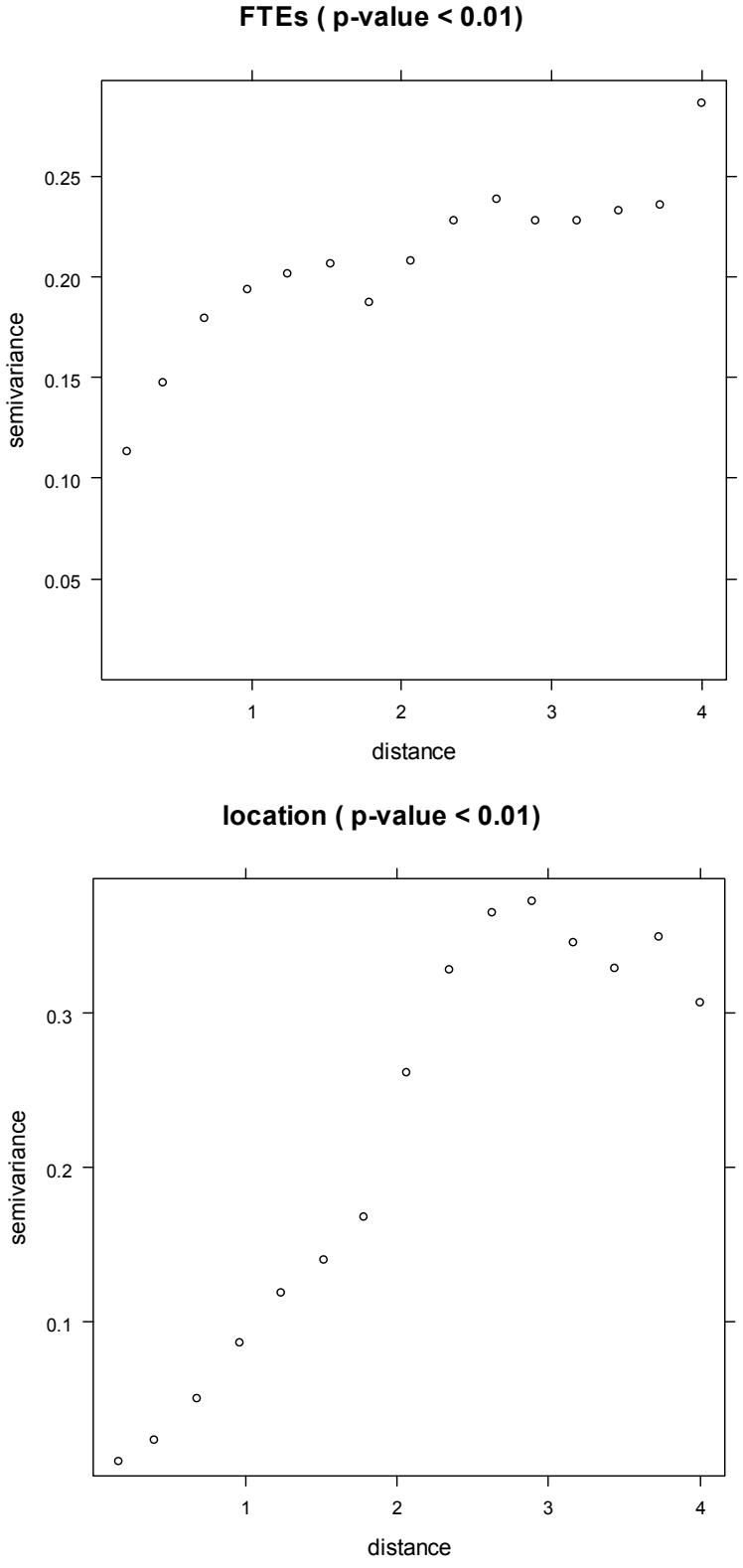


Table 6: GLS Effect Estimator and Its Estimated Variance (STD)*

	Working Matrix	
	Gaussian	Exponential
Estimated Effect	0.3255	0.3396
Naïve Variance	0.0081(3.62)	0.0089(3.60)
Empirical Variance	0.0076(3.73)	0.0085(3.68)

*Standardized effect estimates are in parentheses

Conclusion

This article addresses the problem of estimating exposure (or treatment) effect in a regression models with spatially correlated errors. Considering both OLS and GLS estimators, a new robust variance estimator was presented based on the estimated semivariogram. In order to evaluate the OLS and GLS estimators or their corresponding variance estimators under spatial correlated errors, simulation studies were conducted. Two different spatial correlation models were considered: spatial Gaussian and spatial exponential.

For spatial Gaussian and exponential simulated data, neither the OLS nor GLS estimators showed evidence of bias. When the spatial range increased, the true variance decreased. For the OLS estimator, the bias of the naïve (independence) estimated variance was smallest at spatial range 2 among three spatial ranges. The empirical estimated variance for the OLS estimator was closer to the true value than the other three estimated variances. For the GLS estimator, the naïve estimated variance was closer to the simulated variance than the empirical estimated variance. However, when the GLS estimator used an incorrect working matrix, the naïve estimated variance would be far from the simulated variance (e.g., GLS with an exponential working matrix for spatial Gaussian errors data). In addition, even when the correct working matrix is used, the estimated variance of the GLS estimate sometimes varied substantially from the true (simulation) value. Therefore, estimating exposure effects via ordinary least squares (OLS) with the empirical variance estimator is recommended when the data exhibit spatial patterns.

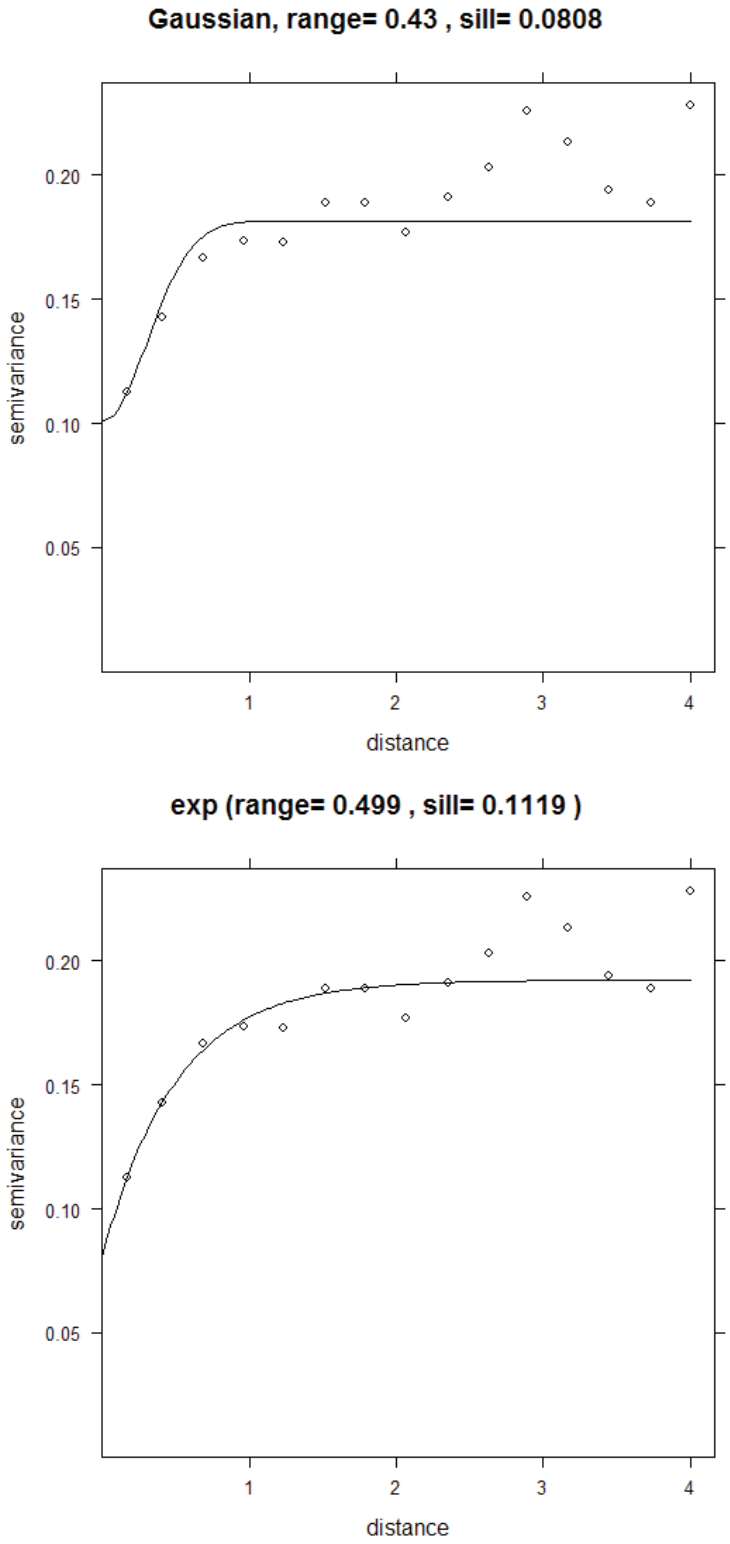
The effect of hospital locations on FTE where both variables exhibited spatial patterns (based on their empirical semivariogram and spatial correlation test) across California in 2004

Table 5: OLS Effect Estimate, Variance Estimates and Standardized Effect Estimates (STD)*

	Effect	Variance			
		Indep**	Empirical	Gaussian	Exponential
Estimate	0.3018	0.0059	0.0044	0.0184	0.0231
STD		3.9291	4.5498	2.2249	1.9857

*STD: the effect estimate divided by the square root of the variance estimate;
**Indep: independence covariance structure

Figure 3: Semivariograms of OLS Residuals



was examined. The linear relationship between hospital location (urban/rural) and full-time-equivalents (FTE) for registered nurse adjusted by the number of occupied beds was assessed via the OLS and the GLS estimators. From the semivariogram of the OLS errors, the OLS errors exhibited a spatial pattern. Therefore, the OLS estimated effect with corresponding empirical variance was preferred. Based on OLS, the estimated difference between urban and rural hospitals was 0.3 FTE. The empirical estimated variance for the OLS estimator was around 0.004 and the ratio of estimated effect to the square root of empirical variance was 4.55. This result, corroborating the previous findings, suggests that there is a significant difference in FTE for urban versus rural hospitals.

The robust approach proposed could be used with the maximum likelihood estimates, though results are expected to be similar to GLS. A limitation of this study is that it assumed the spatial field to be stationary. For a non-stationary field, semivariogram models are not valid as the semivariogram is not defined for non-stationary correlation structures. Another limitation is that the outcome was assumed to be continuous and normally distributed. For a categorical or other non-normally distributed outcome, the linear regression would not be suitable. It will be necessary to use the logistic regression or to do a Box-Cox transformation for such outcomes. In addition, for some extreme values, the Cressie-Hawkins robust estimator could be considered for the estimation of the semivariogram (Cressie & Hawkins, 1980) instead of the Matheron estimator. The empirical covariogram used is a biased estimator of the covariance function; therefore, the problem of the biased estimator of the covariogram will need to be solved in the future.

Acknowledgement

The authors would like to thank Mireya Diaz for directing the first author to this area of research and for helpful comments on an earlier draft.

References

Basu, S., & Reinsel, G. C. (1994). Regression models with spatially correlated errors. *Journal of the American Statistical Association*, *89*, 88-99.

Bloomfield, P., & Watson, G. S. (1975). The inefficiency of least squares. *Biometrika*, *62*, 88-116.

Cressie, N., & Hawkins, D. M. (1980). Robust estimation of the variogram: I. *Journal of the International Association for Mathematical Geology*, *12*, 115-125.

Charnes, A., Frome, E. L., & Yu, P.L. (1976). The equivalence of generalized least squares and maximum likelihood estimates in the exponential family. *Journal of the American Statistical Association*, *71*, 214-222

Diblasi, A., & Bowman, A. W. (2001). On the use of the variogram in checking for independence in spatial data. *Biometrics 57*, 211–218.

Diggle, P. J., Heagerty, P., Liang, K-Y, & Zeger, S. L. (2003). *Analysis of Longitudinal Data* (*2nd Ed.*). Oxford Statistical Science Series.

HCUP State Inpatient Databases (Sid). (2003-2004). *Healthcare Cost And Utilization Project* (*HCUP*). Agency for Healthcare Research and Quality, Rockville, MD. www.hcupus.ahrq.gov/sidoverview.jsp.

Jiang, H. J., Stocks, C., & Wong, C. J. (2006). Disparities between two common data sources on hospital nurse staffing. *Journal of Nursing Scholarship*, *38*, 187-193.

Kovner, C., Jones, C., Zhan, C., Gergen, P. J., & Basu, J. (2002). Nurse staffing and postsurgical adverse events: An analysis of administrative data from a sample of U.S. hospitals, 1990-1996. *Health Services Research*, *37*, 611-629.

Kovner, C., Mezey, M., & Harrington, C. (2000). Research priorities for staffing, case mix and quality of care in the U.S. nursing homes. *Journal of Nursing Scholarship*, *32*, 77-80.

Lee, J., & Lund, R. (2004). Revisiting simple linear regression with autocorrelated errors. *Biometrika*, *91*, 240-245.

Liang, K., & Zeger, S. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, *73*, 13-22.

Mardia, K. V., & Marshall, R. J. (1984). Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika*, *71*,135-46.

Metersky, M. L., Ma, A., Bratzler, D. W., & Houck, P. M. (2004). Predicting bacteremia in patients with community-acquired pneumonia. *American Journal of Respiratory and Critical Care Medicine*, 169, 342-347.

Rosenblatt, R. A., Andrilla, C. H. A., Curtin, T., & Hart, L. G. (2006). Shortatges of medical personnel at community health centers. *Journal of the American Medical Association*, *295*, 1042-1049.

Waterer, G. W., & Wunderink, R. G. (2001). The influence of the severity of community-acquired pneumonia on the usefulness of blood cultures. *Respiratory Medicine*, 95, 78-82.

Journal of Modern Applied Statistical Methods
November 2011, Vol. 10, No. 2, 476-493

1538 – 9472/11/$95.00

Probabilistic Inferences for the Sample Pearson Product Moment Correlation

Jeffrey R. Harring John A. Wasko
University of Maryland,
College Park, MD

Fisher's correlation transformation is commonly used to draw inferences regarding the reliability of tests comprised of dichotomous or polytomous items. It is illustrated theoretically and empirically that omitting test length and difficulty results in inflated Type I error. An empirically unbiased correction is introduced within the transformation that is applicable under any test conditions.

Key words: Correlation coefficients, measurement, test characteristics, reliability, parallel forms, test equivalency.

Introduction

It has been well-established that the sample correlation coefficient, *r*, is a biased estimator of the population correlation coefficient, ρ, for normal populations, and this bias can be as much as 0.05 in absolute value under realistic research conditions (Zimmerman, Zumbo & Williams, 2003). This difference may not be vital if the research question is to simply ascertain whether a non-zero correlation exists. However, if the focus is on a precise estimate of the magnitude of a non-zero correlation in test and measurement procedures, then this discrepancy may be of concern. The Pearson product moment correlation is still commonly used as an index of reliability, exampled with parallel test forms (Coleman, 2001), test-retest conditions (Robinson-Kuropius, 2005), and inter-rater consistency (Lebreton, 2007). In such cases, calculations use a total score comprised of dichotomous or polytomous items (Kline, 2005). With increasing frequency, practitioners working in these contexts recognize sample estimates are insufficient and, therefore, are correctly utilizing the Fisher transformation to provide accompanying probabilistic inferences (Fouladi, 2002).

The motivation for this study centers on the failure of Fisher's transformation to incorporate either test length or test difficulty into confidence interval calculations. Without correction, test statistics and confidence intervals from utilizing the Fisher transformation become increasingly imprecise ultimately resulting in inflated Type I error. To date, research has neither demonstrated the inefficiencies of utilizing this method, nor further advocated a test statistic inclusive of test properties upon which to draw more accurate inferences about the population. In this article, an empirical demonstration of systemic errors between the empirical distribution and the Fisher transformation is presented which can be traced to test properties of length and difficulty. Based on the results, a correction factor inclusive of test properties is introduced and examined using a Monte Carlo simulation study to explore the performance of the corrected statistic to the existing Fisher transformation.

Jeffrey Harring is an Associate Professor in the Department of Measurement, Statistics and Evaluation. Email him at: harring@umd.edu. John Wasko is a Colonel in the U.S. Army. Email him at: john.wasko@us.army.mil.

Methodology

Pearson Correlation

The Pearson's correlation coefficient is a measure of the strength of the linear relation between two continuous variables and is defined as

$$\rho = \rho(\mathbf{x},\mathbf{y}) = \frac{Cov(\mathbf{x},\mathbf{y})}{\sigma_{\mathbf{x}}\sigma_{\mathbf{y}}} \quad (1)$$

where $\mathbf{x}$ and $\mathbf{y}$ are vectors of scores of size $n, Cov(\mathbf{x},\mathbf{y})$ represents the population covariance and $\sigma_{\mathbf{x}}$ and $\sigma_{\mathbf{y}}$ are population standard deviations. Invariably researchers report a point estimate for reliability using the form

$$\hat{\rho} = \rho(\mathbf{x},\mathbf{y}) = r = \frac{s_{\mathbf{xy}}}{s_{\mathbf{x}}s_{\mathbf{y}}},$$

where $s_{\mathbf{xy}}$, $s_{\mathbf{x}}$ and $s_{\mathbf{y}}$ are sample statistics corresponding to the population quantities in (1). For test-retest reliability let,

$$\mathbf{x} = (x_{T1},\ldots,x_{Tn}) \overset{iid}{\sim} N(\mu_{xT},\sigma^2_{xT})$$
$$\mathbf{y} = (y_{T1},\ldots,y_{Tn}) \overset{iid}{\sim} N(\mu_{yT},\sigma^2_{yT})$$

represent the total scores of n respondents administered the same test on different occasions. For parallel forms, let

$$\mathbf{x} = (x_{A1},\ldots,x_{An}) \overset{iid}{\sim} N(\mu_{xA},\sigma^2_{xA})$$
$$\mathbf{y} = (y_{B1},\ldots,y_{Bn}) \overset{iid}{\sim} N(\mu_{yB},\sigma^2_{yB})$$

represent the total scores of n respondents administered different tests on different occasions. By letting A and B represent two raters scoring the same test for n respondents would constitute inter-rater reliability. Particular to test-retest and parallel forms, it is assumed that no learning has occurred as a result of the first exam or in the interim prior to administration of the second exam.

Central Limit Theorem Application

The Pearson's correlation coefficient assumes total scores to be normally distributed; this is made possible by the central limit theorem (CLT) (see Hogg & Craig, 1995 for a full description). Reviewing its application, if $i_1, i_2, \ldots i_J$ represent the scores for a test of J items, independent and identically distributed from any distribution, then their sum

$$i_1 + i_2 + \ldots + i_J = T_0 \sim N(J\mu, J^2\sigma^2)$$

is approximately normal for sufficiently large values of J. Although sufficiently large is not a quantifiable number, this requirement is important given the need for a bivariate normal distribution upon which correlation inferences are predicated (Quereshi, 1971). A rule of thumb of J exceeding 30 items has been suggested. Not to be overlooked are the other requirements for use of the CLT. First is the requirement of independence. Conditional independence is assumed, where the likelihood a respondent answers an item correctly or incorrectly is independent of their response to any other test item. Second is the concept of identically distributed, where the collection of J items should all be dichotomously scored, $i = [0,1]$, or polytomously scored $i = [0,1,\ldots,R]$.

Even if the total score is well approximated by a normal distribution, the total score random variable is still discrete. In such cases, when making probabilistic inferences with a continuous distribution with discrete data, a continuity correction is often applied (Devore, 2000). Recall that Pearson's correlation is designed for continuous random variable pairs that follow a bivariate normal distribution. Without a sufficient number of J items, the total score distributions depart from univariate normality.

This condition is further exacerbated in extremely easy or difficult shorter tests resulting in highly skewed total scores; although this becomes less of an issue as test length increases, test difficulty affects the rate of asymptotic convergence to a normal distribution. Further, the total score variable is not continuous, it is discrete. With all statistics, when underpinning assumptions are violated, the accuracy of the results becomes increasingly questionable. Such inaccuracies are often commensurate with inflated Type I error rates. It is within this framework that the need for an item-type correction encompassing test length and difficulty and a continuity correction may be advocated.

Fisher Transformation

With

$$(x_1, y_1), (x_2, y_2), ..., (x_n, y_n) \sim N(\mu, \Sigma),$$

following a bivariate normal distribution, define a random variable Z as

$$Z = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right),$$

approximated by the following normal distribution characterized by its mean and variance

$$Z \sim N\left(\frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right),\ \frac{1}{n-3}\right).$$

Being normally distributed, these relations can be used in the traditional construction of confidence intervals and hypothesis tests. The transformation of the *r* random variable is called the Fisher transformation; the immediate discussion centers on confidence intervals, presentation of appropriate hypothesis tests are provided later.

A 2-sided $(1-\alpha)\%$ confidence interval for the true correlation, ρ, is obtained via the following steps:

1. Determine the $(1-\alpha)\%$ confidence interval for *Z* such that

   $$(1-\alpha)\%\ CI = (Z_L, Z_U)$$

   where

   $$Z_L = Z + \frac{1}{\sqrt{n-3}}\Phi^{-1}\left(\alpha/2\right)$$

   and

   $$Z_U = Z + \frac{1}{\sqrt{n-3}}\Phi^{-1}\left(1-\alpha/2\right).$$

2. Create a $(1-\alpha)\%$ confidence interval for ρ by transforming these *Z* confidence limits back onto the correlation scale

$$(1-\alpha)\%\ CI = \left(\frac{\exp(2Z_L)-1}{\exp(2Z_L)+1}, \frac{\exp(2Z_U)-1}{\exp(2Z_U)+1}\right).$$

Empirical Demonstration of Theoretical Findings

To illustrate the need to account for the number of test items for asymptotic convergence to a normal distribution, two empirical experiments are conducted. Conditions for the first simulation are a test length of $J = 25$ items, a population correlation of $\rho = 0.8$, administered to $n = 100$ respondents, where each item is an independent dichotomous response with a *p*-value of 0.60.

Conditions for the second simulation are $J = 35$, $\rho = 0.7$, $n = 100$, and a *p*-value of 0.70. For each simulation, responses for *J* items for respondent *i* ($i = 1, 2, \ldots, n$) were created according to a particular *p*-value representing a test. A second set of responses, representing a second test, were created such that each item was correlated with its first test equivalent according to a particular ρ. The item scores were totaled for each test for each respondent, resulting in a paired set of total scores of length *n*. A correlation estimate was calculated and retained for this set of total scores and, using the Fisher transform, two-sided 90% and 95% confidence intervals were calculated. Knowing the true ρ, each interval was evaluated to determine if it encompassed the true value, successes were noted. This was repeated for 10,000 trials for each experimental condition, the percentage of these successes estimates the coverage probability. Success percentages below the $(1-\alpha)\%$ specification indicates an inflated Type I error (the probability of rejecting a correct null hypothesis).

For each simulation, every sample correlation value was transformed to a *Z* random variable. A histogram of the sampling distribution is overlaid with the Fisher transform. Sampling distributions for 3rd and 4th moment statistics are provided on each plot including coverage probabilities.

Clearly, a snapshot exploring just two experimental conditions does not provide

Figure 1: Empirical *Z*-Scaled Histogram with Fisher Transform Overlay
10,000 trials, $\rho = 0.8$, n =100, test length $J = 25$, p-value = 0.6

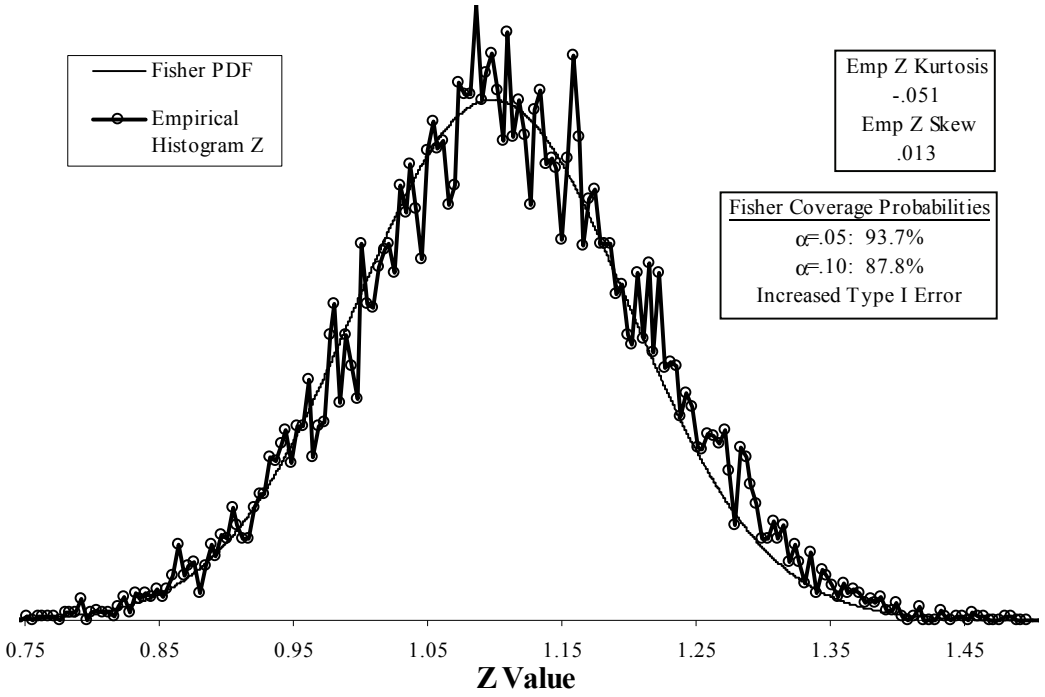
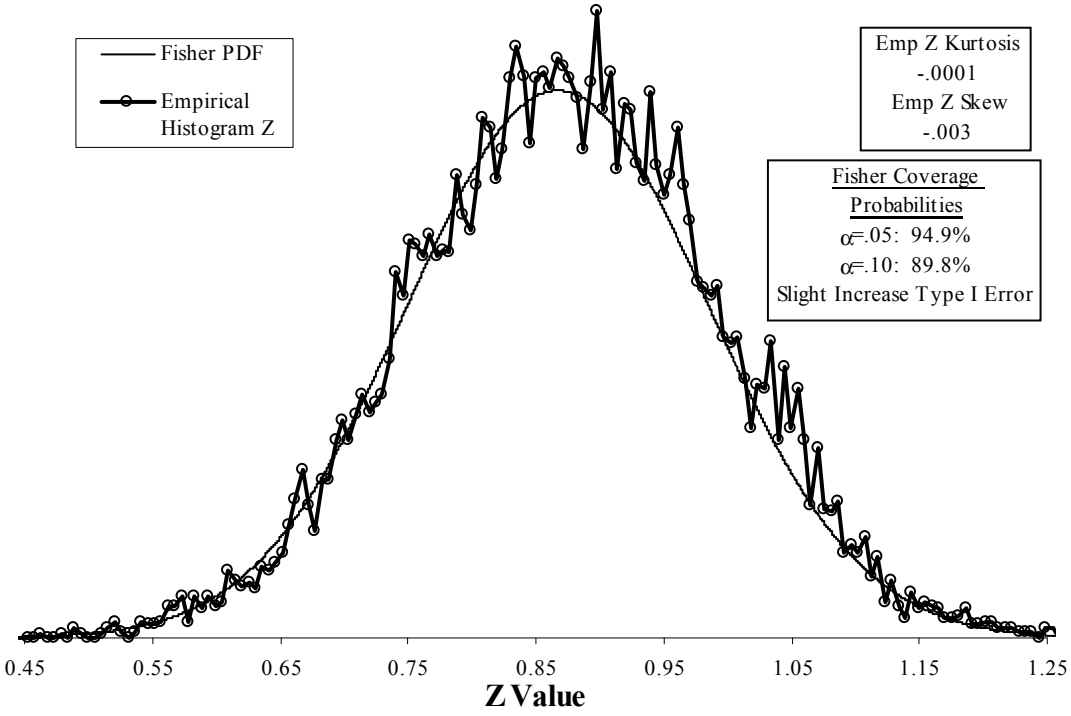


Figure 2: Empirical *Z*-Scaled Histogram with Fisher Transform Overlay
10,000 trials, $\rho = 0.7$, n =100, test length $J = 35$, p-value = 0.7



irrefutable evidence; but results highlight areas requiring further exploration.

1. The transformation of the sample correlation remains well characterized by a normal distribution.

2. There was inflated Type I error in both cases, albeit to different degrees. From these two simulations, it is difficult to tell if the results are due exclusively to sampling error, the coarseness of measurement, or a more systemic problem commensurate with the CLT requirements previously noted. Operating under the assumption the results are indicative of a systemic problem, then:

    a. It would appear that higher levels of skewness and negative kurtosis in the sampling distribution comparatively increased the Type I error. A negative kurtosis is indicative of a platykurtic distribution with larger tails. This finding is commensurate with the requirement for a sufficient number of *J* items under the CLT to subscribe to a normal distribution. Accordingly, insufficient numbers of *J* items are more likely to demonstrate skewness and kurtotic properties in the sampling distribution.

    b. In the case of very small negative kurtosis and skewness, there remains a slight inflation in Type I error. Again, assuming this is a systemic condition above and beyond sampling error, this would coincide with need for a continuity correction.

    c. There is not enough information, however, demonstrating systemic coverage probability error to suggest a parametric form for a correction or adjustment which would result in a more accurate test statistic.

To better evaluate the viability of systemic inflated Type I errors, as well as to explore a functional parametric form as a remedy, a broader, multi-factor simulation study was carried out. Retaining the finding that the Z transform of the sample correlation is reasonably represented by a normal distribution, the estimate of the μ parameter is retained. If these occurrences prove to be systemic, they can be mitigated by developing a correction to the σ parameter specified as part of the Fisher transformation.

Study Design

This multi-factor empirical study was designed to jointly assesses the performance of the Fisher transformation and explore a viable parametric form for a correction. As a result of the theoretical analysis, it was expected that the sampling statistic would be consistently negatively biased. Such a bias corresponds to an increased Type I error rate, thus substantiating the need for a continuity correction. Further, it was additionally expected that the bias would be exacerbated by some function of *J* items as *J* decreased; this would substantiate the need for an item-type correction. Subsequent steps in developing a correction would only be necessary if these expectations are observed.

Using the same factors previously noted, a wide-ranging series of experimental conditions for each factor was used. Table 1 displays the conditions under which independent dichotomous responses were generated.

Table 1: Simulation Study Experimental Conditions and Corresponding Levels

Conditions	Levels
n = number of respondents in the sample	4 levels (25, 50, 100, 200)
J = number of items on the test	4 levels (10, 20, 40, 60)
p = probability of getting the item correct	3 levels (0.50, 0.65, 0.80)
ρ = correlation between two tests	3 levels (0.60, 0.75, 0.90)

The result is $4 \times 4 \times 3 \times 3 = 144$ different experimental conditions using the same simulation process previously described. Again, 10,000 trials were conducted per condition.

As opposed to assessing probability coverage and overall sampling distribution characteristics, the differences between the sampling distribution and the Fisher transformation at various percentiles were investigated. This change was adopted for two reasons. First, the hypothesis that the Fisher transformation is inaccurate necessitates anchoring the empirical sampling distribution as the correct distribution. Second, assessment of differences at various percentiles under various treatment conditions facilitates development of a functional form for a correction. These percentiles are analogous to the most common Type I error controls in confidence interval construction and hypothesis testing, both 1-sided and 2-sided. To evaluate the distributional differences, for each set of 10,000 trials, sample correlation values were numerically ordered where

$$r_i = r_1, r_2, ...r_{10000}$$
$$r_1 \le r_2 \le r_3 . \le ..r_{10000}$$

and the following values were retained

$$(r_{100}, r_{9900}), (r_{250}, r_{9750}), (r_{500}, r_{9500}), (r_{1000}, r_{9000})$$

These are the empirical analogs to Type I error values, α, of 0.01, 0.025, 0.05, and 0.10 respectively. For each treatment condition, knowing ρ and n, corresponding r interval bounds from the Fisher transformation process were calculated corresponding to the particular α. Error was computed as

$$Error = r_{empirical,\%} - r_{Fisher,\alpha}$$

A plot of the error for all treatment conditions is provided in Figure 3. The pattern of errors, with $(1-\alpha)$ yielding positive errors and α negative errors indicates an underestimation of variance at smaller test lengths. Recognition of a pattern also provides sufficient empirical evidence of a systemic problem beyond sampling error. Although this plot shows a pattern, it is does not provide definitive relationships purely as a function of test length, failing to address test difficulty.

Basic statistic textbooks indicate that binomial distributions approximate well to a normal distribution as its expected value, *np*, exceeds some heuristic value. Using that principle, consider the expected total score or total correct as the independent variable. The expected total score is a function encompassing both test length, *J*, and test difficulty, *p*-value. For dichotomous tests,

$$\begin{aligned} \mathrm{E}(T_o) &= \sum_{i=1}^{J} p - value_i \\ &= \bar{p}J \\ &= \frac{\sum_{i=1}^{N} T_{o,i}}{N} \end{aligned}$$

For polytomous scored items, each item must follow the same scale, $r = 0, 1, 2, \ldots, R$.

$$\mathrm{E}(T_o) = \frac{\sum_{i=1}^{N} T_{o,i}}{NR}$$

A reduced number of treatment conditions using the expected total score as the independent variable are displayed in the error plot in Figure 4. Evidently, there is distinctive pattern as the expected total score decreases. This pattern is similar across all treatment conditions. Figure 5 shows another set of treatment conditions illustrating similar findings.

Dotted lines in Figure 5 indicate bias as a result of failure to implement a continuity correction. This correction remains constant regardless of the E(T_o) value. Additionally, there is a systemic increase in error as the expected total number of correct items decreases. This decaying relationship asymptotes to the continuity correction value as E(T_o) increases. These empirical results reinforce the theoretical findings noted when data deviate from required conditions in applying the CLT. Because these graphs are presented as a separate set of

Figure 3: Error versus Test Length across All Treatment Conditions

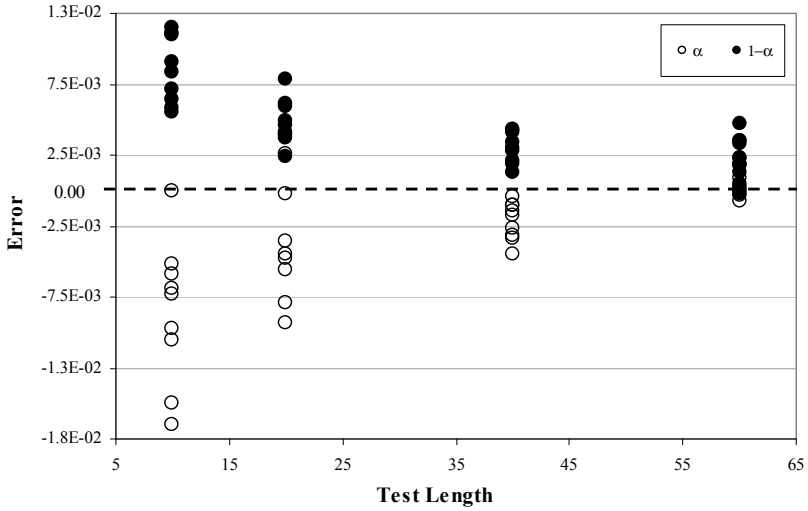


Figure 4: Error versus Expected Total Score across a Reduced Number of Experimental Conditions

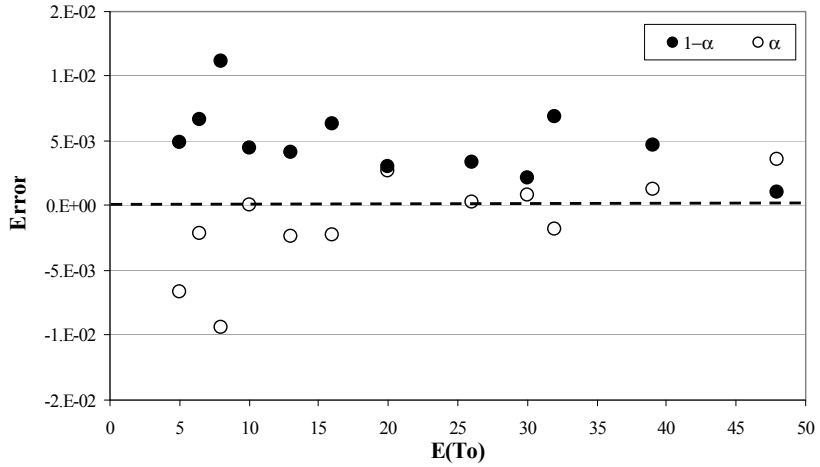
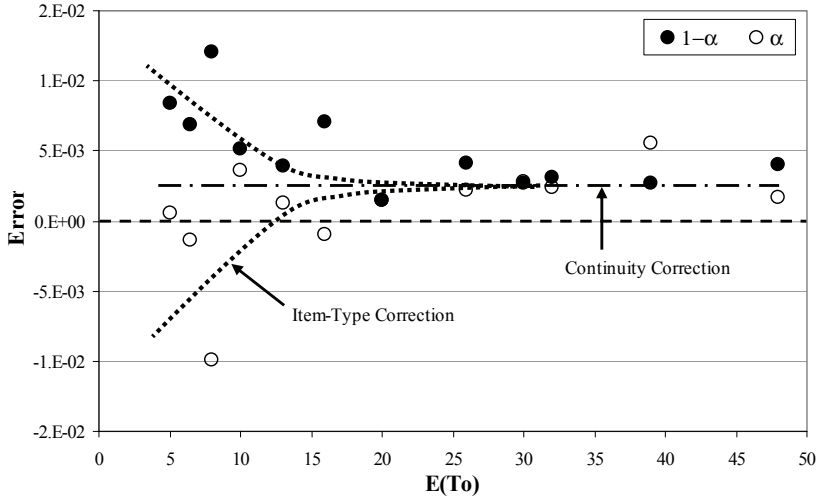


Figure 5: Error versus Expected Total Score Indicating Parametric Corrections



snapshots, there is a third relation observed which cannot be easily illustrated. Although each plot consistently exhibits a decaying relationship as E(T_o) increased, the amount and rate of decay differed conditioned upon the *p*-value or test difficulty treatment conditions. Higher *p*-values exhibited greater errors at lower E(T_o) values and took slightly longer to converge to the continuity correction. These findings are consistent with previous CLT discussions.

Proposed Correction

Though illustrating the need for a correction when applying Fisher's transformation inclusive of test properties is informative, its value is only realized with a corresponding remedy. Thus, the distributional properties of the Fisher transformation with independence of its first two moments are maintained. The item-type correction and continuity correction are independent corrections and can be treated as such in a specified solution. The impact of the *p*-value on the rate of change only affects the item-type correction. Accordingly, Fisher's transform is retained as

$$Z = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)$$

but, as opposed to utilizing the form $\sigma_Z = \frac{1}{\sqrt{n-3}}$, a corrected form is derived as

$$\sigma_Z^* = \left(\frac{\ln\left(\left(\frac{1}{1+a(pval-.5)^2} * bE(T_o)\right)+1\right)}{\ln\left(\frac{1}{1+a(pval-.5)^2} * bE(T_o)\right)} + c\right)\left(\frac{1}{\sqrt{n-3}}\right)$$

where *a*, *b*, and *c* are undetermined constants. The *a* term is associated with the *p*-value's effect on the amount and rate of decay associated with E(T_o). The *b* term is associated with the general rate of decay as the item-type or E(T_o) correction. The *c* term is associated with the continuity correction. Note that the overall correction

limit

$$_{E(To)\to\infty}\left(\frac{\ln\left(\left(\frac{1}{1+a(pval-.5)^2} * bE(T_o)\right)+1\right)}{\ln\left(\frac{1}{1+a(pval-.5)^2} * bE(T_o)\right)} + c\right) = 1 + c$$

is commensurate with the error plots previously presented. More specifically, the term

$$\frac{\ln\left(bE(T_o)+1\right)}{\ln\left(bE(T_o)\right)}$$

represents the decaying relation associated with E(T_o). Because these relations change as a function of the *p*-value, the following is introduced within the logarithm

$$\frac{1}{1+a(pval-.5)^2}$$

Figure 8 displays the correction factor shown for differing *p*-values.

Although the effect on the rate of decay is symmetrical around 0.50, the overall correction is not due to the effect of the *p*-value in the E(T_o) calculation. Figure 9 illustrates this lack of symmetry for 3 different tests lengths under a range of average *p*-values.

Other parametric representations may also be available for the correction. This choice appeared reasonable and parsimonious based on the observations of the errors between the empirical distributions and an uncorrected Fisher transform. Values for these constants were determined via an iterative process minimizing the total squared error across all treatment conditions of the form.

$$Total\ Error = \sum_{n=1}^{4}\sum_{l=1}^{4}\sum_{k=1}^{3}\sum_{j=1}^{4}\sum_{i=1}^{8}\left(r_{empirical,\%,,ijkln} - r_{Fisher^*,ijkln}\right)^2 \quad (3)$$

where *i* corresponds to the values of α, *j* represents the test length, *k* denotes the *p*-values

Figure 8: Z Standard Deviation Correction versus Number of Correct Items for Various *p*-values

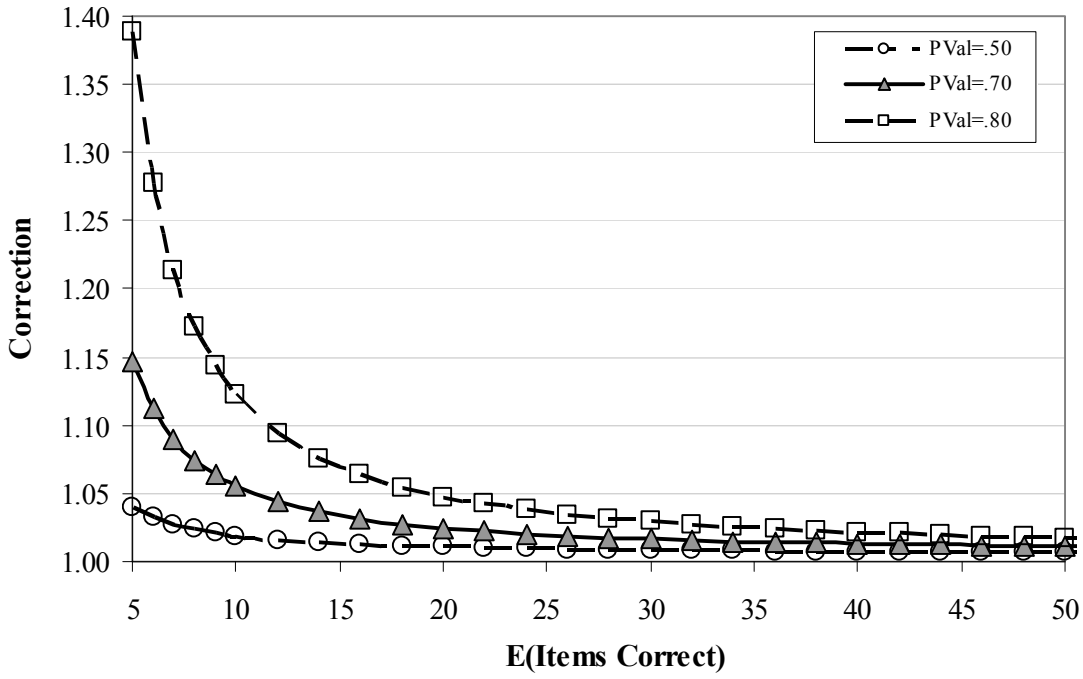
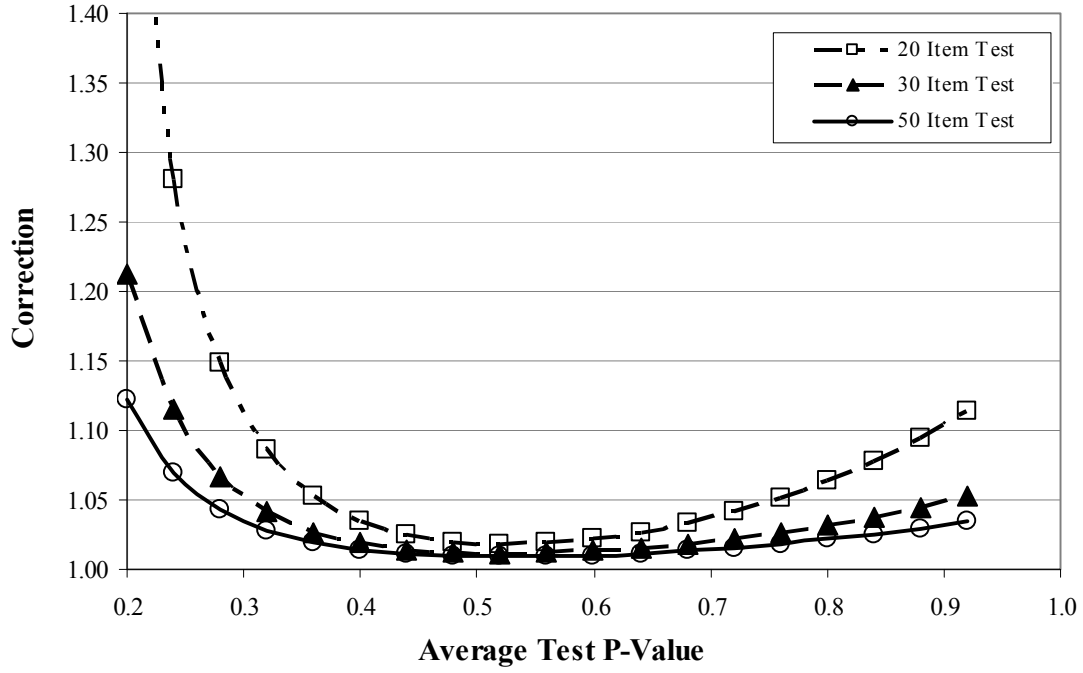


Figure 9: Z Standard Deviation Correction versus *p*-values for Various Test Lengths



for each test item, l represents the true correlation between items on each test, and n denotes the number of examinees. An evolutionary solver add-in to Excel from Frontline systems was utilized searching within a range of acceptable values. This particular solver is well suited to handle this nonlinear, mixed integer optimization problem. The resulting minimized error solution takes the form of

$$\sigma_Z^* = \left(\frac{\ln\left(\left(\frac{1}{1+40(pval-.5)^2} * 2.25E(T_o)\right)+1\right)}{\ln\left(\frac{1}{1+40(pval-.5)^2} * 2.25E(T_o)\right)} + .005 \right)\left(\frac{1}{\sqrt{n-3}}\right) \quad (4)$$

Results

Correct Assessment

Although the strategy in advocating a parametric correction is valid, it suffers from two flaws. First, the constants selected were optimized based on a set of 144 treatment conditions. As a means of cross-validation, this correction should be assessed under a different set of treatment conditions. Second, and more importantly, is the aspect of coverage probability. Reduced distributional errors resulting from an adjusted standard deviation in the Z transform does not necessarily correspond to a definitive improvement in coverage probability.

By utilizing aspects of both previous simulations, both flaws are addressed and a more thorough assessment of the proposed correction is provided. Using the same factors, consider next a broader series of treatments for each factor. Independent dichotomous responses were generated under the following conditions enumerated in Table 2.

The result is $5 \times 4 \times 4 \times 3 = 240$ different treatment conditions using the same process. Using both the Fisher transform and the proposed correction, two-sided 90%, 95%, and 99% confidence intervals were calculated from the sample correlation value used in this study. Knowing the true ρ for each trial an assessment was made as to whether this value was within the Fisher and the corrected interval, noting successes. This was repeated for 10,000 trials for each simulation resulting in an estimate of the coverage probability. Success percentages below the $(1-\alpha)\%$ specification indicate an inflated Type I error.

As formal statistical assessments of these coverage probabilities, performance in terms of bias and mean square error across all conditions was considered. Bias is defined as $Bias(\hat{\theta},\theta) = E(\hat{\theta}-\theta)$, where θ is the specified confidence interval, 99%, 95% or 90%, and $\hat{\theta}$ represents the proportion of intervals containing the true population correlation value separately for the Fisher transformation and the proposed correction.

Mean square error (MSE) is determined by: $MSE = V(\hat{\theta}) + Bias^2$ where $V(\hat{\theta})$ is the variance of the estimates determined across the set of the treatment conditions.

Graphical summaries in Figures 10a, 10b, and 10c are presented as boxplots of coverage probability results from the conditions over each of the 3 test related parameters associated in calculating the proposed formula: sample size of respondents (n), expected number of items correct ($E(T_o)$), and an average test p-value, respectively.

Table 2: Simulation Study Experimental Conditions and Corresponding Levels

Conditions	Levels
n = number of respondents in the sample	5 levels (25, 50, 100, 200, 400)
J = number of items on the test	4 levels (10, 20, 40, 80)
p = probability of getting the item correct	3 levels (0.50, 0.60, 0.70, 0.80)
ρ = correlation between two tests	3 levels (0.65, 0.75, 0.85)

Figure 10a: Side-by-Side Boxplots of Coverage Probability Error Comparison at $\alpha = 0.01$ Over Expected Correct Items across All Conditions

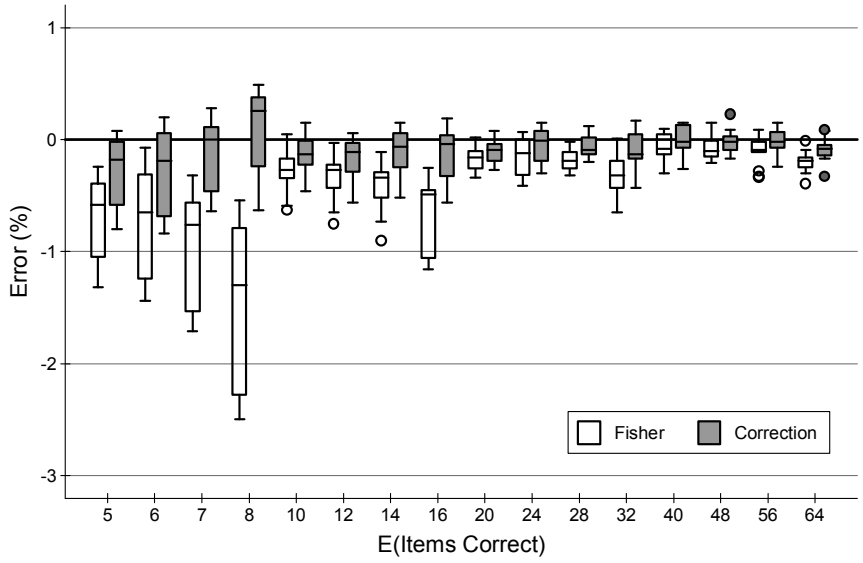
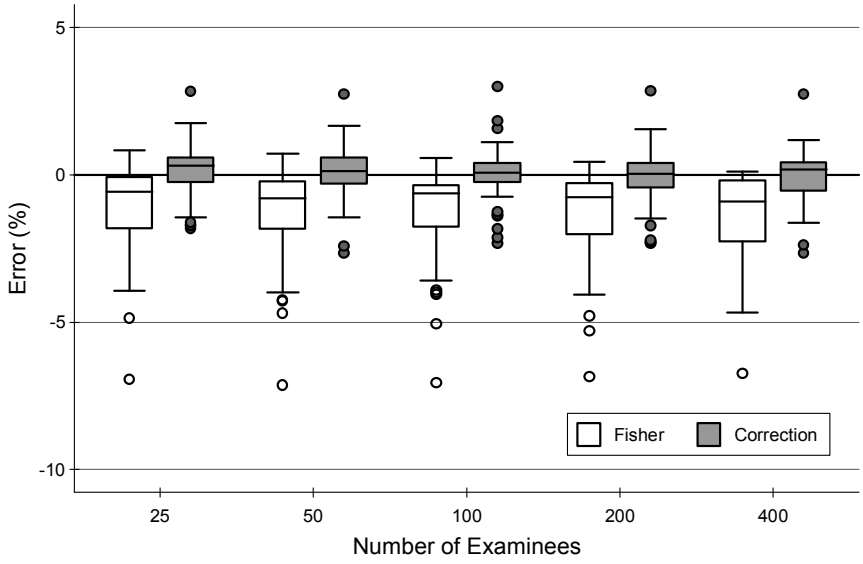


Figure 10b: Side-by-Side Boxplots of Coverage Probability Error Comparison at $\alpha = 0.05$ over average *p*-value across All Conditions



Summary results are shown in Table 3, with bias and mean squared error values provided across all conditions. The results showed improvement over the uncorrected Fisher transformation with 10 times less bias and a total reduction of error exceeding 500% across all conditions. These improvements are also consistent with each of the 28 cross-classified results, outperforming the Fisher transform with smaller bias and mean square error.

Table 3: Bias and MSE for Fisher's Transformation and the Proposed Correction for All Experimental Conditions

Description	Fisher Transformation		Proposed Model	
	Bias	MSE	Bias	MSE
Overall	-0.936	2.285	-0.095	0.443
By Sample Size				
25	-0.887	2.168	-0.060	0.447
50	-0.929	2.267	-0.089	0.451
100	-0.937	2.261	-0.098	0.435
200	-0.965	2.422	-0.120	0.469
400	-0.960	2.348	-0.107	0.422
By *p*-value				
0.50	-0.658	1.206	-0.214	0.432
0.60	-0.739	1.403	-0.223	0.379
0.70	-0.916	2.060	-0.051	0.352
0.80	-1.431	4.494	0.109	0.614
By Alpha				
0.01	-0.423	0.396	-0.096	0.061
0.05	-1.105	2.471	-0.195	0.427
0.10	-1.279	3.999	0.007	0.844
By $E(T_0)$				
5	-1.535	3.605	-0.574	1.123
6	-1.730	4.358	-0.587	1.012
7	-2.116	6.464	-0.082	0.905
8	-3.115	13.403	0.667	1.714
10	-0.703	1.018	-0.276	0.495
12	-0.779	0.963	-0.285	0.327
14	-0.915	1.332	-0.110	0.323
16	-1.612	3.766	-0.151	0.530
20	-0.294	0.148	-0.066	0.060
24	-0.314	0.228	-0.052	0.110
28	-0.420	0.353	-0.030	0.129
32	-0.696	0.702	-0.087	0.154
40	-0.098	0.077	0.060	0.072
48	-0.133	0.086	0.033	0.083
56	-0.214	0.164	0.006	0.091
64	-0.300	0.164	0.006	0.091

Though the proposed correction is empirically unbiased, it cannot be theoretically demonstrated as an unbiased estimator. Given the variety of treatment conditions examined, a theoretical proof becomes difficult without many simplifying assumptions. Some additional comments regarding a theoretical assessment include:

1. Although the need for correction based on the expected total number of items correct and the average *p*-value of the testing instrument has been theoretically and empirically demonstrated, a proper parametric form to implement such correction into probability coverage is not clear. As noted previously, there are other parametric forms which may be considered. Also, recall that the assumption of normality upon transform is still operating, which becomes more tenuous in low number of test items and extreme *p*-values. Other distributional forms can be considered upon which one would make probabilistic inferences. Finally, regarding parametric forms and distributions, this discussion is predicated that there exists a common distribution characterized by respondents and test conditions which results in an unbiased, consistent estimator controlling Type I error.

2. Due to confidence the Fisher transformation is incomplete without inclusion of summary test information in its calculations, the empirical distribution of the sample correlation values were treated as the true distribution. This was also necessary to assess systemic errors in the development of a functional parametric form for a correction. This reference empirical distribution has sampling error, which has been minimized given the large number of trials.

3. Estimates via a complex evolutionary search method were obtained from the Frontline Premium Solver add-in for the Excel Solver. Determining a so-called best set of parameter estimates for a complex nonlinear optimization required parameter constraints and other considerations in order to achieve convergence.

Based on these findings, when reporting sample Pearson product moment correlations for dichotomous and polytomously scored items, the adjustment in (4) is recommended; it is well characterized by a normal distribution. These corrections provide robust results due to violations in the application of the central limit theorem. It further provides a researcher inclusion of summary test information into any inferential statistics. Unfortunately, because of the transformation process, simple reporting of the standard error is uninformative. As such, presented below are two examples which should be used as the proper mechanism for reporting sample correlation properties.

Applications: Parallel Test Forms

Forms A and B of a particular test are each administered to 70 respondents from the same population. Each test consists of 25 items and both test are polytomously scored on a scale of [0, 1, …, 4]. The average score for form A was 41 and 45 for form B. The sample correlation was r = 0.82, and it is desired to report a 95% confidence interval for the population correlation. Z is computed with accompanying standard deviation:

$$Z = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) = \frac{1}{2}\ln\left(\frac{1+.82}{1-.82}\right) = 1.157$$

$$\sigma_z = \frac{1}{\sqrt{n-3}} = \frac{1}{\sqrt{70-3}} = .1222$$

Next, the proposed correction is determined, which takes the form

$$\left(\frac{\ln\left(\left(\frac{1}{1+40(0.43-.5)^2}\cdot 2.25\cdot 10.75\right)+1\right)}{\ln\left(\frac{1}{1+40(0.43-.5)^2}\cdot 2.25\cdot 10.75\right)}+.005\right) = 1.016$$

where

$$E(T_o) = \frac{41+45}{(2)(4)} = 10.75$$

and

$$pval = .5*\left(\frac{41}{100}+\frac{45}{100}\right) = 0.43,$$

therefore the estimate for the standard deviation of the transformation becomes:

$$\sigma_Z^* = 0.1222*1.016 = 0.1242.$$

Because Z follows a normal distribution, a traditional 95% confidence interval for Z can be computed as follows

$$\begin{aligned} Z_L^* &= \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) + .1242*\Phi^{-1}\left(\alpha/2\right) \\ &= 1.157 + .1242(-1.96) = .9136 \end{aligned}$$

$$\begin{aligned} Z_U^* &= \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) + .12441*\Phi^{-1}\left(1-\alpha/2\right) \\ &= 1.157 + .12441(1.96) = 1.40 \end{aligned}$$

which can be back transformed into intervals for the population correlation

$$\begin{aligned} &(1-\alpha)\%\ CI\ for\ \rho = \\ &= \left(\frac{\exp(2Z_L^*)-1}{\exp(2Z_L^*)+1}, \frac{\exp(2Z_U^*)-1}{\exp(2Z_U^*)+1}\right) \\ &= \left(\frac{\exp(2*.9136)-1}{\exp(2*.9136)+1}, \frac{\exp(2*1.40)-1}{\exp(2*1.40)+1}\right) \\ &= (0.723, 0.886). \end{aligned}$$

The uncorrected confidence interval is $(1-\alpha)\%\ CI\ for\ \rho = (0.725, 0.885)$. The reporting should include both the sample correlation estimate and the corresponding interval values.

Applications: Inter-rater Reliability

Suppose two graders score an exam consisting of 20 dichotomous items administered to 125 respondents. The average score for each grader was 17 and the sample correlation was $r = 0.77$. Test the hypothesis the population correlation between the two graders exceeds the minimally desired reliability value of at least 0.70 at significance level of 0.05.

Using a similar process to determine the standard deviation for the proposed correction, the Fisher transformation of the standard deviation is

$$\sigma_z = \frac{1}{\sqrt{125-3}} = \frac{1}{\sqrt{125-3}} = .0905.$$

The corrected standard deviation is

$$\left(\frac{\ln\left(\left(\frac{1}{1+40(0.85-.5)^2}\cdot 2.25\cdot 16.5\right)+1\right)}{\ln\left(\frac{1}{1+40(0.85-.5)^2}\cdot 2.25\cdot 16.5\right)} + .005\right) = 1.08$$

where

$$E(T_o) = 16.5$$

and

$$pval = \left(\frac{17}{20}\right) = .85.$$

Therefore, the estimate for the corrected standard deviation of the transformation becomes

$$\sigma_Z^* = .0905*1.08 = .0978$$

and Z^* is determined via

$$\begin{aligned} Z^* &= \frac{\frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) - \frac{1}{2}\ln\left(\frac{1+\rho_o}{1-\rho_o}\right)}{.0978} \\ &= \frac{\frac{1}{2}\ln\left(\frac{1+.77}{1-.77}\right) - \frac{1}{2}\ln\left(\frac{1+.70}{1-.70}\right)}{.0978} \\ &= \frac{1.0203-.8673}{.0978} = 1.564. \end{aligned}$$

Because

$$Z^* \le Z_{crit,1-\alpha}$$

$$1.564 \le 1.644$$

the null hypothesis H_o is retained. It appears these graders do not meet the minimally acceptable inter-rater reliability. Corrective actions, such as additional grader training,

would be required in such cases. However, the hypothesis test without the correction results in

$$Z^* \geq Z_{crit,1-\alpha}$$
$$1.691 \geq 1.644.$$

In contrast to the results using the correction, the null hypothesis would be incorrectly rejected. Multiple rater comparisons or multiple parallel forms may as well be addressed with this correction using a multiple comparison Type I error adjustment such as Bonferroni or Tukey.

Because the proposed correction occurs within the Z transform (see Figures 8 and 9), it is difficult to interpret its impact in the original correlation scale. The width of a correlation confidence interval is not only a function of *r*, α, and *n*, but this study has demonstrated $E(T_o)$ and the average *p*-value as well. To better understand the effects of this correction in the desired scale, the following 3D plots show the difference in CI widths between the Fisher transformation and this correction, where the proposed correction always result in larger widths in order to maintain an accurate Type I error control. In each plot, *r* was 0.75 and α was 0.05. The range of test items used coincides with test section lengths of the major standardized educational exams such as the SAT, GRE, LSAT, and MCAT.

Conclusion

The Fisher transformation is remarkably efficient, yet was not designed with an intended use of summed dichotomous or polytomous data. This correction accounts for departures from asymptotic convergence under the central limit theorem due to test length and average item difficulty. Further, this correction can be easily applied, providing substantially more accurate results over the Fisher transformation. This study also illustrates the coarseness of dichotomous measures has no effect on the coverage probability results of the true population correlation as this is accounted for in the correction and results from application of the central limit theorem.

For those positing a unidimensional construct, the use of Pearson correlation can be easily extended to allow for items which load differently on the latent dimension. By weighting each item and making an adjustment to the total score, an omnibus reliability measure based on total score can be obtained.

Throughout the study, a homogeneous *p*-value for each test item was used. Because most tests are comprised of items with varying *p*-values, the performance of this correction was examined under a wide range of *p*-value distributions. This robust analysis explored extreme deviations from the simulation conditions, using a highly kurtotic uniform distribution and bi-modal distributions with different expected average *p*-values. The results for this analysis are present in Appendix A and reaffirm the use of this correction under any conditions.

Though the proposed correction is easily implemented with demonstrated efficiency across a wide range of test conditions, a nonparametric alternative is also available. Nonparametric bootstrap methods remain a viable option for researchers desiring confidence interval estimates; whereas such options might also produce robust results, they require both sufficient data and custom coding.

References

Barnette, J. J. (2005). ScoreRel CI: An Excel program for computing confidence intervals for commonly used score reliability coefficients. *Educational and Psychological Measurement*, *65*, 980-983.

Colman, A. M. (2001). *A dictionary of psychology*. Oxford University Press: Great Britain.

Denton, G., Durning, S., & Hemmer, P. (2004). A call for use of confidence intervals with correlation coefficients. *Teaching and Learning in Medicine*, *16*, 111-112.

Devore, J. L. (2000). *Probability and statistics for engineering and the sciences* (*5th Ed.*). Pacific Grove, CA: Duxbury.

Fan, X., & Thompson, B. (2001). Confidence intervals for effect sizes. *Educational and Psychological Measurement*, *61*, 517-531.

Figure 13: Confidence Interval Width Difference between Proposed Correction and Fisher Transform at $n = 150$

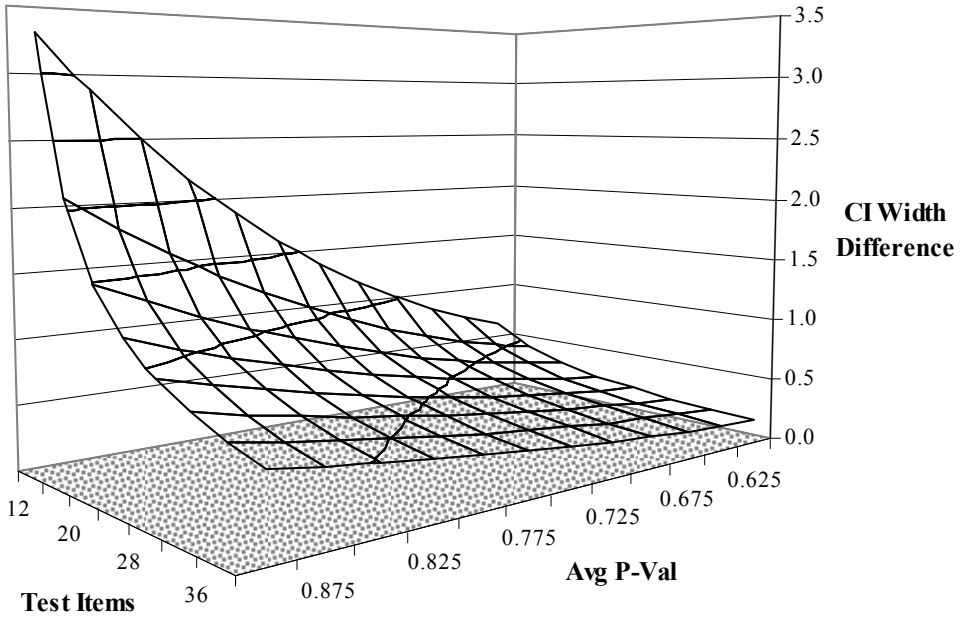
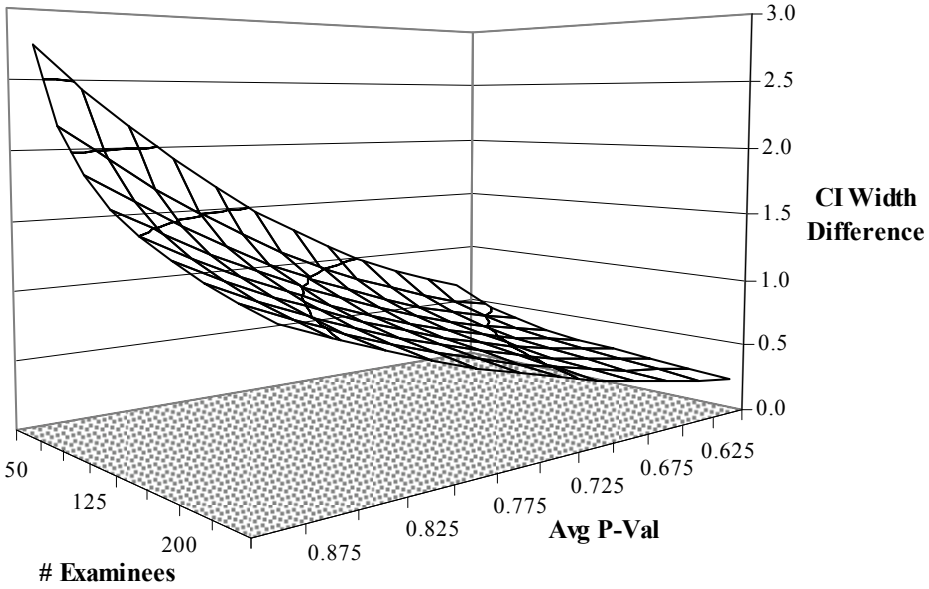


Figure 14: Confidence Interval Width Difference between Proposed Correction and Fisher Transform at $J = 20$



Fouladi, R. T., Marani, S. K., & Steiger, J. H. (2002). Moments of the Fisher transform: applications using small samples. *American Statistical Association Proceedings of the Joint Statistical Meetings [CD-ROM]*, 1032-1037.

Hogg, R. V., & Craig, A. T. (1995). *Introduction to mathematical statistics* (*5th Ed*.). Englewood Cliffs, NJ: Prentice Hall.

Kline, T. J. (2005). *Psychological testing: A practical approach to design and evaluation*. Thousand Oaks, CA: Sage Publications, Inc.

Laubscher, N. F. (1959). Note on Fisher's transformation of the correlation coefficient. *Journal of the Royal Statistical Society, Series B (Methodological*), *21*, 409-410.

Lebreton, J. M., & Senter, J. (2007). Answers to 20 questions about interrater reliability and interrater agreement. *Organizational Research Methods Online First*, 1-38.

Mulry, M. H., & Wolter, K. M. (1981). The effect of Fisher's Z transformation on confidence intervals for the correlation coefficient. *U.S. Bureau of the Census*, 601-608.

Qualls-Payne, A. L. (1992). A comparison of score level estimates of the standard error of measurement. *Journal of Educational Measurement*, *29*, 213-225.

Quereshi, M. Y. (1971). Note on the Pearson r as a function of the bivariate distributional characteristic. *Journal of Educational Measurement*, *8*(*3*), 142-147.

Robinson-Kurpius, S. E., & Stafford, M. E. (2005). *Testing and measurement: A user-friendly guide*. Thousand Oaks, CA: Sage.

Shen, D., & Lu, Z. (2005). Computation of correlation coefficient and its confidence interval in SAS. *SAS Paper*, *170-31*. SAS Institute.

Task Force on Reporting of Research Methods in AERA Publications (2006). *Standards for reporting on empirical social science research in AERA publications*. Washington, DC: American Educational Research Association.

Thompson, B. (2002). What future quantitative social science research could look like: Confidence intervals for effect sizes. *Educational Researcher*, *31*, 25-32.

Traub, R. E. (1994). *Reliability for the social sciences: Theory and applications, Volume 3*. Thousand Oaks, CA: SAGE.

Wilkinson, L. (1999). The American Psychological Association Task Force on Statistical Inference, Statistical Methods in Psychology: Guidelines and Explanations. *American Psychologist*, *54*, 594-604.

Winterbottom, A. (1979). A note on the derivation of Fisher's transformation of the correlation coefficient. *The American Statistician*, *33*, 142-143.

Zimmerman, D. W., Zumbo, B. D., & Williams, R. H. (2003). Bias in estimation and hypothesis testing of correlation. *Psicologica*, *24*, 133-158.

Appendix

As a means of robust analysis, the proposed correction was explored under 4 different sets of varied p-values. Empirical treatments remained unchanged for sample size, population correlation, and test length. However, instead of a homogeneous *p*-value for each item on a test of length *J*, the following were considered:

a. *p-value* = 0.50 per test item to a bimodal distribution of the following form

$$\frac{J}{2}Unif(.2-.4)+\frac{J}{2}Unif(.6-.8)$$

per test. *P*-values were redrawn from this distribution for each trial. The average p-value is 0.50.

b. *p-value* = .60 per item to a distribution of the form

$$Unif(.3-.9)$$

per test, redrawn for each trial. The average *p-value* is 0.60.

c. *p-value* = 0.70 per item to a distribution of the form

$$\frac{J}{2}Unif(.45-.65)+\frac{J}{2}Unif(.75-.95)$$

per test, redrawn for each trial. The average *p*-value is 0.70.

d. *p-value* = 0.80 per item to a distribution of the form

$$Unif\,(.65-.95)$$

per test, redrawn for each trial. The average *p-value* is 0.80.

Collective results are presented in the Table 4. Similar to this validation study, in bias and mean square error, overall and across each of treatment conditions, the proposed correction outperformed the Fisher transformation. Further, the Type I error of the Fisher transformation is comparatively higher compared with a test of items with homogeneous *p*-values. This reaffirms the suitability of this correction under any conditions, regardless of the *p*-value distribution underpinning the test items.

Table 4: Robust Analysis for Extreme *p*-values; Bias and MSE for Fisher's Transformation and the Proposed Model across All Experimental Conditions

Description	Fisher Transformation		Proposed Model	
	Bias	MSE	Bias	MSE
Overall	-1.100	3.081	-0.229	0.698
By Sample Size				
25	-1.020	2.615	-0.169	0.574
50	-1.143	3.231	-0.260	0.727
100	-1.078	3.031	-0.204	0.694
200	-1.097	3.156	-0.220	0.678
400	-1.164	3.423	-0.291	0.837
By P-value				
0.50	-0.929	2.125	-0.461	0.837
0.60	-0.896	2.007	-0.341	0.636
0.70	-1.086	3.005	-0.191	0.612
0.80	-1.490	5.217	0.078	0.718
By Alpha				
0.01	-0.522	0.578	-0.166	0.114
0.05	-1.253	3.284	-0.318	0.721
0.10	-1.526	5.395	-0.202	1.266
By $E(T_0)$				
5	-2.116	6.560	-1.087	2.448
6	-2.027	6.026	-0.760	1.669
7	-2.495	9.539	-0.371	1.786
8	-3.451	16.079	0.493	1.816
10	-1.037	1.530	-0.591	0.659
12	-1.001	1.667	-0.480	0.722
14	-1.163	2.013	-0.331	0.490
16	-1.620	3.918	-0.182	0.744
20	-0.422	0.379	-0.174	0.216
24	-0.427	0.310	-0.163	0.122
28	-0.508	0.428	-0.098	0.133
32	-0.683	0.769	-0.073	0.239
40	-0.140	0.063	0.008	0.052
48	-0.131	0.058	0.040	0.058
56	-0.178	0.095	0.035	0.077
64	-0.244	0.172	0.074	0.116

Journal of Modern Applied Statistical Methods
November 2011, Vol. 10, No. 2, 494-504

1538 – 9472/11/$95.00

Estimation of Parameters of Johnson's System of Distributions

Florence George
Florida International University
Miami, FL

K. M. Ramachandran
University of South Florida
Tampa, FL

Fitting distributions to data has a long history and many different procedures have been advocated. Although models like normal, log-normal and gamma lead to a wide variety of distribution shapes, they do not provide the degree of generality that is frequently desirable (Hahn & Shapiro, 1967). To formally represent a set of data by an empirical distribution, Johnson (1949) derived a system of curves with the flexibility to cover a wide variety of shapes. Methods available to estimate the parameters of the Johnson distribution are discussed, and a new approach to estimate the four parameters of the Johnson family is proposed. The estimate makes use of both the maximum likelihood procedure and least square theory. The new MLE-Least Square approach is compared with other two commonly used methods. A simulation study shows that the MLE-Least square approach provides better results for S_B, S_U and S_L families.

Key words: Johnson distribution, unbouded, bounded, lognormal, estimation.

Introduction

Any data set with finite moments can be fitted by a member of the Johnson families such as S_B, S_U or S_L. The most commonly used methods to estimate the parameters of the Johnson distribution are the percentile approach (Shapiro, 1980) and Quantile method (Wheeler, 1980). A new approach is proposed for the estimation of Johnson parameters and is compard to other methods. For additional reerences, see Drapper (1952), Hill (1976), Hahn and Shapiro (1967), George, et al (2009).

The Johnson Translation System

Given a continuous random variable X whose distribution is unknown and is to be approximated, Johnson proposed three normalizing transformations having the general form:

$$Z = \gamma + \delta f\left(\frac{X-\xi}{\lambda}\right) \tag{2.1}$$

where $f(.)$ denotes the transformation function, Z is a standard normal random variable, γ and δ are shape parameters, λ is a scale parameter and ξ is a location parameter. Without loss of generality, it is assumed that $\delta > 0$ and $\lambda > 0$. The first transformation proposed by Johnson defines the lognormal system of distributions denoted by S_L:

$$\begin{aligned} Z &= \gamma + \delta \ln\left(\frac{X-\xi}{\lambda}\right), X > \xi \\ &= \gamma^* + \delta \ln(X-\xi), X > \xi \end{aligned} \tag{2.2}$$

S_L curves cover the lognormal family.

The bounded system of distributions S_B is defined by

$$Z = \gamma + \delta \ln\left(\frac{X-\xi}{\xi+\lambda-X}\right), \xi < X < \xi + \lambda \tag{2.3}$$

Florence George is an Assistant Professor in the Department of of Mathematics and Statistics, Florida International University. Email: fgeorge@fiu.edu. K. M. Ramachandran is a Professor in the Department of Mathematics and Statistics, University of South Florida. Email: ram@usf.edu.

S_B curves cover bounded distributions. The distributions can be bounded on the lower end, the upper end or both ends. This family covers Gamma distributions, Beta distributions and many others.

The unbounded system of distributions S_U is defined by

$$Z = \gamma + \delta \ln\left[\left(\frac{X-\xi}{\lambda}\right) + \left\{\left(\frac{X-\xi}{\lambda}\right)^2 + 1\right\}^{1/2}\right],$$
$$-\infty < X < \infty$$
$$= \gamma + \delta \sinh^{-1}\left(\frac{X-\xi}{\lambda}\right) \tag{2.4}$$

The S_U curves are unbounded and cover the t and normal distributions, among others. Using the fact that, after the transformation in (2.1), Z follows standard normal distribution, the probability density function (pdf) of each of the family in the Johnson system can be derived. If X follows the Johnson distribution and $Y = \frac{X-\xi}{\lambda}$ then, for S_L family, the pdf is

$$p(y) = \frac{\delta}{\sqrt{2\pi}} \times \frac{1}{y} \times \exp\left\{-\frac{1}{2}[\gamma + \delta.\ln(y)]^2\right\},$$
$$\xi < X < +\infty.$$

similarly, for the S_B family, the pdf is,

$$p(y) = \frac{\delta}{\sqrt{2\pi}} \times \frac{1}{[y/(1-y)]} \times \exp\left\{-\frac{1}{2}\left[\gamma + \delta.\ln\left(\frac{y}{1-y}\right)\right]^2\right\}$$
$$\xi < X < +\xi + \lambda.$$

The pdf for the S_U family is

$$p(y) = \frac{\delta}{\sqrt{2\pi}} \times \frac{1}{\sqrt{y^2+1}} \times \exp\left\{-\frac{1}{2}\left[\gamma + \delta.\ln(y + \sqrt{y^2+1})\right]^2\right\},$$
$$-\infty < X < +\infty.$$

In general the pdf of X is given by,

$$p(x) = \frac{\delta}{\lambda\sqrt{2\pi}} \times g'\left(\frac{x-\xi}{\lambda}\right) \times \exp\left\{-\frac{1}{2}\left[\gamma + \delta.g\left(\frac{x-\xi}{\lambda}\right)\right]^2\right\} \tag{2.5}$$

for all $x \in H$, where

$$g'(y) = \frac{1}{y} \quad \text{for the } S_L \text{ family}$$
$$= \frac{1}{[y(1-y)]} \quad \text{for the } S_B \text{ family}$$
$$= \frac{1}{\sqrt{y^2+1}} \quad \text{for the } S_U \text{ family}$$

and

$$g(y) = \ln(y) \quad \text{for the } S_L \text{ family}$$
$$= \ln(y/(1-y)) \quad \text{for the } S_B \text{ family}$$
$$= \ln[y + \sqrt{y^2+1}] \quad \text{for the } S_U \text{ family.} \tag{2.6}$$

The support H of the distribution is:

$$H = [\xi, +\infty) \quad \text{for the } S_L \text{ family}$$
$$= [\xi, \xi + \lambda] \quad \text{for the } S_B \text{ family}$$
$$= (-\infty, +\infty) \quad \text{for the } S_U \text{ family.}$$

Parameter Estimation of the Johnson System: Percentile Matching

Percentile matching involves estimating k required parameters by matching k selected quantiles of the standard normal distribution with corresponding quantile estimates of the target population. For given percentages $\{\alpha_j : 1 \le j \le k\}$, the corresponding quantiles $\{z_{\alpha_j}\}$ and $\{x_{\alpha_j}\}$ are given respectively by

$$z_{\alpha_j} = \Phi^{-1}(\alpha_j)$$

and

$$x_{\alpha_j} = F^{-1}(\alpha_j)$$

where $\Phi(.)$ is the standard normal distribution function and F is the target distribution function. Once the functional form $f(.)$ among systems given by equations 2.2-2.4 has been identified, the method of percentile matching attempts to solve the k equations

$$z_{\alpha_j} = \gamma + \delta f\left(\frac{\hat{x}_{\alpha_j} - \xi}{\lambda}\right), 1 \le j \le k$$

where $\hat{x}_{\alpha_j}$ is an estimator of the quantile x_{α_j} based on sample data.

Slifker and Shapiro (1980) introduced a selection rule, which is a function of four percentiles for selecting one of the three families, to give estimates of the Johnson parameters. The fit parameters for the transformation are calculated by solving the transformation equation for the chosen distribution type at the four selected percentiles. Choose any fixed value z $(0 < z < 1)$ of a standard normal variate; the four points $\pm z$ and $\pm 3z$ determine three intervals of equal length. Determine the percentile P_ζ corresponding to $\zeta = 3z, z, -z, -3z$ respectively. For example, if $z = 0.5$ then $P_{0.5} = 0.6915 * 100 = 69.15$. Let $x_{3z}, x_z, x_{-z}, x_{-3z}$ be the percentiles of data values corresponding to the four selected percentiles of the normal distribution. The type of Johnson distribution chosen is based on the value of the discriminant d calculated as follows.

$$d = \frac{mn}{p^2}$$

where

$p = x_z - x_{-z}$, $m = x_{3z} - x_z$, $n = x_{-z} - x_{-3z}$.

If the calculated discriminant d is greater than 1.001, then an unbounded distribution is chosen; if the value is less than 0.999, then a bounded distribution is chosen. A discriminant equal to or between the two values results in a lognormal fit. The fit parameters for the transformation are calculated by solving the transformation equation for the chosen distribution type at the four selected percentiles. The parameter estimates for the Johnson S_U distribution are:

$$\hat{\delta} = \frac{2z}{\cosh^{-1}\left[\frac{1}{2}\left(\frac{m}{p} + \frac{n}{p}\right)\right]},$$

$$\hat{\gamma} = \hat{\delta}\sinh^{-1}\left[\frac{\frac{n}{p} - \frac{m}{p}}{2\left(\frac{m}{p}\frac{n}{p} - 1\right)^{1/2}}\right],$$

$$\hat{\lambda} = \frac{2p\left(\frac{m}{p}\frac{n}{p} - 1\right)^{1/2}}{\left(\frac{m}{p} + \frac{n}{p} - 2\right)\left(\frac{m}{p} + \frac{n}{p} + 2\right)^{1/2}},$$

and

$$\hat{\xi} = \frac{x_z + x_{-z}}{2} + \frac{p\left(\frac{n}{p} - \frac{m}{p}\right)}{2\left(\frac{m}{p} + \frac{n}{p} - 2\right)}.$$

The parameter estimates for the S_B distribution are

$$\hat{\delta} = \frac{z}{\cosh^{-1}\left(\frac{1}{2}\left[\left(1 + \frac{p}{m}\right)\left(1 + \frac{p}{n}\right)\right]^{1/2}\right)};$$

$$(\delta > 0),$$

$$\hat{\gamma} = \hat{\delta}\sinh^{-1}\left[\frac{\left(\frac{p}{n}-\frac{p}{m}\right)\left[\left(1+\frac{p}{m}\right)\left(1+\frac{p}{n}\right)-4\right]^{1/2}}{2\left(\frac{p}{m}\frac{p}{n}-1\right)}\right],$$

$$\hat{\lambda} = \frac{p\left[\left\{\left(1+\frac{p}{m}\right)\left(1+\frac{p}{n}\right)-2\right\}^{2}-4\right]^{1/2}}{\frac{p}{m}\frac{p}{n}-1},$$

and

$$\hat{\xi} = \frac{x_z + x_{-z}}{2} - \frac{\lambda}{2} + \frac{p\left(\frac{p}{n}-\frac{p}{m}\right)}{2\left(\frac{p}{m}\frac{p}{n}-1\right)}.$$

The parameter estimates for the Johnson S_L distribution are:

$$\hat{\delta} = \frac{2z}{\ln\left(\frac{m}{p}\right)},$$

$$\hat{\gamma}^{*} = \hat{\delta}\ln\left[\frac{\frac{m}{p}-1}{p\left(\frac{m}{p}\right)^{1/2}}\right],$$

and

$$\hat{\xi} = \frac{x_z + x_{-z}}{2} - \frac{p}{2}\frac{\frac{m}{p}+1}{\frac{m}{p}-1}.$$

Parameter Estimation of the Johnson System: Quantile Estimators

Wheeler (1980) proposed a method to estimate the parameters γ and δ in the Johnson family using five quantiles. Let $p_n = (n-\frac{1}{2})/n$, where n is the sample size. Denote the quantile of the standard normal distribution corresponding to the cumulative probability p_n by z_n. For example, if $n = 100$, then $p_n = 0.995$, so that $z_n = 2.5758$. Choose five quantiles x_p, x_k, x_0, x_m, x_n from data corresponding to standard normal quantiles $z = -z_n,\ -\frac{1}{2}z_n,\ 0,\ \frac{1}{2}z_n, z_n$. The general form of the Johnson system can be written as

$$z = \gamma + \delta \ln f(y)$$

where $f(y) = y$ for S_L, $f(y) = y + (1+y^2)^{1/2}$; for S_U, $f(y) = y/(1-y)$; and for S_B $y = (x-\xi)/\lambda$. Wheeler uses the fact that any quantity of the form

$$\frac{x_i - x_j}{x_r - x_s} = \frac{f^{-1}(\omega_i) - f^{-1}(\omega_j)}{f^{-1}(\omega_r) - f^{-1}(\omega_s)}$$

where $\omega = e^{(z-\gamma)/\delta}$, does not depend on ξ or λ. The parameter estimates for the S_U curves are:

$$\hat{\delta} = \frac{1}{2}z_n / ln b$$

where

$$b = \frac{1}{2}t_u + [(\frac{1}{2}t_u)^2 - 1]^{1/2},$$

and

$$t_u = \frac{x_n - x_p}{x_m - x_k};$$

and

$$\hat{\gamma} = -\delta ln(a)$$

where

$$a^2 = \frac{1-tb^2}{t-b^2} \text{ and } t = \frac{x_n - x_0}{x_0 - x_p}.$$

For S_B curves the parameter estimates are:

$$\hat{\delta} = \frac{1}{2} z_n / lnb,$$

where

$$b = \frac{1}{2} t_b + [(\frac{1}{2} t_b)^2 - 1]^{1/2},$$

and

$$t_b = \frac{(x_m - x_0)(x_n - x_p)}{(x_n - x_m)(x_0 - x_p)},$$

$$\hat{\gamma} = -\delta ln(a),$$

where

$$a = \frac{t - b^2}{1 - tb^2} \text{ and } t = \frac{x_n - x_0}{x_0 - x_p}.$$

For S_L curves,

$$\hat{\delta} = \frac{z_n}{\ln t}$$

where

$$t = \frac{x_n - x_0}{x_0 - x_p}.$$

To differentiate the three types of Johnson curves, the ratio

$$\frac{t_b}{t_u} = \frac{(x_m - x_0)(x_m - x_k)}{(x_n - x_m)(x_0 - x_p)}$$

is used. It is less than 1 for s_U, equal to 1 for S_L and greater than 1 for S_B.

Parameter Estimation of the Johnson System: Proposed MLE-Least Square Approach

A new algorithm to estimate parameters of Johnson's distribution is now proposed; this algorithm is named the MLE-Least Square Approach, because both maximum likelihood and least square approaches were employed to estimate the four parameters. Although the maximum likelihood equations for γ and δ were derived by Storer (1987), there are no closed form solutions for ξ and λ. The idea of combining both a maximum likelihood approach and least square theory makes the derivation of all four parameters more tractable analytically.

The probability density functions of the members of the Johnson family are known. First consider the S_U and S_B family of the Johnson system. Using the general form of Johnson densities (see equation 2.5), the likelihood function is:

$$L(x) = \frac{\delta^n}{\lambda^n (2\pi)^{n/2}} \prod_{i=1}^{n} g'(\frac{x - \xi}{\lambda}) e^{-\frac{1}{2}\sum_{i=1}^{n} (\gamma + \delta g(\frac{x-\xi}{\lambda}))^2},$$

and the log-likelihood is,

$$logL = n \log \delta - n \log \lambda - n/2 \log(2\pi) + \sum_{i=1}^{n} g'(\frac{x - \xi}{\lambda}) - \frac{1}{2} \sum_{i=1}^{n} (\gamma + \delta g(\frac{x - \xi}{\lambda}))^2$$

Setting the partial derivatives with respect to δ to zero,

$$\frac{n}{\delta} - \delta \sum [g(\frac{x - \xi}{\lambda})]^2 - \gamma \sum g(\frac{x - \xi}{\lambda}) = 0$$

which can be written as,

$$\delta^2 \sum [g(\frac{x - \xi}{\lambda})]^2 + \gamma \delta \sum g(\frac{x - \xi}{\lambda}) - n = 0 \quad (3.1)$$

Setting the partial derivatives with respect to γ to zero,

$$n\gamma + \delta \sum g(\frac{x - \xi}{\lambda}) = 0$$

which yields,

$$\hat{\gamma} = \frac{-\delta \sum g(\frac{x - \xi}{\lambda})}{n} = -\delta \bar{g} \quad (3.2)$$

Using (3.3) in (3.2):

$$\hat{\delta}^2 = \frac{n}{\sum[g(\frac{x-\xi}{\lambda})]^2 - \frac{1}{n}[\sum g(\frac{x-\xi}{\lambda})]^2} = \frac{1}{var(g)} \tag{3.3}$$

where $\bar{g}$ is the mean and $var(g)$ is the variance of the values of g defined in (2.6).

The partial derivatives of the log-likelihood with respect to ξ and λ are not simple. Storer (1987) presents a lengthy strategy for obtaining the solutions of these parameters. In the maximum likelihood estimation method, Kamziah, et al. (1999) applied the Newton-Raphson iteration to maximize the log likelihood of the Johnson distribution. They observed that, for some samples, the log likelihood function does not have a local maximum with respect to parameters ξ and λ. This non-regularity of the likelihood function caused occasional non-convergence of the Newton-Raphson iteration that was used to maximize the log-likelihood (Hosking, 1985)

The least squares method is applied herein to estimate parameters ξ and λ. From (2.1), $x = \xi + \lambda f^{-1}(\frac{z-\gamma}{\delta})$ is obtained. For fixed values of γ and δ, this equation may be considered as a linear equation with parameters ξ and λ.

The sum of squares of errors is,

$$S(\xi,\lambda) = \sum[x - \xi + \lambda f^{-1}(\frac{z-\gamma}{\delta})]^2.$$

To determine the value of ξ and λ that minimizes $S(\xi,\lambda)$, the partial derivatives of $S(\xi,\lambda)$ with respect to ξ and λ are calculated and these partial derivatives are equated to zero. The following two equations, called normal equations, are then obtained:

$$\sum x = n\xi + \lambda\sum f^{-1}(\frac{z-\gamma}{\delta}) \tag{3.4}$$

$$\sum x f^{-1}(\frac{z-\gamma}{\delta}) = \xi\sum f^{-1}(\frac{z-\gamma}{\delta}) + \lambda\sum[f^{-1}(\frac{z-\gamma}{\delta})]^2$$

Note that z is a standard normal variate. The quantiles of x and the corresponding quantiles of z can be considered paired observations. If there are 100 or more x values, the percentiles 1 through 99 would be considered. If the number of data points of x is k where k is less than 100, $k-1$ quantiles of x and the corresponding $k-1$ quantiles of z would be considered as paired observations.

Solving the normal equations results in

$$\hat{\lambda} = \frac{n\sum x f^{-1}(\frac{z-\gamma}{\delta}) - \sum f^{-1}(\frac{z-\gamma}{\delta})\sum x}{n\sum[f^{-1}(\frac{z-\gamma}{\delta})]^2 - [\sum f^{-1}\frac{z-\gamma}{\delta}]^2} \tag{3.5}$$

and

$$\hat{\xi} = \bar{x} - \lambda * mean[f^{-1}(\frac{z-\gamma}{\delta})] \tag{3.6}$$

where $\bar{x}$ is the mean of x-quantiles and $\bar{z}$ is the mean of z-quantiles used in the above equations. Starting with some initial values of ξ and λ, these initial values may be taken as the estimates obtained by any one of the previous methods. The estimates of γ and δ are then calculated using equations (3.2) and (3.3). After the estimates of γ and δ are obtained, equations (3.5) and (3.6) can be used to revise the ξ and λ estimates. Now these steps may be repeated, each time using the most recent estimates; the Residual Sum of Squares(RSS) can be tracked and, after a few steps, the estimate with minimum RSS value selected.

For the S_L family, consider the transformation in equation (2.2), so that there are only 3 parameters included. The probability density function can be given by,

$$p(x) = \frac{\delta}{\sqrt{2\pi}}\frac{1}{(x-\xi)}e^{-\frac{1}{2}[\gamma^* + \delta\ln(x-\xi)]^2}$$

The likelihood function is,

$$L(x) = \frac{\delta^n}{(2\pi)^{n/2}} \frac{1}{\prod(x-\xi)} e^{-\frac{1}{2}\sum[\gamma^* + \delta \ln(x-\xi)]^2}$$

Setting the partial derivative of log-likelihood with respect to δ to zero we get,

$$\frac{n}{\delta} - \delta \sum[\ln(x-\xi)]^2 - \gamma^* \sum[\ln(x-\xi)] = 0$$

which can be written as,

$$\delta^2 \sum[\ln(x-\xi)]^2 + \gamma^* \delta \sum[\ln(x-\xi)] - n = 0. \tag{3.8}$$

Setting the partial derivative of log-likelihood with respect to γ^* to zero,

$$n\gamma^* + \delta \sum[\ln(x-\xi)] = 0$$

which gives,

$$\begin{aligned} \hat{\gamma}^* &= -\frac{1}{n}\delta \sum[\ln(x-\xi)] \\ &= -\delta \bar{g}^*. \end{aligned} \tag{3.9}$$

Using (3.9) in (3.8) and solving for δ, results in

$$\begin{aligned} \hat{\delta}^2 &= \frac{n}{\sum[\ln(x-\xi)]^2 - \frac{[\sum \ln(x-\xi)]^2}{n}} \\ &= \frac{1}{var(g^*)} \end{aligned} \tag{3.10}$$

where $g^* = \ln(x-\xi)$. To estimate ξ, as before, use the method of least squares in the equation

$$x = \xi + f^{-1}(\frac{z-\gamma^*}{\delta}).$$

The sum of squares of errors is,

$$S(\xi) = \sum(x - \xi + f^{-1}(\frac{z-\gamma^*}{\delta}))^2$$

To find the value of ξ that minimizes $S(\xi)$, obtain

$$\frac{dS}{d\xi} = -2\sum(x - \xi - f^{-1}(\frac{z-\gamma^*}{\delta}))$$

Setting this derivative equal to zero, results in:

$$\hat{\xi} = \bar{x} - mean[f^{-1}(\frac{z-\gamma^*}{\delta})]$$

Here the same situation arises, the estimate of ξ depends on γ^* and δ and vice versa; as in the case of the S_U and S_B distributions. Thus, start with some initial value of ξ to estimate γ^* and δ, then use these estimated values to estimate ξ. Repeat this procedure, keeping track of RSS, and choose the one with least RSS.

Results

Data of size 2,000 were simulated from the S_U, S_B and S_L distributions to compare different methods of estimation. Twenty samples of size 2,000 were generated from each of the three specified models. The mean and the Mean Square Error (MSE) of the estimated values of the S_B, S_U, and S_L families are shown in Tables 1, 2 and 3. It can be observed that the average of the estimates are close to the true values of the parameters and, in general, the MSE of the estimates are smaller in the proposed method than the other methods.

Conclusion

A new approach that makes use of both the maximum likelihood procedure and least square theory was proposed to estimate the four parameters of the Johnson family of distributions. The new MLE-Least Square approach is compared with two other commonly used methods. The simulation study shows that the MLE-Least square approach gives better results for the S_B, S_U and S_L families. The findings of this study should be useful for applied practitioners.

Table 1: Mean and (Mean Square Error-MSE) of Parameter Estimates for the Johnson S_B Family

Sl. No.	Parameter	True Value	Percentile Method	Quantile Method	MLE-Least Square Approach
1	γ	1	0.998(0.167)	1.063(0.409)	0.997(0.026)
	δ	1	1.001(0.059)	1.024(0.083)	0.997(0.026)
	ξ	10	10.047(0.085)	9.982(0.131)	9.93(0.08)
	λ	10	10.049(5.92)	10.402(14.37)	10.57(4.99)
2	γ	0.5	0.503(0.009)	0.503(0.0493)	0.494(0.007)
	δ	0.5	0.505(0.003)	0.519(0.023)	0.507(0.001)
	ξ	10	9.11(4.038)	9.97(0.077)	10.004(0.004)
	λ	10	10.005(0.285)	10.094(1.614)	9.868(2.056)
3	γ	1	1.032(0.065)	1.01(0.015)	1.016(0.017)
	δ	0.5	0.507(0.0039)	0.5006(0.0013)	0.509(0.002)
	ξ	10	9.698(.488)	10.001(0.001)	10.001(0.001)
	λ	10	10.355(4.63)	10.085(0.69)	9.86(0.70)
4	γ	0.5	0.558(0.287)	0.539(0.136)	0.561(0.165)
	δ	1	1.013(0.191)	1.024(0.108)	1.055(0.115)
	ξ	10	9.82(1.097)	9.94(0.55)	9.91(0.52)
	λ	10	10.31(15.4)	10.30(8.2)	9.83(0.50)

Table 2: Mean and (Mean Square Error-MSE) of Parameter Estimates for the Johnson S_U Family

Sl. No.	Parameter	True Value	Percentile Method	Quantile Method	MLE-Least Square Approach
1	γ	0	0.04(0.32)	0.015(0.05)	0.015(0.05)
	δ	2	1.41(3.3)	2.08(0.34)	2.05(0.29)
	ξ	10	10.24(8.9)	10.1(1.5)	10.1(1.4)
	λ	10	12.3(99.9)	10.5(12.6)	10.3(10.1)
2	γ	0.5	0.82(2.9)	0.52(0.11)	0.51(0.09)
	δ	2	2.47(3.23)	2.08(0.45)	2.06(0.37)
	ξ	10	11.51(64.6)	10.06(2.79)	10.04(2.59)
	λ	10	12.07(56.5)	10.35(12.6)	10.25(11.22)
3	γ	0	-0.003(0.003)	0.005(0.002)	0.003(0.002)
	δ	1	1.033(0.006)	0.99(0.003)	0.99(0.002)
	ξ	10	10.03(.43)	10.05(0.25)	10.06(0.25)
	λ	10	10.45(1.43)	9.82(0.7)	9.75(0.73)
4	γ	0.5	0.514(0.009)	0.488(0.006)	0.487(0.007)
	δ	1	1.008(0.006)	0.999(0.006)	0.996(0.006)
	ξ	10	10.243(1.203)	9.95(0.9)	9.94(1.05)
	λ	10	10.06(0.96)	10.06(1.13)	10.02(1.43)

Table 3: Mean and (Mean Square Error-MSE) of Parameter Estimates for the Johnson S_L Family

Sl. No.	Parameter	True Value	Percentile Method	Quantile Method	MLE-Least Square Approach
1	γ^* (γ,λ)	1.303 (1,10)	-1.353(0.051)	-1.29(0.027)	1.303(0.04)
	δ	1	1.012(0.006)	0.97(0.008)	1.012(0.008)
	ξ	0	-0.98(0.14)	0.53(0.057)	0.53(0.057)
2	γ^* (γ,λ)	-2.3 (0,10)	-2.24(0.04)	-2.26(0.01)	-2.21(0.07)
	δ	1	0.98(0.003)	0.98(0.002)	0.98(0.007)
	ξ	0	0.18(0.41)	0.22(0.36)	0.33(0.28)
3	γ^* (γ,λ)	-5.91 (1,10)	-6.53(22.9)	-5.26(18.13)	-5.47(12.36)
	δ	3	3.18(2.28)	2.66(3.66)	2.87(1.42)
	ξ	0	-0.503(15.28)	0.72(18.3)	0.504(7.17)
4	γ^* (γ,λ)	-3.45 (1,10)	-3.78(3.26)	-3.45(0.99)	-3.45(1.63)
	δ	2	2.06(0.35)	1.88(0.35)	1.97(0.21)
	ξ	0	-0.13(4.12)	0.43(4.41)	0.29(1.67)

Acknowledgment

The authors are grateful to Dr. B. M. Golam Kibria for his valuable and constructive comments which improved the presetantion of the study.

References

Draper, J. (1952). Properties of distributions resulting from certain simple transformations of the normal distribution. *Biometrika*, *39*, 290-301.

George, F., Ramachandran, K. M., & Lihua, L. (2009). Gene selection with Johnson distribution. *Journal of Statistical Research*, *43*, 117-125.

Hahn, J. G., & Shapiro S. S. (1967). *Statistical models in engineering*. John Wiley & Sons, New York.

Hill, I. D., Hill, R., & Holder, R. L. (1976). Fitting Johnson curvesby moments. *Applied Statistics*, *25* 180-189.

Hosking, J. R. M., Wallis J. R., & Wood E. F. (1985). Estimation of the generalized extreme-value distribution by the method of probability-weighted moments. *Technometrics*, *27* 251-261.

Johnson, N. L. (1949). Systems of frequency curves generated by methods of translation. *Biometrika*, *36*, 149-176.

Kamziah, A. K., Ahmad, M. I, & Jaffirin, L. (1999). Nonlinear regresion approach to estimating Johnson SB parameters for diameter data. *Canadian Journal of Forestry Resources*, *29*, 310-314.

Slifker, J., & Shapiro, S. (1980). The Johnson system: Selection and parameter estimation. *Technometrics*, *22*, 239-247.

Storer, R. H. (1987). *Adaptive estimation by maximum likelihood: Fitting of Johnson distributions*. Unpublished Ph.D. thesis, School of Industrial and Systems Engineering, Georgia Institue of Technology.

Wheeler, R. (1980). Quantile Estimators of Johnson curve Parameters. *Biometrika*, *67*, 725-728.

Journal of Modern Applied Statistical Methods
November 2011, Vol. 10, No. 2, 505-512

1538 – 9472/11/$95.00

Error Analysis on the Generalized Negative Binomial Distribution

Felix Famoye
Central Michigan University,
Mt. Pleasant, MI

Oluwakemi Aremu
University of Lagos,
Akoka-Yaba, Lagos, Nigeria

The generalized negative binomial distribution characterized by three parameters, has been used to fit data from various fields of study. The distribution can model data for which the variance is larger or smaller than the mean, however, it becomes truncated under certain conditions. This truncation error is investigated via a detailed error analysis that determines the parameter space when the model can be used in place of the truncated generalized negative binomial distribution. The fitting of a generalized negative binomial distribution to a data set of absenteeism among shift-workers in a steel industry is re-analyzed.

Key words: Truncation error, dispersion, maximum likelihood estimates.

Introduction

A generalized negative binomial distribution (GNBD) was defined and studied by Jain and Consul (1971). The probability mass function of the GNBD is given by

$$P_x = P(X = x) = \begin{cases} \dfrac{m}{m+\beta x}\binom{m+\beta x}{x}\theta^x(1-\theta)^{m+\beta x-x}, & x = 0,1,2,\ldots \\ 0, & \text{for } x > k \text{ when } \beta < 0 \text{ or } 0 < \beta < 1, \end{cases} \quad (1.1)$$

and zero otherwise, where $0 < \theta < 1$, $m > 0$ and $\beta = 0$ or $0 < \beta < 1/\theta$ and k is the largest positive integer for which $m + 1 + (\beta - 1)k > 0$ when $\beta < 0$ or $0 < \beta < 1$. The GNBD in (1.1) reduces to the binomial distribution when $\beta = 0$ and m is an integer, and to the negative binomial distribution when $\beta = 1$. For the non-truncated GNBD, the mean and variance are

$$\mu = m\theta / (1 - \theta\beta)$$

and

$$\sigma^2 = m\theta(1-\theta) / (1-\theta\beta)^3 . \quad (1.2)$$

The moments in (1.2) exist when $\theta\beta < 1$.

Famoye and Consul (1993) defined and studied the truncated GNBD. The advantage of the truncated GNBD is that the distribution is defined for all values of β. However, the truncated GNBD is more difficult to estimate than the ordinary GNBD. The major difficulty is in finding suitable initial estimates for the model parameters.

All the estimation methods suggested by Famoye and Consul (1993) involve iterative procedure like the Newton-Raphson method. Because no estimation technique can be done without iteration, it is difficult to determine an initial estimate for the iteration. One way to obtain an initial estimate is to use the moment estimate of the non-truncated GNBD as the initial estimate; however, the moment estimates of non-truncated GNBD may not provide satisfactory initial estimates.

Famoye (1997) discussed parameter estimation for the GNBD. The asymptotic relative efficiencies of the estimators were compared. The method of first two moments and proportion of zeros (MOZE) has good efficiency when compared to the maximum likelihood estimates. From the simulation results, the MOZE method performed very well when both

Felix Famoye is a Professor in the Department of Mathematics. Email him at: felix.famoye@cmich.edu. Oluwakemi Aremu is a student in the Department of Mathematics. Email him at: chemmy413@yahoo.com.

bias and variance of the estimators were considered.

Nelson (1975) noted that the GNBD as first defined by Jain and Consul (1971) is truncated on the right hand side when $\beta < 0$. Also, the distribution gets truncated when $0 < \beta < 1$. Nelson (1975) remarking on GNBD stated that "A rigorous error analysis has not been performed, but it appears that for $n > -3\beta$, the error resulting from having negative value of β should be tolerable for most applications" (p. 136). The parameter n was replaced with m in (1.1), and to the best of our knowledge, no such error analysis has been conducted for the GNBD. One motivation for this study is to examine the error analysis for the GNBD when $\beta < 0$ and when $0 < \beta < 1$.

Due to the truncation described above, the sum of the probabilities in (1.1) may differ from unity. The difference between 1 and the sum of the probabilities (ΣP_x) is the truncation error. The percentage truncation error is computed as $100(1 - \Sigma P_x)$. Some illustrative examples for $k \leq 3$ are presented in Table 1. For two classes only, the truncation leads to only two probabilities P_0 and P_1, and the sum of the two probabilities could be very small or very large as shown in Table 1. As the values of θ decrease, the truncation error decreases. In general, the sum of the non-negative probabilities is much closer to 1 for small values of θ. As m increases, the value of k increases and, as the value of k increases, the truncation error decreases.

Other parameter sets can be used to illustrate the same phenomena. When $\beta < 1$ many of the cases shown in Table 1 satisfy the condition $m > -3\beta$, however, these values produce the sums of probabilities that are not close to 1. The statement that the error may be tolerable when $m > -3\beta$ does not seem to hold; more conditions than this are required. This study seeks to determine these other conditions such that the error will be tolerable or negligible. For example, in row 7 for $k = 1$, the sum of the probabilities is more than 3 on the account that the $P(X = 1)$ leads to $1 - \theta$ being raised to a negative power (see Table 1).

Review of the GNBD Dispersion Property

The GNBD model in (1.1) is over-dispersed (the variance is larger than the mean) when $\theta < (2\beta - 1)/\beta^2$, under-dispersed (the variance is smaller than the mean) when $\theta > (2\beta - 1)/\beta^2$ and equi-dispersed (the variance is equal to the mean) when $\theta = (2\beta - 1)/\beta^2$. These conditions differ from those given by Jain and Consul (1971), which involve the square root of $1 - \theta$. When $\beta \geq 1$, it is known that $\theta\beta < 1$ for the existence of the moments, therefore the condition for over-dispersion is always satisfied; hence, the GNBD is over-dispersed when $\beta \geq 1$. The GNBD model is under-dispersed whenever $\beta \leq 0.5$. When $0.5 < \beta < 1$, the GNBD is over-dispersed for all values of θ satisfying $0 < \theta < (2\beta - 1)\beta^{-2}$ and under-dispersed for values of θ satisfying $(2\beta - 1)\beta^{-2} < \theta < 1$. These results for the GNBD model can be summarized as follows:

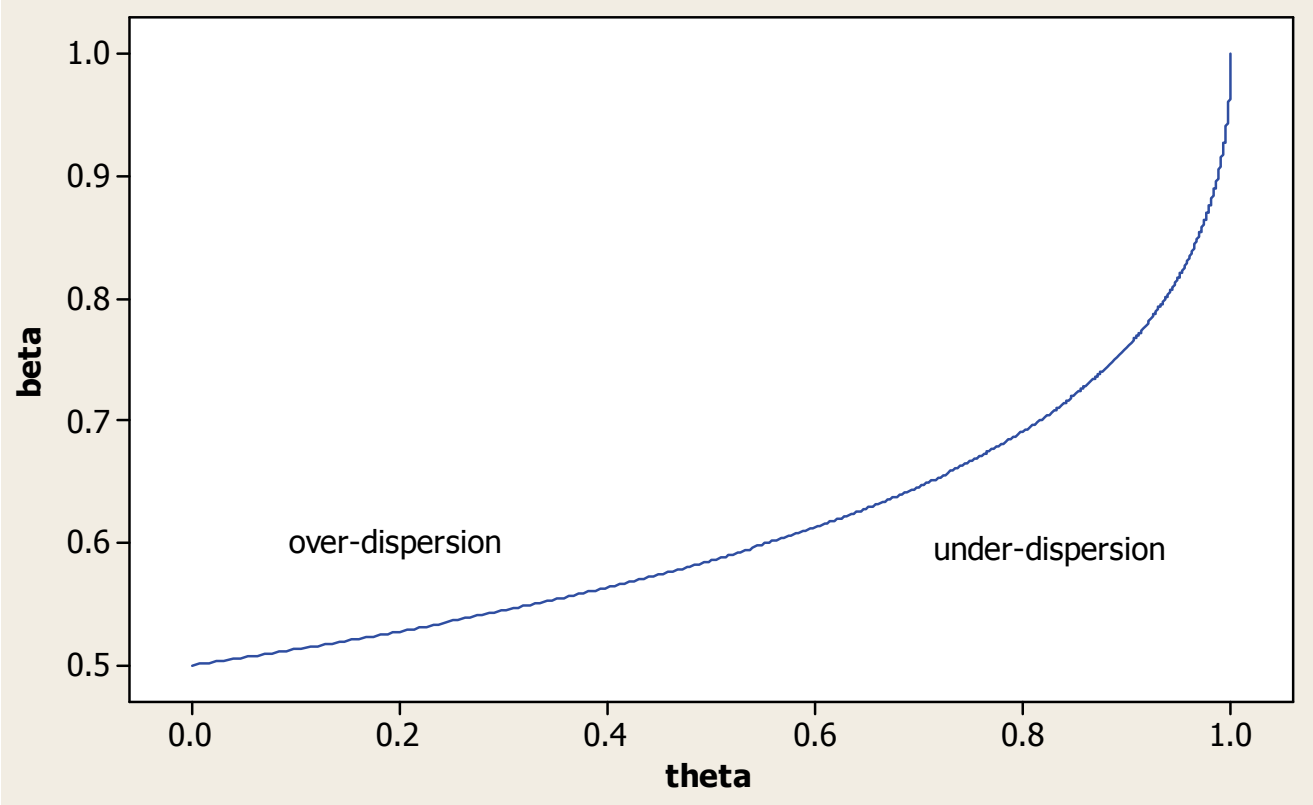
- It is over-dispersed (i) when $\beta \geq 1$ and (ii) when $0.5 < \beta < 1$ and $0 < \theta < (2\beta - 1)\beta^{-2}$.
- It is under-dispersed (i) when $\beta \leq 0.5$ and (ii) when $0.5 < \beta < 1$ and $(2\beta - 1)\beta^{-2} < \theta < 1$.
- It is equi-dispersed when $\theta = (2\beta - 1)\beta^{-2}$.
- The GNBD dispersion is independent of the parameter m.

Figure 1 shows the dispersion regions for the GNBD model: All points above the line $\theta = (2\beta - 1)/\beta^2$ represent the region where the GNBD model is over-dispersed, all points below the line represent the region where the model is under-dispersed, and all points on the line are where the GNBD model is equi-dispersed.

Table 1: Sum of Probabilities for Some GNBD Parameter Sets

k	Parameters			Probabilities				ΣP_x
	θ	β	m	P_0	P_1	P_2	P_3	
1	.95	–2	4.0	0.0000	0.1900			0.1900
	.50	–2	4.0	0.0625	1.0000			1.0625
	.05	–2	4.0	0.8145	0.1900			1.0045
	.95	–.5	1.6	0.0083	1.1265			1.1348
	.50	–.5	1.6	0.3299	0.7464			1.0763
	.05	–.5	1.6	0.9212	0.0796			1.0008
	.95	–.1	0.5	0.2236	2.8662			3.0898
	.50	–.1	0.5	0.7071	0.3789			1.0860
	.05	–.1	0.5	0.9747	0.0258			1.0005
	.95	.1	0.5	0.2236	1.5744			1.7980
	.50	.1	0.5	0.7071	0.3299			1.0370
	.05	.1	0.5	0.9747	0.0255			1.0002
2	.95	–2	7.0	0.0000	0.0000	0.3159		0.3159
	.50	–2	7.0	0.0078	0.2188	0.8750		1.1016
	.05	–2	7.0	0.6983	0.2851	0.0166		1.0000
	.95	–.5	2.6	0.0004	0.0915	2.3332		2.4251
	.50	–.5	2.6	0.1649	0.6065	0.2573		1.0287
	.05	–.5	2.6	0.8751	0.1229	0.0020		1.0000
	.95	–.1	1.5	0.0112	0.4299	1.6533		2.0944
	.50	–.1	1.5	0.3535	0.5684	0.0914		1.0133
	.05	–.1	1.5	0.9259	0.0735	0.0006		1.0000
	.95	.4	0.5	0.2236	0.6409	0.5571		1.4156
	.50	.4	0.5	0.7071	0.2679	0.0305		1.0055
	.05	.4	0.5	0.9747	0.0251	0.0002		1.0000
3	.95	–.5	3.6	0.0000	0.0063	0.4307	0.8388	1.2758
	.50	–.5	3.6	0.0825	0.4199	0.4750	0.0154	0.9928
	.05	–.5	3.6	0.8314	0.1616	0.0070	0.0000	1.0000
	.95	–.1	2.5	0.0006	0.0358	0.5970	0.9419	1.5753
	.50	–.1	2.5	0.1768	0.4737	0.3300	0.0218	1.0023
	.05	–.1	2.5	0.8796	0.1163	0.0040	0.0000	0.9999
	.95	.6	0.5	0.2236	0.3520	0.3880	0.2269	1.1905
	.50	.6	0.5	0.7071	0.2332	0.0639	0.0066	1.0008
	.05	.6	0.5	0.9747	0.0249	0.0004	0.0000	1.0000

Figure 1: Dispersion Region for the GNBD



Error Analysis of the GNBD

Re-writing the GNBD in (1.1), $P_x = m\theta^x(1-\theta)^{m+\beta x-x}[\Pi_{i=1}^{x-1}(m+\beta x-i)]/x!$. When $\beta < 0$ or $0 < \beta < 1$, it is required that $m + \beta x - x + 1 \geq 0$. If this condition is not satisfied, then P_x is set to 0 as shown in (1.1). Thus, the largest x value can be obtained from $0 \leq m + 1 + (\beta - 1)x \Rightarrow (1 - \beta)x \leq m+1 \Rightarrow x \leq (m + 1)/(1 - \beta)$ because $1 - \beta > 0$. The largest x value, k, is given by the integer part of $(m + 1)/(1 - \beta)$. Through computation, a detailed error analysis can be conducted on the GNBD model when $\beta <$ 0 and $0 < \beta < 1$. This analysis considers the values of m and θ in the parameter space of the model and the values of β when the truncation occurs; the values of $m > 0$, $0 < \theta < 1$, $\beta < 0$ and $0 < \beta < 1$. Observe that $\theta\beta$ is always less than 1 when truncation occurs. In the analysis, the values of $P(X = x)$ are computed for $x = 0, 1, 2, \ldots, k$, where k is such that $k \leq (m+1)/(1-\beta)$, and where $\beta < 0$ or $0 < \beta < 1$. In addition to these probabilities, the mean and variance of the truncated model are computed using the formulas $\mu_* = \Sigma x P_x / \Sigma P_x$ and $\sigma_*^2 = \Sigma x^2 P_x / \Sigma P_x - (\mu_*)^2$. After obtaining these values, percentage truncation errors in the sum of probabilities, the means and the variances are calculated using the formulas $100(1 - \Sigma P_x)$, $100(1 - \mu_*/\mu)$, and $100(1 - \sigma_*^2/\sigma^2)$, respectively.

In fitting the GNBD to an observed data set, the three parameters θ, β, and m must be estimated. In order to have at least 1 degree of freedom for the Chi-square goodness-of-fit test, at least five non-zero probability classes are needed. Thus, it is necessary that the smallest value of x be 4; therefore, in all analyses, the smallest x value is required to be 4. The

percentage error of truncation will be said to be tolerable or negligible if it is below 0.5%; in other words, the difference between 1 and the sum of all non-negative probabilities is below 0.005. This value was used by Consul and Shoukri (1985) in their error analysis for the generalized Poisson distribution. In view of this, the error analysis for $k \geq 4$ was conducted.

The maximum truncation error for the different values of m, θ, and β are provided in Table 2. Because at least five non-zero probability classes are needed, the different errors for cases where x is at least 4 are examined. In the error analysis the values of θ = 0.01(0.01)0.99, β = (–2.0)(0.01)(–0.01) and m = 0.1(0.1)(15.0) are considered.

Table 2 shows the ranges for the parameters that produce the maximum percentage error in the sum of the non-zero probabilities and specific parameter values at which the maximum truncation error occurs. The corresponding percentage errors in means and variances are also reported. For example, when $0 < \theta \leq 0.71$, $0.01 \leq \beta \leq 0.99$ and $0.1 \leq m \leq 0.5$, the maximum truncation error with at least 5 non-zero probability classes is –0.4799. When $0 < \beta < 1$, the percentage error in the means and percentage error in variances decrease as m increases. As m values increase, the range of θ values decreases in order to have a maximum truncation error of less than 0.5%. As the number of non-zero probability classes increases, the truncation error decreases.

When $0 < \beta < 1$ and $k \geq 4$, the GNBD can be used in general when $0 < \theta \leq 0.57$ for any value of $m > 0$. If $m < 1$, the range of θ values increases to $0 < \theta \leq 0.65$. When $\beta < 0$ and $k \geq 4$, the GNBD can be used in general when $0 < \theta \leq 0.36$ for $m \geq 4$. When $–1 < \beta < 0$ and $k \geq 4$, the range of θ values increases to $0 < \theta \leq 0.46$ for $4 \leq m \leq 10$.

Application to the Absenteeism Numbers among Shift-Workers

Gupta and Ong (2004) defined a new generalization of the negative binomial distribution by mixing the mean of the Poisson distribution with that of a generalized gamma distribution. The probability mass function of their generalized negative binomial distribution,

Table 2: Maximum Percentage Error and Corresponding Percentage Errors in Means and Variances (k = 5)

Range of Parameter Values			% Error (θ, β, m)	Percentage Errors	
θ	β	m		Means	Variances
[.01, .71]	[.01, .99]	[0.1, 0.5]	–0.4799 (0.71, 0.63, 0.5)	–3.2261	–13.8517
[.01, .65]	[.01, .99]	[0.1, 1.0]	–0.4761 (0.66, 0.53, 1.0)	–1.8264	–8.1959
[.01, .61]	[.01, .99]	[0.1, 2.0]	–0.4547 (0.61, 0.32, 2.0)	–0.9883	–4.8586
[.01, .57]	[.01, .99]	[0.1, 5.0]	–0.4536 (0.57, 0.01, 3.5)	–0.6274	–3.4805
[.01, .57]	[.01, .99]	[3.6, 5.0]	–0.4440 (0.57, 0.01, 3.6)	–0.5878	–3.1860
[.01, .57]	[.01, .99]	[5.0, 15]	–0.0947 (0.57, 0.01, 5.5)	–0.1105	–0.8318
[.01, .54]	[–.99, –.01]	[4.0, 5.0]	–0.4656 (0.54, –0.3, 5.0)	–0.4952	–3.0429
[.01, .46]	[–.99, –.01]	[5.0, 10]	0.4329 (0.46, –0.99, 7.0)	0.4981	4.1317
[.01, .39]	[–2.0, –.01]	[4.0, 10]	0.4397 (0.39, –1.66, 10)	0.4597	3.9250
[.01, .36]	[–2.0, –.01]	[10, 15]	0.4543 (0.36,–2.0,11.6)	0.4400	3.5627

characterized by four parameters, is in terms of the confluent hypergeometric function of the second kind. This new distribution is fitted to a data set on absenteeism among shift-workers in a steel industry. The data comes from Arbous and Sichel (1954). Gupta and Ong (2004) also fitted the data to the GNBD in (1.1) and obtained the following maximum likelihood estimates (MLE): $\hat{\theta}$ = 0.00010775, $\hat{\beta}$ = 5978.5288 and $\hat{m}$ = 29337.08391. They remarked that, because the parameter θ is small and both β and m are large, the fit by the GNBD corresponds to the fit by the generalized Poisson distribution. These large values of β and m and the small value of θ piqued our curiosity to re-analyze the data.

Famoye (1997) stated that the MOZE estimators are better than the moment estimators and they have good efficiency when compared to the MLE. In view of this, the moment estimates and the MOZE estimates of the GNBD in (1.1) were computed. The moment estimates of θ, β and m are respectively 0.9443, 0.9582, and 0.9058. The corresponding results for the MOZE method are $\overline{\theta}$ = 0.4590, $\overline{\beta}$ = 1.5323 and $\overline{m}$ = 5.8071.

Using the moment estimates as the initial for MLE and the Newton-Raphson method in SAS PROC NLMIXED, the ML estimates for the parameters did not reach acceptable convergence. After reaching convergence, the SAS warning that at least one of the gradients is more than 1.0e–3 (i.e. 0.001) was noted. In this analysis, two of the gradients were over 0.001 and the greater value is 0.0072. However, when the initial estimates are taken to be the MOZE estimates, there was proper convergence to the MLE (see Table 3). The maximum gradient was 1.141e–8. The MLEs in Table 3 are very far from the values given by Gupta and Ong (2004). Gupta and Ong did not report what they took as the initial estimates in finding the MLE. It appears the initial estimates might have caused their estimates to be too small or too large.

Based on the MLE result for parameter β, the negative binomial distribution (NBD) should provide an adequate fit to the data. Table 3 shows the fit by the GNBD and the NBD. Exact MLEs reported by Gupta and Ong (2004) for the NBD were not obtained in this study, however, estimates are not far from their results.

Although Gupta and Ong (2004) found that their new GNBD provided an adequate fit to the data, the GNBD in (1.1) also provides an adequate fit. In this example, the MLEs of β ($\hat{\beta}$ = 1.0824) is in the parameter region when the sum of the probabilities is 1. This parameter estimate for β is not significantly different from β = 1.0, for which the GNBD reduces to the NBD. The log-likelihood for both the GNBD and NBD are respectively equal to –793.91 and –794.00. This also shows that the NBD provides an adequate fit to the data.

Conclusion

When $\beta < 0$ or $0 < \beta < 1$, the truncated GNBD can be used. However, due to estimation problems with the truncated GNBD, the non-truncated GNBD should be considered if the truncation error is negligible. This study provides the region of the parameter space for which the truncation error is below 0.5%. It is important to ensure that the number of non-zero probability classes is at least five (that is, $k \geq 4$). By using the parameter region specified in Table 2, it can be determined whether the estimated parameter values are in the region where the truncation error is negligible.

Jain and Consul (1971) applied the non-truncated GNBD to four data sets. The number of non-zero frequency classes and the parameter estimates given by Jain and Consul (1971) are provided in Table 4. In all data sets, the estimated values of β are between 0 and 1. For data sets 1, 2 and 3, the number of non-zero frequency classes is over 5 and the truncation error is expected to be negligible. In data set 4, there are exactly 5 non-zero frequency classes. However, in comparing the parameter estimates with the regions in Table 2, the maximum truncation error is –0.4547. Computed truncation errors for these data sets are: 0.0351%, 0.2616%, 0.0053% and 0.0182% for data sets 1 through 4 respectively. Thus, the truncation error is negligible for all data sets considered by Jain and Consul (1971).

Table 3: Absenteeism Numbers among Shift-Workers

Count	Observed Frequency	NBD	New GNB by GO[a]	GNBD by JC[b]
0	7	11.13	9.23	10.02
1	16	15.74	16.18	15.70
2	23	17.77	19.86	18.39
3	20	18.36	21.06	19.20
4	23	18.10	20.50	18.89
5	24	17.32	18.78	17.94
6	12	16.24	16.46	16.66
7	13	15.01	14.02	15.22
8	9	13.72	11.79	13.76
9	9	12.43	9.95	12.33
10	8	11.19	8.55	10.99
11	10	10.01	7.54	9.74
12	8	8.91	6.84	8.61
13	7	7.90	6.33	7.58
14	2	6.98	5.94	6.67
15	12	6.14	5.61	5.85
16	3	5.40	5.29	5.13
17	5	4.73	4.97	4.49
18	4	4.13	4.64	3.92
19	2	3.61	4.28	3.43
20	2	3.14	3.92	2.99
21	5	2.73	3.55	2.61
22	5	2.37	3.19	2.28
23	2	2.06	2.84	1.99
24	1	1.78	2.50	1.74
25 – 48	16	11.10	14.13	11.87
Total	248	248.00		248.00
$\hat{\theta}$		0.8525 (0.0157)		0.7435 (0.3284)
$\hat{m}$		1.6792 (0.1775)		2.3580 (2.4079)
$\hat{\beta}$				1.0824 (0.3264)
[c]Chi-Square		15.97	8.27	13.27
df		17	15	16
p-value		0.5260	0.9125	0.6529

[a]Gupta and Ong (2004); [b]Jain and Consul (1971); [c]Adjacent classes for Chi-square values were combined as in Gupta and Ong (2004)

Table 4: Parameter Estimates for Data Sets Analyzed by Jain and Consul (1971)

Data Set	Number of Non-Zero Frequency Classes	Parameter Estimates		
		$\tilde{\theta}$	$\tilde{\beta}$	$\tilde{m}$
1 (in Table 1 of JC[a])	6	0.6013	0.8020	0.4006
2 (in Table 2 of JC)	8	0.7806	0.8549	0.4886
3 (in Table 3 of JC)	11	0.3531	0.0389	11.3188
4 (in Table 4 of JC)	5	0.3171	0.5496	1.5884

[a]Jain and Consul (1971)

Acknowledgements

This work was conducted while Felix Famoye, Central Michigan University, was on sabbatical leave at the Department of Mathematics, University of Lagos, Nigeria. The author gratefully acknowledges the support received from the U.S. Department of State, Bureau of Education and Cultural Affairs under the grant #09-78737.

References

Arbous, A. G., & Sichel, H. S. (1954). New techniques for the analysis of absenteeism data. *Biometrika*, *41*, 77-90.

Consul, P. C., & Shoukri, M. M. (1985). The generalized Poisson distribution when the sample mean is larger than the sample variance. *Communications in Statistics – Simulation and Computation*, *14*(*3*), 667-681.

Famoye, F. (1997). Parameter estimation for generalized negative binomial distribution. *Communications in Statistics – Simulation and Computation*, *26*(*1*), 269-279.

Famoye, F., & Consul, P. C. (1993). The truncated generalized negative binomial distribution. *Journal of Applied Statistical Science*, *1*(*2*), 141-157.

Gupta, R. C., & Ong, S. H. (2004). A new generalization of the negative binomial distribution. *Computational Statistics and Data Analysis*, *45*, 287-300.

Jain, G. C., & Consul, P. C. (1971). A generalized negative binomial distribution. *SIAM Journal of Applied Mathematics*, *21*(*4*), 501-513.

Nelson, D. L. (1975). Some remarks on generalization of the negative binomial and Poisson distributions. *Technometrics*, *17*(*1*), 135-136.

Journal of Modern Applied Statistical Methods
November 2011, Vol. 10, No. 2, 513-527

1538 – 9472/11/$95.00

Ordinal Regression Analysis: Predicting Mathematics Proficiency Using the Continuation Ratio Model

Xing Liu
Eastern Connecticut State University,
Willimantic, CT

Ann A. O'Connell
The Ohio State University,
Columbus OH

Hari Koirala
Eastern Connecticut State University,
Willimantic, CT

One commonly used model to analyze ordinal response data is the proportional odds (PO) model. However, if research interest is focused on a particular category and if an individual must pass through lower categories before achieving a higher level, the continuation ratio (CR) model is a more appropriate choice than the PO model. In addition, statistical software, such as Stata and SAS, may use different techniques to estimate the parameters. The CR model is used to illustrate the analysis of ordinal data in education using Stata and SAS and compares the results of fitting the CR model between these two packages.

Key words: Continuation ratio models, proportional odds models, ordinal regression analysis, mathematics proficiency, Stata, SAS, comparison.

Introduction

Ordinal data are abundantly collected in educational research. For example, it is common for data on student's SES to be ordered from low to high, responses to a survey item scaled from strongly disagree to strongly agree, children's reading proficiency scored from level 0 to 5 or students' educational proficiency levels in a state test ranging from fail to pass to proficient. One commonly used model to analyze ordinal data is the proportional odds (PO), or cumulative odds, model (Agresti, 1996, 2002, 2007; Armstrong & Sloan, 1989; Hilbe, 2009; Liu; 2009; Long, 1997, Long & Freese, 2006; McCullagh, 1980; McCullagh & Nelder, 1989; O'Connell, 2000, 2006; O'Connell & Liu, 2011; Powers & Xie, 2000).

The PO model is used to estimate the cumulative probability of being at or below a particular level of a response variable, or being beyond a particular level, which is the complementary direction. However, when research is focused on a particular category, rather than at or below that category, given that an individual has achieved a higher level, the continuation ratio (CR) model (Fienberg, 1980; Hardin & Hilbe, 2007; Long & Freese, 2006) is a more appropriate choice than the PO model. In particular, the CR model is more appealing than other models when analyzing educational attainment data (Allison, 1999). The CR model is very useful in analyzing data such as student academic proficiency levels that are measured annually or frequently using a mastery test as under the No Child Left Behind Act (NCLB).

In a CR model, the ordinal categories represent successive stages, or proficiency levels, through which an individual can progress; for example, faculty ranks from assistant professor to associate professor to full professor, or educational attainment from high school diploma to Bachelor's degree, Master's degree and to doctorate degree. In both of these examples, individuals must pass through lower stages or levels in order to reach higher stages or

Xing Liu is an Associate Professor of Research and Assessment in the Education Department. Email him at: liux@easternct.edu. Ann A. O'Connell is a Professor in the Program in Quantitative Research, Evaluation, and Measurement (QREM) in the School of Educational Policy and Leadership within the College of Education and Human Ecology. Email her at: aoconnell@ehe.osu.edu. Hari Koirala is a Professor in the Education Department. Email: koiralah@easternct.edu.

levels. A CR model estimates the odds of being in a certain category relative to being beyond that category. In terms of probability, this model estimates the probability of being in a category, given that an individual has been in that category or beyond. In addition, because these two conditional probabilities are complementary, the model estimates the conditional probability of being beyond a category given a person has attained that particular category.

Although the PO model is commonly used, the CR model seems to be overlooked. In addition, not all general-purpose statistical software packages have developed procedures to directly estimate a CR model, and for those packages which are capable of conducting a CR analysis, they may use different parameterizations to estimate the model. However, no study has been conducted to identify these differences and clarify misunderstandings.

Ignoring these differences may result in erroneous interpretations of results. Therefore, it is critical for researchers to understand this model and apply it correctly. To fill this gap, this study was conducted to demonstrate the use of the continuation ratio (CR) model to predict the mathematics proficiency of high school students using Stata and SAS, and to compare the results of fitting the continuation ratio model between these two packages. Ordinal regression analyses were based on the data from the Educational Longitudinal Study of 2002 (ELS:2002) in which the ordinal outcome of students' mathematics proficiency was predicted from a set of students' classroom activities, such as, reviewing work from the previous day in math class, listening to teachers' lectures, copying notes from the board, using books besides textbooks, doing problem solving in class, using general and graphing calculators, using computers, explaining work orally and participating in student-led discussions.

Theoretical Framework: General Logistic Regression Model and the Proportional Odds Model

The binary logistic regression model predicts an outcome variable with two categories, with 1 = experiencing the event, and 0 = not experiencing the event. This model estimates the log odds of the outcome, and thus the probability of success on a set of predictors. The logistic regression model has the following form:

$$\begin{aligned} \ln(Y') &= \text{logit}\,[\pi(x)] \\ &= \ln\left(\frac{\pi(\underline{x})}{1-\pi(\underline{x})}\right) \\ &= \alpha+\beta_1 X_1+\beta_2 X_2+\ \ldots+\beta_p X_p \end{aligned} \tag{1}$$

An ordinal logistic regression model is a generalization of a binary logistic regression model when the outcome variable has more than two ordinal levels. It estimates the probability of being at or below a specific outcome level, conditional on a collection of explanatory variables. The ordinal logistic regression model can be expressed as a latent variable model (Agresti, 2002; Greene, 2003; Long, 1997, Long & Freese, 2006; Powers & Xie, 2000; Wooldridge & Jeffrey, 2001). Assuming a latent variable, Y* exists, Y* can be defined as a function of a set of predictor variables and a random error. Let Y* be divided by some cut points (thresholds): α_1, α_2, α_3, …, α_j, and $\alpha_1 < \alpha_2 < \alpha_3 \ldots < \alpha_j$. The values of the observed ordinal variable, Y, fall within the regions divided by these cut points (thresholds). For example, Y = 0, if Y* $\leq \alpha_1$. The observed mathematics proficiency level is the ordinal outcome, y, ranging from 0 to 5, is defined as follows:

$$y=\begin{cases} 0 & if\ y^* \le \alpha_1 \\ 1 & if\ \alpha_1 < y^* \le \alpha_2 \\ 2 & if\ \alpha_2 < y^* \le \alpha_3 \\ 3 & if\ \alpha_3 < y^* \le \alpha_4 \\ 4 & if\ \alpha_4 < y^* \le \alpha_5 \\ 5 & if\ \alpha_5 < y^* \le \infty \end{cases} \tag{2}$$

Therefore, the probability of a student achieving each proficiency level and the cumulative probabilities as can both be predicted by: $P(Y \leq j) = F(\alpha_j - \mathbf{x\beta})$, where j = 1, 2, …, J–1.

Because different software packages utilize different parameterizations in estimating logit coefficients, the ordinal logistic regression model can be expressed in different forms (Liu, 2009). In Stata, it is expressed in logit form as follows:

$$\begin{aligned}\ln(Y_j') &= \text{logit}\,[\pi(x)] \\ &= \ln\left(\frac{\pi_j(x)}{1-\pi_j(x)}\right) \\ &= \alpha_j + \left(-\beta_1 X_1 - \beta_2 X_2 - \ \ldots - \beta_p X_p\right),\end{aligned} \tag{3}$$

where $\pi_j(\underline{x}) = \pi(Y \le j|x_1, x_2, \ldots, x_p)$, which is the probability of being at or below category j, given a set of predictors; j = 1, 2, …,J–1. α_j are the cut points, and β_1, β_2, …, β_p are logit coefficients. To estimate the ln (odds) of being at or below the j^{th} category, the PO model can be rewritten as:

$$\begin{aligned}&\text{logit}\,[\pi(Y \le j \mid x_1, x_2, ..., x_p)] \\ &= \ln\left(\frac{\pi(Y \le j \mid x_1, x_2, ..., x_p)}{\pi\left(Y > j \mid x_1, x_2, ..., x_p\right)}\right) \\ &= \alpha_j + \left(-\beta_1 X_1 - \beta_2 X_2 - \ldots - \beta_p X_p\right).\end{aligned} \tag{4}$$

This is the form of the proportional odds (PO) model because it assumes that the logit coefficients of any predictor are identical across all comparisons; this equal logit slope assumption can be assessed by the Brant test (Brant, 1990). Similar to the binary logistic regression, the PO model estimates the logit, or the log of the odds of being at or below a particular category versus being beyond that category. Thus, this model predicts cumulative logits across J–1 response categories. Methods of model diagnostics for the ordinal logistic regression models are provided by O'Connell and Liu (2011).

Just as Stata, the ordinal logit model is also based on the latent continuous outcome variable for SPSS PLUM, and it takes the same form. However, SAS uses a different ordinal logit model for estimating the parameters from Stata. For SAS PROC LOGISTIC (the ascending option), the ordinal logit model has the following form:

$$\begin{aligned}&\text{logit}\,[\pi(Y \le j \mid x_1, x_2, ..., x_p)] \\ &= \ln\left(\frac{\pi(Y \le j \mid x_1, x_2, ..., x_p)}{\pi\left(Y > j \mid x_1, x_2, ..., x_p\right)}\right) \\ &= \alpha_j + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p.\end{aligned} \tag{5}$$

Using SAS with the descending option, the ordinal logit model can be expressed as:

$$\begin{aligned}&\text{logit}\,[\pi(Y \ge j \mid x_1, x_2, ..., x_p)] \\ &= \ln\left(\frac{\pi(Y \ge j \mid x_1, x_2, ..., x_p)}{\pi\left(Y < j \mid x_1, x_2, ..., x_p\right)}\right) \\ &= \alpha_j + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p.\end{aligned} \tag{6}$$

where, in both equations, α_j are the intercepts, and β_1, β_2, β_p are logit coefficients.

Theoretical Framework: The Continuation Ratio Model

As notes, statistical software packages, such as Stata, SAS and SPSS, use different techniques to estimate the parameters in the proportional odds (PO) models (Liu, 2009). This is also true for the continuation ratio (CR) model: they use different formulations, estimate parameters differently, and produce different output results. When estimating the conditional probability of being beyond a category, given that individual has attained that particular category (e,g., $\pi(Y > j \mid Y \ge j \mid)$, the CR model can be expressed as (Allison, 1999; O'Connell, 2006):

$$\begin{aligned}&\ln\left(\frac{\pi\left(Y > j|x_1, x_2, \ldots x_p\right)}{\pi\left(Y = j|x_1, x_2, \ldots x_p\right)}\right) \\ &= \alpha_j + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p,\end{aligned} \tag{7}$$

where $\pi(Y > j \mid x_1, x_2, \ldots, x_p)$ is the conditional probability of being beyond a category j, conditional on being in that category, given a set of predictors. j =1, 2, …, J–1 and where α_j are the cut points and β_1, β_2, β_p are logit coefficients. SAS follows this form in estimating the continuation ratio model with the PROC LOGISTIC command. Before the model is fitted, the data set must be restructured following a series of steps (Allison, 1999; O'Connell, 2006).

First, separate sub-data set must be constructed with the binary outcome variable being beyond a category coded as 1 and 0 otherwise. Individuals who have not advanced to a particular proficiency level are dropped at each stage. If the ordinal dependent variable has j categories, J–1 sub-data sets should be created, these data sets are then combined into one data set with a new binary outcome variable with 1 = beyond a particular category. Finally, the CR model is fitted using the SAS PROC LOGISTICS with the descending option.

The CR models also estimates the odds of being in a particular category *j* relative to being beyond that category. In this situation, the CR model can be formulated as (Ananth & Kleinbaum, 1997; Armstrong & Sloan, 1989; Fienberg, 1980; Long & Freese, 2006):

$$\ln\left(\frac{\pi\left(Y = j|x_1, x_2, \ldots x_p\right)}{\pi\left(Y > j|x_1, x_2, \ldots x_p\right)}\right) = \alpha_j + \left(-\beta_1 X_1 - \beta_2 X_2 - \ldots - \beta_p X_p\right) \quad (8)$$

where $\pi(Y = j \mid x_1, x_2, \ldots, x_p)$ is the conditional probability of being in category j, conditional on being that category or beyond, given a set of predictors, and j =1, 2, …, J–1, α_j are the cut points, and β_1, β_2 …β_p are logit coefficients. Different from SAS, Stata follows this form to fit the CR model, which is known as the forward CR model (Bender & Bender, 2000). Another distinctive difference is that Stata does not require data restructuring before model fitting; this makes data analysis of the CR model much easier. The following analyses demonstrate how to fit a CR model using Stata; results of model fitting between Stata and SAS are also compared.

Methodology

Sample

Data were from the Educational Longitudinal Study (ELS, 2002). The ELS:2002 study was conducted by the National Center for Educational Statistics (NCES) and was designed to provide longitudinal data regarding the transitions of high school sophomores in 2002 to postsecondary school education and their future careers. In the 2002 base year of the study, more than 15,000 high school sophomores from a national sample of 752 public and private high schools participated in the study by taking cognitive tests and responding to surveys.

The outcome variable of interest was students' mathematics proficiency levels in high school, which was an ordinal categorical variable with five levels (1 = students can do simple arithmetical operations on whole numbers; 2 = students can do simple operations with decimals, fractions, powers and root; 3 = students can do simple problem solving; 4 = students can understand intermediate-level mathematical concepts and/or find multi-step solutions to word problems; and 5 = students can solve complex multiple-step word problems and/or understand advanced mathematical material) (Ingels, Pratt, Roger, Siegel & Stutts, 2004, 2005). The five proficiency domains were hierarchically structured: mastery of higher proficiency level indicated mastery of all previous levels. Students had to pass through the first four levels of proficiency before achieving the final fifth level; those students who failed to pass through level 1 were assigned to level 0. Table 1 shows the frequency of the six mathematics proficiency levels.

Data Analysis

The continuation ratio model is first fitted with a single explanatory variable using the Stata *ocratio* command (Wolfe, 1998) with the link functions of logit and CLOG-LOG, a proportional odds (PO) model was fitted next, and finally, a full-model with all 11 explanatory variables was fitted. The *eform* option was used to estimate the odds ratios and corresponding standard errors and the confidence intervals. The

ologit command in Stata was used to fit the proportional odds models. The results from both the CR models and the PO models were compared and interpreted. For comparison, the same model was fitted using SAS (V. 9.1.3).

Model fit statistics in the CR model, such as likelihood ratio test and Pseudo R^2, were reported. Other fit statistics, such as Hosmer-Lemeshow GoF test, and Pulkstenis-Robinson (2004) modification, are currently unavailable in the CR model. Following a suggestion by Hilbe (2009), the Stata AIC command was also used to compare model fit.

The log likelihood ratio Chi-Square test with 1 degree of freedom, LR $\chi^2_{(1)} = 38.90$, $p < 0.001$, indicated that the logit regression coefficient of the predictor, gender was statistically different from 0, therefore, the model with one predictor provides a better fit than the null model with no independent variables in predicting conditional probabilities for mathematics proficiency level. The Pseudo R^2=.0008, which is the likelihood ratio R^2_L, suggested that the relationship between the response variable, mathematics proficiency and the predictor (gender) was small: the AIC statistic was 0.922.

Results

Continuation Ratio Model with a Single Explanatory Variable

A continuation ratio model with a single predictor, gender, was fitted first. The Stata *ocratio* command with the logit function as default was used. Figure 1 displays the Stata output for the single predictor continuation ratio model.

Table 1: Proficiency Categories and Frequencies (Proportions) for the Study Sample, ELS 2002 (N = 15,976)

Proficiency Category	Description	Frequency
0	Did not pass level 1	842 (5.27%)
1	Can do simple arithmetical operations on whole numbers	3882 (24.30%)
2	Can do simple operations with decimals, fractions, powers, and root	3422 (21.42%)
3	Can do simple problem solving	4521 (28.30%)
4	Can understand intermediate-level mathematical concepts and/or find multi-step solutions to word problems	3196 (20.01%)
5	Can solve complex multiple-step word problems and/or understand advanced mathematical material	113 (0.71%)

Figure 1: Stata Continuation Ratio Model with Logit Link: Single Predictor, Gender

```
. ocratio Profmath BYGENDER, link (logit)

Continuation-ratio logit Estimates                    Number of obs =   51353
                                                      chi2(1)       =   38.90
                                                      Prob > chi2   =  0.0000
Log Likelihood =  -23683.4                            Pseudo R2     =  0.0008

------------------------------------------------------------------------------
    Profmath |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    BYGENDER |   .1416361   .0227235     6.23   0.000     .0970989    .1861732
------------------------------------------------------------------------------
 _cut1   |  -2.790613   .0372137            (Ancillary parameters)
 _cut2   |  -.9961043   .0219305
 _cut3   |  -.7736138   .0238228
 _cut4   |    .368887    .026111
 _cut5   |   3.392331   .0966743
------------------------------------------------------------------------------



. ocratio Profmath BYGENDER, link (logit) eform

Continuation-ratio logit Estimates                    Number of obs =   51353
                                                      chi2(1)       =   38.90
                                                      Prob > chi2   =  0.0000
Log Likelihood =  -23683.4                            Pseudo R2     =  0.0008

------------------------------------------------------------------------------
    Profmath | Odds ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    BYGENDER |   1.152157    .026181     6.23   0.000     1.101969    1.204631
------------------------------------------------------------------------------
 _cut1   |  -2.790613   .0372137            (Ancillary parameters)
 _cut2   |  -.9961043   .0219305
 _cut3   |  -.7736138   .0238228
 _cut4   |    .368887    .026111
 _cut5   |   3.392331   .0966743
------------------------------------------------------------------------------
. aic
AIC Statistic =   .9224153
```

The estimated logit regression coefficient, $\beta = 0.1416$, $z = 6.23$, $p < 0.001$, indicated that gender had a significant effect on mathematics proficiency. Substituting the value of the coefficient into the formula (8), logit $[\pi(Y= j \mid Y \geq j, \text{gender})] = \alpha_j + (-\beta_1 X_1)$, the logit $[\pi(Y= j \mid Y \geq j, \text{gender})] = \alpha_j - 0.1416$ (gender), OR = $e^{(-.1416)} = 0.8680$, was calculated indicating that male students were 0.8680 times the odds for female students of being in any category compared to being in higher categories, that is, female students were more likely than male students to drop out in a particular category, because males are coded as 1 and females are coded as 0.

To estimate the conditional probability of being beyond a category of mathematics proficiency, which is the complement of the conditional probability of being at a category, the signs before the cutpoints and the estimated logits in the equation (8) are changed and the logit $[\pi(Y>j \mid Y\geq j, \text{gender})] = -\alpha_j + 0.1416$ (gender) calculated. Exponentiating 0.1416, results in the OR = 1.152, which indicated that male students were 1.152 times more likely to be beyond a particular mathematics proficiency level than female students.

The CR model could also be fitted using the complementary log-log link (clog-log) with the cumulative option within the Stata *ocratio* command. The CR model with the complementary log-log link is actually the discrete-time proportional hazards model for the event history analysis or survival analysis (Allison, 1999; O'Connell, 2006). It estimates the hazard ratio (HR) rather than the odds ratio (OR) of being in a particular category relative to advancing to a higher category. Figure 2 displays the Stata output for the clog-log continuation model.

The log likelihood ratio Chi-Square test with 1 degree of freedom, LR $\chi^2_{(1)} = 51.38$, $p < 0.001$, indicating that the full model with one predictor provides a better fit than the null model with no independent variables. The Pseudo R^2=0.0011, suggested that the relationship between the response variable, mathematics proficiency, and the predictor, gender was small. The AIC statistic was 0.922

The estimated clog-log coefficient, $\beta = 0.1257$, $z = 7.17$, $p < 0.001$, indicating that gender had a significant effect on mathematics proficiency. Since Clog-log $[\pi(Y=j \mid Y\geq j, \text{gender})] = \log(-\log(1-\pi)) = \alpha_j + (-\beta_1 X_1)$, we calculated $\log(-\log(1-\pi)) = \alpha_j - 0.1257$ (gender). By exponentiating –0.1257, the hazard ratio, HR = $e^{(-.1257)} = 0.8819$ was obtained, indicating that the hazard of being in a particular proficiency level rather than beyond for male students was 0.8819 times the hazard for female students, that is, the hazard for female students of stopping out in a particular category was 1.134 times as great as that for male students.

Proportional Odds Model with a Single Explanatory Variable

Next, for comparison purposes, a proportional odds model analysis with the same single predictor, gender was conducted using the Stata *ologit* procedure. Figure 3 displays the Stata output for the one-predictor proportional odds model.

LR $\chi^2_{(1)} = 28.13$, $p < 0.001$, indicating that the one-predictor PO model provided a better fit than the null model with no independent variables in predicting cumulative probabilities for mathematics proficiency level. The Pseudo $R^2 = 0.0006$, which was as small as that in the continuation ratio model.

The estimated logit regression coefficient, $\beta = 0.1527$, $z = 5.30$, $p < 0.001$. Because the PO model estimates the cumulative odds and cumulative probabilities of being at or below a particular category of the ordinal response outcome, logit $[\pi(Y\leq j \mid \text{gender})] = \alpha_j - 0.1527$ (gender) was calculated. By exponentiating the logit, –0.1527, the odds ratio (OR), $e^{(-.1527)} = 0.8584$ was obtained, indicating that the odds of being at or below a mathematics proficiency level were 0.8584 times as great for male students as they were for female students, thus, female students were more likely than male students to be at or below a particular proficiency level.

The PO model can estimate J–1 cumulative probabilities of being at or below a category of the ordinal response variable with *j* levels. When the ordinal response variable, mathematics proficiency, has six levels from 0 to 5, the proportional odds model estimates five cumulative probabilities: $P(Y \leq 0)$, $P(Y \leq 1)$,

Figure 2: Stata Continuation Ratio Model with Clog-log Link: Single Predictor, Gender

```
. ocratio Profmath BYGENDER, link (cloglog) cumulative

Ordered cloglog Estimates                              Number of obs =   51353
                                                       chi2(1)       =   51.38
                                                       Prob > chi2   =  0.0000
Log Likelihood = -23677.16                             Pseudo R2     =  0.0011

------------------------------------------------------------------------------
    Profmath |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    BYGENDER |   .1256615   .0175265     7.17   0.000     .0913103    .1600128
-------------+----------------------------------------------------------------
 _cut1       |  -2.826367   .0356499             (Ancillary parameters)
 _cut2       |  -.9834265    .022463
 _cut3       |  -.2817271   .0217445
 _cut4       |   .5087509   .0202158
 _cut5       |   1.663668   .0274349
------------------------------------------------------------------------------

. aic
AIC Statistic =   .9221723



. ocratio Profmath BYGENDER, link (cloglog) eform cumulative

Ordered cloglog Estimates                              Number of obs =   51353
                                                       chi2(1)       =   51.38
                                                       Prob > chi2   =  0.0000
Log Likelihood = -23677.16                             Pseudo R2     =  0.0011

------------------------------------------------------------------------------
    Profmath | Haz. ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    BYGENDER |   1.133898   .0198732     7.17   0.000     1.095609    1.173526
-------------+----------------------------------------------------------------
 _cut1       |  -2.826367   .0356499             (Ancillary parameters)
 _cut2       |  -.9834265    .022463
 _cut3       |  -.2817271   .0217445
 _cut4       |   .5087509   .0202158
 _cut5       |   1.663668   .0274349
------------------------------------------------------------------------------
```

P(Y ≤ 2), P(Y ≤ 3) and P(Y ≤ 4). The cumulative probabilities of being beyond a category can also be estimated because they are the complementary probabilities of the being at or below a particular category.

Different from cumulative probabilities in the PO model, the logit CR model estimates conditional probabilities. In the gender-only CR model, it estimates conditional probabilities of being in category j, conditional on being at or beyond that category, that is, P (Y = j | Y ≥ j, gender). This CR model can also estimate the conditional probability of being beyond a category given that individual has achieved that particular category, because P (Y > j | Y ≥ j, gender) is the complementary form of P (Y = j | Y ≥ j, gender).

Another difference between the CR model and the PO model is the change in sample size. In the gender-only PO model, the sample size was 15,325, however, the number of observations increased to 51,353 in the CR model due to different comparisons between proficiency levels, which included level 0 versus levels 1, 2, 3, 4 and 5; level 1 versus levels 2, 3, 4 and 5; level 2 versus 3, 4 and 5; level 3 versus 4 and 5; and level 4 versus level 5 (Table 2 shows the comparisons between the six proficiency levels). Fitting the CR model using SAS required a restructured data set from the J−1concatenated sub-data sets from the comparisons between proficiency levels (Allison, 1999; O'Connell, 2006), though Stata can fit the CR model directly without the data restructuring procedure.

Figure 3: Stata Proportional Odds Model: Single Predictor, Gender

```
ologit Profmath BYGENDER

Iteration 0:   log likelihood = -23702.845
Iteration 1:   log likelihood = -23688.779
Iteration 2:   log likelihood = -23688.778

Ordered logistic regression                     Number of obs   =      15325
                                                LR chi2(1)      =      28.13
                                                Prob > chi2     =     0.0000
Log likelihood = -23688.778                     Pseudo R2       =     0.0006

------------------------------------------------------------------------------
    Profmath |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    BYGENDER |   .1527419   .0288057     5.30   0.000     .0962839       .2092
-------------+----------------------------------------------------------------
       /cut1 |  -2.785918   .0381689                     -2.860728   -2.711108
       /cut2 |  -.7893203   .0224898                     -.8333995   -.7452411
       /cut3 |   .1072826   .0214844                      .065174    .1493911
       /cut4 |   1.402499   .0246227                     1.354239    1.450758
       /cut5 |   4.981085    .095611                     4.793691    5.168479
------------------------------------------------------------------------------
```

Continuation Ratio Model with 11 Explanatory Variables

A CR model was fitted with 11 explanatory variables; this was referred to as the full model. Table 3 displays the results for the fitting of the full model with all the predictors.

The log likelihood ratio Chi-Square test, LR $\chi^2_{(11)} = 3069.32$, $p < 0.001$, indicating that the full model with 11 predictor provides a better fit than the null model with no independent variables in predicting conditional probability for mathematics proficiency. Although the likelihood ratio $R^2_L = 0.0777$, was much larger than that of the gender-only model, it was still fairly small, suggesting that the relationship between the response variable, mathematics proficiency and 11 predictors, was small. AIC Goodness-of-fit statistics were used for model comparisons using the AIC command (Hilbe, 2009). Compared with the gender-only model (0.9224), the AIC statistic indicated that the full-model fit the data better (0.8483).

Using the *eform* option, odds ratios could be obtained for all the predictors. Overall, these predictors, such as, being male students (gender), reviewing work from the previous day in math class (review), listening to teachers' lectures (listen), doing problem solving in class (probsolv), using general calculators (usecalcu), using graphing calculators (usegraph), and explaining work orally (explain), were positively associated with the odds of being beyond a particular mathematics proficiency level. Copying notes from board in class (copynote), using books besides textbooks (usebooks), using computers (usecompu), and participating in student-lead discussions (participate) were less likely to advance to a higher proficiency level, that is, they were more likely to stop out in a particular proficiency level.

In terms of odds ratios, male students had 1.359 times greater odds than female students to be beyond a given proficiency level (OR = 1.359), after controlling for the effects of other predictors in the full model. The odds of being beyond a particular proficiency level relative to being in that level were 1.166 times greater with one unit increase in the frequency of reviewing work from the previous day (OR = 1.166). Similarly, listening to teachers' lectures (OR = 1.192), doing problem solving in class (OR = 1.077), using general calculators (OR = 1.179), using graphing calculators (OR = 1.173), and explaining work orally (OR =1.066) were more likely to be in a higher proficiency level. Conversely, for every one unit increase in copying notes from board in class, the odds of being beyond a particular category decreased by a factor of 0.96 (OR = 0.96). In other words, the more the students copied notes from board, the more likely they would stop out in a mathematics proficiency level. Similarly, the odds decreased by a factor of 0.785 (OR = 0.785), for a unit increase in using textbooks besides the mathematics textbook, they decreased by a factor of 0.833 for a unit increase in using computers in math classes, and they decreased by a factor of 0.892 in participating in student-led discussions, holding the effects of the other variables constant.

Table 3 also provides the results of the multiple regression (MR) analysis. Although the results of MR analysis looked similar to those estimated by the CR model, they were different in nature: the former estimates the linear effects the classroom practices on mathematics proficiency level, while the latter estimates the conditional probability of being in a proficiency level relative to being beyond, or its complement, the probability of advancing to a higher proficiency level rather than being in that particular level. The MR analysis could be used as a preliminary analysis before the CR model fitting.

Comparison of Results of a Single Variable CR Logit Model Using Stata and SAS

When fitting CR models with logit link, Stata and SAS use different procedures to restructure data, estimate parameters differently and produce different outputs. It is, therefore, important to understand how data sets are restructured and how to interpret these estimates. Before using the LOGISTIC procedure, SAS requires a process of data restructuring in order to estimate conditional probabilities of not advancing to a higher proficiency level. If there are j categories, J−1 sub-data sets are needed. Because the mathematics proficiency includes six levels, five sub-data sets are created. Corresponding to the category comparisons indicated in Table 2 (i.e., level 0 versus level 1

Table 2: Category Comparisons for the Continuation Odds Model with Six Mathematics Proficiency Levels (j = 0, 1, 2, …, 5).

Proficiency Category	Conditional Probability P(Y= j | Y≥j)	Odds Ratio	Probability Comparisons
0	P(Y= 0 | Y≥ 0)	$\frac{P(Y=0)}{P(Y\geq 0)}$	Category 0 vs. all categories above
1	P(Y= 1 | Y≥ 1)	$\frac{P(Y=1)}{P(Y\geq 1)}$	Category 1 vs. Categories 2 through 5
2	P(Y= 2 | Y≥ 2)	$\frac{P(Y=2)}{P(Y\geq 2)}$	Category 2 vs. Categories 3 - 5
3	P(Y= 3 | Y≥ 3)	$\frac{P(Y=3)}{P(Y\geq 3)}$	Category 3 vs. Categories 4 and 5
4	P(Y= 4 | Y≥ 4)	$\frac{P(Y=4)}{P(Y\geq 4)}$	Category 4 vs. 5

Table 3: Results of the Continuation Ratio Model and the OLS Regression Model (Full Model), n = 42,992

Variable	Continuation Ratio Model (logit)		OLS Model
	b (se(b))	OR	
α_1	–1.50 (0.08)		1.15 (0.06)
α_2	0.49 (0.08)		
α_3	0.89 (0.08)		
α_4	2.27 (0.08)		
α_5	5.64 (0.13)		
Gender$^{\delta}$	0.31 (0.03)**	1.36	0.21 (0.02)**
Review	0.15 (0.01)**	1.17	0.12 (0.01)**
Listen	0.18 (0.01)**	1.19	0.13 (0.01)**
Copynote	–0.04 (0.01)**	0.96	–0.02 (0.01)*
Usebooks	–0.24 (0.01)**	0.79	–0.18 (0.01)**
Probsolv	0.07 (0.01)**	1.08	0.05 (0.01)**
Usecalcu	0.16 (0.01)**	1.18	0.12 (0.01)**
Usegraph	0.16 (0.01)**	1.17	0.11 (0.01)**
Usecompu	–0.18 (0.01)**	0.83	–0.14 (0.01)**
Explain	0.06 (0.01)**	1.06	0.05 (0.01)**
Participate	–0.11 (0.01)**	0.89	–0.09 (0.01)**
R^2	$R^2_L = 0.078$		$R^2 = 0.221$
Model Fit[a]	$\chi^2_{11} = 3039.32$ (p < 0.0001)		$F(11, 12768) = 329.24^{**}$

$^{\delta}$ gender: male=1; [a] Likelihood ratio test; *Significant at p<0.05; ** p<0.01

and above; level 1 versus level 2, and above; level 2 versus 3, 4 and 5; level 3 versus 4 and 5; and level 4 versus level 5), observations for students who did not make to the given proficiency level were dropped out of the concatenated data sets. These sub-data sets were merged into one data set with each individual having as many observations as the number of proficiency levels to which she/her could advance. A new binary variable was created in each data set with being beyond a category coded as 1 and 0 otherwise (see O'Connell, 2006 for details on data restructuring). Different from SAS, the Stata *ocratio* procedure does not require the above process because it restructures the data internally and produces the same sample size as that of the restructured data in SAS.

Table 4 presents a comparison of the results of fitting the single-variable CR model with logit link using both Stata *ocratio* and SAS PROC LOGISTIC with the descending option. In Stata, the CR model estimates the odds of being a particular category versus beyond; while this model in SAS with the descending option estimates the odds of being beyond a given category relative to being in that category, which are the reciprocal. Using Stata and SAS descending, the estimated coefficients are the same in both magnitude and sign. Using the Stata CR model equation (8), logit $[\pi(Y= j \mid Y \geq j, \text{gender})] = \alpha_j + (-\beta_1 X_1)$, logit $[\pi(Y= j \mid Y \geq j, \text{gender})] = \alpha_j - 0.1416$ (gender) was calculated, and OR = $e^{(-.1416)} = 0.8680$, indicating that male students were 0.8680 times the odds for female students of being in any category compared to being in higher categories.

To estimate the conditional probability of being beyond a category of mathematics proficiency using Stata, it is necessary to negate the signs before the cutpoints and the estimated logits in the equation (8) to get the complementary probability of being in a category conditional on being beyond, i.e., logit $[\pi(Y>j \mid Y\geq j, \text{gender})] = -\alpha_j + \beta_1 X_1$. Substituting the coefficient into the equation results in logit $[\pi(Y>j \mid Y\geq j, \text{gender})] = -\alpha_j + 0.1416$ (gender). Exponentiating 0.1416, resulted in the OR of 1.152, which indicated that male students were 1.152 times more likely to be beyond a particular mathematics proficiency level than female students. Using equation (7) for the SAS CR logit model, it was found that logit $[\pi(Y>j \mid Y\geq j, \text{gender})] = \alpha_j + 0.1416$ (gender). Exponentiating the logit coefficient 0.1416 resulted in the same odds ratio, 1.152.

The CR model using Stata also estimates the cutpoints based on different logit comparisons; these are useful to calculate the conditional probabilities. From the left to the right direction, five cutpoints were –2.791, –0.996, –0.774, 0.369, and 3.392. The results of the CR model using SAS descending as shown in Table 4 provide the estimated intercept, and dumcr0 through dumcr3, which are dummy coded variables for logit comparisons with the final comparison as the reference group. The intercept, –3.392, was the fifth cutpoint, α_5 , because it was used to find the odds of being beyond the proficiency level 4 relative to being in that level. The first cutpoint = intercept + dumcr0 = –3.392 + 6.182 = 2.790. The second cutpoint = intercept + dumcr1 = –3.392 +4.388 = 0.996. Using the same method resulted in the third, 0.773, and the fourth cutpoints, –0.369, respectively. Comparing the results of the cutpoints estimated by the CR model using Stata and SAS descending, it was found that they were the same in magnitude but had opposite signs. SAS does not provide direct estimates of these cutpoints, but they can be calculated from the estimated intercept and dummy variables.

Although the omnibus likelihood ratio tests for the CR model using Stata and SAS indicated that the single-variable model had better fit than the null model, their degrees of freedom (df) were different because SAS estimated four extra parameters: an intercept and three dummy variables. Accordingly, the log likelihood $R^2_L = 0.254$ estimated using SAS, was much larger than that using Stata, $R^2_L = .0008$. Both CR models had the same sample size when SAS restructured the data (N = 51,353). Feature comparisons of fitting the CR model with the logit link are provided in Table 5.

Table 4: Results of the CR Logit Models with a Single Variable Using Stata and SAS: A Comparison, n = 51,353 (Restructured Data

Model Estimates	STATA	SAS (Descending)
	P(Y= j | Y≥j)	P(Y> j | Y≥j)
Cutpoints (Stata)/ Intercept (SAS)	$\alpha_1 = -2.791$	Intercept= –3.392
	$\alpha_2 = -0.996$	Dumcr0 = 6.182
	$\alpha_3 = -0.774$	Dumcr1 = 4.388
	$\alpha_4 = 0.369$	Dumcr2 = 4.165
	$\alpha_5 = 3.392$	Dumcr3 = 3.023
BYGENDER[δ]	0.142 (0.023) **	0.142 (0.023) **
LR R^2	$R^2_L = 0.0008$	$R^2_L = 0.254$
Model Fit[a]	$\chi^2_1 = 38.90$ (p < 0.0001)**	$\chi^2_5 = 15040.557$ (p < 0.0001)**

[δ]BYGENDER: male=1; [a]Likelihood ratio test; Results are incomparable due to data restructuring using SAS; *Significant at p<0.05; ** p<0.01

Table 5: Feature Comparisons of the CR Model with Logit Link Using Stata and SAS

	STATA	SAS
Model Specification		
Cutpoints/ thresholds	√	
Intercept		√
Test hypotheses of logit coefficients	√	√
Maximum Likelihood Estimates		
Odds Ratio	√	√
z-statistic or Wald test for Parameter Estimate	√	
Chi-square Statistic for Parameter Estimate		√
Confidence Interval for Parameter Estimate	√	
Fit Statistics		
Log likelihood	√	√
Goodness-of-fit Test	√	√
Pseudo R-Square	√	√
Association of Predicted Probabilities and Observed Responses		√

Conclusion

This article illustrated the use of continuation ratio models to estimate high school students' mathematics proficiency from a set of predictors of classroom practices. Model fitting started from a single-variable CR with both logit and clog-log links and then progressed to a PO model, and finally a full CR logit model with 11 predictor variables.

Results from the CR models suggested that some classroom practices, such as reviewing work from the previous day in math class, listening to teachers' lectures, doing problem solving in class, using general calculators, using graphing calculators and explaining work orally, had positive effects on the odds of being beyond a particular mathematics proficiency level relative to being in that level; while other classroom practices, such as, copying notes from board, using books besides textbooks, using computers in class and participating in student-led discussions were associated with odds of stopping out in a particular proficiency level rather than advancing to a higher proficiency level.

Comparing Stata and SAS, it was found that both packages used different formulations to estimate the CR model and the requirements for data restructuring were also different. Compared to SAS, Stata could estimate the CR model directly without data restructuring. Compared to Stata, SAS produced different model fit statistics, because it estimated more parameters in the CR model, such as dummy coding variables. The estimated logit coefficients were the same using both packages. However, regarding the CR cutpoints, SAS provided different results in the output from those estimated by Stata. Equivalent cutpoints in magnitude could be obtained after further calculations, but they were reversed in sign, because the conditional probabilities estimated by the CR model using Stata and SAS with the descending option were complementary.

In educational research, the demand for ordinal response data analysis is increasing tremendously, it is therefore crucial for researchers to understand different statistical methods for analyzing ordinal response variables. Although comparisons have been made between statistical software packages, a preference of one package over the other is not suggested; this is left to researchers to choose. It is our hope that this article will help researchers become familiar with continuation ratio models and utilize them correctly in their research.

References

Agresti, A. (1996). *An introduction to categorical data analysis*. New York: John Wiley & Sons.

Agresti, A. (2002). *Categorical data analysis* (*2nd Ed*.). New York: John Wiley & Sons.

Agresti, A. (2007). *An introduction to categorical data analysis* (*2nd Ed*.). New York: John Wiley & Sons.

Allison, P. D. (1999). *Logistic regression using the SAS system: Theory and application*. Cary, NC: SAS Institute, Inc.

Ananth, C. V., & Kleinbaum, D. G. (1997). Regression models for ordinal responses: A review of methods and applications. *International Journal of Epidemiology*, *26*, 1323-1333.

Armstrong, B. B., & Sloan, M. (1989). Ordinal regression models for epidemiological data. *American Journal of Epidemiology*, *129*(*1*), 191-204.

Bender, R., & Benner, A. (2000). Calculating ordinal regression models in SAS and S-Plus. *Biometrical Journal*, *42*(*6*), 677-699.

Brant. (1990). Assessing proportionality in the proportional odds model for ordinal logistic regression. *Biometrics*, *46*, 1171-1178.

Clogg, C. C., & Shihadeh, E. S. (1994). *Statistical models for ordinal variables*. Thousand Oaks, CA: Sage.

Fienberg, S. E. (1980). *The analysis of cross-classified categorical data.* Cambridge, MA: The MIT Press.

Greene, W. H. (2003*). Econometric analysis* (*5th Ed*.). Upper Saddle River, NJ: Prentice Hall.

Hardin, J. W., & Hilbe, J. M. (2007). *Generalized linear models and extensions* (*2nd Ed*.). Texas: Stata Press.

Hilbe, J. M. (2009). *Logistic regression models*. Boca Raton, FL: Chapman & Hall/CRC.

Hosmer, D. W., & Lemeshow, S. (2000). *Applied logistic regression* (*2nd Ed*.). New York: John Wiley & Sons.

Ingels, S. J., Pratt, D. J., Roger, J., Siegel, P. H., & Stutts, E. (2004). *ELS: 2002 base year data file user's manual.* Washington, DC: NCES (NCES 2004-405).

Ingels, S. J., Pratt, D. J., Roger, J., Siegel, P. H., & Stutts, E. (2005). *Education Longitudinal Study: 2002/04 public use base-year to first follow-up data files and electronic codebook system*. Washington DC: NCES (NCES 2006-346).

Liu, X. (2009). Ordinal regression analysis: Fitting the proportional odds model using Stata, SAS and SPSS. *Journal of Modern Applied Statistical Methods*, *8*(*2*), 632-645.

Long, J. S. (1997). *Regression models for categorical and limited dependent variables*. Thousand Oaks, CA: Sage.

Long, J. S., & Freese, J. (2006). *Regression models for categorical dependent variables using Stata* (*2nd Ed*.). Texas: Stata Press.

McCullagh, P. (1980). Regression models for ordinal data (with discussion). *Journal of the Royal Statistical Society*, *Series B*, *42*, 109-142.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models* (*2nd Ed*.). London: Chapman and Hall.

Menard, S. (1995). *Applied logistic regression analysis*. Thousand Oaks, CA: Sage.

O'Connell, A. A. (2000). Methods for modeling ordinal outcome variables. *Measurement and Evaluation in Counseling and Development*, *33*(*3*), 170-193.

O'Connell, A. A. (2006). *Logistic regression models for ordinal response variables*. Thousand Oaks: SAGE.

O'Connell, A. A., & Liu, X. (2011). Model diagnostics for proportional and partial proportional odds models. *Journal of Modern Applied Statistical Methods*, *10*(*1*), 139-175.

Powers, D. A., & Xie, Y. (2000). *Statistical models for categorical data analysis*. San Diego, CA: Academic Press.

Pulkstenis, E., & Robinson, T. J. (2004). Goodness-of-fit tests for ordinal response regression models. *Statistics in Medicine*, *23*(*6*), 999-1014.

Wooldridge, J. M. (2001). *Econometric analysis of cross section and panel data*. Cambridge, MA: The MIT Press.

Wolfe, R. (1998). Continuation-ratio models for ordinal response data. *Stata Technical Bulletin*, *44*, 18-21.

Journal of Modern Applied Statistical Methods
November 2011, Vol. 10, No. 2, 528-538

1538 – 9472/11/$95.00

Higher Order Markov Structure-Based Logistic Model and Likelihood Inference for Ordinal Data

Soma Chowdhury Biswas
University of Chittagong,
Chittagong, Bangladesh

M. Ataharul Islam
University of Dhaka,
Dhaka, Bangladesh

Jamal Nazrul Islam
University of Chittagong,
Chittagong, Bangladesh

Azzalini (1994) proposed a first order Markov chain for binary data. Azzalini's model is extended for ordinal data and introduces a second order model. Further, the test statistics are developed and the power of the test is determined. An application using real data is also presented.

Key words: Markov chain, serial correlation, longitudinal data, ordinal data, covariate dependence, repeated measures.

Introduction

The Markov chain model is one of the most important and effective model classes for the assessment of probability for time dependent processes. A number of models have been proposed for analyzing repeated categorical and ordinal data. Muenz and Rubinstein (1985) employed a logistic regression model to analyze the transitional probabilities from one state to another. Azzalini (1994) introduced a Markov chain model that incorporated serial dependence and facilitated expression of covariate effects on marginal features. Raftery & Tavare (1994) suggested a Markov chain model of order higher than one that involves only one parameter for each extra lag variable: Heagerty and Zeger (2000) and Heagerty (2002) extended that work to a q^{th}-order marginalized transition model (MTM). These models are based on binary data and do not address the more general issue of ordinal data that arises in many biomedical studies. Islam and Chowdhury (2007) reviewed the first order model of Muenz and Rubinstem (1985) and developed a general procedure based on the Chapman-Kolmogorov equation for transition, reverse transition and repeated transition. Lee and Daniels (2007) extended Heagerty's (2002) MTM to accommodate longitudinal ordinal data. Ching, Fung and Ng (2004) generalized the Raftery and Tavare (1994) model by allowing $Q = \{q_{ij}\}$ to vary with different lags; they also developed an efficient method to estimate the model parameters. Ching, Ng and Fung (2007) extended their 2004 results (Ching, Fung & Ng, 2004) and proposed a higher-order multivariate Markov model for multiple categorical data sequences.

Azzalini's (1994) Markov structure based regression model for ordinal data is extended here, and a second order model is proposed. Likelihood based inferences are possible because the model is fully specified so that resulting estimators are consistent and fully efficient. The proposed methods are applied to real data collected at successive time points from diabetic patients registered at Bangladesh Institute of Research and Rehabilitation in Diabetes, Endocrine and Metabolic disorders (BIRDEM) in Bangladesh.

Soma Chowdhury Biswas is a Professor in the Department of Statistics. Email him at: soma_chow@hotmail.com. M. Ataharul Islam is a Professor in the Department of Statistics, Biostatistics and Informatics. Email him at: mataharul@yahoo.com. Jamal Nazrul Islam is Professor Emeritus, RCMPS. Email him at: jnislam@yahoo.com.

First Order Covariate Dependent Markov Model

Consider a stationary process $\{Y_{ij}\}$ for individual i (i = 1, 2, …, n) at follow-up j (j = 1, 2, …, n) representing past and present responses

where at time t_j the response $Y_j = k\,(k = 0,1,2)$. If the transition models for which the conditional distribution of $\boldsymbol{Y_{ij}}$ given the prior observations $Y_{ij-1} \ldots Y_{ij-r}$ is considered the model of order r, then the first order Markov model can be expressed as:

$$\Pr\left(Y_{ij} \middle| Y_{ij-1}\right) = \Pr\left(Y_{ij} \middle| Y_{i,j-r},\ \ldots, Y_{i,j-1}\right).$$

For a three state Markov chain the corresponding transition probability matrix is given by

$$P = \begin{pmatrix} p_{00} & p_{01} & p_{02} \\ p_{10} & p_{11} & p_{12} \\ p_{20} & p_{21} & p_{22} \end{pmatrix}$$

and the transition probabilities are:

$$P_0 = P_{it,0} = P\left(Y_{it} \middle| Y_{it=1} = 0\right)$$

$$P_1 = P_{it,1} = P\left(Y_{it} \middle| Y_{it-1} = 1\right)$$

$$P_2 = P_{it,2} = P\left(Y_{it} \middle| Y_{it-1} = 2\right).$$

Consider a vector of covariates for the i^{th} person $X'_{ij} = [1, X_{i1}, \ldots, X_{ip}]$ and the corresponding vector of parameters $\beta'_k = [\beta_{k0}, \beta_{k1}, \ldots, \beta_{kp}]$. The transition probabilities can be expressed in terms of conditional probabilities (Hosmer & Lemeshow, 1989) as follows:

$$p_{ij} = P\left(Y_{i,j} = j \middle| Y_{i,j-1} = j-1, X_{i,j-1}\right) = \frac{e^{g_j X}}{\sum_{k=0}^{m-1} e^{g_j X}},$$

where

$$g_k(X) = \ln\left[\frac{p(Y = k / X)}{p(Y = 0 / X)}\right] = \beta_{k0} + \beta_{k1}X_1 + \ldots + \beta_{kp}X_p \quad (1)$$

Azzalini (1994) searched for a parameterization such that $\theta = E(Y_t)$, which is free from a parameter and that regulates serial dependence. The odds ratio Ψ is a quantity that measures the dependence between successive observations. A technical reason in favor of this choice was provided by Fitzmaurice and Laird (1993) who stated that, when the association between observations is modeled using an odds ratio, the estimates of the mean are relatively insensitive to changes of the association parameter. Moreover, the range of feasible values for Ψ is independent of the θ value.

The above three stated Markov models for non-stationary cases are parameterized as

$$\theta_t = P_1\theta_{t-1} + P_0(1 - \theta_{t-1}) \quad (2)$$

and

$$\theta_t = P_2\theta_{t-1} + P_0(1 - \theta_{t-1}) \quad (3)$$

where, P_0, P_1 and P_2 will vary with t and (t = 2, 3, 4, 5, …, T). For t = 1, $E(Y_1) = \Pr(Y_1 = 1) = \theta_1$. and odds ratios:

$$\psi_1 = \frac{P_1 \big| (1 - P_1)}{P_0 \big| (1 - P_0)} \quad (4)$$

$$\psi_2 = \frac{P_2 \big| (1 - P_2)}{P_0 \big| (1 - P_0)} \quad (5)$$

For a given value of β the sequence of θ can be determined by (1) and solving (1) and (3); (2) and (4) and after algebraic manipulation, results in,

$$P_{;} = \begin{cases} \theta_t, \text{ for } \psi_i = 1 \\ \dfrac{\delta - 1 + (\psi_i - 1)(\theta_t - \theta_{t-1})}{2(\psi_i - 1)(1 - \theta_t - 1)} \\ + j\dfrac{1 - \delta + (\psi_i - 1)(\theta_t + \theta_{t-1} - 2\theta_t\theta_{t-1})}{2(\psi_i - 1)\theta_{t-1}(1 - \theta_{t-1})} \end{cases} ; \psi \neq 1, \tag{6}$$

where

$$\delta^2 = 1 + (\Psi_i - 1)\begin{Bmatrix} (\theta_t - \theta_{t-1})^2 \Psi_i \\ -(\theta_t - \theta_{t-1})^2 + 2(\theta_t + \theta_{t-1}) \end{Bmatrix}$$

and

$$P_{y_{t-1}} = \begin{Bmatrix} \dfrac{\delta - 1 + (\psi - 1)(\theta_t - \theta_{t-1})}{2(\psi - 1)(1 - \theta_{t-1})} \\ + y_{t-1}\dfrac{1 - \delta + (\psi - 1)(\theta_t + \theta_{t-1} - 2\theta_t\theta_{t-1})}{2(\psi - 1)\theta_{t-1}(1 - \theta_{t-1})} \end{Bmatrix}$$

for t > 1 and for t = 1, $p_{y_{it}} = \theta_1$.

Second Order Covariate Dependent Markov Model

A second order Markov model assumes that the current response variable is dependent on the history not only through the immediate previous response but also on the previous two responses, that is,

$$\Pr\left(Y_{it} \middle| Y_{it-2}, Y_{ij=1}\right) = \Pr\left(Y_{it} \middle| Y_{it-1}, Y_{it-2}, \ldots, Y_{it-n}\right).$$

The transition probabilities for the three state second order Markov Chain can be written as:

$$P_0 = p_{it,0} = \Pr\left(y_{it} \mid y_{it-1} = 0, y_{t-2} = 0\right)$$

$$P_1 = p_{it,1} = \Pr\left(y_{it} \mid y_{it-1} = 0, y_{t-2} = 1\right)$$

$$P_2 = p_{it,2} = \Pr\left(y_{it} \mid y_{it-1} = 0, y_{t-2} = 2\right).$$

Let the parameterization of mean and odds ratio for second order can be extended as

$$\begin{aligned} \theta_{it} &= p_{it,1}\,\theta_{it-2} + p_{it,0}\,(1 - \theta_{it-2}) \\ &= \begin{pmatrix} P_{001} + P_{101} + P_{201} + P_{011} + P_{111} \\ + P_{211} + P_{021} + P_{121} + P_{221} \end{pmatrix}\theta_{t-2} \\ &\quad + \begin{pmatrix} P_{000} + P_{100} + P_{200} + P_{010} + P_{110} \\ + P_{210} + P_{020} + P_{120} + P_{220} \end{pmatrix}(1 - \theta_{t-2}) \end{aligned} \tag{7}$$

$$\begin{aligned} \theta_{it} &= p_{it,2}\,\theta_{it-2} + p_{it,0}\,(1 - \theta_{it-2}) \\ &= \begin{pmatrix} P_{002} + P_{102} + P_{202} + P_{012} + P_{112} \\ + P_{212} + P_{022} + P_{122} + P_{222} \end{pmatrix}\theta_{t-2} \\ &\quad + \begin{pmatrix} P_{000} + P_{100} + P_{200} + P_{010} + P_{110} \\ + P_{210} + P_{020} + P_{120} + P_{220} \end{pmatrix}(1 - \theta_{t-2}) \end{aligned} \tag{8}$$

$$\psi_1 = \frac{p_{it,1} \left| (1 - P_{it,1}) \right.}{p_{it,0} \left| (1 - P_{it,0}) \right.} \tag{9}$$

$$\psi_2 = \frac{p_{it,2} \mid (1 - P_{it,2})}{p_{it,0} \mid (1 - P_{it,0})} \tag{10}$$

and

$$P_j = P_{y_{t-2}} = \frac{\delta - 1 + (\Psi_i - 1)(\theta_i - \theta_{t-2})}{2(\Psi_i - 1)(1 - \theta_{t-2})} + j\frac{(\delta - 1) + (\Psi_i - 1)(\theta_i + \theta_{t-2} - 2\theta_t\theta_{t-2})}{2(\Psi_i - 1)\theta_{t-2}(1 - \theta_{t-2})}$$

for $\Psi_i \neq 1$,

where

$$\delta^2 = 1 + (\Psi_i - 1)\left\{(\theta_t - \theta_{t-2})^2 \Psi_i - (\theta_t - \theta_{t-2})^2 + 2(\theta_t + \theta_{t-2})\right\}$$

and $\log \Psi_i = \lambda_1$ and $\log \Psi_2 = \lambda_2$.

These relationships generate a process having the desired properties. Upon taking $P_r(y_t = 1) = \theta_1$ and generating $y_2, \ldots, y_t$ via a non-homogeneous Markov chain with transition probabilities P_j, a sequence is obtained such that

$E(y_t) = \theta_t$ for t = 1, 2, ..., t and the odds ratios for (y_{t-2}, y_t) are equal to ψ.

Estimation

The conditional likelihood function for a sample of n independent observations is:

$$L(\beta) = \prod_{i=1}^{n} \left[\left[P(0|x)\right]^{y_{0i}} \left[P(1|x)\right]^{y_{1i}} \left[P(2|x)\right]^{y_{2i}} \right] \tag{11}$$

The log-likelihood function can be written as

$$l_t(\beta) = \sum_{i=1}^{n} \begin{bmatrix} y_{01} \log P(0|x) \\ +y_{1i} \log P(1|x) \\ +y_{2i} \log P(2|x) \end{bmatrix}$$

$$= \sum_{i=1}^{n} \begin{bmatrix} y_{0i} \log \left(\dfrac{\mathrm{P}(y=0)}{(1-\mathrm{P}(y=0))} \right) \\ +y_{1i} \log \dfrac{\mathrm{P}(y=1)}{1-P(y=1)} \\ +y_{2i} \log \dfrac{\mathrm{P}(y=2)}{1-P(y=2)} \end{bmatrix} \tag{12}$$

and

$$= \sum_{i=1}^{n} \left[y_{0i} g_0(x) + y_{1i} g_1(x) + y_{2i} g_0(x) \right]$$

$$= \sum_{i=1}^{n} \begin{bmatrix} y_{1i} g_1(x) + y_{2i} g_2(x) \\ -\log\left(1 + y_{1i} g_1(x) + y_{2i} g_2(x)\right) \end{bmatrix} \tag{13}$$

where

$$(y_{0i} + y_{1i} + y_{2i} = 1)]$$

and

$$\log \psi_1 = \lambda_1 \Rightarrow \psi_1 = e^{\lambda_1},$$

$$\log \psi_1 = \lambda_2 \Rightarrow \psi_2 = e^{\lambda_2}.$$

In the case of repeated measures, because dependence between successive observations on the same individual must be taken into account, it is plausible that adjacent data are more strongly correlated than data that are separated by time and that different individuals behave independently on the log likelihood, this is given by:

$$l(\beta, \lambda) = \sum_{T=1}^{4} l_t(\beta, \lambda)$$

where score vectors are:

$$\frac{\partial l}{\partial \beta} = \sum_{t=1}^{T} \sum_{i=1}^{n} \frac{\partial l_{it}}{\partial \beta}$$

and

$$\frac{\partial l}{\partial \lambda} = \sum_{t=1}^{T} \sum_{i=1}^{n} \frac{\partial l_{it}}{\partial \lambda}$$

and the variance of the estimate is approximated by

$$V\left(\hat{\beta}, \hat{\lambda}\right) = \left[\sum_{i=1}^{n} \begin{pmatrix} \dfrac{\partial l_t}{\partial \beta} \\ \dfrac{\partial l_t}{\partial \lambda} \end{pmatrix} \begin{pmatrix} \dfrac{\partial l_t}{\partial \beta} \\ \dfrac{\partial l_t}{\partial \lambda} \end{pmatrix}^{T} \right]^{-1}_{\beta=\hat{\beta}, \lambda=\hat{\lambda}}$$

The quantity inside the square brackets approximates the Fisher Information for large n. Similarly, the model can be generalized to third, fourth and up to n^{th} order.

Methodology

Test Procedure

To test the null hypothesis that all slope coefficients are simultaneously equal to zero is the usual likelihood ratio statistic used and it follows the χ^2 distribution with degrees of freedom (df) equal to the number of explanatory variable(s). For an r^{th} order Markov model the df for Chi square is $m^r(m-1)p$, where m is the number of states. Therefore, to test the null hypothesis $H_0 : \beta = 0$, the usual likelihood ratio test can be employed.

$$-2[\ln L(\beta_0) - \ln L(\beta)] \approx X^2_{m^r(m-1)p}$$

where the vectors $m^r(m-1)$ set a parameters for the r[th] order Markov model.

For a first order Markov model with dichotomous transition outcomes and an independent variable, the likelihood ratio Chi square is

$$-2\left[\ln L(\beta_0) - \ln L(\beta)\right] \approx \chi^2_{3^1(3-1)p} \approx \chi^2_{6p}$$

where

$$\ln L(\beta) = \log l_t(\beta,\lambda) = \sum_{t-2}^{T}\left\{y_t\, logit\left(P_{y_{t=t2}}\right) + \log\left(1 - P_{y_{t-2}}\right)\right\}.$$

Similarly, for a second order model with binary outcomes and p independent variables, the null hypothesis of the null parameter vector can be tested by using the test statistic:

$$-2\left[\ln L(\beta_0) - \ln L(\beta)\right] \approx \chi^2_{3^2(3-1)p} = \chi^2_{18p}.$$

The Wald test statistic for the null hypothesis H_o: $\beta_j = 0$ can be written as the multivariate analogue of this test, which follows a Chi square distribution and is given by

$$W_i = \left[(\hat{\beta} - \beta_0)\right]' \left[I(\hat{\beta}, o)\right]^{-1} \left[(\hat{\hat{\beta}} - \beta_0)\right].$$

Data and Variables

The proposed model is illustrated using Diabetes mellitus data. This data was collected by the Bangladesh Institute of Research and Rehabilitation in Diabetes, Endocrine and Metabolic disorders (BIRDEM). After registration, a patient visits the BIRDEM for regular check-ups and treatment. During registration each patient answers a detailed questionnaire and a comprehensive record sheet is maintained for each patient until death of the patient or loss of follow up over time. The patients experience impaired glucose tolerance (IGT) levels at that time of registration, so the number of follow-ups for each patient is not equal. For convenience of analysis 999 patients were randomly selected for this study that had four consecutive follow-up visits to BIRDEM. Consider a few selected variables such as, age (AGE), sex (SEX: coded as 0 = female, 1= male) and blood glucose level in each visit. Some clinical variables were not used in the study due to a high percentage of missing values over the four consecutive visits. In the record sheet, the age of first registration of the patient is logged. The age at different consecutive follow-ups visits from date of registration of the patients was calculated using SPSS, thus making age a continuous variable that was used in the study directly to observe the time effect.

The blood glucose level of each patient (two hours after 75 grams of glucose load) was assumed to be indicative of the patient's diabetic health status and was therefore considered the dependent variable. Specifically, a person having blood glucose concentration level in venous plasma after 2 hours of 75 grams glucose load greater than or equal to 11.1 mmol/liter was considered a confirmed diabetic and coded as 2; a person with a blood glucose level between 4.4 and 11.1 mmol/l was considered as border line diabetic and coded as 1; and a person with a blood glucose concentration level less than 4.4 mmol/liter was considered a controlled diabetic and coded as 0. The response variable Y_{it} of interest can be defined as:

$$Y_{it} = \begin{cases} 0, \text{Controlled diabetic} \\ \quad if \text{ blood glucose level} < 4.4 \\ 1, \text{ Borderline diabetic} \\ \quad if \ 4.4 \le \text{blood glucose level} < 11.1 \\ 2, \text{Confirmed diabetic} \\ \quad if \text{ blood glucose level} \ge 11.1 \end{cases}$$

where t = 1, 2, 3, 4. The response Y_{it} to be generated by a ordinal Markov chain with values 0, 1 and 2 and with transition probabilities p_{ij} = $\Pr\left(Y_{it} = i \middle| Y_{it-1} = j\right)$ for j = 0, 1, 2 for the first order, and p_{ijk}= Pr $\left(y_{it} = i / y_{it-1} = j, y_{t-2} = k\right)$ for i, j, k = 0, 1, 2 for the second order. This study concentrates on modeling the mean value θ_{it} via a covariate, which can be obtained by using (1).

Results

First Order Markov Model

The dependence between successive observations y_{it} t = 1, 2, 3, 4 is measured by the odds ratio Ψ_{it} which is defined in (4) and (5). A sequence of mean values of the process, such that, $\theta_{it} = E(y_{it})$ for t = 1, 2, 3, 4 and odds ratio Ψ_{it} for $(y_{i(t-1)}y_{it})$ could be obtained by taking $Pr(y_{i1} = 1) = \theta_{i1}$ and $Pr(y_{i2} = 2) = \theta_{i2}$, generating y_{i2}, y_{i3}, y_{i4} via a homogeneous Markov Chain with transition probabilities p_{t0}, p_{t1} and p_{t2} obtained from (6). Because an objective is to determine the effects of covariates on the risk of confirmed and borderline diabetics, the marginal probabilities of confirmed and borderline diabetic for each given value of the covariates is the quantity of interest. The log-likelihood function for β and λ = $\log \Psi t$ for T = 4 time points can be described by (12) and also for repeated data. The model in (1) for the marginal probability of the event is fitted to the data. The parameters of the model were estimated using the maximum likelihood method of estimation and the Newton-Raphson iteration method. All the calculations were performed by programming in R. Table 1 summarizes the results of fitting of the 1st order model.

As shown in Table 1, the likelihood ratio test value, 6759.67, is highly significant, thus, it can be used to identify the effect of the covariates on the disease status of the patient. For model $0 \rightarrow 1$, age is a significant factor and has a positive association with the transition of the disease from controlled (0) to borderline (1); sex is not significant a factor although it has a positive association with the transition of disease, that is, female patients are less likely to transition from controlled to borderline diabetes compared to male patients. For model $0 \rightarrow 2$, both covariates have a positive association with the transition of the disease from controlled to confirmed diabetes. The risk of transition from controlled (0) to confirmed (2) diabetes increases with the increase of age. Female patients are more likely to transition from borderline diabetes to confirmed diabetes compared to male patients, although the difference is not significant.

Figure 1 illustrates the power comparison of the Wald test for testing the hypothesis $H_0 : \beta_i = 0$ versus alternative hypothesis $H_1 : \beta_i \neq 0$ for a first order model. Both graphs show that the power obtained from the $0 \rightarrow 2$ model (controlled to confirmed) is higher than that of the $0 \rightarrow 1$ (controlled to borderline) model for the parameters sex and age.

Second Order Markov Model

The dependence between successive observations Y_{it}, t = 1, 2, 3, 4 is measured by the odds ratio Ψ_{it} which is defined in (9) and (10). A sequence of mean values of the process, such that $\theta_{it} = E(Yit)$ for t = 1, 2, 3, 4, and odds ratio Ψ_{it} for $(Y_{i(t-2)}Y_{it})$ could be obtained by taking $Pr(Y_{it=1} = 1) = \theta_{1t}$ and $Pr(Y_{it-2} = 2) = \theta_{2t}$. This will generate Y_{i3}, Y_{i4} via a non-homogeneous Markov Chain with transition probability P_{t2} *and* P_{t3} obtained in (11). The log-likelihood function for β and λ = $\log \Psi t$ for T = 4 time points is described by (12). Model (13) for the marginal probability of the event was fitted to the data and the parameters of the Markov based second order model were estimated using the maximum likelihood and Newton Raphson Iteration methods; all calculations were performed with R.

Table 2 summarizes the results of the fitted second order model and shows that the likelihood ratio for the overall model is 6780349.580, which is significant and follows a Chi-square distribution with 5 df; thus, the null hypothesis may be rejected and significance for at least one of the covariates may be concluded. To reveal the significance of individual parameters, the Wald test was performed. For model $0 \rightarrow 1$, age and sex show a positive association with the response variable. The risk of transition from controlled (0) to borderline (1) diabetes increases as age increases; both variables have a significant effect on the transition from controlled to borderline diabetes.

Table 1: Estimates and Associated Wald Test from First Order Markov Model

Variable	Estimated Coefficient	Standard Error	Wald χ^2	P-value	Odds Ratio
$0 \rightarrow 1$					
Constant	0.3579	0.1906	3.5247	NA	--
Sex	0.1944	0.01848	110.61	0.0000	1.21
Age	0.1733	0.2079	0.69441	0.4046	1.19
λ	-1.1965	0.00329	13204.62	0.0000	0.31
$0 \rightarrow 2$					
Constant	0.3662	0.22831	2.5731	NA	--
Sex	0.1984	0.00651	927.688	0.0000	1.22
Age	0.1013	0.12993	0.60825	0.4354	1.11
λ	-2.1997	0.00145	22994.55	0.0000	0.11

Likelihood Ratio = 6759.67; p-value = 0.000

Figure 1: Power Curves for Covariates Sex and Age for First Order Markov Model

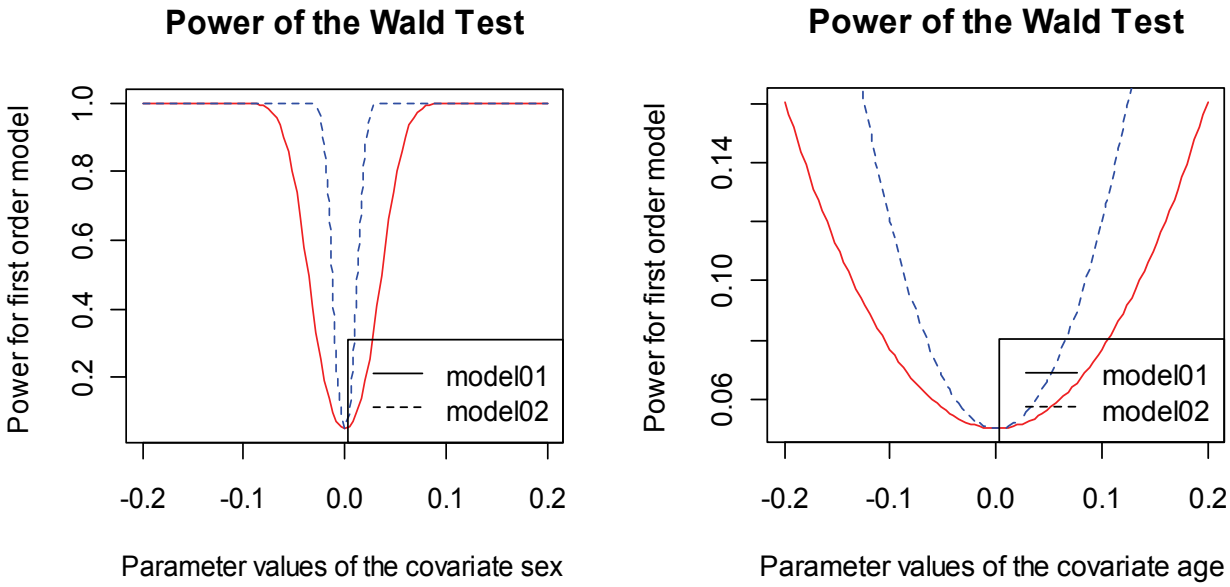
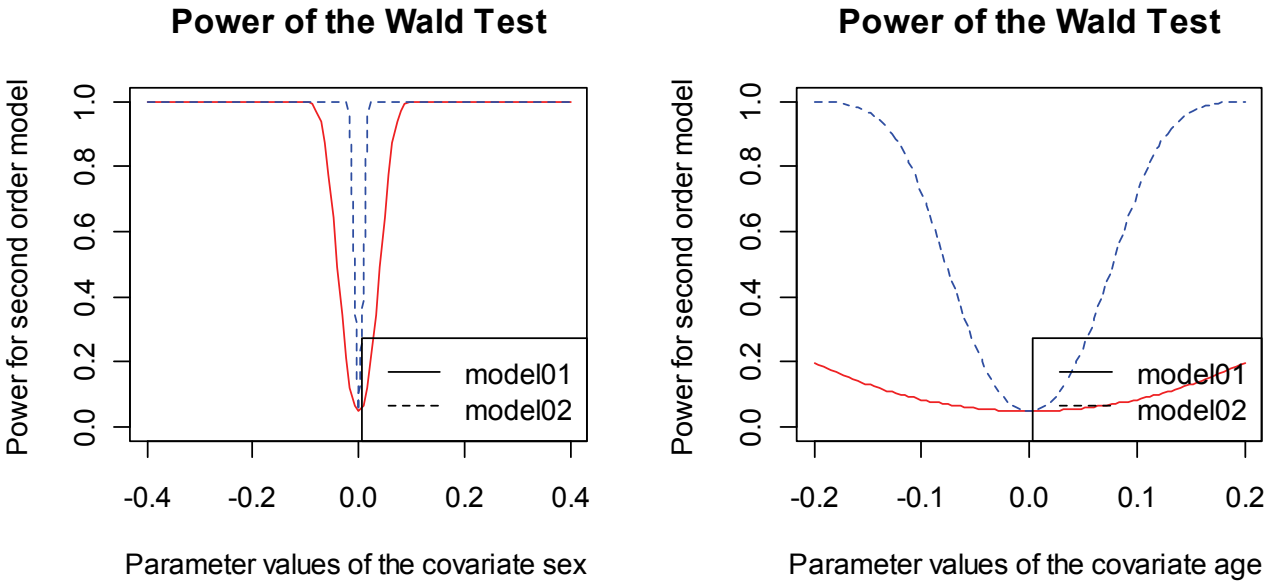


Table 2: Estimates and Associated Test from Second Order Markov Model

Variable	Estimated Coefficient	Standard Error	Wald χ^2	P-value	Odds Ratio
$0 \rightarrow 1$					
Constant	0.3962	0.0498	63.31	NA	--
Sex	0.4001	0.0187	456.71	0.0000	1.49
Age	0.1864	0.0479	151194	0.0001	1.20
λ	-0.9001	0.00007	1364683.0	0.0000	0.40
$0 \rightarrow 2$					
Constant	0.7127	0.0596	142.56	NA	--
Sex	0.3695	0.0044	702.64	0.0000	1.44
Age	0.0718	0.0394	3.3334	0.0067	1.07
λ	-1.2001	0.0001	121102.9	0.0000	0.30

Overall Chi square = 6780349.580; p-value = 0.0000

Figure 2: Power Curves for Covariates Sex and Age for Second Order Markov Model



For model $0 \rightarrow 2$, similar to $0 \rightarrow 1$, age and sex show a positive association with the response variable and the association is significant with the subject's transition both from controlled (0) to boarderline (1) and controlled (0) to confirmed (2) diabetes. The risk of transition in both cases increases with a unit increase in age level. Male patients are more likely to transition to borderline (1) and to confirmed (2) diabetes compared to female patients. Finally, it can be concluded that an increase in age increases the risk of transition of the disease to a higher stage. The value of likelihood ratio also has a noticeable increase when considering a higher order Markov model.

Figure 2 shows the power curves of the Wald test for testing the hypothesis $H_0 : \beta_i = 0$ against the alternative hypothesis $H_1 : \beta_i \neq 0$ for a second order model. The charts illustrate that the power obtained from the $0 \rightarrow 2$ model is higher than that of the $0 \rightarrow 1$ model.

Conclusion

This study extended Azzalini's (1994) model for ordinal data up to second order, which can then be generalized to any order in the same setting as Islam and Chowdhury (2007). The proposed model was applied to repeated measures of diabetes mellitus testing and it was observed that the variables age and sex show significant contributions to the diabetes status of a patient. Comparison of the estimates of first and second order and power curve are displayed for the Wald Chi square test, which shows a significant improvement in power for the $0 \rightarrow 2$ transition model. Based on results of this study it is reasonable to conclude that, for analyzing the repeated measures data of diabetes mellitus, a higher order Markov model approach can be conveniently employed for any number of states and for any order with any number of covariates; this may prove valuable for health policy makers. Further research could be conducted using a continuous time Markov model for estimation and testing in other settings.

References

Azzalini, A. (1994). Logistic regression for auto correlated data with application to repeated measures. *Biometrika*, *81*, 767-775.

Ching, W. K., Ng, M. K., & Fung, E. S. (2008). Higher order multivariate Markov chains and their applications. *Linear Algebra and its Applications*, *428*(*2-3*), 492-507.

Ching, W. K., Fung, E. S., & Ng, K. (2004). Higher order Markov chain models for categorical data sequences. *International Journal of Naval Research Logistics*, *51*, 557-574.

Fitzmanurice, G. M., & Laird, N. M. (1993). A likelihood based method for analyzing binary responses. *Biometrika*, *80*, 141-151.

Hosmer, D. W., & Lemeshow, S. (1989). *Applied logistic regression*. New York, NY: John Wiley and Sons.

Heagerty, P. J., & Zeger, S. L. (2000). Marginalized multilevel models and likelihood inference (with discussion). *Statistical Science*, *15*, 1-26.

Heagerty, P. J. (2002). Marginalized transition models and likelihood inference for longitudinal categorical data, *Biometrics 58, 342-35*.

Islam, M. A., & Chowdhury, R. I. (2006). A higher order Markov model of analyzing covariate dependence. *Applied Mathematical Modeling*, *30*, 477-488.

Islam, M. A., & Chowdhury, R. I. (2008). First and higher order transition models with covariate dependence. In *Progress in Applied Mathematical Modeling*, 153-196. New York, NY: Nova Science Publishers, Inc.

Islam, M. A., Chowdhury, R. I., & Huda, S. (2009). *Markov models with covariate dependence for repeated measures*. New York, NY: Nova Science Publishers, Inc.

Lee, K., & Daniels, M. J. (2007). A class of Markov models for longitudinal ordinal data. *Biometrics, 64, 4,1060-1067*.

Muenz, K.R., & Rubinstein, L.V. (1985). Markov chain for covariance dependence of binary sequences. *Biometrics*, *41*, 91-101.

Phillips, P. C. B. (1986). The exact distribution of the Wald statistic. *Econometrica*, *54*(*4*), 881-895.

Raftery, A., & Tavare, S. (1994). Estimating and modeling repeated patterns in higher order Markov chains with the mixture transition distribution model. *Applied Statistics*, *43*(*1*), 179-199.

Rahman, M. S., & Islam, M. A. (2007). Markov structure based logistic regression for repeated measures: An application to diabetes mellitus data. *Statistical Methodology*, *4*, 448-460.

Appendix

The derivatives $\frac{\partial l}{\partial \beta}$ and $\frac{\partial l}{\partial \lambda}$ were computed using a chain rule, giving elements of score vectors of a first order model. The parameters β and λ were estimated by maximum likelihood method and using chain rule of differentiation.

$$\frac{\partial l_t}{\partial \beta} = \frac{\partial l_t}{\partial p_{y_{t-1}}}\left(\frac{\partial p_{yt-1}}{\partial \theta_t}.\frac{\partial \theta_t}{\partial \beta} + \frac{\partial p_{yt-1}}{\partial \theta_{t-1}}.\frac{\partial \theta_{t-1}}{\partial \beta}\right)$$
$$(t = 1, 2, \ \ldots \ T)$$

$$\frac{\partial l_t}{\partial \lambda} = \frac{\partial l_t}{\partial p_{y_{t-1}}}.\frac{\partial p_{y_{t-1}}}{\partial \psi}.\frac{\partial \psi}{\partial \lambda}$$
$$(t = 2, 3, \ \ldots \ T)$$

where,

$$\log \psi_1 = \lambda_1 \Rightarrow \psi_1 = e^{\lambda_1};$$

$$\log \psi_2 = \lambda_2 \Rightarrow \psi_2 = e^{\lambda_2}$$

$$\frac{\partial \mathrm{Py}_{t-1}}{\partial \theta_t} = \frac{1}{\mathrm{A}}(-(2\mathrm{y}_{t-1} - 1)\frac{\partial \delta}{\partial \theta_t} + \psi - 1)$$

$$\frac{\partial \theta_t}{\partial \beta} = \theta_t(1 - \theta_t)\mathrm{x}_t; \text{where } \theta_t = \frac{e^{\mathrm{x}_{t\beta}}}{1 + e^{\mathrm{x}_{t\beta}}}$$

$$\frac{\partial Py_{t-1}}{\partial \theta_{t-1}} = \frac{1}{A^2}\begin{Bmatrix}(-(2y_{t-1} - 1)(\psi - 1 - \frac{\partial \delta}{\partial \theta_{t-1}})\\ A - 2(\psi - 1)(2y_{t-1} - 1)\mathrm{B}\end{Bmatrix}$$

where

$$B = (2y_{t-1})\{(1 - \delta + (\psi - 1)\theta_t\}(\psi - 1)\theta_t - 1$$

and

$$A = 2(\psi - 1)(1 - y_{t-1}) + (2y_{t-1} - 1)\theta_t$$

$$\frac{\partial \theta_{t-1}}{\partial \beta} = \theta_{t-1}(1 - \theta_{t-1})\mathrm{x}_{t-1}$$

$$\frac{\partial p_{yt-1}}{\partial \theta_{t-1}} = \frac{1}{A^2}\begin{bmatrix}((2y_{t-1} - 1)(-\frac{\partial \delta}{\partial \psi} + \theta_{t-1}) + \theta_t\}\\ A - 2\mathrm{B}\{1 - y_{t-1} + (2y_{t-1})\theta_{t-1}\}\end{bmatrix}$$

$$\frac{\partial s}{\partial \theta} = \frac{1}{\delta}\begin{bmatrix}(\psi - 1)\{\psi(\theta_t - \theta_{t-1})\\ -(\theta_t - \theta_{t-1}) + 1\}\end{bmatrix}$$

$$\frac{\partial s}{\partial \theta_{t-1}} = \frac{1}{\delta}\begin{bmatrix}(\psi - 1)\{-\psi(\theta_t - \theta_{t-1})\\ -(\theta_t - \theta_{t-1}) + 1\}\end{bmatrix}$$

$$\frac{\partial s}{\partial \psi} = \frac{1}{2\delta}\begin{bmatrix}(\theta_t - \theta_{t-1})^2(2\psi - 1)\\ -(\theta_t + \theta_{t-1})^2 + 2(\theta_t + \theta_{t-1})\}\end{bmatrix}$$

$$\frac{\partial \psi}{\partial \lambda} = \psi$$

and

$$\frac{\partial p_{y_{t-1}}}{\partial \psi} = \frac{1}{A2}\begin{bmatrix}\left\{(2y_{t-1} - 1)\left(-\frac{\partial \delta}{\partial \psi} + \theta_{t-1}\right) + \theta_t\right\}\\ A - 2B(1 - y_{t-1} + (2y_{t-1} - 1)\theta_{t-1}\end{bmatrix}$$

Second order model elements of score vectors were computed as:

$$\frac{\partial p_{\gamma_{t-2}}}{\partial \theta_t} = \frac{1}{A}(-2(2y_{t-2} - 1)\frac{\partial \delta}{\partial \theta_t} + \psi - 1$$

where

$$\mathrm{A} = 2(\psi\text{-}1)\ \{(1\text{-}y) + (2y\text{-}1)\ \theta_{t\text{-}2}\}$$

$$\frac{\partial \theta_t}{\partial \beta} = \theta_t(1 - \theta_t)\chi_t$$

where

$$\theta_t = \frac{e^{\mathrm{x}}\beta}{1 + e^{\mathrm{x}}\beta}$$

$$\frac{\partial p_{y-2}}{\partial \theta_{t-2}} = \frac{1}{A^2}\begin{Bmatrix}(2y_{t-2}-1)\left(\psi - 1 - \dfrac{\partial \delta}{\partial \theta_{t-2}}\right)\\ A - 2(\psi-1)(2y_{t-2}-1)\mathrm{B}\end{Bmatrix}$$

where

$$\mathrm{B} = (2\mathrm{y}-1)(1-\delta(\Psi-1)\times\theta_{\mathrm{t}-2} + (\Psi-1)\theta_{\mathrm{t}}\}$$

$$\frac{\partial \theta_{t-2}}{\partial \beta} = \theta_{t-2}(1-\theta_{t-2})\chi_{t-2}$$

$$\frac{\partial p_{\gamma t-2}}{\partial \psi} = \frac{1}{A^2}\begin{bmatrix}(2y_{t-1}-1)\left(-\dfrac{\partial \delta}{\partial \psi} + \theta_{t-2}\right) + \theta_t \;)\\ A - 2B(1-y_{t-2}) + (2y_{t-2}-1)\theta_{t-2}\end{bmatrix}$$

$$\frac{\partial \delta}{\partial \theta_t} = \frac{1}{\delta}\left[(\psi-1)\begin{Bmatrix}\psi(\theta_t - \theta_{t-2})\\ -(\theta_t + \theta_{t-2}) + 1\end{Bmatrix}\right]$$

$$\frac{\partial \delta}{\partial \theta_{t-2}} = \frac{1}{\delta}\left[(\psi-1)\begin{Bmatrix}-\psi(\theta_t - \theta_{t-2})\\ -(\theta_t + \theta_{t-2}) + 1\end{Bmatrix}\right]$$

$$\frac{\partial \delta}{\partial \psi} = \frac{1}{2\delta}\begin{bmatrix}(\theta_t - \theta_{t-2})^2(2\psi-1)\\ -(\theta_t + \theta_{t-2})^2 + 2(\theta_t + \theta_{t-2})\;\}\end{bmatrix}$$

Journal of Modern Applied Statistical Methods
November 2011, Vol. 10, No. 2, 539-548

1538 – 9472/11/$95.00

Estimation and Hypothesis Testing in LAV Regression with Autocorrelated Errors: Is Correction for Autocorrelation Helpful?

Terry E. Dielman
Texas Christian University
Fort Worth, TX

Using the Prais-Winsten correction and adding a lagged variable provides improved estimates (smaller MSE) in least absolute value (LAV) regression when moderate to high levels of autocorrelation are present. When comparing empirical levels of significance for hypothesis tests, adding a lagged variable outperforms other approaches but has a relative high empirical level of significance.

Key words: Monte Carlo simulation, serial correlation, Cochrane-Orcutt, Prais-Winsten, lagged variable.

Introduction

Least absolute value (LAV) regression is one technique often suggested for robust regression (see Dielman, 2005 for a review of LAV research). LAV estimates are less strongly affected by extreme observations compared to their least squares counterparts. The use of regression to model time-series data often results in the violation of the assumption of independent disturbances. The Prais-Winsten (PW) and Cochrane-Orcutt (CO) methods are two procedures used for correcting for autocorrelation in time-series regression models: Both methods transform the data using a differencing transformation to remove autocorrelation. LAV estimation applied to the transformed observations yields estimators that are asymptotically more efficient than LAV applied to the original data. The two methods are essentially equivalent except for the treatment of the first observation in the data set. The CO method omits the first observation; the PW method transforms and retains the observation. Asymptotically, no difference exists in the efficiency of estimators produced by the two methods. In previous studies of small sample behavior, however, the PW procedure has been found to produce more efficient estimates; using the CO procedure results in estimators that can be much less efficient in small samples.

Koenker and Bassett (1982) suggested the WALD, likelihood ratio (LR), and Lagrange multiplier (LM) tests for coefficient significance when using LAV estimation. Stangenhaus (1987), Dielman and Pfaffenberger (1990, 1992), Dielman and Rose (1996), and Koenker (1987) have studied inference for regression using LAV estimation when disturbances are independent but not necessarily normal.

Some research has considered LAV estimation when errors are not independent. Dielman and Rose (1994a, 1995b) examined the accuracy of estimation for model coefficients using LAV regression with autocorrelation correction, and Dielman and Rose (1994b) considered the accuracy of forecasts from LAV estimated regressions with autocorrelation correction. Dielman and Rose (1997) examined both estimation and inference in autocorrelated models.

A simulation study was conducted to address questions of estimation and inference in the presence of serial correlation. The PW and CO corrections for autocorrelation are considered and compared to the performance of a model with a lagged dependent variable added. Estimation accuracy after correction for autocorrelation is compared using mean square

Terry E. Dielman is a Professor of Decision Sciences in the Information Systems and Supply Chain Management Department in the M. J. Neeley School of Business. Email him at: t.dielman@tcu.edu.

estimation error. The performance of hypothesis tests for the slope coefficient is assessed using observed significance levels, and alternative estimators of the scale parameter used in the test procedures are considered. In addition, performance in small samples is considered due to the practical importance of smaller sample sizes - particularly for applications in business and economics - and the inability to rely upon asymptotic results under such circumstances.

Methodology

A simple regression model is considered:

$$y_t = \beta_0 + \beta_1 x_t + \varepsilon_t,$$

with

$$\varepsilon_t = \rho\varepsilon_{t-1} + \eta_t \tag{1}$$

for $t = 1, 2, ..., T$. In (1), y_t and x_t are the t^{th} observations on the dependent and explanatory variables, respectively, and ε_t is a random disturbance for the t^{th} observation and may be subject to autocorrelation. The η_t represents disturbance components that are assumed to be independent and identically distributed, although not necessarily normal. The parameters β_0 and β_1 are unknown and must be estimated. The parameter ρ is the autocorrelation coefficient, with $|\rho|<1$.

Using matrix notation, the model can be written as:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{2}$$

where

$$\mathrm{Y} = \begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ y_T \end{bmatrix}, \ \mathrm{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ . & . \\ . & . \\ 1 & x_T \end{bmatrix}, \ \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ . \\ . \\ \varepsilon_T \end{bmatrix}, \ \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}. \tag{3}$$

Two well-known procedures employed to correct for autocorrelation are the Prais-Winsten (1954) and Cochrane-Orcutt (1949) procedures. Both transform the data using the autocorrelation coefficient, ρ, after which the transformed data are used in estimation. The procedures differ in their treatment of the first observation, (x_1, y_1). The PW transformation matrix is:

$$\mathrm{M_{PW}} = \begin{bmatrix} (1-\rho^2)^{1/2} & 0 & . & . & . & 0 & 0 \\ -\rho & 1 & . & . & . & 0 & 0 \\ . & . & . & & & . & . \\ . & . & & . & & . & . \\ . & . & & & . & & \\ 0 & 0 & . & . & . & -\rho & 1 \end{bmatrix}. \tag{4}$$

Pre-multiplying the model in (2) by $\mathbf{M}_{\mathrm{PW}}$ yields

$$\mathrm{M_{PW}Y} = \mathrm{M_{PW}X}\beta + \mathrm{M_{PW}}\varepsilon \tag{5}$$

or

$$\mathbf{Y}^* = \mathbf{X}^*\boldsymbol{\beta} + \boldsymbol{\eta} \tag{6}$$

where $\mathbf{Y}^*$ contains the transformed dependent variable values and $\mathbf{X}^*$ is the matrix of transformed independent variable values, thus:

$$\mathbf{Y}^* = \left[(1-\rho^2)^{1/2} y_1, y_2 - \rho y_1, ..., y_T - \rho y_{T-1}\right] \tag{7}$$

and

$$\mathbf{X}^* = \begin{bmatrix} (1-\rho^2)^{1/2} & (1-\rho^2)^{1/2} x_1 \\ 1-\rho & x_2 - \rho x_1 \\ . & . \\ . & . \\ . & . \\ 1-\rho & x_T - \rho x_{T-1} \end{bmatrix} \tag{8}$$

In (6), $\boldsymbol{\eta}$ is the vector of serially uncorrelated η_t errors.

The CO transformation matrix is the $(T-1) \times 1$ matrix obtained by removing the first row of the $\mathbf{M_{PW}}$ transformation matrix. The use of the CO transformation means that $(T-1)$ observations, rather than T, are used to estimate the model. In the CO transformation, the first

observation is omitted, whereas it is transformed and included in the estimation in the PW transformation. Asymptotically, the loss of this single observation is of minimal concern. However, for small samples, omitting the first observation may result in an estimator inferior to that obtained when the first observation is retained and transformed as shown in Maeshiro (1979), Park and Mitchell (1980) and Dielman and Pfaffenberger (1984) for least squares and in Dielman and Rose (1994a) for LAV. The two methods described are referred to as LAVPW and LAVCO when combined with LAV estimation.

In practice, the value of ρ will be unknown. In this case it must be estimated from sample data. The estimator of ρ is as follows:

$$\hat{\rho}_{PW} = \frac{\sum_{t=2}^{T} \hat{\varepsilon}_t \hat{\varepsilon}_{t-1}}{\sum_{t=2}^{T} \hat{\varepsilon}_t^2} \qquad (9)$$

when PW correction is used, and

$$\hat{\rho}_{CO} = \frac{\sum_{t=2}^{T} \hat{\varepsilon}_t \hat{\varepsilon}_{t-1}}{\sum_{t=1}^{T-1} \hat{\varepsilon}_t^2} \qquad (10)$$

when CO correction is used, where $\hat{\varepsilon}_t$ represents LAV residuals from the uncorrected LAV regression. These are the estimators suggested by Park and Mitchell (1980) when using least squares estimation and are also typical of those that have been used in the LAV context.

An alternative approach suggested by Mizon (1995) is to include a lagged dependent variable as an explanatory variable and view this as part of the data generating process (DGP). No other testing for autocorrelation or correction for autocorrelation would be used. The model suggested can be written

$$y_t = \beta_0 + \beta_1 x_t + \beta_2 y_{t-1} + \eta_t, \qquad (11)$$

for $t = 2, ..., T$ (note that t = 1 is not used due to the inclusion of the lagged variable). In (11), y_t and x_t are the t^{th} observations on the dependent and explanatory variables, respectively. The η_t represents disturbance components, which are assumed to be independent and identically distributed, although not necessarily normal. The parameters β_0, β_1 and β_2 are unknown and must be estimated; however, in this application it is β_1 that is of interest. This method is referred to as LAVLAG.

Referring to the model in (2), Bassett and Koenker (1978) showed that the LAV coefficient estimator has an asymptotic distribution that converges to $N(\boldsymbol{\beta}, \lambda^2(\mathbf{X}'\mathbf{X})^{-1})$ where $\frac{\lambda^2}{T}$ is the asymptotic variance of the sample median for a sample of size T from the disturbance distribution. The scale parameter, λ, is defined as $\lambda = 1/[2\, f(m)]$, where $f(m)$ is the probability density function (pdf) of the disturbance distribution evaluated at the median. These same results are obtained when $\mathbf{X}$ is replaced by $\mathbf{X}^*$ for the model in (6) (Weiss, 1990).

The test considered in this study is the basic test for slope coefficient significance, i.e., H_0: $\beta_1 = 0$.

Three test statistics were examined: the WALD, the Likelihood Ratio (LR) and the Lagrange Multiplier (LM). The WALD, LR and LM statistics each have, asymptotically, a Chi-square distribution with k_2 degrees of freedom. (See Koenker and Bassett (1982) for further details on these test statistics.) The small sample properties of the test statistics are analytically intractable. Examination of the empirical level of significance of the test statistics in small samples was performed using a simulation.

Both the WALD and LR test statistics require the estimation of the scale parameter λ, whereas the LM test statistic does not. One often-suggested estimator for λ can be computed as follows:

$$\hat{\lambda} = \frac{(T')^{1/2}[e_{(T'-m-1)} - e_{(m)}]}{z_{\alpha/2}}$$

where

$$m = \frac{T'+1}{2} - z_{\alpha/2}\left(\frac{T'}{4}\right)^{1/2} \qquad (12)$$

where the $e_{(.)}$ are ordered residuals from the LAV-fitted model, and $T' = T - r$ where r is the number of zero residuals. A value of $\alpha = 0.05$ is typically suggested. This estimator is referred to as the SECI estimator. McKean and Schrader (1984) used Monte Carlo simulation to compare several methods of studentizing the sample median in which the SECI performed well and the value of $\alpha = 0.05$ produced the best results.

Sheather (1987) summarized the results of a Monte Carlo simulation to compare the SECI estimator and several other estimators for λ, including some that do not extend easily to the regression application. The conclusion was that the SECI estimator provides a good, quick point estimate of the standard error. Dielman and Pfaffenberger (1992) and Dielman and Rose (1996) also noted that this estimator performs reasonably well when used to compute the LR test statistic.

In this study, four different options in constructing the estimator of λ were considered. as follows:

1. SECI1: $\hat{\lambda}_1$ uses $z = 1.96$ (the $\alpha = 0.05$ value) and T' = total number of observations (T).

2. SECI2: $\hat{\lambda}_2$ uses $t_{0.025}$ with T degrees of freedom rather than the z value and T' = total number of observations (T).

3. SECI3: $\hat{\lambda}_3$ uses $z = 1.96$ (the $\alpha = 0.05$ value) and $T' = T - r$ where r is the number of zero residuals.

4. SECI4: $\hat{\lambda}_4$ uses $t_{0.025}$ with $T - r$ degrees of freedom rather than the z value and $T' = T - r$ where r is the number of zero residuals.

The notation W1, W2, W3 and W4 is used to indicate the WALD test using variance estimator 1, 2, 3 or 4, and L1, L2, L3 and L4 indicate the LR test using variance estimator 1, 2, 3 or 4. Most literature in this area recommends using the estimator SECI3. These options were considered in Dielman (2006) for models with independent errors and SECI1 and SECI2 were found to produce improved results over SECI3 in small samples. As noted, the LM test does not require the use of an estimate of λ.

The model considered in this study is described in (1). The explanatory variable values were generated as follows:

1. Autoregressive independent variable: $x_t = ax_{t-1} + u_t$ for $t = 1, 2, ..., T$ with u_t chosen from the N(0, 2) distribution. The values of a used were 0.0, 0.4 and 0.8

2. Stochastic trend: $x_t = at + u_t$ for $t = 1, 2, ..., T$ with the u_t chosen from the N(0, 2) distribution. The values of a used were 0.4 and 0.8.

3. Linear time trend: $x_t = t$ for $t = 1, 2, …., T$

After being generated, the independent variable values are held fixed throughout the experiment. The disturbances, η_t, were chosen from one of the following disturbance distributions:

1. Normal (0, 1);

2. Laplace with mean 0 and variance 2;

3. Contaminated normal with disturbances drawn from the standard normal distribution 85% of the time, and a normal distribution with mean 0 and variance 25 the other 15% of the time; and

4. Cauchy with median 0 and scale parameter 1.

Finally, after generating the η_t, the ε_t values are created as $\varepsilon_t = \rho\varepsilon_{t-1} + \eta_t$ where $\varepsilon_0 = \frac{\eta_0}{1-\rho^2}$ and η_0 is an initial draw from the disturbance distribution. The values of ρ used were 0.0, 0.1, 0.3, 0.5, 0.7 and 0.9.

The disturbances were generated independently of the explanatory variables. All random numbers were generated using IMSL subroutines and the simulation was written in FORTRAN.

The parameter β_0 was set equal to zero (without loss of generality). To determine

empirical levels of significance, the parameter β_1 is set equal to zero, the test of H_0: $\beta_1 = 0$ is performed, and the number of rejections of the true hypothesis is recorded.

The sample size used was $T = 20$. For each factor level combination in the experimental design, 10,000 Monte Carlo trials were used to evaluate estimates and assess levels of significance. (Each factor level combination is determined by the disturbance distribution, type of independent variable and the value of the autocorrelation coefficient for a total of 144 factor level combinations).

Results

Estimation

Table 1 shows mean square error (MSE) ratios for the estimates of the coefficient of the explanatory variable. The ratios are of the MSE of each estimation method to the MSE of the LAV estimator. MSE ratios less than one favor each of the estimator types over LAV; MSE ratios greater than one favor LAV. These are medians of the results over the four error distributions (Cauchy, Laplace, Contaminated Normal, Normal). Each of the six explanatory variable types is listed in a separate panel of the table. Panels A, B and C are for autoregressive explanatory variables with Lambda = 0.0, 0.4 and 0.8 respectively.

For example, in Panel A the explanatory variable is autoregressive with Lambda = 0.0 (that is, a normally distributed explanatory variable). The MSE ratio of LAVPW to LAV when Rho = 0.0 is 1.01. Thus, LAV is favored over LAVPW (barely) in this instance. However, little is lost by performing the correction for autocorrelation. For the autoregressive independent variable, this is true in all cases when Rho = 0.0. Although LAV is never unfavorable, there is often little or no difference, so the option to always correct for autocorrelation results in little loss in estimator efficiency. When the explanatory variable is autoregressive, there is little difference in whether the LAVPW or LAVCO correction is used. The LAVLAG alternative results in a larger loss in efficiency when Rho is small, for example the MSE ratio of LAVLAG to LAV is 1.05. As Rho increases, the relative efficiency of LAVLAG to LAV increases, but not as quickly as LAVPW or LAVCO when Lambda is 0.0 or 0.4. However, when Rho is large and Lambda is 0.8, the LAVLAG alternative results in greater efficiency than LAV and, in fact, greater efficiency than the other alternatives.

When the independent variable follows a stochastic trend (Panels D and E) it is also true that little is lost by performing the correction for autocorrelation. In this case, however, LAVPW is slightly better than LAVCO. The LAVLAG alternative shows a larger loss in efficiency when Rho is small than in the autoregressive case. For example the MSE ratio of LAVLAG to LAV is 1.16 for Lambda = 0.4 and 1.07 for Lambda = 0.8. As Rho increases, the relative efficiency of LAVLAG to LAV increases faster than LAVPW and the LAVLAG alternative soon provides greater efficiency than LAV and greater efficiency than the other alternatives.

The results for the fixed trend are similar to those for the stochastic trend, except that the LAVCO method fails miserably once Rho reaches 0.5. The LAVLAG MSE ratio is 1.2 when Rho is zero, but this approach recovers quickly and is more efficient than any of the other approaches when Rho is 0.3 or greater. The primary conclusion from examination of MSEs is to avoid the LAVCO correction. A secondary conclusion is that LAVLAG compares favorably to LAVPW.

Hypothesis Testing

Empirical significance levels of the test for coefficient significance were examined. Due to the poor estimation performance of the LAVCO method, that procedure is eliminated from consideration. All tests were performed using a nominal level of 0.05, thus, it is desirable to have the resulting empirical level close to this value. As a result, for purposes of this analysis a test is considered well-behaved if the empirical level is 0.06 or less.

Table 2 shows the number of times each method had an empirical significance level of 0.06 or less. Tests with larger numbers in Table 2 are viewed as more reliable because they do not overly reject true null hypotheses. The LR2, W2, LR1, LR3 and LM tests (in that order) had the highest total incidences of empirical levels that were at or below 0.06 over all the experimental design points.

Table 1: MSE Ratios for the Estimates of the Coefficient of the Explanatory Variable

Panel A: Autoregressive with Lambda = 0.0

	Rho					
	0.0	0.1	0.3	0.5	0.7	0.9
LAVPW	1.01	1.00	0.91	0.77	0.61	0.46
LAVCO	1.00	0.98	0.90	0.76	0.60	0.46
LAVLAG	1.05	1.04	0.99	0.86	0.70	0.51

Panel B: Autoregressive with Lambda = 0.4

	Rho					
	0.0	0.1	0.3	0.5	0.7	0.9
LAVPW	1.02	0.97	0.92	0.81	0.67	0.48
LAVCO	1.01	0.99	0.91	0.81	0.66	0.48
LAVLAG	1.05	1.02	0.92	0.82	0.68	0.50

Panel C: Autoregressive with Lambda = 0.8

	Rho					
	0.0	0.1	0.3	0.5	0.7	0.9
LAVPW	1.01	1.00	0.93	0.80	0.65	0.48
LAVCO	1.01	1.00	0.92	0.81	0.66	0.48
LAVLAG	1.07	1.01	0.87	0.71	0.53	0.37

Panel D: Stochastic Trend with Lambda = 0.4

	Rho					
	0.0	0.1	0.3	0.5	0.7	0.9
LAVPW	1.00	1.01	0.96	0.86	0.76	0.83
LAVCO	1.05	1.04	1.01	0.92	0.79	0.84
LAVLAG	1.16	1.07	0.90	0.70	0.51	0.35

Panel E: Stochastic Trend with Lambda = 0.8

	Rho					
	0.0	0.1	0.3	0.5	0.7	0.9
LAVPW	1.01	1.00	0.93	0.80	0.65	0.48
LAVCO	1.01	1.00	0.92	0.81	0.66	0.48
LAVLAG	1.07	1.01	0.87	0.71	0.53	0.37

Panel F: Linear Trend

	Rho					
	0.0	0.1	0.3	0.5	0.7	0.9
LAVPW	1.01	1.01	0.99	0.94	0.88	0.88
LAVCO	1.06	1.08	5.88	1351	1952	3455
LAVLAG	1.20	1.10	0.92	0.73	0.56	0.45

Notes: The ratios are of the MSE of each result to the MSE of the LAV estimator. MSE ratios less than one favor each of the estimator types over LAV; MSE ratios greater than one favor LAV. These are medians of the results over four error distributions. Each of the six explanatory variable types is listed in a separate panel of the table.

Considering estimation procedures, the LAVLAG procedure had the most instances overall, 668, at or below 0.06. Combinations of test and estimation procedure that have the largest number of empirical significance levels at or below 0.06 are (in order): LAVLAG/LR1, LAVLAG/LR3, LAVLAG/W1 and LAVLAG/LR2. Note that LAVPW does not perform particularly well. LAVPW is the autocorrelation correction procedure typically recommended in previous studies. Also, LR3 is the test used in many previous studies, but LR1 or LR2 could be viewed as preferred in this study. This is consistent with the findings of Dielman (2006) in models without autocorrelation.

Table 3 provides detail on specific empirical levels of significance for estimation method/test combinations for selected values of the autocorrelation coefficient, Rho (panels in the table correspond to Rho = 0.0, 0.1, 0.3, 0.5, 0.7, 0.9). The values in the table represent the median percentage of rejections for estimation method/test combinations with median taken over the four error distributions and over the six explanatory variable types. In the first panel of the table, for example, empirical levels of significance for Rho = 0.0 are shown.

The LAV method had empirical significance level of 0.06 or less for several of the tests: W2, LM, LR1, LR2, and LR3. The level for LAVPW was 0.06 or less for W2 and LR2. The LAVLAG method had level of 0.06 or less for W1, W3, LR1, LR2, LR3 and LR4.

When autocorrelation is at a moderate level of 0.5, there are two combinations with empirical level of significance below 0.06: LAVLAG/LR1 and LAVLAG/LR3. All levels for LAV and LAVPW are above 0.06 and are similar for these two methods, even though LAVPW supposedly corrects for autocorrelation.

When Rho is 0.9 (a high level of autocorrelation), there are no cases when the empirical level of significance is below 0.06. The closest values are 0.09 for LAVLAG/W1, LAVLAG/LR1 and LAVLAG/LR3. Note that the LAVPW method, one of the traditional corrections for autocorrelation, had very high empirical levels in a case when it might be expected to perform well. The levels are better than the uncorrected LAV, but still very high.

Table 2: Number of Times Each Method Had Empirical Significance Level of 0.06 or Less

Method	Test									Totals
	W1	W2	W3	W4	LM	LR1	LR2	LR3	LR4	
LAV	21	85	17	9	67	56	84	46	24	409
LAVPW	0	91	0	0	22	1	74	1	1	190
LAVLAG	101	45	76	18	43	107	92	101	85	668
Totals	122	221	93	27	132	164	250	148	110	

Table 3: Empirical Levels of Significance (Proportion of Rejections) for Estimation Method/Test Combination for Selected Values of the Autocorrelation Coefficient, Rho

Rho	Method	Test								
		W1	W2	W3	W4	LM	LR1	LR2	LR3	LR4
0	LAV	0.08	0.03	0.09	0.10	0.05	0.06	0.03	0.06	0.07
	LAVPW	0.10	0.05	0.12	0.13	0.08	0.10	0.06	0.10	0.11
	LAVLAG	0.05	0.07	0.06	0.08	0.12	0.05	0.05	0.05	0.06
0.1	LAV	0.09	0.04	0.10	0.12	0.07	0.07	0.04	0.08	0.09
	LAVPW	0.11	0.05	0.12	0.14	0.08	0.10	0.06	0.11	0.12
	LAVLAG	0.05	0.07	0.06	0.08	0.12	0.05	0.06	0.05	0.06
0.3	LAV	0.13	0.06	0.14	0.16	0.10	0.11	0.07	0.12	0.13
	LAVPW	0.14	0.06	0.15	0.17	0.09	0.12	0.08	0.13	0.14
	LAVLAG	0.06	0.07	0.06	0.08	0.13	0.05	0.06	0.06	0.07
0.5	LAV	0.17	0.10	0.19	0.21	0.15	0.17	0.12	0.17	0.19
	LAVPW	0.16	0.09	0.17	0.20	0.11	0.15	0.10	0.15	0.17
	LAVLAG	0.07	0.08	0.07	0.09	0.14	0.06	0.07	0.06	0.07
0.7	LAV	0.25	0.16	0.27	0.30	0.22	0.25	0.20	0.26	0.28
	LAVPW	0.19	0.11	0.20	0.23	0.13	0.18	0.13	0.19	0.20
	LAVLAG	0.08	0.09	0.08	0.10	0.15	0.07	0.08	0.08	0.09
0.9	LAV	0.35	0.26	0.37	0.40	0.32	0.37	0.31	0.38	0.40
	LAVPW	0.25	0.17	0.26	0.29	0.15	0.25	0.20	0.26	0.27
	LAVLAG	0.09	0.11	0.10	0.12	0.18	0.09	0.10	0.09	0.11

Note: These are medians of the results over the four error distributions and over the six explanatory variable types.

Conclusion

The following conclusions are derived from the simulation study. Regarding estimation:

1. The LAVCO correction should be avoided due to possible extreme loss in efficiency.

2. The option to always correct for autocorrelation using the LAVPW correction never results in much efficiency loss.

3. Adding a lagged dependent variable rather than using the LAVPW correction is a viable option. The LAVLAG alternative typically results in a larger loss in efficiency than LAVPW when there is little autocorrelation, but an increase in efficiency when autocorrelation is more severe.

For hypothesis testing, the LAVLAG method had empirical levels of significance that were acceptable more often than LAVPW so is preferred in this sense. Both LAVPW and LAVLAG provide better protection against type one errors than LAV. However, the empirical levels of both are still high in some cases.

When estimating a regression with independent disturbances, Dielman and Rose (1995a, 2002) compared bootstrap tests to traditional tests in a LAV regression with independent errors and found that the bootstrap tests were generally competitive with LR tests that also perform well when disturbances are independent. It would be prudent to examine a bootstrap test in the context of autocorrelated errors as well; however, care must be taken in designing the bootstrap resampling process to preserve the autocorrelation structure.

References

Bassett, G., & Koenker, R. (1978). Asymptotic theory of least absolute error regressions. *Journal of the American Statistical Association*, *73*, 618-622.

Cochrane, D., & Orcutt, G. (1949). Application of least squares regression to relationships containing autocorrelated error terms. *Journal of the American Statistical Association*, *44*, 32-61.

Dielman, T. (2005). Least absolute value regression: Recent contributions. *Journal of Statistical Computation and Simulation*, *75*, 263-286.

Dielman, T. (2006). Variance estimates and hypothesis tests in least absolute value regression. *Journal of Statistical Computation and Simulation*, *76*, 103-114.

Dielman, T., & Pfaffenberger, R. (1984). Small sample properties of estimators in the autocorrelated error model: A review and some additional simulations. *Statistical Papers/Statistiche Hefte*, *30*, 163-183.

Dielman, T., & Pfaffenberger, R. (1990). Tests of linear hypotheses in LAV regression. *Communications in Statistics - Simulation and Computation*, *19*, 1179-1199.

Dielman, T., & Pfaffenberger, R. (1992). A further comparison of tests of hypotheses in LAV regression, *Computational Statistics and Data Analysis*, *14*, 375-384.

Dielman, T., & Rose, E. (1994a). Estimation in least absolute value regression with autocorrelated errors. *Journal of Statistical Computation and Simulation*, *50*, 29-43.

Dielman, T., & Rose, E. (1994b). Forecasting in least absolute value regression with autocorrelated errors: a small-sample study. *International Journal of Forecasting*, *10*, 539-547.

Dielman, T., & Rose, E. (1995a). A bootstrap approach to hypothesis testing in least absolute value regression. *Computational Statistics and Data Analysis*, *20*, 119-130.

Dielman, T., & Rose, E. (1995b). Estimation after pre-testing in least absolute value regression with autocorrelated errors. *Journal of Business and Management*, *2*, 74-95.

Dielman, T., & Rose, E. (1996). A note on hypothesis testing in LAV multiple regression: A small sample comparison. *Computational Statistics and Data Analysis*, *21*, 463-470.

Dielman, T., & Rose, E. (1997). Estimation and testing in least absolute value regression with serially correlated disturbances. *Annals of Operations Research*, *74*, 239-257.

Dielman, T., & Rose, E. (2002). Bootstrap versus traditional hypothesis testing procedures for coefficients in least absolute value regression. *Journal of Statistical Computation and Simulation*, *72*, 665-675.

Koenker, R. (1987). A comparison of asymptotic testing methods for L_1-regression. In: Y. Dodge (Ed.), *Statistical data analysis based on the L_1-norm and related methods*, 287-295. Amsterdam: North-Holland.

Koenker, R., & Bassett, G. (1982). Tests of linear hypotheses and L_1 estimation. *Econometrica*, *50*, 1577-1583.

Maeshiro, A. (1979). On the retention of the first observation in serial correlation adjustment of regression models. *International Economic Review*, *20*, 259-265.

Mizon, G. (1995). A simple message for autocorrelation correctors: Don't. *Journal of Econometrics*, *69*, 267-288.

McKean, J., & Schrader, R. (1984). A comparison of methods for studentizing the sample median. *Communications in Statistics - Simulation and Computation*, *13*, 751-773.

Park, R., & Mitchell, G. (1980). Estimating the autocorrelated error model with trended data. *Journal of Econometrics*, *13*, 185-201.

Prais, S., & Winsten, C. (1954). *Trend estimators and serial correlation*. Cowles Commission Discussion Paper: Stat. No. 383, Chicago.

Sheather, S. (1987). Assessing the accuracy of the sample median: Estimated standard errors versus interpolated confidence intervals. In: Y. Dodge (Ed.), *Statistical data analysis based on the L_1-norm and related methods*, 203-215. Amsterdam: North-Holland.

Stangenhaus, G. (1987). Bootstrap and inference procedures for L_1 regression. In: Y. Dodge (Ed.) *Statistical data analysis based on the L_1-norm and related methods*, 323-332. Amsterdam: North-Holland.

Weiss, A. (1990). Least absolute error estimation in the presence of serial correlation. *Journal of Econometrics*, *44*, 127-158.

Journal of Modern Applied Statistical Methods
November 2011, Vol. 10, No. 2, 549-570

1538 – 9472/11/$95.00

A Comparison of Factor Rotation Methods for Dichotomous Data

W. Holmes Finch
Ball State University,
Muncie, IN

Exploratory factor analysis (EFA) is frequently used in the social sciences and is a common component in many validity studies. A core aspect of EFA is the determination of which observed indicator variables are associated with which latent factors through the use of factor loadings. Loadings are initially extracted using an algorithm, such as maximum likelihood or weighted least squares, and then transformed - or rotated - to make them more interpretable. There are a number of rotational techniques available to the researcher making use of EFA. Prior work has discussed the advantages of a number of these criteria from a theoretical perspective, but few previous studies compare their performance across a broad range of conditions. This simulation study compared eight factor rotation criteria in terms of how well they were able to group dichotomous indicator variables correctly on the same factor, order the indicators by the magnitude of the factor loadings (identifying those indicators that were most strongly associated with the factors) and estimate the inter-factor correlations. Results reveal a mixed pattern of performance among the various rotations with the orthogonal Equamax consistently near the top in terms of correctly grouping and ordering indicator variables, and the orthogonal Facparsim performing well with more observed indicators. Advice regarding possible rotations to use for researchers conducting EFA with dichotomous indicators is provided.

Key words: Factor rotation, dichotomous data, exploratory factor analysis, EFA.

Introduction

Exploratory Factor Analysis (EFA) of items on an instrument is a tool employed by psychometricians in the investigation of validity evidence for cognitive and affective measures (Zumbo, 2007; McDonald, 1999). In conjunction with subject matter expertise regarding the purpose of the instrument and its assumed structure, EFA can be used to identify the latent constructs underlying the observed items (McLeod, Swygert & Thissen, 2001). When items are found to group in conceptually meaningful ways based on content, instrument developers can conclude that the traits the scale is intended to measure are actually being represented. Conversely, when individual items are found to load on multiple factors - or to group in ways that do not conform to their content or intent - developers may target them for revision or removal from the instrument (Sass & Schmitt, 2010). Given its role in validity assessment, psychometricians must have a full understanding regarding the performance of EFA in the context of item level data under a variety of conditions. The objective of this simulation study was to investigate one important aspect of the EFA analysis process: factor rotation. A variety of factor rotation methods were compared with respect to how well they recovered the underlying latent structure for a set of dichotomous indicators like those that might comprise a psychological or educational scale. (Readers interested in learning more about the basic factor analysis model are encouraged to read one of several excellent references including: Gorsuch, 1983; Thompson, 2004; McLeod, Swygert & Thissen, 2001.)

W. Holmes Finch is a Professor of Psychology in the Department of Educational Psychology, and Educational Psychology Director of Research in the Office of Charter School. Email him at: whfinch@bsu.edu.

Factor Analysis of Dichotomous Data

The original EFA model was based on the presumption that observed indicators were continuous variables, calling into question its applicability for dichotomous data such as that from item responses (Gorsuch, 1983). Early analyses applying the standard linear EFA model to dichotomous item response data consistently identified a factor reflecting item difficulty, having nothing to do with substantive dimensions related to item content (Hattie, 1985; Guilford, 1941; Spearman, 1927). Furthermore, the use of linear factor analysis with dichotomous items was found to produce distorted factor loading estimates for very difficult and very easy items (Hattie, 1985).

In response to these problems, McDonald introduced nonlinear factor analysis based on the normal ogive (McDonald, 1967; 1962). In the case of dichotomous variables such as item responses, this factor model takes the form

$$P\{U_j = 1 \mid \theta\} = N(\beta_{j0} + \beta_{j1}\theta + \beta_{21}\theta + ... + \beta_{jm}\theta) \quad (1)$$

where U_j is the response to item with a 1 indicating correct, β_{j0} is the intercept for item j and β_{j1} is the factor loading for item j with latent trait m. Parameter estimation in this Normal Ogive Harmonic Analysis Robust Method (NOHARM) is conducted using unweighted least squares (ULS), allowing for analysis of large sets of items exhibiting high dimensionality (McDonald, 1981; 1967). This model was implemented in the NOHARM software package (Fraser & McDonald, 1988) and features both Varimax and Promax rotations.

Bock and Aitkin (1981) developed an alternative model for the factor analysis of dichotomous item response data that takes the form:

$$P\left(x_{ij} = 1 \mid \theta_j\right) = \frac{1}{\sqrt{2\pi}} \int_{-z_i(\theta_j)}^{\infty} e^{-\frac{t^2}{2}} dt \quad (2)$$

where $z_i\left(\theta_j\right) = a_j\left(\theta_{ti} - b_j\right)$, a_j is the slope for item j, b_j is the threshold for item j, and θ_{tj} is the latent trait for subject i on item j. In this conceptualization of the model, a_j corresponds to item discrimination and b_j corresponds to item difficulty, in the context of item response theory. This full information factor model underlies the TESTFACT software (Bock, et al., 2003) and is estimated using marginal maximum likelihood (MML), in contrast to the ULS used with NOHARM. Researchers comparing these approaches have found that ULS tends to provide more accurate parameter estimation for a smaller number of items, although MML is generally more accurate for more items (Gosz & Walker, 2002). As with NOHARM, TESTFACT allows for either VARIMAX or PROMAX rotations.

Christofferson (1975) also introduced a factor model for item response data based on the normal ogive model, as was McDonald's approach. The Christofferson model is expressed as

$$P\left(u_i = 1\right) = \int_{z_i}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt \quad (3)$$

where z_i is the threshold for item i. This model was expanded upon by Muthén (1978) and has been shown to be equivalent to McDonald's model (McDonald, 1997).

Another approach to factor analysis for dichotomous data, such as item responses, is based on robust weighted least squares (RWLS). Weighted least squares (WLS) estimation has been shown to perform poorly for categorical variables in the context of factor analysis with small to moderate sample sizes (Flora & Curran, 2004). Muthén, du Toit and Spisic (1997) and Muthén (1993) extended the WLS approach in the form of RWLS, which does not require the inversion of the weight matrix used in the standard WLS approach, leading to very stable parameter estimation for samples as small as 100 with dichotomous indicator variables (Flora & Curran, 2004). The RWLS approach can also be used in the context of EFA with the MPLus software package (Muthén & Muthén, 2007) as was done herein.

Factor Rotation

The estimation of factor loadings in EFA typically occurs in two stages, the first of

which - factor extraction - involves the initial estimation of loadings based on the covariance matrix for the indicator variables. The second step in an EFA - factor rotation - involves the transformation of the initial factor loadings in order to make them more interpretable in terms of (ideally) clearly associating an indicator variable with a single factor (Gorsuch, 1983). Although a large number of rotation algorithms have been described in the literature, these criteria all have the common goal of reducing a complexity function, $f(\Lambda)$, so that the loadings approximate a simple structure and are thus more interpretable in practice.

The notion of simple structure has been discussed extensively in the factor analysis literature, and though there is a common sense as to its meaning, there is no agreement regarding exact details. Thurstone (1947) first described simple structure as occurring when each row in the factor loading matrix has at least one zero, where rows represent indicator variables and columns represent factors. He also included 4 other rules that were initially intended to yield over-determination and stability of the factor loading matrix, but which were subsequently used by other researchers to define simple structure for methods of rotation (Browne, 2001). Subsequent to Thurstone's work, others varying definitions of simple structure have been provided. For example, Jennrich (2007) defined perfect simple structure as occurring when each indicator has only one nonzero factor loading and compared it to Thurstone simple structure in which there are a fair number of zeros in the factor loading matrix, but not as many as in perfect simple structure. Conversely, Browne (2001) defined the complexity of a factor pattern as the number of nonzero elements in the rows of the loading matrix. These many varying definitions of simple structure have led to the development of a number of rotational criteria with the overarching goal of obtaining the most interpretable solution possible for a set of data (Asparouhov & Muthén, 2009).

Factor rotations are broadly classified as either: (1) orthogonal, in which the factors are constrained to be uncorrelated, or (2) oblique, in which this constraint is relaxed. Within each of these classes, several options are available. Browne (2001) provides an excellent discussion of a number of these rotational criteria along with a history of their development and concluded that, when the factor pattern in the population conformed to what is termed above as pure simple structure, most of the rotation methods reviewed produced acceptable solutions. However, when there was greater complexity in the factor pattern, the rotations did not all perform equally well and - in some cases - the majority of them produced unacceptable results (Browne, 2001). For this reason, he argued for the need of educated human judgment in the selection of the best factor rotation solution for a given problem. In a similar vein, Yates (1987) stated that some rotations are designed to find a perfect simple structure solution in all cases, even when this may not be appropriate for the data at hand.

Several excellent discussions of these rotation criteria are available in the literature, including two recently published manuscripts which provide detailed descriptions for interested readers (Sass & Schmitt, 2010; Asparouhov & Muthén, 2009). The rotations included in this study are summarized in Table 1. Many of these methods are readily available in common statistical software packages such as MPlus (Muthén & Muthén, 2007), which is featured in this study, as well as SAS and SPSS. Perhaps the most popular method in applied practice is the orthogonal Varimax rotation (Kaiser, 1958), which is a member of a larger group of criteria known collectively as the Orthomax family of rotations. The goal in Varimax rotation is to create simple structure by maximizing differences among loadings within factors across variables. Other notable Orthomax rotations include Quartimax, Equamax, Parsimax and Factor Parsimony. Promax is a two-stage oblique Procrustean rotation in which loadings are first obtained from the orthogonal Varimax rotation and then transformed based upon a target matrix of loadings raised to a particular power (typically the 4^{th} power), after which a transformation matrix is obtained using least squares (Hendrickson & White, 1964). Other Procrustean rotations include Promaj (Trendafilov, 1994) and Promin (Lorenzo-Seva, 1999). Another group of factor rotations is the Crawford-Ferguson (CF) family (Crawford &

Table 1: Summary of Studied Rotation Methods*

Rotation Criteria	Definition	Comments
Varimax	$f(\Lambda)=\left[p\sum_{i=1}^{p}\left(\lambda_{ij}^2\right)^2-\left(\sum_{i=1}^{p}\left(\lambda_{ij}^2\right)^2\right)^2\right]/\,p^2$	Spreads variance across factors
Quartimin	$f(\Lambda)=\sum_{i=1}^{p}\sum_{j=1}^{m}\sum_{l\neq j}^{m}\lambda_{ij}^2\lambda_{il}^2$	Designed to minimize complexity of loadings across indicator variables.
Quartimax	$f(\Lambda)=\sum_{i=1}^{p}\sum_{j=1}^{m}\lambda_{ij}^4+\sum_{i=1}^{p}\sum_{j=1}^{m}\sum_{l\neq j}^{m}\lambda_{ij}^2\lambda_{il}^2$	Spreads variance across indicators
Equamax	$\left(1-\frac{m}{2p}\right)\sum_{i=1}^{p}\sum_{j=1}^{m}\sum_{l\neq j}^{m}\lambda_{ij}^2\lambda_{il}^2+\frac{m}{2p}\sum_{i=1}^{m}\sum_{i=1}^{p}\sum_{l\neq j}^{p}\lambda_{ij}^2\lambda_{ij}^2$	Combines Quartimax and Varimax criteria
Parsimax	$f(\Lambda)=\left(1-\frac{m-1}{p+m-2}\right)\sum_{i=1}^{p}\sum_{j=1}^{m}\sum_{l\neq j}^{m}\lambda_{ij}^2\lambda_{il}^2+\frac{m-1}{p+m-2}\left(\sum_{i=1}^{m}\sum_{i=1}^{p}\sum_{l\neq j}^{p}\lambda_{ij}^2\lambda_{il}^2\right)$	Equal weight is given to factor and indicator complexity.
Geomin	$f(\Lambda)=\sum_{i=1}^{p}\left[\prod_{j=1}^{m}\left(\lambda_{ij}^2+\in\right)\right]^{\frac{1}{m}}$	Accommodates factor complexity while still providing interpretable solution.
Promax	Raise loadings from Varimax to some power (e.g., 4) and rotate the resulting matrix allowing for correlated factors.	Based on Varimax rotation, but allows for correlated factors.
Facparsim	$f(\Lambda)=\sum_{i=1}^{m}\sum_{i=1}^{p}\sum_{l\neq j}^{p}\lambda_{ij}^2\lambda_{il}^2$	Minimizes loading complexity across factors.

*p=Number of indicators, m=Number of factors, λ=Extracted factor loading

Ferguson, 1970). This criterion accounts for complexity across both the indicator variables and the factors. Members of the CF family differ in terms of a parameter, *k*, ranging between 0 and 1, where larger values of *k* place greater weight on minimizing factor complexity, whereas lower values emphasize the minimization of indicator variable complexity (Crawford & Ferguson, 1970). Other rotations that have been discussed widely in the literature are oblique Quartimin (Carroll, 1957), which seeks to minimize complexity only within the indicator variables, and Geomin (Yates, 1987) which also was designed to minimize variable complexity, but which allows for more such complexity than does Quartimin. There are a number of other rotation criteria extant in the literature. However, given that the current study is focused on comparing methods that are available to practitioners in commonly available software, they will not be discussed here. The interested reader is invited to read Mulaik (2010) and Browne (2001) for excellent descriptions of these alternative methods of rotation.

Prior Research on Factor Rotations

As noted, a large number of rotational criteria are available to a researcher interested in using EFA. Some of these, such as Varimax and Promax, are well known and frequently used, while others may be less well known but offer statistical advantages over the more commonly used approaches (Asparouhov & Muthén, 2009). Despite the abundance of available rotational methods, a great deal of empirical research has not been conducted regarding which might be best in a given research context (Sass & Schmitt, 2010). In addition, virtually none of the prior work examining the performance of these various rotation methods has been conducted with dichotomous indicator variables (the focus of this study). Therefore, earlier work using continuous indicators provides the only extant evidence regarding the comparative behavior of factor rotation methods, all of which can be applied to both EFA with continuous or dichotomous indicators. Thus, although they did not utilize dichotomous indicators, earlier studies provide researchers with some insights into what might be expected with regard to the performance of these rotation methods in general. Nevertheless, it is not clear to what extent earlier research with continuous indicators may be applicable. Therefore, this article builds upon this earlier research in an attempt to extend these results based on continuous variables to the case of dichotomous indicators.

One recent Monte Carlo study (Sass & Schmitt, 2010) compared the ability of four rotational methods in terms of their abilities to reproduce the population factor loadings used to generate the data. This study involved 30 standard normally distributed observed indicators with 2 factors, and 4 different types of factor structure including perfect simple, approximate simple, complex and general (a single common factor) structures; note that the variables used in this study were continuous and not categorical. Sass and Schmidt focused on the performance of these rotation methods for normally distributed indicator variables; however, their study is relevant to this research with dichotomous indicators in that it is one of the few to systematically compare multiple rotational criteria. Furthermore, several of the rotations considered by Sass and Schmidt are also included in this study. Therefore, although their results with continuous, normally distributed variables may not be directly applicable to situations involving dichotomous indicators, their study does provide some potential insights into the performance of the rotational criteria that may in turn inform this research.

Sass and Schmidt generated a sample of 300, with correlations between the factors (0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6 and 0.7) and used four oblique single stage rotational criteria, including Quartimin, CF-Equamax, CF-Facparsim and Geomin. They found that in the perfect simple structure condition all of the methods performed equally well, echoing Browne (2001). In the more complex cases, however, CF-Equamax and CF-Facparsim demonstrated somewhat less bias in factor loading estimates than did the other rotations. These authors concluded that researchers must be careful not to think of a particular rotational solution as inherently right or wrong, given that model fit does not change based on rotation. Echoing Browne (2001), Sass and Schmitt argued that the selection of the best

rotation must be made by the researcher using informed judgment, and cannot be done deterministically based solely on statistical results.

A similar finding was reported by Asparouhov & Muthén (2009), who stated that based on their own simulated comparisons of the Geomin and Quartimin rotation criteria with loading bias as the primary outcome variable, the researcher in the end is responsible for determining what constitutes a simple and interpretable solution. Consistent with Sass and Schmitt (2010), they found that for simple factor patterns the rotation criteria performed similarly, but for more complex patterns the results across rotational methods (and even for the same method using different settings) might differ substantially. As noted, although the previous simulation research comparing factor rotation performance was focused on continuous indicator variables, it remains relevant for this study in that it provides the only published evidence regarding the behavior of these rotation criteria, all of which can be used with dichotomous indicators.

The goal of this simulation study was to extend upon this earlier work by comparing the performance of several methods of factor rotation with dichotomous, rather than normally distributed continuous, indicator variables, and by including several more rotation criteria, including the very popular Varimax and Promax methods as well as others that have been shown to be effective previously. Furthermore, the current study extends upon these earlier efforts by including a broader range of conditions with respect to number of indicator variables, sample sizes and number of factors. Finally, the focus of this study in terms of outcomes is different than that of the previously mentioned research.

Methodology

A Monte Carlo simulation study was conducted to compare the performance of several methods of factor rotation in four areas: (1) proportion of correctly grouped indicator variables, (2) proportion of incorrectly grouped indicator variables, (3) proportion of indicator variables correctly ordered based on their population factor loading values, and (4) for oblique rotations, bias in the estimates of inter-factor correlations. Outcome 1 was the proportion of all item pairs that should have been grouped together that actually were, and outcome 2 was the proportion of all item pairs that should have been kept separate that actually were. Outcome 3 was the proportion of cases in which the item with the larger factor loading in the population also had the larger loading in the sample. Outcome 4 was the degree of accuracy of the inter-factor correlation estimate, which was calculated as $r_o - r_p$, where r_o = sample estimate of inter-factor correlation between two factors and r_p = population inter-factor correlation used in data simulation. In addition, the standardized bias of the correlation estimates was also calculated as the bias defined previously divided by the standard deviation of the correlation estimates.

These outcomes were selected because they reflect issues that applied researchers might be interested in; that is, how accurately are the factors defined by appropriately grouped variables, how well ordered are the indicators in terms of the magnitude of their relationships to the factors and how well estimated are the correlations among the factors. Although all of these outcomes may be important in specific contexts, one could argue that the ability to accurately identify the factor structure by correctly grouping the items together may be the most crucial. Given that validity assessment is typically based on the extent to which the empirically identified factors reflect what would be expected for the constructs in question based on substantive content of the items, the accuracy of an EFA solution from a sample to reproduce the population factor structure would seem to be paramount. However, in certain circumstances each of these outcomes would be important to researchers using EFA.

For each combination of the simulation conditions, 1,000 replications were generated using MPlus, version 5.1 (Muthén & Muthén, 2007) and all study conditions were completely crossed with one another. Dichotomous indicators were generated in MPlus using threshold values of 0.25 and were held constant across the observed variables. The relationship between the threshold (τ) value and the probability (P_i) of a respondent endorsing a

dichotomous item is $P_i = \frac{1}{1+e^{-\tau}}$. The threshold value of 0.25 corresponds to a probability of endorsing an item of 0.56 and was selected because it has been used in other simulation research involving factor analysis of dichotomous data (French & Finch, 2006; Meade & Lautenschlager, 2004).

For each replication, exploratory factor analysis with Robust Weighted Least Squares (WLSMV) extraction was conducted using the MPlus software because it has been supported for use with categorical data in prior research (e.g., Muthén & Muthén, 2007; Flora & Curran, 2004). In conducting EFA with dichotomous data, MPlus first calculates the tetrachoric correlation matrix among the variables and then uses it to estimate the factor analysis parameters (factor loadings, inter-factor correlations). The commands to run the analysis requested the extraction of the correct number of factors (2 or 4) for a given replication but because the analysis was EFA, individual indicators were not linked to specific factors as they would have been in a confirmatory factor analysis. For example, when the data generated were from a 2 factor condition, the MPlus commands to run the EFA on the sample requested the extraction of 2 factors, but the individual indicators were not linked to a given factor.

Data were generated for either 2 or 4 factors in the population, and for each factor there were either 6 or 12 observed indicator variables, leading to the following combinations: 2 factors with 6 indicators each, 2 factors with 12 indicators each, 4 factors with 6 indicators each and 4 factors with 12 indicators each. Four inter-factor correlation conditions were simulated: 0.1, 0.3, 0.5 and 0.7. All pairs of factors were correlated at the same level for a given combination of study conditions. For example, in the 4 factor, 6 indicator condition with r = 0.3, each pair of the 4 factors were generated with a correlation of 0.3. Four sample size conditions were simulated, 100, 200, 500 and 1,000. Prior research studying the minimum sample size necessary for EFA to provide reliable results with continuous indicators has found that when communalities are relatively high (e.g., 0.5), and most of the factors have a large number of indicators population factor are recovered well with samples as small as 100 subjects (MacCallum, et al., 1999).

Conversely, MacCallum, et al. (1999), found that for low communalities and many factors, each of which has a small number indicators, samples of 500 or more are necessary. Preacher and MacCallum (2002) found that for sample sizes as low as 30, factor structure recovery was good (low root mean square error) provided that communalities were high (e.g., 0.8), the number of factors retained was 4 or fewer and the total number of indicators was 25 or more.

Subsequently, other researchers investigating the impact of sample size on factor analysis have reported similar findings with regard to the need for larger samples with relatively poorly conditioned solutions (fewer indicators with low factor loadings, low communalities and many factors), and the positive performance with smaller samples (fewer than 50) when factors are well conditioned (de Winter, Dodou & Wieringa, 2009; Gagné & Hancock, 2006; Mundfrom, Shaw & Ke, 2005). Of particular interest given the inclusion of non-simple structure conditions in the current research are the results of de Winter, et al., who found that in the presence of non-simple structure, EFA performs worse with relatively smaller samples in terms of factor structure recovery, particularly when factors are correlated at 0.5 or greater. Given these earlier studies, sample sizes selected for the current research range from what might be considered somewhat small (100) to very large (1,000).

Finally, the data were generated with 4 levels of factor structure complexity, reflecting different degrees to which individual indicators cross-loaded with a secondary factor. Table 2 provides an example of these patterns for each level of structural complexity in the 2 factor 6 indicator condition. For example, in complexity condition 1 each indicator has non-zero loadings for only one factor, whereas in the other 3 conditions, each indicator has an additional non-zero loading on one other factor with complexity conditions differing based upon the magnitude of these non-zero loadings. In the 4 factor conditions, each indicator variable had only 2 non-zero loadings, one for its primary factor and

Table 2: Example Factor Loading Patterns Used In the Simulations

Complexity Condition 1			Complexity Condition 2		
Indicator	Factor 1	Factor 2	Indicator	Factor 1	Factor 2
Y1	0.8	0	Y1	0.8	0.1
Y2	0.8	0	Y2	0.8	0.1
Y3	0.6	0	Y3	0.6	0.1
Y4	0.6	0	Y4	0.6	0.1
Y5	0.4	0	Y5	0.4	0.1
Y6	0.4	0	Y6	0.4	0.1
Y7	0	0.8	Y7	0.1	0.8
Y8	0	0.8	Y8	0.1	0.8
Y9	0	0.6	Y9	0.1	0.6
Y10	0	0.6	Y10	0.1	0.6
Y11	0	0.4	Y11	0.1	0.4
Y12	0	0.4	Y12	0.1	0.4
Complexity Condition 3			Complexity Condition 4		
Indicator	Factor 1	Factor 2	Indicator	Factor 1	Factor 2
Y1	0.8	0.2	Y1	0.8	0.3
Y2	0.8	0.2	Y2	0.8	0.3
Y3	0.6	0.2	Y3	0.6	0.3
Y4	0.6	0.2	Y4	0.6	0.3
Y5	0.4	0.2	Y5	0.4	0.3
Y6	0.4	0.2	Y6	0.4	0.3
Y7	0.2	0.8	Y7	0.3	0.8
Y8	0.2	0.8	Y8	0.3	0.8
Y9	0.2	0.6	Y9	0.3	0.6
Y10	0.2	0.6	Y10	0.3	0.6
Y11	0.2	0.4	Y11	0.3	0.4
Y12	0.2	0.4	Y12	0.3	0.4

the other for a single secondary factor. For example, in complexity condition 2 with 4 factors and 12 indicators for each, indicator 1 had a loading of 0.8 for factor 1, a loading of 0.1 for factor 2 and loadings of 0 for factors 3 and 4. On the other hand, indicator 48 had a loading of 0.4 for factor 4, a loading of 0.1 for factor 3 and 0 loadings for factors 1 and 2. The decision to allow indicators in the 4 factor conditions to cross load with only one other factor was made to avoid confounding the number of cross loadings with the number of factors, making it impossible to directly compare results in the 2 and 4 factors cases. Similar factor loading patterns were used with the other factor and indicator combinations included in this study. Although a very large number of different such factor patterns could have been simulated using the number of factors and indicators included in this study, these patterns were selected because it was felt that they represented a range of non-simple structure conditions, were few enough so as to keep the study manageable and allowed for investigation of the impact of progressively greater factor complexity.

The methods of factor rotation included the study were Quartimin (oblique), Varimax (orthogonal), Quartimax (orthogonal), Equamax (orthogonal), Parsimax (oblique), Geomin (oblique), Promax (oblique) and Facparsim (oblique). The selection of these particular rotations was made based upon a combination of prior research results, popularity in use and availability in statistical software. Again, though prior research comparing performance of rotational criteria used continuous indicators, these are the only available studies examining this issue; therefore, it was determined that these earlier studies did provide some insights into which rotations should be used. Sass and Schmitt (2010) used only oblique rotations, including Quartimin, oblique CF-Equamax, CF-Facparsim and Geomin, and found that Geomin and Quartimin performed slightly better in a pure simple structure condition (Complexity condition 1 in the current study), whereas oblique CF-Equamax and CF-Facparsim were somewhat better in the more complex cases. Asparouhov and Muthén (2009) compared Quartimin with Geomin using two values of the constant ε, 0.01 and 0.0001 and reported that Geomin with $\varepsilon = 0.001$ consistently produced the least bias in factor loading estimates. Based on these results, the current study included Geomin with $\varepsilon = 0.001$, Quartimin, and Facparsim. In addition, three orthogonal rotations (i.e., Varimax, Quartimax and Equamax) were included because heretofore their performance has not been investigated in such a study and they are very commonly used in practice. Similarly, Promax was included in the study because of its popularity and ubiquity in statistical software, and the fact that it was not included in the earlier work. For each included rotation criterion, except for Geomin as noted above, the default settings in MPlus were used in conducting the analyses in order to mimic what researchers are likely to do in practice.

In addition to the Monte Carlo simulation, this study also included the use of EFA with item responses from a sample of 1,000 examinees who took the Law School Admissions Test (LSAT). These data, which have been discussed previously in the literature, have been shown to contain 4 separate factors corresponding to the 4 reading passages contained in the exam (Stout, et al., 1996). For these data, EFA using the RWLS method of extraction was followed with each of the rotations included in the simulation study. Note that analysis was conducted on the raw binary data.

Results

Because an initial examination of the simulation outcomes revealed that the results for factors 1, 2, 3 and 4 were similar in terms of the grouping of indicators and the ordering of indicators by factor loading magnitude, results are presented for the first factor only. Similarly, estimates of the inter-factor correlation between factors 1 and 2 were similar to those for the other factor pairs (where applicable), thus, only the results for this correlation will be presented.

Factor Grouping

A repeated measures Analysis of Variance (ANOVA) was used to identify which of the manipulated conditions and their interactions were significantly associated with the proportion of item pairs correctly grouped together, which served as the dependent

variable. These conditions included type of rotation, number of observed indicators per factor, number of factors, factor complexity, sample size and inter-factor correlation. Assumptions of equality of variance and normality of errors were assessed using Levene's test and QQ plots, respectively, and were found to have been met. The results of the ANOVA indicated that the highest order significant ($\alpha = 0.05$)interaction was type of rotation by number of factors, number of indicators and factor complexity ($\eta^2 = 0.112$). In addition, the interaction among type of rotation, inter-factor correlation and factor complexity was also significantly related to the proportion of indicators correctly grouped ($\eta^2 = 0.482$), as was the main effect of sample size ($\eta^2 = 0.801$). All other significant main effects and interactions were subsumed in one of these three terms and will therefore not be discussed.

Table 3 shows the proportion of observed indicator variables correctly and incorrectly grouped by the number of factors, number of indicators per factor, factor complexity and method of rotation. An examination of these results reveals that across methods of rotation the proportion of variables correctly grouped declined as the factor structure became more complex, but the proportion incorrectly grouped together increased. (Note that the numbers for complexity conditions presented in subsequent tables correspond to the numbers in Table 2). This decrease in indicator grouping accuracy with increased structural complexity was less marked for the Quartimin (QMIN) rotation across the number of factors and number of indicators, and the Facparsim (FAC) when there were 12 indicators per factor, regardless of the number of factors. Indeed, when there were 12 indicators per factor the decline in grouping accuracy for QMIN was very small, 0.04 for 2 factors and 0.02 for 4 factors. By contrast, QMIN also demonstrated a much higher rate of incorrectly grouping indicator variables together for more complex factor patterns, across numbers of factors and indicators. The other rotations generally demonstrated comparable levels of grouping accuracy across the conditions contained in Table 3. The only exceptions to this general result were for Varimax (VAR) and Parsimax (PAR) with 4 factors, both of which had somewhat larger declines in the proportion of correctly grouped indicators than the other approaches in the presence of 4 factors, and for Equamax (EQU), which consistently demonstrated among the lowest rates of incorrectly grouping indicators together, and comparable rates of correctly grouping indicators with one another.

Table 4 presents the proportions of correctly and incorrectly grouped indicators by method of rotation, inter-factor correlation and factor complexity. As evident in Table 3, with increasing model complexity QMIN displayed a smaller decline in the proportion of correctly grouped indicators and a greater increase in the proportion of incorrectly indicators, than did the other rotation methods. Of particular interest is that two of the orthogonal rotations, VAR and EQU, did not show any greater diminution in the proportion of correctly grouped indicators than the oblique rotations as the inter-factor correlations increased, nor did they have greater increases in the proportion of incorrectly grouped items. By contrast, the orthogonal method QUA exhibited among the highest rates of incorrectly grouped indicators for the more complex factor patterns when the inter-factor correlation was 0.5 or 0.7. EQU and PAR consistently demonstrated among the lowest rates of incorrect indicator grouping, while being comparable to the other rotational methods (except QMIN) in terms of correctly grouped indicator variables.

The impact of the factor pattern on correct indicator grouping was essentially the same regardless of the inter-factor correlation, with decreases in the proportion of correctly grouped item pairs and increases in the proportion of correctly grouped item pairs. For all methods of rotation, the proportion of correctly grouped indicator variables increased concomitantly with increases in sample size, whereas the proportion of incorrectly grouped indicators declined (see Table 5).

Factor Loading Magnitudes

As with the proportion of correctly grouped items, repeated measures ANOVA was used to determine which of the study conditions

Table 3: Proportion of Variables Correctly | Incorrectly Grouped into Factors by Number of Factors (F), Number of Indicators per Factor (I) and Population Factor Complexity (C)

F	I	C	EQU*	GEO	PAR	PRO	QUA	QMIN	VAR	FAC
2	6	1	.94|.10	.94|.11	.94|.10	.91|.10	.94|.11	.93|.14	.94|.10	.88|.10
		2	.91|.16	.90|.18	.91|.16	.87|.17	.90|.18	.91|.31	.91|.16	.79|.17
		3	.86|.27	.85|.31	.86|.28	.82|.29	.85|.33	.88|.57	.85|.27	.69|.26
		4	.78|.45	.77|.51	.77|.45	.74|.48	.78|.56	.87|.83	.75|.46	.66|.49
2	12	1	.97|.02	.97|.03	.97|.02	.95|.02	.97|.03	.96|.12	.97|.02	.99|.03
		2	.95|.04	.95|.05	.95|.04	.93|.05	.95|.05	.96|.32	.95|.04	.98|.06
		3	.89|.10	.88|.11	.89|.09	.86|.10	.89|.22	.92|.66	.88|.10	.98|.13
		4	.80|.24	.80|.28	.80|.21	.77|.23	.83|.48	.92|.96	.78|.22	.95|.29
4	6	1	.92|.13	.91|.14	.91|.13	.90|.14	.91|.14	.91|.21	.90|.13	.82|.15
		2	.90|.17	.89|.17	.89|.17	.87|.16	.89|.18	.90|.30	.88|.16	.73|.19
		3	.86|.25	.86|.26	.85|.25	.83|.24	.86|.27	.90|.43	.83|.25	.63|.26
		4	.82|.38	.82|.41	.79|.38	.82|.41	.85|.42	.90|.59	.73|.42	.51|.43
4	12	1	.96|.05	.95|.05	.95|.05	.95|.15	.95|.05	.95|.07	.95|.14	.99|.06
		2	.94|.06	.94|.07	.94|.06	.94|.19	.94|.06	.95|.13	.94|.18	.96|.08
		3	.89|.13	.92|.18	.88|.12	.92|.32	.90|.16	.94|.28	.93|.31	.95|.18
		4	.82|.22	.88|.31	.79|.20	.88|.45	.85|.28	.93|.41	.83|.45	.93|.31

*EQU = Equamax, GEO = Geomin, PAR = Parsimax, PRO = Promax, QUA = Quartimax, QMIN = Quartimin, VAR = Varimax, FAC = Facparsim.

Table 4: Proportion of Variables Correctly | Incorrectly Grouped into Factors by Inter-Factor Correlations (r) and Population Factor Complexity (C)

r	C	EQU*	GEO	PAR	PRO	QUA	QMIN	VAR	FAC
0.1	1	.97|.04	.97|.04	.97|.04	.95|.08	.97|.04	.97|.04	.97|.07	.98|.03
	2	.96|.05	.95|.05	.95|.05	.94|.08	.95|.05	.95|.05	.95|.07	.95|.05
	3	.93|.10	.92|.11	.92|.10	.91|.13	.92|.10	.92|.11	.92|.12	.92|.11
	4	.85|.21	.84|.24	.84|.21	.83|.30	.84|.23	.91|.60	.85|.30	.80|.26
0.3	1	.96|.05	.96|.05	.96|.05	.95|.07	.96|.05	.96|.05	.96|.07	.96|.05
	2	.94|.07	.94|.08	.94|.08	.92|.10	.94|.08	.94|.08	.94|.09	.92|.08
	3	.91|.13	.90|.15	.90|.14	.88|.18	.89|.15	.93|.40	.89|.17	.86|.16
	4	.83|.26	.84|.31	.84|.26	.82|.36	.82|.30	.91|.68	.86|.36	.84|.37
0.5	1	.95|.08	.94|.08	.94|.08	.93|.10	.94|.08	.94|.08	.94|.09	.95|.08
	2	.94|.10	.93|11	.93|.10	.91|.13	.93|.11	.92|.22	.93|.12	.91|.10
	3	.88|.19	.87|.23	.87|.20	.85|.27	.87|.22	.93|.66	.87|.27	.86|.24
	4	.80|.34	.85|.41	.78|.33	.83|.42	.85|.55	.94|.73	.84|.41	.80|.48
0.7	1	.91|.14	.90|.16	.90|.15	.88|.17	.90|.16	.94|.38	.89|.16	.87|.16
	2	.87|.21	.87|.24	.87|.21	.84|.27	.87|23	.94|.75	.86|.26	.85|.24
	3	.82|.32	.85|.38	.81|.31	.82|.38	.85|.52	.92|.80	.83|.37	.79|.51
	4	.76|.49	.75|.55	.74|.43	.74|.49	.77|.69	.88|.82	.74|.48	.73|.50

*EQU = Equamax, GEO = Geomin, PAR = Parsimax, PRO = Promax, QUA = Quartimax, QMIN = Quartimin, VAR = Varimax, FAC = Facparsim.

Table 5: Proportion of Variables Correctly | Incorrectly Grouped into Factors by Sample Size

N	EQU*	GEO	PAR	PRO	QUA	QMIN	VAR	FAC
100	.83|.31	.85|.33	.82|.30	.82|34	.84|.33	.88|.45	.84|.33	.84|.32
200	.87|.21	.88|.24	.86|.21	.86|.25	.88|.24	.92|.40	.88|.24	.86|.23
500	.92|.11	.92|.14	.91|.11	.90|.17	.93|.17	.95|.37	.92|.16	.89|.13
1000	.94|.07	.94|.09	.94|.07	.93|.13	.94|.14	.97|.37	.94|.13	.94|.12

*EQU = Equamax, GEO = Geomin, PAR = Parsimax, PRO = Promax, QUA = Quartimax, QMIN = Quartimin, VAR = Varimax, FAC = Facparsim.

and their interactions were significantly related to the proportion of correctly ordered factor indicators based on their loading magnitudes in the sample. The highest order significant interaction was the rotation by inter-factor correlation by factor pattern ($\eta^2 = 0.201$). In addition, the 2-way interactions of rotation by number of indicators per factor ($\eta^2 = 0.236$) and rotation by number of factors ($\eta^2 = 0.275$) were also statistically significant, as was the main effect of sample size ($\eta^2 = 0.858$).

For all of the rotations, results demonstrate (see Table 6) that the proportion of correctly ordered factor indicators by loading magnitude declines with increases in the inter-factor correlation and with increased factor complexity (reflected through higher numbers for the factor complexity condition). In addition, the deleterious impact of greater factor complexity was more pronounced for larger values of the inter-factor correlation. For example, in the simple structure condition (C = 1) with correlations of 0.1 and 0.3, the rotations performed similarly with respect to correct ordering of the factor indicators by loading magnitude, whereas for r = 0.5 FAC displayed a higher proportion of correctly ordered factor loadings, and for r = 0.7, FAC, QMIN and VAR all had somewhat higher proportions of correctly ordered loadings. On the other hand, for the greatest factor complexity (C = 4) VAR consistently had the highest proportion of correctly ordered loadings, with a variety of other rotations performing comparably for a given inter-factor correlation. For example, QMIN performed similarly to VAR in the most complex case for inter-factor correlations of 0.1, 0.3 and 0.7, and FAC had similar values to VAR for proportion of correctly ordered loadings in the most complex case when r = 0.3.

Results in Table 7 show that all of the rotations were more accurate in terms of correctly ordering indicators by the magnitude of factor loadings for 12 indicators, for 2 factors and for larger sample sizes. FAC was the rotation method whose performance was most strongly influenced by the number of indicators. For 6 indicators per factor, it performed the worst in terms of correctly ordering loadings, whereas for 12 indicators it performed the best. QMIN and VAR consistently produced among the most accurate ordering of loadings by magnitude across all of the conditions contained in Table 7. The performances of the other rotation methods were generally similar to one another, and somewhat worse than that of QMIN and VAR.

Inter-Factor Correlation Bias

A repeated measures ANOVA identified the 3-way interaction of rotation method by

Table 6: Proportion of Factor Loadings Correctly Ordered by Magnitude by Inter-Factor Correlations (r) and Population Factor Complexity (C)

r	C	EQU*	GEO	PAR	PRO	QUA	QMIN	VAR	FAC
0.1	1	.94	.94	.94	.93	.94	.94	.94	.96
	2	.93	.92	.93	.91	.93	.93	.93	.94
	3	.90	.89	.89	.88	.91	.93	.90	.92
	4	.83	.81	.81	.81	.81	.84	.84	.81
0.3	1	.93	.92	.93	.91	.92	.93	.93	.94
	2	.91	.90	.90	.89	.90	.91	.91	.91
	3	.87	.85	.85	.84	.85	.87	.87	.84
	4	.78	.76	.75	.77	.76	.81	.81	.82
0.5	1	.90	.89	.89	.88	.89	.90	.90	.95
	2	.89	.87	.88	.86	.87	.89	.90	.92
	3	.81	.79	.79	.79	.78	.84	.83	.81
	4	.73	.70	.68	.68	.73	.70	.77	.70
0.7	1	.83	.81	.81	.81	.80	.84	.85	.84
	2	.79	.75	.75	.77	.75	.82	.81	.80
	3	.72	.69	.67	.71	.70	.76	.76	.69
	4	.65	.63	.57	.65	.64	.70	.70	.64

*EQU = Equamax, GEO = Geomin, PAR = Parsimax, PRO = Promax, QUA = Quartimax, QMIN = Quartimin, VAR = Varimax, FAC = Facparsim.

Table 7: Proportion of Factor Loadings Correctly Ordered by Magnitude by Number of Indicators per Factor (I), Number of Factors (F), and Sample Size

I	EQU*	GEO	PAR	PRO	QUA	QMIN	VAR	FAC
6	.75	.72	.72	.74	.72	.76	.77	.66
12	.93	.92	.91	.90	.92	.94	.93	.97

F	EQU	GEO	PAR	PRO	QUA	QMIN	VAR	FAC
2	.89	.86	.87	.86	.88	.91	.90	.86
4	.78	.77	.76	.79	.76	.79	.80	.78

N	EQU	GEO	PAR	PRO	QUA	QMIN	VAR	FAC
100	.69	.67	.66	.67	.67	.71	.70	.68
200	.80	.78	.77	.78	.78	.82	.82	.79
500	.91	.89	.89	.90	.89	.94	.93	.90
1000	.95	.94	.94	.94	.94	.96	.96	.92

*EQU = Equamax, GEO = Geomin, PAR = Parsimax, PRO = Promax, QUA = Quartimax, QMIN = Quartimin, VAR = Varimax, FAC = Facparsim.

inter-factor correlation by factor complexity ($\eta^2 = 0.049$) as the highest order significant term. In addition, the main effects of number of factors ($\eta^2 = 0.313$), number of indicators per factor ($\eta^2 = 0.041$), and sample size ($\eta^2 = 0.021$) were also statistically significant. Table 8 contains the mean raw bias and the standardized bias values across replications by the inter-factor correlation and the degree of model complexity. For $r = 0.1$, the sample correlation estimates displayed a positive bias across rotations, except for the simple structure condition (C = 1). In addition, as the degree of complexity increased, so did both raw and standardized bias, except for PRO. When $r = 0.3$, the negative bias in the simple structure condition was greater than for $r = 0.1$, and the positive bias for more complex models was lower, across rotation methods. For $r = 0.5$ and 0.7, bias was uniformly negative across levels of factor complexity, with greater negative bias associated with the largest population correlation. In addition, for $r = 0.5$ all rotation methods, except PAR, displayed greater negative bias for simple structure data (C = 1) or for the most complex structure (C = 4). In contrast, when $r = 0.7$, bias was generally higher for simple structure than for the next level of factor complexity (C = 2), after which bias increased concomitantly with increased model complexity. None of the rotation criteria consistently produced the least raw or standardized biased estimates.

Table 9 shows that inter-factor correlation bias was more pronounced (and negative) when more indicators were present. In addition, the degree of bias for most of the rotation methods was slightly greater (and negative) for 4 factors as compared to 2, where the bias was positive. Finally, bias in the inter-factor correlation estimates declined with increased sample size, and across all conditions PAR produced somewhat more negatively biased estimates than the other criteria. Otherwise, differences in estimation accuracy across the conditions were relatively minor.

Analysis of LSAT Data

In order to demonstrate the relative performance of the rotation criteria on an actual, well studied data set, EFA was run on the LSAT data described in Stout, et al. (1996). Given that these authors, and others, reported the presence of 4 stable dimensions, 4 factors were extracted in this analysis, and each rotation was applied. Table 10 contains the factor loadings only for the primary factor for each item in order to save space. There were no cross-loaded items for any of the rotation criteria, defined as having multiple factors for which the loading values were great than 0.32 (Tabachnick & Fidell, 2007). A perusal of these results demonstrates that across items and factors, the loading values for the 8 different rotations were very similar to one another. There is no discernible pattern of difference in loadings by rotation, suggesting that a researcher using any of these criteria would reach the same substantive conclusions regarding both how items grouped together, and the strength of relationships between items and factors.

Table 11 includes the correlation estimates for the 4 factor solution of the LSAT data for each of the oblique rotations studied here, and their standard errors with the exception of PROMAX, for which standard errors are not calculated in MPlus. These results demonstrate a greater degree of variation across rotation criteria than was evident for the factor loadings. For example, PROMAX had much larger inter-factor correlation estimates than the other methods for factor 1 with 3, 1 with 4 and 3 with 4. By contrast, PARSIMAX had much lower correlation estimates than the other methods for factors 1 with 3, 1 with 4, 2 with 4 and 3 with 4. GEOMIN, QUARTIMIN and FACPARSIM had very similar inter-factor correlation estimates to one another for this sample.

Conclusion

This study extends previous research comparing rotations in EFA, which focused on continuous factor indicator variables by comparing the performance of 8 factor rotation criteria with dichotomous indicator variables using the WLSMV initial extraction method in MPlus across a variety of conditions. Among the rotations included were some that had previously been found to be promising in terms of accuracy of factor loading estimates such as Geomin and Facparsim, and others that had not been studied before but which are very commonly used in practice, including Varimax and Promax. The outcomes of interest included

Table 8: Inter-Correlation Bias (Standardized Bias) by Inter-Factor Correlations (r) and Population Factor Complexity (C)

r	C	GEO*	PAR	PRO	QMIN	FAC
0.1	1	-0.04 (-0.43)	-0.05 (-0.75)	-0.02 (-0.23)	-0.04 (-0.44)	-0.03 (-0.33)
	2	0.08 (0.45)	0.05 (0.34)	0.12 (0.69)	0.08 (0.49)	0.09 (0.50)
	3	0.18 (0.65)	0.14 (0.55)	0.22 (0.84)	0.19 (0.73	0.19 (0.73)
	4	0.24 (0.73)	0.21 (0.73)	0.17 (0.52)	0.25 (0.80)	0.26 (0.79)
0.3	1	-0.12 (-0.84)	-0.16 (-0.93)	-0.10 (-0.71)	-0.11 (-0.79)	-0.11 (-0.78)
	2	-0.01 (-0.04)	-0.07 (-0.38)	0.03 (0.13)	-0.01 (-0.02)	0.01 (0.01)
	3	0.07 (0.22)	0.01 (0.05)	0.08 (0.30)	0.08 (0.27)	0.09 (0.29)
	4	0.09 (0.25)	0.07 (0.23)	-0.02 (-0.07)	0.12 (0.35)	0.11 (0.32)
0.5	1	-0.21 (-0.95)	-0.27 (-1.53)	-0.18 (-0.94)	-0.20 (-0.92)	-0.21 (-0.92)
	2	-0.09 (-0.34)	-0.17 (-0.77)	-0.08 (-0.31)	-0.09 (-0.32)	-0.09 (-0.33)
	3	-0.08 (-0.22)	-0.12 (-0.46)	-0.14 (-0.43)	-0.06 (-0.18)	-0.08 (-0.19)
	4	-0.13 (-0.35)	-0.09 (-0.29)	-0.20 (-0.58)	-0.19 (-0.60)	-0.20 (-0.59)
0.7	1	-0.31 (-1.00)	-0.37 (-1.65)	-0.31 (-1.20)	-0.30 (-1.07)	-0.32 (-1.06)
	2	-0.26 (-0.78)	-0.31 (-1.15)	-0.32 (-1.00)	-0.25 (-0.76)	-0.27 (-0.79)
	3	-0.30 (-0.80)	-0.28 (-0.80)	-0.35 (-1.06)	-0.36 (-1.06)	-0.36 (-1.08)
	4	-0.38 (-0.99)	-0.31 (-0.81)	-0.33 (-1.05)	-0.33 (-1.54)	-0.36 (-1.44)

*GEO = Geomin, PAR = Parsimax, PRO = Promax, QMIN = Quartimin, FAC = Facparsim.

Table 9: Inter-Correlation Bias by Magnitude by Number of Indicators Per Factor (I), Number of Factors (F), and Sample Size

I	GEO*	PAR	PRO	QMIN	FAC
6	0.03	-0.03	0.02	0.04	0.06
12	-0.13	-0.15	-0.15	-0.14	-0.13

F	GEO	PAR	PRO	QMIN	FAC
2	0.17	0.10	0.16	0.11	0.12
4	-0.16	-0.17	-0.17	-0.14	-0.15

N	GEO	PAR	PRO	QMIN	FAC
100	-0.11	-0.12	-0.08	-0.10	-0.11
200	-0.10	-0.12	-0.08	-0.09	-0.10
500	-0.06	-0.09	-0.08	-0.06	-0.08
1000	-0.03	-0.08	-0.08	-0.04	-0.06

*GEO = Geomin, PAR = Parsimax, PRO = Promax, QMIN = Quartimin, FAC = Facparsim.

the proportion of accurately grouped indicator variables, the proportion of indicators correctly ordered by the magnitude of their loading values and, for the oblique methods, the accuracy of inter-factor correlation estimates. It is hoped that this study builds upon earlier work by focusing on dichotomous indicators (i.e., items), by including outcomes that would be of interest to practitioners interested in using these methods to identify potential latent variables in existing measures and by expanding the range of conditions under which the rotations are examined, including the rotations themselves.

Implications for Practice

One implication of this study for researchers using EFA with categorical indicator variables is that when they know, or suspect, that the correlations among the factors will be upwards of 0.5, they should expect to have problems not only with appropriately grouping variables together, but also with accurately ordering variables in terms of the importance of their relationships with the factors. These problems are likely to be particularly acute if the factor pattern structure is very complex. It does seem however, that having a larger sample may ameliorate these problems to some extent, so that when it is likely the factors will be highly correlated and/or the factor pattern may be complex in nature, researchers should ideally try to obtain samples of 500 or more. These results are similar to those reported in de Winter, Dodou and Wieringa (2009) for continuous data.

A second implication is that - for the oblique methods of rotation studied - there may be problems with accurately estimating inter-factor correlations across conditions like those simulated here. When these correlations were greater than 0.3, all of the criteria produced underestimates of *r*, whereas for lower correlations *r* was overestimated for more complex factor patterns and underestimated for the less complex patterns. These correlation estimation bias results are similar to those

Table 10: Rotated Factor Loading Matrices for LSAT Data

Item	EQU*	GEO	PAR	PRO	QUA	QMIN	VAR	FAC
Factor 1								
1	0.35	0.33	0.33	0.33	0.34	0.33	0.35	0.32
2	0.40	0.40	0.40	0.41	0.40	0.40	0.40	0.39
3	0.43	.045	0.45	0.47	0.43	0.45	0.43	0.45
4	0.36	0.39	0.38	0.40	0.36	0.38	0.36	0.38
5	0.40	0.38	0.38	0.39	0.39	0.38	0.39	0.38
6	0.51	0.53	0.52	0.55	0.51	0.53	0.51	0.53
7	0.33	0.30	0.31	0.30	0.31	0.30	0.33	0.30
Factor 2								
8	0.52	0.54	0.54	0.56	0.51	0.54	0.52	0.53
9	0.38	0.40	0.40	0.41	0.37	0.39	0.38	0.39
10	0.52	0.55	0.55	0.57	0.51	0.55	0.53	0.54
11	0.28	0.27	0.28	0.28	0.27	0.27	0.28	0.27
12	0.37	0.40	0.39	0.42	0.37	0.40	0.37	0.39
13	0.38	0.37	0.37	0.39	0.38	0.38	0.37	0.38
Factor 3								
14	0.54	0.55	0.54	0.58	0.54	0.56	0.54	0.55
15	0.53	0.54	0.53	0.56	0.53	0.54	0.53	0.54
16	0.44	0.46	0.45	0.48	0.44	0.46	0.44	0.46
17	0.16	0.15	0.15	0.15	0.16	0.15	0.16	0.15
18	0.48	0.48	0.45	0.49	0.49	0.49	0.47	0.49
19	0.51	0.50	0.47	0.51	0.52	0.51	0.50	0.51
Factor 4								
20	0.42	0.41	0.38	0.41	0.43	0.41	0.42	0.41
21	0.56	0.56	0.53	0.57	0.57	0.57	0.55	0.56
22	0.59	0.60	0.56	0.61	0.60	0.60	0.58	0.60
23	0.47	0.48	0.45	0.49	0.48	0.48	0.47	0.48
24	0.50	0.52	0.49	0.53	0.50	0.52	0.50	0.52

*EQU = Equamax, GEO = Geomin, PAR = Parsimax, PRO = Promax, QUA = Quartimax, QMIN = Quartimin, VAR = Varimax, FAC = Facparsim.

Table 11: Inter-Factor Correlation (Standard Error) Estimates for LSAT Data by Oblique Rotations

Factor Pair	GEO*	PAR	PRO	QMIN	FAC
1 with 2	0.35 (0.05)	0.30 (0.04)	0.32 (NA)	0.34 (0.06)	0.34 (0.06)
1 with 3	0.28 (0.05)	0.20 (0.04)	0.42 (NA)	0.28 (0.05)	0.29 (0.05)
1 with 4	0.26 (0.05)	0.18 (0.04)	0.35 (NA)	0.26 (0.05)	0.26 (0.06)
2 with 3	0.32 (0.05)	0.35 (0.04)	0.36 (NA)	0.33 (0.05)	0.31 (0.05)
2 with 4	0.42 (0.05)	0.23 (0.04)	0.38 (NA)	0.42 (0.05)	0.42 (0.05)
3 with 4	0.30 (0.04)	0.20 (0.03)	0.50 (NA)	0.32 (0.04)	0.33 (0.05)

*EQU = Equamax, GEO = Geomin, PAR = Parsimax, PRO = Promax, QMIN = Quartimin, FAC = Facparsim.

reported by Sass and Schmitt (2010) for the case of continuous indicators.

A third implication for practitioners is that including more indicator variables (assuming that they are of good quality) will yield better solutions both in terms of correctly grouping the indicators and accurately ordering them in terms of their relationships to the factors. This result seems reasonable given that including more indicators for each factor provides a greater amount of information for the EFA extraction algorithm as well as for the rotations. The number of indicators was particularly important for the FAC technique, particularly in the case of a more complex factor pattern structures with more factors. Based on these results, researchers may consider using FAC when they have at least 12 indicators per factor, as it demonstrated better performance in terms of grouping the variables as well as ordering them, particularly in the 4 factor case. On the other hand, FAC would not appear to be optimal with fewer indicators per factor.

A final implication of these results is that, in terms of both indicator grouping and ordering of importance in terms of factor relationships, researchers may generally find orthogonal and oblique rotations will produce similar results. Indeed, one of the consistently best performers in this study was the orthogonal rotation EQU. This result is not completely surprising, as EQU was designed to spread loading variation more equally across factors than several of the other rotations studied here (Saunders, 1962) by combining the VAR and QUA criteria. Thus, although VAR seeks to maximize the variation of loadings for factors, and QUA seeks to simplify loadings for the observed variables, EQU combines these two goals. This is not to suggest that researchers should only use EQU as the rotation of choice for all problems. When factors are thought to be correlated, the choice of an orthogonal rotation may not be appropriate, regardless of how well it performs. However, when the inter-factor correlation is low and the primary goal of a study is to identify which indicators are associated with which factors, EQU would be a reasonable choice.

When a researcher is interested in estimating inter-factor correlations, or they believe that these correlations may be fairly large (greater than 0.5), several of the oblique rotations studied here would appear to be appropriate. In particular, PAR and FAC (for situations with a larger number of indicator

variables) demonstrated consistently strong performance in terms of correctly grouping and ordering indicator variables. On the other hand, QMIN may not be reliable for researchers interested in finding the correct groupings of factor indicators, as it (or the equivalent methods of oblique Quartimax and Oblimin) appears to reduce dimensionality in the sample too much by grouping most of the variables into a single factor. As a consequence, researchers using QMIN may come to the conclusion that, based on the sample there are a smaller number of factors present than is actually true for the population.

Limitations

As with any research effort, limitations to this study that must be considered when interpreting the results. First, for all of the rotations the MPlus system defaults were used. This was a decision made for two reasons: (1) It was desired to mimic what might be most commonly done in practices, and (2) In many cases there are a very large number of alternative settings that could have been used for some of the rotations. Therefore, in order to keep the study to a manageable size and the interpretation of the results fairly straightforward, it was felt that only a limited number of options could be used. Nonetheless, in practice researchers can choose from a broader range of settings when using many of these rotational criteria.

A second limitation relates to the conditions simulated, including the factor patterns used and the number of indicators. In both cases, the selections made for this study were designed to mimic what would be seen in practice. However, clearly many other factor patterns and numbers of indicators could have been included, which may well have provided different results. Future studies should focus on both of these issues in order to expand upon what was learned here.

Finally, these results were based on dichotomous indicator variables, which may not translate directly to ordinal data, such as that commonly found in many psychological scales. It should be noted that because rotations focus on loadings rather than the raw data, it is not clear how important this issue might be. Nonetheless, future research should verify to what extent the nature of the categorical data has an impact on the performance of rotational criteria.

Summary

In the final analysis, the admonition offered by Browne (2001) for researchers to use their expert judgment in conjunction with statistical results is definitely supported by these results. It is clearly not possible to state that any single rotational criterion will fit all EFA problems adequately, although in practice researchers often appear to use favorites regardless of the context. However, these results do suggest that certain features of the data will support the use of one or more such methods studied here. Clearly the ubiquitous VAR and PRO rotations must be used with caution when at all, as often they do not produce optimal results in terms of accurately reflecting the underlying factor structure. With the increased availability of other rotations in software packages such as MPlus, researchers are no longer limited to a small number of available options, and can thus experiment with a broader array of tools than could be done previously.

References

Asparouhov, T., & Muthén, B. (2009). Exploratory structural equation modeling. *Structural Equation Modeling*, *16*, 397-438.

Bock, R. D. & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*(*4*), 443-459.

Bock, R. D., Gibbons, R., Schilling, S. G., Muraki, E., Wilson, D. T. & Wood, R. (2003). *TESTFACT 4*. Lincolnwood, IL: Scientific Software International.

Browne, M. W. (2001). An overview of analytic rotations in exploratory factor analysis. *Multivariate Behavioral Research*, *36*(*1*), 111-150.

Carroll, J. B. (1957). Biquartimin criterion for rotation to oblique simple structure in factor analysis. *Science*, *126*, 1114-1115.

Christofferson, A. (1975). Factor analysis of dichotomized variables. *Psychometrika*, *40*, 5-32.

Crawford, C. B., & Ferguson, G. A. (1970). A general rotation criterion and its use in orthogonal rotation. *Psychometrika*, *35*, 321-332.

de Winter, J. C. F., Dodou, D., & Wieringa, P. A. (2009). Exploratory factor analysis with small sample sizes. *Multivariate Behavioral Research*, *44*, 147-181.

Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods*, *9*, 466-491.

Fraser, C., & McDonald, R. P. (1988). NOHARM: Least squares item factor analysis. *Multivariate Behavioral Research*, *23*, 267-269.

French, B. F. & Finch, W. H. (2006). Confirmatory factor analytic procedures for the determination of measurement invariance. *Structural Equation Modeling*, *13*(*3*), 378-402.

Gagné, P., & Hancock, G. R. (2006). Measurement model quality, sample size, and solution propriety in confirmation factor models. *Multivariate Behavioral Research*, *41*, 65-83.

Gorsuch, R. L. (1983). *Factor analysis*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Gosz, J. K., & Walker, C. M. (2002). *An empirical comparison of multidimensional item response data using TESTFACT and NOHARM*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Guilford, J. P. (1941). The difficulty of a test and its factor composition. *Psychometrika*, *6*, 67-78.

Hattie, J. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement*, *9*(*2*), 139-164.

Jennrich, R. I. (2007). Rotation methods, algorithms, and standard errors. In R. Cudek & R. C. MacCallum (Eds.), *Factor analysis at 100: Historical developments and future directions*, 315-335. Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Kaiser, H. F. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, *23*(*3*), 187-200.

Lorenzo-Seva, U. (1999). Promin: A method for oblique factor rotation. *Multivariate Behavioral Research*, *34*, 347-365.

MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods*, *4*(*1*), 84-99.

McDonald, R. P. (1962). A general approach to nonlinear factor analysis. *Psychometrika*, *27*, 297-415.

McDonald, R. P. (1967). Nonlinear factor analysis. *Psychometric Monographs*, *15*.

McDonald, R. P. (1981). The dimensinality of tests and items. *British Journal of Mathematical and Statistical Psychology*, *34*, 100-117.

McDonald, R. P. (1997). Normal-ogive multidimensional model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of Modern Item Response Theory*, 257-269. New York: Springer.

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Lawrence Erlbaum Associates.

McLeod, L. D., Swygert, K. A., & Thissen, D. (2001). Factor analysis for items scored in two categories. In D. Thissen & H. Wainer (Eds.), *Test scoring*, 189-216. Mahwah, NJ: Lawrence Erlbaum Associates.

Meade, A. W., & Lautenschlager, G. J. (2004). A Monte-Carlo study of confirmatory factor analytic tests of measurement equivalence/invariance. *Structural Equation* Modeling, *11*, 60-72.

Mulaik, S. A. (2010). *Foundations of factor analysis*. Boca Raton, FL: Chapman & Hall/CRC.

Mundfrom, D. J., Shaw, D. G., & Ke, T. L. (2005). Minimum sample size recommendations for conducting factor analyses. *International Journal of Testing*, *5*, 159-168.

Muthén, B. O. (1978). Contributions to factor analysis of dichotomous variables. *Psychometrika*, *43*, 531-560.

Muthén, B. O. (1993). Goodness of fit with categorical and other non-normal variables. In K. A. Bollen, & J. S. Long (Eds.), *Testing structural equation models*, 205-243. Newbury Park, CA: Sage.

Muthén, B. O., du Toit, S. H. C., & Spisic, D. (1997). *Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes*. Unpublished manuscript.

Muthén, L. K., & Muthén, B. O. (2007). *MPlus, version 5: User's guide*. Los Angeles: Author.

Preacher, K. J. & MacCallum, R. C. (2002). Exploratory factor analysis in behavior genetics research: Factor recovery with small sample sizes. *Behavior Genetics*, *32*(*2*), 153-161.

Sass, D. A., & Schmitt, T. A. (2010). A comparative investigation of rotation criteria within exploratory factor analysis. *Multivariate Behavioral Research*, *45*, 73-103.

Saunders, D. R. (1962). *Trans-varimax: Some properties of the Ratiomax and Equamax criteria for blind orthogonal rotation*. Paper presented at the Meeting of the American Psychological Association, St. Louis, MO, September.

Spearman, C. (1929). *The abilities of man.* New York: Macmillan.

Stout, W., Habing, B., Douglas, J., Kim, H. R., Roussos, L., & Zhang, J. Conditional covariance-based nonparametric multidimensionality assessment. *Applied Psychological Measurement*, *19*(*4*), 331-354.

Tabachnick, B. G., & Fidell, L. S. (2007). *Using Multivariate Statistics*. Boston: Pearson.

Thompson, B. (2004). *Exploratory and confirmatory factor analysis: Understanding concepts and applications*. Washington, DC: American Psychological Association.

Thurstone, L. L. (1947). *Multiple factor analysis.* Chicago: University of Chicago press.

Trendafilov, N. T. (1994). A simple method for procrustean rotation in factor analysis using majorization theory. *Multivariate Behavioral Research*, *29*, 385-408.

Yates, A. (1987). *Multivariate exploratory data analysis: A perspective on exploratory factor analysis.* Albany: State University of New York Press.

Zumbo, B. D. (2007). Validity: Foundational issues and statistical methodology. In C. R. Rao (Series Ed.) & S. Sinharay (Volume Ed.), *Handbook of statistics: Vol 25 Psychometrics*, 45-80. Amsterdam: Elsevier.

Journal of Modern Applied Statistical Methods
November 2011, Vol. 10, No. 2, 571-582

1538 – 9472/11/$95.00

Discriminant Analysis for Repeated Measures Data: Effects of Mean and Covariance Misspecification on Bias and Error in Discriminant Function Coefficients

Tolulope T. Sajobi Lisa M. Lix Longhai Li William Laverty
University of Saskatchewan,
Saskatoon, Canada

Discriminant analysis (DA) procedures based on parsimonious mean and/or covariance structures have been proposed for repeated measures (RM) data. Bias and means square error of discriminant function coefficients (DFCs) for DA procedures are investigated when the mean and/or covariance structures are correctly specified and misspecified.

Key words: Multivariate, model misspecification, discriminant function coefficient, mean square error, bias.

Introduction

Linear discriminant analysis (DA) is a multivariate procedure, originally proposed by Fisher (1936), for predicting group membership (predictive discriminant analysis; PDA) and/or describing group separation (descriptive discriminant analysis; DDA) (Huberty & Olejnik, 2006) on multiple variables. The classical linear PDA procedure has been applied to repeated measures (RM) data (Feighner & Sverdlov, 2002; Levesque, Ducharme, Zarit, Lachance & Giroux, 2008), in which study participants are measured on a single variable at two or more occasions. Classical linear DA will not result in an efficient classification rule in multivariate or RM data when there is a large number of variables or measurement occasions relative to sample size. In recent years, a number of PDA procedures for RM data have been proposed (Marshall & Baron, 2000; Roy & Khatree, 2005a, 2005b, 2007; Tomasko, Helms & Snappin, 1999).

Roy and Khattree (2005a, 2005b) developed DA procedures based on parsimonious mean and covariance structures for both univariate (measurements on one outcome variable) and multivariate (measurements on two or more outcome variables) RM data to address the issue of classification efficiency when sample size is small. For univariate RM data, they proposed procedures based on constant RM mean vectors and either a compound symmetric (CS) or first-order autoregressive (AR-1) covariance. Though these procedures can result in efficient classification rules in high-dimensional data (Roy & Khatree, 2007), they may also result in inflated misclassification error rates (MERs) when the mean and/or covariance structure is/are incorrectly specified.

Although these procedures were originally developed for PDA, the discriminant function coefficients (DFCs) produced can be used for DDA, that is, to quantify the relative importance of the measurement occasions for discriminating among groups (Thomas, 1992). In classical linear DA, it is known that bias and error variation of DFCs is influenced by a

Tolulope T. Sajobi completed his Ph.D. in the School of Public Health. Email him at: tolusajobi920@gmail.com. Lisa M. Lix is an Associate Professor & Centennial Research Chair in the School of Public Health. Email her at: lisa.lix@usask.ca. Longhai Li is an Assistant Professor in Department of Mathematics and Statistics. Email him at: longhai@math.usask.ca. William Laverty is an Associate Professor in Department of Mathematics & Statistics. Email him at: laverty@math.usask.ca.

variety of data characteristics, including degree and pattern of separation between groups (group mean vectors) and magnitude of correlation among the outcome variables (Williams & Titus, 1998; Williams, Titus & Hines, 1991). However, to date, there has been little – if any – research, regarding the effects of misspecifying the mean and/or covariance structure on DDA procedures for RM data. Thus, the purpose of this study is to investigate the effects of RM mean and/or covariance misspecification on bias and error in DFCs of DDA procedures based on constant mean vectors and/or structured covariance matrices in univariate RM data.

Estimation of DFCs in DA Procedures for RM Data

Consider the case of $g = 2$ groups (which can be generalized to $g > 2$). In general, the number of uncorrelated DFC vectors is equal to $g - 1$. Let $\mathbf{y}_{ij}$ be the $p \times 1$ random vector of observed measurements for the i^{th} study participant ($i = 1, ..., n_j$; $N = n_1 + n_2$) in the j^{th} group ($j = 1, 2$). It is assumed that $\mathbf{y}_{ij} \sim N_p(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$, where $\boldsymbol{\mu}_j$ and $\boldsymbol{\Sigma}_j$ are the population mean vector and covariance for the j^{th} group and are estimated by $\hat{\boldsymbol{\mu}}_j$ and $\hat{\boldsymbol{\Sigma}}_j$, respectively. The linear DFC vector is estimated by

$$\hat{\mathbf{a}} = \hat{\boldsymbol{\Sigma}}^{-1}(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2). \tag{1}$$

For Fisher's (1936) linear DA procedure,

$$\hat{\boldsymbol{\Sigma}} = \frac{(n_1 - 1)\hat{\boldsymbol{\Sigma}}_1 + (n_2 - 1)\hat{\boldsymbol{\Sigma}}_2}{n_1 + n_2 - 2}, \tag{2}$$

and

$$\hat{\boldsymbol{\mu}}_j = \bar{\mathbf{y}}_j, \tag{3}$$

where

$$\bar{\mathbf{y}}_j = \frac{\sum_{i=1}^{n_j} \mathbf{y}_{ij}}{n_j}.$$

These quantities are estimated using the least-squares approach.

Roy and Khatree (2005a) proposed a DA procedure based on constant RM mean vectors and CS covariance structure. With a CS structure, $\boldsymbol{\Sigma}$ has diagonal elements σ^2 and off-diagonal elements $\sigma^2\rho$. For constant RM mean vectors, $\hat{\boldsymbol{\mu}}_j = c_j \mathbf{1}_p$, the maximum likelihood (ML) estimate of c_j is

$$\hat{c}_j = \frac{\mathbf{1}_p^{\mathrm{T}} \bar{\mathbf{y}}_j}{p}, \tag{4}$$

where $\mathbf{1}_p$ is a $p \times 1$ vector of ones, T is the transpose operator, and $\bar{\mathbf{y}}_j$ is the sample mean vector for the j^{th} group. The ML estimates of σ^2 and ρ can be obtained by simultaneously solving the following system of equations.

$$\begin{aligned} 0 = &-Np(1-\rho)(1+(p-1)\rho)\sigma^2 \\ &+(1+(p-1)\rho)(a_1+a_2) \\ &-\rho(b_1+b_2), \end{aligned} \tag{5}$$

and

$$\begin{aligned} 0 = &-N(p-1)p(1+(p-1)\rho)(1-\rho)\rho\sigma^2 \\ &-(a_1+a_2)(1+(p-1)\rho)^2 \\ &+(b_1+b_2)(\rho^2(p-1)+1), \end{aligned} \tag{6}$$

where $a_1 = \mathrm{tr}(\mathbf{W}_1)$, $a_2 = \mathrm{tr}(\mathbf{W}_2)$, $b_1 = \mathrm{tr}(\mathbf{J}\mathbf{W}_1)$, $b_2 = \mathrm{tr}(\mathbf{J}\mathbf{W}_2)$, $\mathbf{J} = \mathbf{1}_p\mathbf{1}_p^{\mathrm{T}}$,

$$\mathbf{W}_j = \sum_{i=1}^{n_j} (\mathbf{y}_{ij} - \bar{\mathbf{y}}_j)(\mathbf{y}_{ij} - \bar{\mathbf{y}}_j)^{\mathrm{T}}, \tag{7}$$

and tr is the trace operator. The DFCs are estimated by substituting the ML estimates of $\boldsymbol{\Sigma}$ and $\boldsymbol{\mu}_j$ in (1).

Roy and Khattree (2005a) proposed a DA procedure based on constant RM mean vectors and AR-1 covariance structure. With an AR-1 structure, Σ has diagonal elements σ^2, and off-diagonal elements $\sigma^2\rho^l$, where l is the number of lags between measurement occasions. Estimates of c_j, σ^2, and ρ are obtained by simultaneously solving

$$0 = (p-2)\rho c_j - pc_j + pm_{j1} - (p-2)\rho m_{j2}, \tag{8}$$

$$\begin{aligned} 0 = {} & Np\sigma^2(1-\rho^2) - (\beta_1\rho^2 - 2\gamma_1\rho + \alpha_1) \\ & + n_1c_1(\beta_2\rho^2 - 2\gamma_2\rho + \alpha_2) \\ & + n_2c_2(\beta_3\rho^2 - 2\gamma_3\rho + \alpha_3) \\ & - (n_1c_1^2 + n_2c_2^2)((p-2)\rho^2 \\ & - 2(p-1)\rho + p), \end{aligned} \tag{9}$$

and

$$\begin{aligned} 0 = {} & N(p-1)\sigma^2\rho - N(p-1)\sigma^2\rho^3 \\ & - \{\rho(\alpha_1 + \beta_1) - \gamma_1\rho^2 - \gamma_1\} \\ & + n_1c_1\{\rho(\alpha_2 + \beta_2) - \gamma_2\rho^2 - \gamma_2\} \\ & + n_2c_2\{\rho(\alpha_3 + \beta_3) - \gamma_3\rho^2 - \gamma_3\} \\ & - (n_1c_1^2 + n_2c_2^2)\{\rho(2p-2) \\ & - (p-1)\rho^2 - (p-1)\}. \end{aligned} \tag{10}$$

Details of these equations are provided in the Appendix. The estimates of the DFCs are obtained by substituting the ML estimates of $\boldsymbol{\Sigma}$ and $\boldsymbol{\mu}_j$ in (1).

For the DA procedure based on constant RM mean vectors and unstructured covariance, the ML estimate of $\boldsymbol{\mu}_j$ is as shown in equation 3 and $\boldsymbol{\Sigma}$ is estimated as

$$\hat{\boldsymbol{\Sigma}} = \frac{\sum_{j=1}^{2} \mathbf{W}_j}{N}, \tag{11}$$

where $\mathbf{W}_j$ is obtained from (7).

Methodology

The investigated procedures in the Monte Carlo study were: (a) DA procedure based on unstructured mean vectors and unstructured covariances (UN), (b) DA procedure based on constant mean vectors and unstructured covariances (STUN), (c) DA procedure based on constant mean vectors and CS covariances (STCS), and (d) DA based on constant mean vectors and AR-1 covariances (STAR).

The following conditions were manipulated in the study: (a) number of repeated measurements (p), (b) total sample size (N), (c) group sizes, (d) pattern and magnitude of correlation among the repeated measurements, and (e) RM mean vector configuration. The number of groups (g = 2) and the population distribution (normal) were fixed.

The number of RMs was set at p = 3, 5, 7 and 9. Previous studies have considered values of p ranging from 3 to 10 (Roy & Khattree, 2005a; 2005b; Williams & Titus, 1988). Total sample sizes of N = 60, 90 and 120 were investigated, giving an N/p ranging from 6.6 to 40.0.

Although previous simulation studies about DA procedures for RM data have primarily focused on equal group size conditions (Roy & Khattree, 2005a, 2005b), unequal group sizes have also been investigated for multivariate designs (Baron, 1991; He & Fung, 2000). Based on the research of Baron (1991) and Lei and Koehly (2003), the unequal group sizes selected for this study were (n_1, n_2) = (24, 36) for N = 60, (36, 54) for N = 90, and (48, 72) for N = 120.

The standard error of DFCs is known to be influenced by the magnitude of correlation among the variables (Thomas & Zumbo, 1996). Six population correlation structures were investigated: (1) $\mathbf{Q}_1$: CS structure with parameter ρ = 0.3, (2) $\mathbf{Q}_2$: CS structure with ρ = 0.7, (3) $\mathbf{Q}_3$: AR-1 structure with ρ = 0.3, (4) $\mathbf{Q}_4$: AR-1 structure with ρ = 0.7, (5) $\mathbf{Q}_5$: unstructured with average correlation amongst the off-diagonal elements of 0.3, and (6) $\mathbf{Q}_6$: unstructured with average correlation amongst the off-diagonal elements of 0.7.

Pseudorandom observation vectors $\mathbf{y}_{ij}$ were generated from a multivariate normal distribution with mean $\boldsymbol{\mu}_j$ and correlation matrix $\mathbf{Q}_{mj} = \mathbf{Q}_m$ (m = 1 ,…, 6). A vector of standard normal deviates, $\mathbf{C}_{ij}$, was transformed to a vector of multivariate observations via $\mathbf{y}_{ij} = \boldsymbol{\mu}_j + \mathbf{L}\mathbf{C}_{ij}^{\mathrm{T}}$. The Cholesky decomposition was used to obtain $\mathbf{L}$, an upper triangular matrix of dimension p satisfying the equality $\mathbf{L}^{\mathrm{T}}\mathbf{L} = \mathbf{Q}_{mj}$ and then $\mathbf{y}_{ij}$ was multiplied by $\mathbf{V}_j$, a diagonal matrix with elements σ_j to obtain multivariate observations with the desired

variances and covariances, such that $\mathbf{\Sigma}_j = \mathbf{V}_j \mathbf{Q}_{mj} \mathbf{V}_j^{\mathrm{T}}$. For all investigated conditions $\sigma_1^2 = \sigma_2^2 = 1$ was selected. The RANNOR function in SAS (SAS Institute Inc., 2008) was used to generate the standard normal deviates.

A variety of mean vector conditions have been investigated in previous research (Titus & Williams, 1988; Roy & Khattree, 2005a). In this study, three configurations for $\boldsymbol{\mu}_1$ were selected for each value of p (see Table 1); for all conditions, $\boldsymbol{\mu}_2$ was the null vector. Configuration I had constant means for all RM occasions in both groups. Configuration II had non-constant RM mean with a quadratic, cubic or polynomial pattern for the RM occasions in the first group and constant means in the second group. For configuration III, a monotonic decreasing linear pattern was specified for the means in the first group and the means in the second group were constant.

Overall, 1,493 combinations of simulation conditions were investigated with 5,000 replications for each combination. The study was conducted using SAS/IML software (SAS Institute Inc., 2008).

Two measures of performance were used to evaluate the DFCs, namely: mean square error (MSE) and norm of the average bias (Crouxe & Dehon, 2001). The norm of the average bias is

$$b = \| \frac{1}{M} \sum_{k=1}^{M} (\hat{\mathbf{a}}_k - \mathbf{a}) \|, \qquad (12)$$

and the MSE is

$$e = \frac{1}{M} \sum_{k=1}^{M} \| \hat{\mathbf{a}}_k - \mathbf{a} \|^2, \qquad (13)$$

where $\mathbf{a}$ is the population vector of DFCs, $\|\mathbf{x}\|$ is the norm of $\mathbf{x}$ and M is the number of replications (M = 5,000). Both measures take values on the interval [0, ∞) and the smaller the bias or error in the DFCs the better. To adjust for the confounding effect of degree of separation between the two group means on bias and error, the bias and MSE in the DFCs were standardized using the distance between the two group mean vectors. Therefore,

$$b_{st} = \frac{b}{\| \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 \|}, \qquad (14)$$

and

$$e_{st} = \frac{e}{\| \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 \|}. \qquad (15)$$

Results

The average standardized MSE and bias values are summarized in Tables 2 - 5 for the four investigated values of p. As Table 2 shows for p = 3, when the observations in both groups are sampled from populations with constant mean vectors (configuration I), the MSE was smallest (and similar) for both the STCS and STAR DA procedures, and largest for the UN procedure.

Table 1: Configurations of μ_1 Investigated in the Simulation Study

p	I	II	III
3	(0.5, 0.5, 0.5)	(0.5, 1, 0.5)	(0.5, 0.25, 0)
5	(0.5, 0.5, 0.5, 0.5, 0.5)	(0.5, 1, 1.5, 1, 0.5)	(1, 0.75, 0.5, 0.25, 0)
7	(0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5)	(0.5, 1, 1.5, 2, 1.5, 1, 0.5)	(1.5, 1.25, 1, 0.75, 0.5, 0.25, 0)
9	(0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5)	(0.5, 1, 1.5, 2, 2.5, 2, 1.5, 1, 0.5)	(2, 1.75, 1.5, 1.25, 1, 0.75, 0.5, 0.25, 0)

Note: $\boldsymbol{\mu}_2$ was equal to the null vector for all conditions

When the data were sampled from a population with a non-constant mean configuration (configurations II or III), MSE and bias were smallest for either UN or STCS procedure and were substantially larger for STUN and STAR procedures. For example, under a CS covariance structure when $\rho = 0.7$ and $p = 3$, the UN and STAR procedures had the smallest and largest average MSE, respectively, when data were sampled from a population with mean configuration II, whereas the UN and STUN procedures had the smallest and largest MSE, respectively, when data were sampled from a population with mean configuration III.

For DA procedures based on constant mean vectors STUN, STCS and STAR, the average MSE decreased as the correlation among the RMs increased when the mean and covariance structure were correctly specified. This finding was observed regardless of the number of RMs, however, when either the covariance or mean structure was misspecified, the average MSE increased as the correlation among the repeated measurements increased. For example, when $p = 3$ and under AR-1 population covariance structure, the average MSE for UN procedure was 0.35 and 0.64 when $\rho = 0.3$ and $\rho = 0.7$, respectively, whereas the average MSE of STAR procedure were 0.07 and 0.05 when $\rho = 0.3$ and $\rho = 0.7$, respectively, when data were sampled from a population with constant mean configuration (see Table 2).

For DA procedures based on structured covariances, the average MSE and bias increased when the covariance structure was misspecified and the mean structures were correctly specified, regardless of the number of RMs. For example, under an AR-1 population covariance structure and when $\rho = 0.3$ and $p = 3$, the average MSE and bias of STCS procedure were 1.3 and 2.0 times the average MSE of STAR procedure, respectively, when the data were sampled from a population with mean configuration I. Similarly, the average MSE and bias of DA procedures based on structured covariances increased under a correctly specified population covariance but a misspecified mean structure. For example, when $p = 3$ and $\rho = 0.3$ under an AR-1 population covariance structure, the average MSE and bias of the STAR procedure when the data were sampled from a population with mean configuration II were 6.4 and 7.0 times the average MSE and bias of STAR procedure under a constant mean configuration, respectively.

For the STUN procedure, the average bias increased when the mean and covariance structures were misspecified, but STCS procedure had the smallest MSE when the data were sampled from a population with a constant mean configuration, regardless of the number of RM. For example, when $p = 7$, under an unstructured population covariance structure and when $\rho = 0.3$ and $p = 7$, the average MSE and bias of STUN procedure were 0.70 and 2.75 times the average MSE and bias of STCS procedures, respectively, when the data were sampled from a population with a constant mean configuration (see Table 4).

Moreover, for each DA procedure, the average MSE and bias due to misspecification of the covariance structure increased as the magnitude of correlation and number of RMs increased. For example, when $p = 5$ and under a CS population covariance structure, the average MSEs of STAR procedure were 2.6 and 5.5 times the average MSE of STCS procedure for $\rho = 0.3$ and $\rho = 0.7$, respectively, when data were sampled from a population with a constant mean configuration (see Table 3). The corresponding bias values for STAR procedure were 4.2 and 10.7 times the bias of STCS procedure when $\rho = 0.3$ and $\rho = 0.7$, respectively. Similarly, when $p = 9$, the average MSEs of STCS procedure were 8,3 and 11.0 times the average MSE of STAR for $\rho = 0.3$ and $\rho = 0.7$, respectively, whereas the corresponding average bias values were 11.0 times the average bias of STCS procedure when $\rho = 0.3$ and $\rho = 0.7$ (see Table 5).

Finally, analyses revealed that the average MSE for each of the DA procedures decreased as the sample size increased. For example, the average MSEs of UN procedure were 7.82, 3.77, and 2.50 when $N = 60$, 90 and 120 respectively. By contrast, the average bias for each DA procedure remained largely unchanged as the sample size increased, regardless of the mean configuration and number of RM. For example, the overall average bias of STAR procedure were 2.12, 2.10 and 2.10 when $N = 60$, 90 and 120, respectively.

Table 2: Average standardized MSE and Bias by Covariance Structure, Magnitude of Correlation and Mean Configuration for $p = 3$

Covariance Structure	ρ	Mean Configuration	MSE			
			UN	STUN	STCS	STAR
CS	0.3	I	0.34	0.11	**0.07**	0.09
		II	0.31	0.45	0.38	0.52
		III	0.52	0.64	0.61	0.63
	0.7	I	0.65	0.12	**0.05**	0.09
		II	0.65	1.89	1.81	2.38
		III	1.16	3.00	2.95	2.99
AR(1)	0.3	I	0.35	0.14	0.09	**0.07**
		II	0.30	0.56	0.33	0.44
		III	0.48	0.43	0.41	0.41
	0.7	I	0.64	0.13	0.08	**0.05**
		II	0.66	3.29	2.44	3.10
		III	1.01	1.11	1.06	1.06
UN	0.3	I	0.38	**0.13**	0.08	0.16
		II	**0.34**	0.33	0.41	0.53
		III	**0.61**	1.20	1.25	1.31
	0.7	I	0.67	**0.12**	0.05	0.12
		II	**0.66**	1.47	1.52	2.03
		III	**1.29**	4.34	4.41	4.48
Covariance Structure	ρ	Mean Configuration	Bias			
			UN	STUN	STCS	STAR
CS	0.3	I	0.08	0.08	**0.07**	0.15
		II	0.09	0.52	0.52	0.61
		III	0.13	0.98	0.98	0.98
	0.7	I	0.06	0.05	**0.05**	0.21
		II	0.14	1.20	1.20	1.38
		III	0.25	2.27	2.27	2.29
AR(1)	0.3	I	0.08	0.08	0.15	**0.08**
		II	0.09	0.59	0.47	0.56
		III	0.11	0.75	0.77	0.75
	0.7	I	0.06	0.06	0.22	**0.06**
		II	0.16	1.61	1.40	1.58
		III	0.16	1.34	1.36	1.34
UN	0.3	I	0.08	**0.08**	0.15	0.27
		II	**0.10**	0.42	0.54	0.60
		III	**0.18**	1.40	1.45	1.47
	0.7	I	0.06	**0.05**	0.08	0.27
		II	**0.13**	1.05	1.10	1.27
		III	**0.32**	2.77	2.81	2.83

Notes: See Table 1 for a description of the mean configurations; CS = compound symmetric; AR-1 = first-order autoregressive; UN = unstructured; ρ = correlation parameter; UN = unstructured mean and covariance; STUN = structured mean and unstructured covariance; STCS = structured mean and CS covariance; STAR = structured mean and AR-1 covariance. Numbers in bold correspond to bias and error values of DA procedures for which the mean and covariance structures are correctly specified.

Table 3: Average standardized MSE and Bias by Covariance Structure, Magnitude of Correlation and Mean Configuration for $p = 5$

Covariance Structure	ρ	Mean Configuration	MSE			
			UN	STUN	STCS	STAR
CS	0.3	I	0.56	0.14	**0.05**	0.13
		II	0.53	0.96	0.80	1.09
		III	0.63	1.21	1.13	1.16
	0.7	I	1.10	0.16	**0.02**	0.11
		II	1.35	4.40	4.19	5.20
		III	1.80	6.06	5.95	6.00
AR(1)	0.3	I	0.56	0.20	0.08	**0.05**
		II	0.46	0.76	0.37	0.48
		III	0.55	0.57	0.47	0.45
	0.7	I	1.06	0.21	0.08	**0.04**
		II	0.96	2.42	1.51	2.01
		III	1.08	0.86	0.76	0.72
UN	0.3	I	0.66	**0.20**	0.14	0.20
		II	**0.64**	2.26	1.33	1.67
		III	**0.75**	1.61	1.63	1.61
	0.7	I	1.15	**0.17**	0.03	0.10
		II	**1.40**	4.81	4.44	5.35
		III	**2.04**	7.57	7.66	7.76
Covariance Structure	ρ	Mean Configuration	Bias			
			UN	STUN	STCS	STAR
CS	0.3	I	0.06	0.06	**0.05**	0.21
		II	0.09	0.60	0.60	0.69
		III	0.12	0.89	0.89	0.91
	0.7	I	0.04	0.04	**0.03**	0.23
		II	0.18	1.39	1.39	1.54
		III	0.27	2.08	2.08	2.09
AR(1)	0.3	I	0.09	0.09	0.14	**0.07**
		II	0.09	0.48	0.38	0.45
		III	0.10	0.55	0.56	0.55
	0.7	I	0.05	0.05	0.22	**0.04**
		II	0.11	0.99	0.83	0.95
		III	0.10	0.72	0.74	0.72
UN	0.3	I	0.08	**0.08**	0.31	0.35
		II	**0.11**	0.96	0.77	0.86
		III	**0.15**	1.03	1.08	1.07
	0.7	I	0.04	**0.03**	0.07	0.22
		II	**0.18**	1.45	1.42	1.56
		III	**0.30**	2.33	2.36	2.37

Notes: See Table 1 for a description of the mean configurations; CS = compound symmetric; AR-1 = first-order autoregressive; UN = unstructured; ρ = correlation parameter; UN = unstructured mean and covariance; STUN = structured mean and unstructured covariance; STCS = structured mean and CS covariance; STAR = structured mean and AR-1 covariance. Numbers in bold correspond to bias and error values of DA procedures for which the mean and covariance structures are correctly specified.

Table 4: Average standardized MSE and Bias by Covariance Structure, Magnitude of Correlation and Mean Configuration for $p = 7$

Covariance Structure	ρ	Mean Configuration	MSE			
			UN	STUN	STCS	STAR
CS	0.3	I	0.78	0.19	**0.03**	0.17
		II	0.90	1.67	1.37	1.77
		III	0.97	1.96	1.81	1.83
	0.7	I	1.60	0.22	**0.02**	0.11
		II	2.72	7.68	7.31	8.56
		III	3.29	9.78	9.55	9.61
AR(1)	0.3	I	0.84	0.31	0.08	**0.04**
		II	0.87	1.16	0.43	0.58
		III	0.83	0.87	0.59	0.58
	0.7	I	1.56	0.31	0.08	**0.03**
		II	1.39	2.26	1.09	1.51
		III	1.42	0.96	0.70	0.70
UN	0.3	I	1.23	**0.33**	0.23	0.45
		II	**2.18**	4.70	7.21	7.64
		III	**2.54**	15.77	11.50	11.56
	0.7	I	1.73	**0.24**	0.03	0.15
		II	**2.94**	7.95	7.98	9.36
		III	**4.40**	14.93	15.59	15.84
Covariance Structure	ρ	Mean Configuration	Bias			
			UN	STUN	STCS	STAR
CS	0.3	I	0.06	0.06	**0.04**	0.27
		II	0.11	0.64	0.64	0.72
		III	0.15	0.86	0.86	0.87
	0.7	I	0.04	0.03	**0.02**	0.23
		II	0.22	1.48	1.48	1.60
		III	0.31	2.00	2.00	2.00
AR(1)	0.3	I	0.10	0.10	0.14	**0.07**
		II	0.10	0.44	0.34	0.41
		III	0.11	0.48	0.48	0.48
	0.7	I	0.06	0.06	0.20	**0.04**
		II	0.09	0.72	0.57	0.67
		III	0.09	0.51	0.54	0.51
UN	0.3	I	0.04	**0.04**	0.11	0.29
		II	**0.24**	1.51	1.55	1.68
		III	**0.39**	2.48	2.55	2.58
	0.7	I	0.05	**0.05**	0.05	0.34
		II	**0.14**	0.85	0.85	0.94
		III	**0.19**	1.20	1.20	1.22

Notes: See Table 1 for a description of the mean configurations; CS = compound symmetric; AR-1 = first-order autoregressive; UN = unstructured; ρ = correlation parameter; UN = unstructured mean and covariance; STUN = structured mean and unstructured covariance; STCS = structured mean and CS covariance; STAR = structured mean and AR-1 covariance. Numbers in bold correspond to bias and error values of DA procedures for which the mean and covariance structures are correctly specified.

Table 5: Average standardized MSE and Bias by Covariance Structure, Magnitude of Correlation and Mean Configuration for $p = 9$

Covariance Structure	ρ	Mean Configuration	MSE			
			UN	STUN	STCS	STAR
CS	0.3	I	1.33	0.31	**0.03**	0.25
		II	1.54	2.56	2.04	2.51
		III	1.64	2.88	2.53	2.59
	0.7	I	2.18	0.29	**0.01**	0.11
		II	5.14	11.58	10.97	12.40
		III	6.12	14.07	13.66	13.72
AR(1)	0.3	I	1.19	0.47	0.07	**0.04**
		II	0.98	1.41	0.51	0.75
		III	1.40	1.38	0.74	0.78
	0.7	I	2.17	0.46	0.07	**0.02**
		II	2.05	2.51	0.86	1.22
		III	2.03	1.27	0.69	0.70
UN	0.3	I	1.95	**0.47**	0.09	0.33
		II	**4.73**	10.85	12.28	12.84
		III	**6.85**	35.01	30.47	30.74
	0.7	I	2.86	**0.37**	0.01	0.12
		II	**8.52**	24.32	23.45	25.40
		III	**10.07**	32.21	31.44	32.00
Covariance Structure	ρ	Mean Configuration	Bias			
			UN	STUN	STCS	STAR
CS	0.3	I	0.07	0.07	**0.03**	0.33
		II	0.13	0.66	0.66	0.74
		III	0.16	0.84	0.84	0.85
	0.7	I	0.03	0.03	**0.02**	0.22
		II	0.29	1.54	1.54	1.64
		III	0.37	1.96	1.96	1.96
AR(1)	0.3	I	0.12	0.12	0.13	**0.07**
		II	0.09	0.41	0.31	0.40
		III	0.13	0.44	0.44	0.47
	0.7	I	0.07	0.07	0.19	**0.03**
		II	0.09	0.58	0.43	0.51
		III	0.10	0.41	0.43	0.41
UN	0.3	I	0.08	**0.07**	0.22	0.41
		II	**0.32**	1.46	1.63	1.67
		III	**0.43**	2.40	2.26	2.27
	0.7	I	0.04	**0.03**	0.06	0.23
		II	**0.43**	2.26	2.25	2.35
		III	**0.56**	2.98	2.97	2.99

Notes: See Table 1 for a description of the mean configurations; CS = compound symmetric; AR-1 = first-order autoregressive; UN = unstructured; ρ = correlation parameter; UN = unstructured mean and covariance; STUN = structured mean and unstructured covariance; STCS = structured mean and CS covariance; STAR = structured mean and AR-1 covariance; Numbers in bold correspond to bias and error values of DA procedures for which the mean and covariance structures are correctly specified.

Conclusion

This research investigated the effects of RM mean and/or covariance structure misspecification on bias and error in DFCs for DA procedures based on parsimonious mean and/or covariance structures. As expected, the bias and error in the DFCs of the investigated procedures increased when the RM mean and/or covariance structures were misspecified. The average bias and error variation due to misspecification of the RM mean structure was greater than the average bias and error variation due to RM covariance structure misspecification for all of the investigated procedures. Although DA procedures based on parsimonious RM mean and covariance structures had negligible bias when the mean and covariances are correctly specified, UN DA procedure had the smallest bias when the data were sampled from a population with non-constant mean configuration.

Based on the study findings, adopting a DA procedure based on unstructured mean vectors and covariance matrices when the researcher has prior knowledge to suggest that the mean longitudinal profile for each group will change across the repeated measures occasions is recommended. If the mean longitudinal profile in each group is not expected to increase or decrease across the measurement occasions, then either the STCS or STAR procedure are recommended because they require estimation of the fewer number of parameters, although any of the procedures can be expected to perform well in terms of both bias and error variation.

To reduce the effect of mean and/or covariance structure misspecification on bias and error in the DFCs, preliminary tests of model fit could be undertaken before adopting a DDA procedure for RM data. Graphical exploration of the data, likelihood ratio tests, or penalized log-likelihood measures like the Akaike information criterion have all been proposed to guide the specification of mean and covariance structures (Fitzmaurice, Laird & Ware, 2004)

Study Limitations

This research focused on normally distributed data. The impact of mean and/or covariance misspecification on bias and error in the DFCs when data are sampled from non-normal distribution has not been investigated. Although mild departures from multivariate non-normality are known to have little effect on classification accuracy of classical DA procedure (Ashikaga & Chang, 1981), classification accuracy can be severely affected under large departures (Lachenbruch, Sneeringer & Revo, 1973; Baron, 1991; McLachlan, 1992). Inferences about DFCs of the linear DA procedures may also be affected by the degree of departure from the assumption of multivariate normality (McLachlan, 1992).

The DA procedures considered in this manuscript also focused only on complete data, an assumption which may not be satisfied in RM studies, which are often characterized by missing observations and unbalanced measurements occasions (Fairclough, et al., 1998). In the simulation study, the RM variances were assumed to be constant across variables and groups. Linear DA procedures rest on the assumption of covariance homogeneity (Huberty & Olejnik, 2006). Departures from this assumption may result in reduced classification accuracy (Solberg, 1988). DFCs have been shown to be relatively robust to violation of this assumption when the data are normally distributed (Owen & Chmielewski, 1985), but it is not known if this robustness will continue to be evident when the covariance and/or mean vector is misspecified.

Future Research

A number of opportunities for future research exist in the development of DDA procedures for RM data. Although several studies have examined the effects of population distribution on classification accuracy, there is limited investigation of the effects of population distribution and other data characteristics on bias and error in DFCs. Existing studies in this area have only focused on the effects of sample size, number of outcome variables, and mean configuration on bias and variation in DFCs when data were sampled from normally distributed data (Williams & Titus, 1991; Owen & Chmielewski, 1985). This study investigated DA procedures based on constant mean vectors and/or structured covariances. However, the assumption of a constant repeated measures

group mean structure may not be tenable when the interest is in the assessment of the relative importance of measurement occasions that discriminate between groups. DA procedures based on non-constant mean vectors and CS or AR-1 covariance structures can be further investigated. These procedures which assume non-constant mean configurations and parsimonious structures will be useful for assessing the relative importance of information collected at each measurement occasions in univariate repeated measures studies.

Summary

Although the adoption of a DA procedure based on a parsimonious mean and/or covariance structure can reduce the number of parameters to estimate, which is beneficial when sample size is small (Roy & Khattree, 2005a), this study shows that bias and error variation in the DFCs can be large, particularly when there is misspecification of the RM mean structure. A researcher's choice of a DA procedure for RM data is dependent, in part, on the trade-off between parsimony in parameter estimation and bias and/or error in the DFCs.

Acknowledgements

This research was supported by a Canadian Institutes of Health Research (CIHR) Vanier Graduate Scholarship to the first author, and a CIHR New Investigator Award to the second author.

References

Ashikaga, T., & Chang, P. C. (1981). Robustness of Fisher's linear discriminant function under two-component mixed-normal models. *Journal of the American Statistical Association*, *76*, 375-676.

Baron, A. E. (1991). Misclassification among methods used for multiple group discrimination: The effects of distributional properties. *Statistics in Medicine*, *10*, 757-766.

Beaumont, J. L., Lix, L. M., Yost, K. J., & Hahn, E. A. (2006). Application of robust statistical methods for sensitivity analysis of health-related quality of life outcomes. *Quality of Life Research*, *15*, 349-356.

Croux, C., & Dehon, C. (2001). Robust linear discriminant analysis using S-estimators. *Canadian Journal of Statistics*, *29*, 473-493.

Fairclough, D. L., Peterson, H. F., Cella, D., Bonomi, P. (1998). Comparison of several model-based methods for analysing incomplete quality of life data in cancer clinical trials. *Statistics in Medicine*, *17*, 781-796.

Feighner, J. P., & Sverdlov, L. (2002). The use of discriminant analysis to separate a study population by treatment subgroups in a clinical trial with a new pentapeptide antidepressant. *Journal of Applied Research*, *2*, 17-18.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, *7*, 179-188.

Fitzmaurice, G., Laird, N. M., & Ware, J. H. (2004). *Applied longitudinal analysis*. New Jersey: Wiley.

He, X., & Fung, W. K. (2000). High breakdown estimation for multiple populations with applications to discriminant analysis. *Journal of Multivariate Analysis*, *72*, 151-162.

Huberty, C. J, & Wisenbaker, J. M. (1992). Variable importance in multivariate group comparisons. *Journal of Educational Statistics*, *17*, 75-91.

Huberty, C. J., & Olejnik, S. (2006). *Applied MANOVA and discriminant analysis*. New York: Wiley.

Lachenbruch, P. A., Sneeringer, C., & Revo, L. T. (1973). Robustness of the linear and quadratic discriminant function to certain types of non-normality. *Communications in Statistics*, *1*, 39-57.

Lei, P., & Koehly, L. M. (2003). Linear discriminant analysis versus logistic regression: a comparison of classification errors in the two-group case. *Journal of Experimental Education*, *72*, 25-49.

Levesque, L., Ducharme, F., Zarit, S. H., Lachance, L., & Giroux, F. (2008). Predicting longitudinal patterns of psychological distress in older husband caregivers: further analysis of existing data. *Aging Mental Health*, *12*, 333-343.

Marshall, G., & Baron, A. E. (2000). Linear discriminant models for unbalanced longitudinal data. *Statistics in Medicine*, *19*, 1969-1981.

McLachlan, G. J. (1992). *Discriminant analysis and statistical pattern recognition*. New York: Wiley.

Owen, J. G., Chmielewski, M. A. (1985). On canonical variates analysis and the construction of confidence ellipses in systematic studies. *Systematic Zoology*, *34*, 366-374.

Roy, A., & Khattree, R. (2005a). Discrimination and classification with repeated measures data under different covariance structures. *Communications in Statistics – Simulation and Computation*, *34*, 167-178.

Roy, A., & Khattree, R. (2005b). On discrimination and classification with multivariate repeated measures data. *Journal of Statistical Planning and Inference*, *134*, 462-485.

Roy, A., & Khattree, R. (2007). Classification of multivariate repeated measures data with temporal autocorrelation. *Advances in Data Analysis and Classification*, *1*, 175-199.

SAS Institute Inc. (2008). *SAS/IML user's guide, version 9.2*. Cary, NC: SAS Institute, Inc.

Solberg, H. E. (1988). Discriminant analysis. *Critical Reviews in Clinical Laboratory Sciences*, *9*, 209-242.

Thomas, D. R. (1992). Interpreting discriminant functions: a data analytic approach. *Multivariate Behavioral Research*, *27*, 335-362.

Thomas, D. R., & Zumbo, B. D. (1996). Using a measure of variable importance to investigate the standardization of discriminant coefficients. *Journal of Educational and Behavioral Statistics*, *21*, 110-130.

Tomasko, L., Helms, R. W., & Snappin, S. M. (1999). A discriminant analysis extension to mixed models. *Statistics in Medicine*, *18*, 1249-1260.

Williams, B. K., & Titus, K. (1988). Assessment of sampling stability in ecological applications of discriminant analysis. *Ecology*, *69*, 1275-1285.

Williams, B. K., Titus, K., Hines, J. E. (1991). Stability and bias of classification rates in biological applications of discriminant analysis. *The Journal of Wildlife Management*, *54*, 331-341.

Appendix

As described, more details about ML estimation of the coefficients of STAR procedure is provided here. In (8),

$$m_{j1} = \frac{\mathbf{1}_p^{\mathrm{T}} \bar{\mathbf{y}}_j}{p}, \qquad \text{(A-1)}$$

$$m_{j2} = \frac{\mathbf{1}_p^{\mathrm{T}} \bar{\mathbf{y}}_j - \bar{\mathbf{y}}_{j1} - \bar{\mathbf{y}}_{jp}}{(p-2)}, \qquad \text{(A-2)}$$

nd $\bar{\mathbf{y}}_{j1}$ and $\bar{\mathbf{y}}_{jp}$, are respectively, the first and p^{th} elements of the vector $\bar{\mathbf{y}}_j$. In (9) and (10),

$$\beta_1 = \mathrm{tr}(\mathbf{W}_0) - \mathbf{W}_{0,11} - \mathbf{W}_{0,pp},$$
$$\beta_2 = \alpha_1 - \mathbf{W}_{5,11} - \mathbf{W}_{5,pp}, \text{ and}$$
$$\beta_3 = \alpha_3 - \mathbf{W}_{6,11} - \mathbf{W}_{6,pp}.$$

Further,

$$\alpha_1 = \mathrm{tr}(\mathbf{W}_0 + \mathbf{W}_1 + \mathbf{W}_2 + \mathbf{W}_3 + \mathbf{W}_4),$$
$$\alpha_2 = \mathrm{tr}(\mathbf{W}_5), \text{ and}$$
$$\alpha_3 = tr(\mathbf{W}_6); \mathbf{W}_0 = \mathbf{W} + \mathbf{W}_3 + \mathbf{W}_4.$$

Also,

$$\gamma_1 = \sum_{k=2}^{p} \mathbf{W}_{0,k-1k}, \qquad \text{(A-3)}$$

$$\gamma_2 = \sum_{k=2}^{p} \mathbf{W}_{5,k-1k}, \qquad \text{(A-4)}$$

and

$$\gamma_3 = \sum_{k=2}^{p} \mathbf{W}_{6,k-1k} \qquad \text{(A-5)}$$

where $\mathrm{Wu}_{,k\text{-}1k}$ is the $(k\text{-}1,k)^{\text{th}}$ element of $\mathbf{W}_u$ (u = 0,…,6) and k = 1,…,*p*.

In these equations,

$$\mathbf{W} = \sum_{j=1}^{2} \sum_{i=1}^{n_j} (\mathbf{y}_{ij} - \bar{\mathbf{y}}_j)(\mathbf{y}_{ij} - \bar{\mathbf{y}}_j)^{\mathrm{T}}, \qquad \text{(A-6)}$$

$\mathbf{W}_3 = \bar{\mathbf{y}}_1 \bar{\mathbf{y}}_1^{\mathrm{T}}$, $\mathbf{W}_4 = \bar{\mathbf{y}}_2 \bar{\mathbf{y}}_2^{\mathrm{T}}$, $\mathbf{W}_5 = \mathbf{1}_p^{\mathrm{T}} \bar{\mathbf{y}}_1 + \bar{\mathbf{y}}_1 \mathbf{1}_p^{\mathrm{T}}$, and $\mathbf{W}_6 = \mathbf{1}_p^{\mathrm{T}} \bar{\mathbf{y}}_1 + \bar{\mathbf{y}}_1 \mathbf{1}_p^{\mathrm{T}}$.

Journal of Modern Applied Statistical Methods
November 2011, Vol. 10, No. 2, 583-598

1538 – 9472/11/$95.00

Indeterminacy of Factor Score Estimates In Slightly Misspecified Confirmatory Factor Models

André Beauducel
University of Bonn,
Bonn, Germany

Two methods to calculate a measure for the quality of factor score estimates have been proposed. These methods were compared by means of a simulation study. The method based on a covariance matrix reproduced from a model leads to smaller effects of sampling error.

Key words: Confirmatory factor analysis, structural equation modeling, indeterminacy, factor score estimates.

Introduction

Factor score estimates are computed when individual scores representing the factors of a model are interesting. This can be the case in personnel selection or in educational settings where individuals are to be compared with respect to their scores. Thus, although latent variables might be of interest in factor analysis and structural equation modeling, some applications are still based on the concrete scores of individuals; it is for this reason that factor score estimates are of interest for applied researchers. It should be noted that although factor score estimates are termed estimates, they are not estimates in the usual sense because there are no true values that may be approximated by the estimates (Schönemann & Steiger, 1976).

The term factor score estimates denotes the aim to construct scores that represent the unknown factors in an optimal way. It follows from this reasoning that it is necessary to evaluate the quality of the factor score estimates (Gorsuch, 1983). There are two well-known indices that allow for an evaluation of factor score indeterminacy: The multiple correlation ρ or the squared multiple correlation ρ^2 of the factor with the measured variables and the minimum correlation between two sets of factor score estimates of the same solution, $2\rho^2 - 1$ (Grice, 2001; Green, 1976; Guttman, 1955; Schönemann, 1971). Additional interesting possibilities for the evaluation of different factor score estimates with respect to their determinacy can be found in Krijnen (2006).

Although the computation of factor score estimates is also possible for confirmatory factor analysis (CFA) and specific methods have been developed for this purpose (Beauducel & Rabe, 2009), most applications and discussions of factor score indeterminacy occur in the context of exploratory factor analysis. Beauducel and Rabe (2009) present a new type of factor score estimate representing specific aspects of a CFA model (e.g., parts of a loading matrix), whereas this present study investigates two different methods to calculate factor score indeterminacy.

A difference between exploratory factor analysis and CFA is that in CFA the loadings of the variables and the correlations between factors can be specified according to theoretical assumptions. When the model assumptions are correct, fit indices would indicate that the model fits the data. However, small amounts of model-misspecification do not lead to model rejection according to many general rules (Barrett, 2007;

André Beauducel is a Professor at the Institute of Psychology in the Department of Philosophy. Email him at: beauducel@uni-bonn.de. Address for correspondence: Kaiser-Karl-Ring 9, 53111 Bonn, Germany.

Fan & Sivo, 2007; Beauducel & Wittmann, 2005; Marsh, Hau & Wen, 2004; Hu & Bentler, 1999). As a consequence, model parameters can be over- and/or under-estimated not only because of sampling error, but also because of a difference between the model parameters and the population parameters.

There is a discussion on the size of difference between model and data that might be regarded as acceptable (Marsh, et al., 2004; Barrett, 2007), but a small difference between the covariance matrix implied by the model and the empirical covariance matrix is accepted by many researchers in structural equation modeling. A difference between model and data could also occur in exploratory factor analysis, but the only way to obtain model misspecification in this context is over- or under-extraction of factors. Nevertheless, this article focuses on factor score indeterminacy as it is calculated from CFA with correctly and misspecified model parameters, because indeterminacy has rarely been evaluated in this context. A simulation study was performed in order to investigate the effects of sampling error and model misspecification on factor score indeterminacy.

The Calculation of ρ or ρ²

It should be noted that there are two different ways to calculate indeterminacy, often referred to as ρ, the correlation between the variables and the factor (Grice, 2001). In order to present the calculation of ρ or ρ², the common factor model is described first. The common factor model assumes that the observations are generated by

$$\boldsymbol{X} = \boldsymbol{\Lambda F} + \boldsymbol{E}, \qquad (1)$$

where $\boldsymbol{X}$ is the random vector of observations of order p, $\boldsymbol{F}$ the random vector with factor scores of order q, $\boldsymbol{E}$ the unobservable random error vector of order p, and $\boldsymbol{\Lambda}$ the factor pattern matrix of order p by q. The observations $\boldsymbol{X}$, the factor scores $\boldsymbol{F}$, and the error vectors $\boldsymbol{E}$ are assumed to have an expectation zero ($\varepsilon[\boldsymbol{X}] = 0$, $\varepsilon[\boldsymbol{F}] = 0$, $\varepsilon[\boldsymbol{E}] = 0$). The covariance between the factor scores and the error scores is assumed to be zero ($\mathrm{Cov}[\boldsymbol{F}, \boldsymbol{E}] = 0$). The standard deviation of $\boldsymbol{F}$ is one, the expectation of the covariance of the observed variables is $\boldsymbol{\Sigma}$ ($\varepsilon[\boldsymbol{XX}'] = \boldsymbol{\Sigma}$). The covariance matrix $\boldsymbol{\Sigma}$ can be decomposed by

$$\boldsymbol{\Sigma} = \boldsymbol{\Lambda\Phi\Lambda}' + \boldsymbol{\Psi}^2, \qquad (2)$$

where $\boldsymbol{\Phi}$ represents the q by q factor correlation matrix and $\boldsymbol{\Psi}^2$ the p by p covariance matrix between the observed variables $\boldsymbol{X}$ and the error scores $\boldsymbol{E}$ ($\mathrm{Cov}[\boldsymbol{X}, \boldsymbol{E}] = \boldsymbol{\Psi}^2$) and $\boldsymbol{\Psi}^2$ also represents the covariance matrix of the error scores $\boldsymbol{E}$ ($\mathrm{Cov}[\boldsymbol{E}, \boldsymbol{E}] = \boldsymbol{\Psi}^2$). $\boldsymbol{\Psi}^2$ is generally assumed to be a diagonal matrix and it will be assumed herein that it contains only positive values. In order to investigate CFA modelling as it often occurs in empirical research, it was, however, decided also to allow for some non-diagonal elements of $\boldsymbol{\Psi}^2$.

The factor score indeterminacy ρ, the multiple correlation of the variables with the factor can be described on the basis of Thurstone's (1935) regression score estimate, which is the best linear factor score estimate (Krijnen, Wansbeek & Ten Berge, 1996). The covariances of the factors with the best linear factor score estimates are given by

$$\mathbf{diag}(\boldsymbol{F}\hat{\boldsymbol{F}}') = \mathbf{diag}(\boldsymbol{FX}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda\Phi}). \qquad (3)$$

It follows from equation 1 that it is possible to insert $\boldsymbol{\Phi\Lambda}'$ for $\boldsymbol{FX}'$ into equation 3. Moreover, it is possible to standardize the covariances of the factors with the best linear factor score estimates in order to obtain the correlations. This yields

$$\begin{aligned} &\mathbf{diag}(\boldsymbol{F}\hat{\boldsymbol{F}}') \\ &= \mathbf{diag}(\boldsymbol{\Phi\Lambda}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda\Phi})\mathbf{diag}(\boldsymbol{\Phi\Lambda}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda\Phi})^{-1/2} \\ &= \mathbf{diag}(\boldsymbol{\Phi\Lambda}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda\Phi})^{1/2} \end{aligned} \qquad (4)$$

so that the diagonal elements in the left hand side of equation 4 contain the correlations of the best linear factor score estimates with the factors. Standardizing $\boldsymbol{F}$ is not necessary, because it has by definition a standard deviation of one. Because the best linear factor score estimate is the best linear combination of the

measured variables in order to estimate the factor, the correlations in equation 4 also represent the multiple correlations of the measured variables with the factors.

When a factor model has a perfect fit, $\boldsymbol{\Sigma}$, the expectation of the covariance matrix of observed variables, which is calculated as the covariance matrix reproduced from the model parameters, and **S**, the empirical covariance matrix of the observed variables, are equal. Nevertheless, in the context of CFA, small differences between **S** and $\boldsymbol{\Sigma}$ regularly occur. This is always the case when the Root Mean Square Residual (RMR) is greater than zero, because this index describes the difference between these two covariance matrices. When a relevant difference between **S** and $\boldsymbol{\Sigma}$ occurs, one has to choose between these two covariance matrices for the calculation of factor score indeterminacy. The choice is to calculate indeterminacy according to equation 4 or to use the empirical covariance matrix **S** as in

$$\mathbf{diag}(\boldsymbol{F}\hat{\boldsymbol{F}}') = \mathbf{diag}(\boldsymbol{\Phi\Lambda' S^{-1}\Lambda\Phi})^{1/2}. \quad (5)$$

The calculation of factor score indeterminacy by means of the sample covariance matrix S has been presented by Heermann (1963), Gorsuch (1983) and Grice (2001). The calculation of indeterminacy by means of the reproduced covariance matrix, which is based on the estimated population parameters of the model, is presented in Mulaik and McDonald (1978) and in McDonald (1981). Because both ways to calculate indeterminacy are referred to in the literature and no discussion of the possible differences is currently available, this study compares the two ways to calculate indeterminacy on the basis of a simulation study. The comparison of the coefficients of indeterminacy is especially relevant to CFA, where small amounts of model misspecification are sometimes accepted (Hu & Bentler, 1999). As in other studies (Grice, 2001), the results for the squared validity coefficients (ρ^2) were presented in the following, because ρ^2 can be interpreted as the common variance between the factor and the corresponding factor score estimate.

Methodology

The aim of the simulation study was to compare the two above-mentioned coefficients of indeterminacy (equations 4 and 5) with respect to model misspecification and effects of sampling error. Therefore, the two versions of ρ^2 were first compared for the population CFA models and then for the corresponding CFA models based on samples derived from the population.

Generation of Population CFA Models

Population models based on 2, 4 and 8 factors, moderate (0.40/0.60) and large (0.60/0.80) salient loadings, with orthogonal and oblique factors (with interfactor correlations of 0.30) were investigated. The population models were chosen in order to represent CFA models as they are often found in applied research. This explains why 2-, 4- and 8-factor models were investigated, as well as the size of the loadings and the moderate size of the interfactor correlations for the oblique models. In order to perform CFA modeling like in empirical research, it is necessary to investigate not only correctly specified models but also models with small amounts of model-misspecification. A common type of model-misspecification is the omission of correlated residuals (correlated error terms of observed variables). This type of model-misspecification is interesting in the present context, because it could be expected to have an impact on the loading size and thereby on the coefficients of indeterminacy.

In the first step, the parameters of the correctly specified population models including correlated residuals were fixed to their intended values, then the corresponding population covariance matrices were reproduced from the model parameters (according to equation 2). For simplicity, the size of the model parameters was chosen in a way that ensures that the reproduced covariance matrices were correlation matrices. Finally, the population covariance matrices were used for CFA modeling in order to estimate the misspecified model parameters. The CFA modeling was performed with Mplus 3.11 by means of maximum likelihood estimation. The salient loadings were freely estimated, the non-salient loadings were fixed to zero, the variances of the factors were fixed to one and the

correlations of all residuals were fixed to zero in the misspecified models (the variance of the residuals was freely estimated). For the orthogonal models the correlations between the factors were fixed to zero, for the oblique models they were freely estimated.

Table 1 contains the correctly specified and the misspecified population loadings for the 0.40/0.60 (moderate loadings) condition and for the 0.60/0.80 (large loadings) condition for the orthogonal two-factor models based on the population covariance matrices including correlated residuals (the correlations between the residuals are presented at the bottom of Table 1).

Table 2 contains the corresponding parameters for the oblique models. The misspecified models would be accepted according to conventional cut-off criteria for fit indices (e.g., Hu & Bentler, 1999). It was intended to generate small and generally accepted amounts of model-misspecification, so that even the misspecified models investigated here represent models as they might be published in empirical research. Nevertheless, the omission of the correlated residuals leads to small errors with respect to the loading size both in the orthogonal and in the oblique model (see Tables 1 and 2). The population parameters for the orthogonal and oblique four- and eight-factor models would be identical to the corresponding parameters presented in Table 1 and 2 so that they are not presented.

Another type of model misspecification with an impact on the loading size and thereby on the coefficients of indeterminacy occurs when equality constraints are imposed on loadings that are unequal in the population. In order to base the results of the present simulation study on more than one type of model misspecification, misspecifications resulting from equality constraints on the loadings were also investigated. Again, the parameters of the correctly specified models were fixed in the first step and then the corresponding population covariance matrices were calculated from the model parameters. Finally, these population covariance matrices were used for CFA modeling with misspecified parameters. Again, the model parameters were chosen in a way to ensure that the reproduced covariance matrices were correlation matrices. The misspecified models were again estimated by means of maximum likelihood estimation.

The variances of the factors were constrained to be one, the non-salient loadings were fixed to zero, the unconstrained salient loadings were freely estimated, and the covariance matrix of the error terms was constrained to be diagonal (there were no correlated residuals in these models, but the variances of the residuals were freely estimated). For the orthogonal models the correlations between the factors were fixed to zero, for the oblique models they were freely estimated. The misspecification for the two-factor model was introduced by means of equality constraints for each of the smaller loadings of the variables v_1-v_4 on the first factor with each of the larger loadings v_{13}-v_{16} on the second factor. For the four- and eight-factor models, similar equality constraints were imposed on the loadings of each pair of factors.

Table 3 contains the correctly specified and the misspecified population loadings for the 0.40/0.60 (moderate loadings) condition and for the 0.60/0.80 (large loadings) condition for the orthogonal two-factor models. The equality of loadings resulting from the equality constraints was not perfect in the completely standardized solutions (it was perfect in the unstandardized solutions). Not surprisingly, the fit of the correctly specified population models was perfect, but even the misspecified models fit the data very well (see Table 3). The misspecified population model would not be rejected according to conventional fit criteria (Hu & Bentler, 1999). The population loadings were the same for the four- and eight-factor models and are therefore not presented.

The population loadings for the correctly specified and the misspecified oblique two-factor models are presented in Table 4. As before, the model misspecification was introduced by means of equality constraints on loadings that were not equal in the population (see Table 4). Again an evaluation of the model fit of the misspecified models would not lead to model rejection for conventional criteria (Hu & Bentler, 1999).

Table 1: Population Loadings for the Orthogonal Two-Factor Models
(Completely Standardized Solution)

	Moderate Loadings				Large Loadings			
	Without Model Misspecification [a]		With Model Misspecification [b]		Without Model Misspecification [a]		With Model Misspecification [c]	
	F1	F2	F1	F2	F1	F2	F1	F2
x_1	.400	-	.414	-	.600	-	.607	-
x_2	.400	-	.414	-	.600	-	.607	-
x_3	.400	-	.392	-	.600	-	.596	-
x_4	.400	-	.392	-	.600	-	.596	-
x_5	.600	-	.584	-	.800	-	.793	-
x_6	.600	-	.584	-	.800	-	.793	-
x_7	.600	-	.637	-	.800	-	.816	-
x_8	.600	-	.637	-	.800	-	.816	-
x_9	-	.400	-	.414	-	.600	-	.607
x_{10}	-	.400	-	.414	-	.600	-	.607
x_{11}	-	.400	-	.392	-	.600	-	.596
x_{12}	-	.400	-	.392	-	.600	-	.596
x_{13}	-	.600	-	.584	-	.800	-	.793
x_{14}	-	.600	-	.584	-	.800	-	.793
x_{15}	-	.600	-	.637	-	.800	-	.816
x_{16}	-	.600	-	.637	-	.800	-	.816
Correlated Residuals								
x_1 with x_2	.126		.000		.096		.000	
x_7 with x_8	.096		.000		.054		.000	
x_9 with x_{10}		.126		.000		.096		.000
x_{15} with x_{16}		.096		.000		.054		.000

Notes: [a] The model fit for the population model without misspecification is perfect by definition: $\chi^2(100) = 0.00$; [b] The χ^2-test for the misspecified model with moderate loadings is non-significant even for the largest sample size used in the simulation study (N=750): $\chi^2(104) = 50.93$; Comparative Fit Index = 0.99; Root Mean Square Error of Approximation = 0.026; Standardized Root Mean Square Residual = 0.012. [c] The χ^2-test for the misspecified model with large loadings is non-significant even for the largest sample size used in the simulation study (N=750): $\chi^2(104)= 51.18$; Comparative Fit Index = 0.99; Root Mean Square Error of Approximation = 0.026; Standardized Root Mean Square Residual = 0.012.

Table 2: Population Loadings for the Oblique Two-Factor Models
(Completely Standardized Solution)

	Moderate Loadings				Large Loadings			
	Without Model Misspecification [a]		With Model Misspecification [b]		Without Model Misspecification [a]		With Model Misspecification [c]	
	F1	F2	F1	F2	F1	F2	F1	F2
x_1	.400	-	.414	-	.600	-	.607	-
x_2	.400	-	.414	-	.600	-	.607	-
x_3	.400	-	.392	-	.600	-	.596	-
x_4	.400	-	.392	-	.600	-	.596	-
x_5	.600	-	.585	-	.800	-	.793	-
x_6	.600	-	.585	-	.800	-	.793	-
x_7	.600	-	.636	-	.800	-	.815	-
x_8	.600	-	.636	-	.800	-	.815	-
x_9	-	.400	-	.414	-	.600	-	.607
x_{10}	-	.400	-	.414	-	.600	-	.607
x_{11}	-	.400	-	.392	-	.600	-	.596
x_{12}	-	.400	-	.392	-	.600	-	.596
x_{13}	-	.600	-	.585	-	.800	-	.793
x_{14}	-	.600	-	.585	-	.800	-	.793
x_{15}	-	.600	-	.636	-	.800	-	.815
x_{16}	-	.600	-	.636	-	.800	-	.815
Interfactor-Correlation	.300		.289		.300		.297	
Correlated Residuals								
x_1 with x_2	.126		.000		.096		.000	
x_7 with x_8	.096		.000		.054		.000	
x_9 with x_{10}		.126		.000		.096		.000
x_{15} with x_{16}		.096		.000		.054		.000

Notes: [a] The model fit for the population model without misspecification is perfect by definition: $\chi^2(99) = 0.00$; [b] The χ^2-test for the misspecified model with moderate loadings is non-significant even for the largest sample size used in the simulation study (N=750): $\chi^2(103) = 51.40$; Comparative Fit Index = 0.97; Root Mean Square Error of Approximation = 0.026; Standardized Root Mean Square Residual = 0.017. [c] The χ^2-test for the misspecified model with large loadings is non-significant even for the largest sample size used in the simulation study (N=750): $\chi^2(103)= 51.35$; Comparative Fit Index = 0.99; Root Mean Square Error of Approximation = 0.026; Standardized Root Mean Square Residual = 0.012.

Table 3: Population Loadings for the Orthogonal Two-Factor Models
(Completely Standardized Solution)

	Moderate Loadings				Large Loadings			
	Without Model Misspecification [a]		With Model Misspecification [b]		Without Model Misspecification [a]		With Model Misspecification [c]	
	F1	F2	F1	F2	F1	F2	F1	F2
x_1	.400	-	**.491**	-	.60	-	**.668**	-
x_2	.400	-	**.491**	-	.60	-	**.668**	-
x_3	.400	-	**.491**	-	.60	-	**.668**	-
x_4	.400	-	**.491**	-	.60	-	**.668**	-
x_5	.600	-	.622	-	.80	-	.826	-
x_6	.600	-	.622	-	.80	-	.826	-
x_7	.600	-	.622	-	.80	-	.826	-
x_8	.600	-	.622	-	.80	-	.826	-
x_9	-	.400	-	.384	-	.60	-	.569
x_{10}	-	.400	-	.384	-	.60	-	.569
x_{11}	-	.400	-	.384	-	.60	-	.569
x_{12}	-	.400	-	.384	-	.60	-	.569
x_{13}	-	.600	-	**.535**	-	.80	-	**.765**
x_{14}	-	.600	-	**.535**	-	.80	-	**.765**
x_{15}	-	.600	-	**.535**	-	.80	-	**.765**
x_{16}	-	.600	-	**.535**	-	.80	-	**.765**

Notes: [a] The model fit for the population model without misspecification is perfect by definition: $\chi^2(104) = 0.00$; [b] The χ^2-test for the misspecified model without sampling error and moderate loadings is non-significant even for the largest sample size used in the simulation study (N=750): $\chi^2(108) = 40.13$; Comparative Fit Index = 1.00; Root Mean Square Error of Approximation = 0.000; Standardized Root Mean Square Residual = 0.051. [c] The χ^2-test for the misspecified model without sampling error and large loadings is non-significant even for the largest sample size used in the simulation study (N=750): $\chi^2(108) = 38.37$; Comparative Fit Index = 1.00; Root Mean Square Error of Approximation = 0.000; Standardized Root Mean Square Residual = 0.085. The loadings resulting from an equality constraint are given in bold face. The values in brackets at the bottom of the Table are the differences between ρ^2 based on the unbiased loadings and the corresponding ρ^2 based on the biased loadings from the misspecified model.

Table 4: Population Loadings for the Oblique Two-Factor Models
(Completely Standardized Solution)

	Moderate Loadings				Large Loadings			
	Without Model Misspecification [a]		With Model Misspecification [b]		Without Model Misspecification [a]		With Model Misspecification [c]	
	F1	F2	F1	F2	F1	F2	F1	F2
x_1	.400	-	**.491**	-	.60	-	**.668**	-
x_2	.400	-	**.491**	-	.60	-	**.668**	-
x_3	.400	-	**.491**	-	.60	-	**.668**	-
x_4	.400	-	**.491**	-	.60	-	**.668**	-
x_5	.600	-	.620	-	.80	-	.825	-
x_6	.600	-	.620	-	.80	-	.825	-
x_7	.600	-	.620	-	.80	-	.825	-
x_8	.600	-	.620	-	.80	-	.825	-
x_9	-	.400	-	.385	-	.60	-	.570
x_{10}	-	.400	-	.385	-	.60	-	.570
x_{11}	-	.400	-	.385	-	.60	-	.570
x_{12}	-	.400	-	.385	-	.60	-	.570
x_{13}	-	.600	-	**.534**	-	.80	-	**.765**
x_{14}	-	.600	-	**.534**	-	.80	-	**.765**
x_{15}	-	.600	-	**.534**	-	.80	-	**.765**
x_{16}	-	.600	-	**.534**	-	.80	-	**.765**
Interfactor-Correlation	.300		.295		.300		.293	

Notes: [a] The model fit for the population model without misspecification is perfect by definition: $\chi^2(103) = 0.00$; [b] The χ^2-test for the misspecified model without sampling error and moderate loadings is non-significant even for the largest sample size used in the simulation study (N=750): $\chi^2(107) = 41.53$; Comparative Fit Index = 1.00; Root Mean Square Error of Approximation = .000; Standardized Root Mean Square Residual = 0.050. [c] The χ^2-test for the misspecified model without sampling error and large loadings is non-significant even for the largest sample size used in the simulation study (N=750): $\chi^2(107) = 40.66$; Comparative Fit Index = 1.00; Root Mean Square Error of Approximation = 0.000; Standardized Root Mean Square Residual = 0.084. The loadings resulting from an equality constraint are given in bold face. The values in brackets at the bottom of the Table are the differences between ρ^2 based on the unbiased loadings and the corresponding ρ^2 based on the biased loadings from the misspecified model.

Generation of Populations of Cases

In order to generate populations of cases corresponding to the population correlation matrices implied by the correctly specified population models, four population data sets of variables each containing normally distributed, z-standardized random numbers for 375,000 cases were computed and aggregated with SPSS Version 14.

The first set of 375,000 cases was computed for the orthogonal models with correlated residuals and the second set was computed for the oblique models with correlated residuals. The third set was computed for the orthogonal models without correlated residuals and the fourth set for the oblique models without correlated residuals. In all population data sets, the random variables were orthogonalized by means of principal component analysis with subsequent Varimax-rotation before aggregation in order to exclude that even small sampling errors might affect the population parameters.

Eight orthogonal variables were fixed as orthogonal population factor scores f_i for the orthogonal models, 64 orthogonal variables were fixed as residual or error variances e_j and 16 variables were fixed as common variables c_k representing the correlated residuals. From these orthogonal random variables eight correlated variables per factor were generated. The generation of the variables x_1 and x_2 for the orthogonal models with moderate factor loadings can be described by means of

$$x_j = .40^{0.5} f_i + .60^{0.5}(.85e_j + .15c_k), \quad \text{for } i = 1; j = 1, 2, k = 1. \tag{6}$$

As observed from equation 6, variables x_1 and x_2 share the common variable c_1 and therefore have correlated residuals (the error term is in brackets). Moreover, the weights in equation 6 correspond to the square-root of the (moderate) factor loadings presented in Table 1. Thus, the population loadings presented in Table 1 are the (squared) weights for the aggregation of the population factor scores in order to compute the population variables. The corresponding weights of the population residuals were computed from the communalities (h^2) by means of $w = (1 - h^2)^{0.5}$; because each variable x_j has only one non-zero population loading on one factor f_i, the weight for f_i in equation 6 represents h, the square-root of the communality. Accordingly, the weight w for the residual in equation 6 was calculated as $w = (1 - (0.40^{0.5})^2)^{0.5} = 0.60^{0.5}$. The generation of the variables x_3 and x_4 without correlated residuals can be described by means of

$$x_j = 0.40^{0.5} f_i + 0.60^{0.5} e_j, \quad \text{for } i = 1; j = 3, 4. \tag{7}$$

The equation for the generation of the variables x_5 and x_6 is

$$x_j = 0.60^{0.5} f_i + 0.40^{0.5} e_j, \quad \text{for } i = 1; j = 5, 6; \tag{8}$$

and the equation for the variables x_7 and x_8 is

$$x_j = 0.60^{0.5} f_i + 0.40^{0.5}(.85e_j + .15c_k), \quad \text{for } i = 1; j = 7, 8, k = 2. \tag{9}$$

Equations 6-9 describe the generation of the eight variables loading on the first factor (see Table 1). The equations for the remaining variables loading on factors 2-8 contain the same weights (and different subscripts) and are therefore not presented here. By this procedure 64 variables with moderate loadings on eight factors were generated. The equations describing the generation of variables with large loadings on orthogonal factors and variables with correlated residuals are

$$x_j = 0.60^{0.5} f_i + 0.40^{0.5}(0.85e_j + 0.15c_k), \quad \text{for } i = 1; j = 1, 2, k = 1, \tag{10}$$

$$x_j = 0.60^{0.5} f_i + 0.40^{0.5} e_j, \quad \text{for } i = 1; j = 3, 4, \tag{11}$$

$$x_j = 0.80^{0.5} f_i + 0.20^{0.5} e_j, \quad \text{for } i = 1; j = 5, 6, \tag{12}$$

and

$$x_j = 0.80^{0.5} f_i + 0.20^{0.5}(0.85e_j + 0.15c_k), \quad \text{for } i = 1; j = 7, 8, k = 2. \tag{13}$$

For the oblique models correlated factor scores were computed by means of aggregation of orthogonal random variables. The computation of the eight oblique population factor scores o_i from the z-standardized random variables z_i and

a z-standardized common random variable v can be described as

$$o_i = 0.30^{0.5}\, v + 0.70^{0.5}\, z_i, \qquad \text{for } i = 1 \text{ to } 8. \tag{14}$$

Eight oblique population factor scores were computed as a basis for the oblique two-, four- and eight-factor models. It follows from equation 14 that the interfactor-correlations were 0.30 in the population, according to the weight of the common variable v (see Beauducel & Wittmann, 2005 for more details on the aggregation of random variables). The oblique factor scores o_i were inserted instead of the orthogonal factor scores f_i into equations 6-9 in order to generate the variables for the oblique factor models with moderate loadings and correlated residuals and in equations 10-13 in order to generate the variables for the oblique models with large loadings and correlated residuals.

The two-factor models were based on o_1 and o_2, the four-factor models on o_1-o_4, and the eight factor models on o_1-o_8. For the orthogonal models without correlated residuals, the 64 variables were generated only on the basis of f_i and e_i, without the common terms c_k, so that the equations for the models contained only the weights as in equations 7, 8, 11 and 12 (see Table 3, for the corresponding loadings). For the oblique models without correlated residuals the equations were based on the random variables o_i and e_i and they had also the same weights as equations 7, 8, 11 and 12 (see Table 4, for the corresponding loadings).

Subsamples of variables were analyzed for the two- and four-factor models. The two-factor models were based on the variables x_1-x_{16} (see Table 1), the four-factor models were based on the variables x_1-x_{32} and the eight-factor models were based on the 64 variables. The two types of models and their corresponding misspecifications (omitted correlations between residuals, specification of equal loadings) were analyzed separately, in order to allow for a separate interpretation of the results.

For the analysis of the correctly and misspecified models based on population data with correlated residuals, the results from the population data sets 1 and 2 were combined in order to allow for a combined analysis of orthogonal and oblique models. The conditions for this analysis were computation method of indeterminacy (according to equations 4 and 5), orthogonality (orthogonal versus oblique), number of factors (2, 4 and 8 factors), loading size (moderate versus large loadings), and number of cases or sample size (250, 500 and 750 cases).

For each of these 36 conditions 500 samples were analyzed by means of CFA so that the first simulation study was based on 18,000 samples. For each sample one CFA with correct model specification and one CFA with incorrect model specification was performed. For analysis of the correctly and misspecified models based on population data without correlated residuals, the population data sets 3 and 4 were combined in order to allow for a combined analysis of orthogonal and oblique models. The conditions (computation method, orthogonality, number of factors, loading size and number of cases) were exactly as in the analysis of the models with correlated residuals.

For the correctly specified models, the difference between the population ρ^2 of the correctly specified models and the samples ρ^2 of the corresponding correctly specified models (same number of factors, same loading size, etc.) was calculated and averaged across factors.

For the misspecified models, the difference between the population ρ^2 of the misspecified models and the samples ρ^2 of the corresponding misspecified models (same number of factors, same loading size, etc.) was calculated and averaged across factors. The ρ^2-differences were calculated for both computation methods (see equation 4 and 5) and entered into repeated measures ANOVA.

In order to limit the results to those that are interesting in the present context, only main-effects and interactions involving the factor Computation-method are reported. Due to the very large sample size (6,000 cases) all reported effects were significant at $p < 0.001$ and only effects with large effect sizes (partial $\eta^2 > 0.20$) are reported. The effect sizes of the within-subjects effects were based on Greenhouse-Geisser corrected univariate effects.

Results

Table 5 contains the mean coefficients of indeterminacy for the different population models. The coefficients of indeterminacy were averaged for the factors with odd and even numbers, because the model misspecification based on equality constraints imposed on the loading pattern had different effects on factors with odd and even numbers. The coefficients of indeterminacy were different for the correctly and the misspecified population models (see Table 5).

For the population models based on correlated residuals the coefficients of indeterminacy were larger for all misspecified models than for the correctly specified models. For these models, the effect of misspecification on ρ^2 was identical for factors with odd and even numbers. For the models without correlated residuals, the effects of model-misspecification on ρ^2 were different for factors with odd and even numbers: For factors with odd numbers ρ^2 was larger than in the correctly specified models and for factors with even numbers ρ^2 was smaller than in the correctly specified models. Overall, the population models show some variation of ρ^2, which might be regarded as a basis for an investigation of ρ^2 in the samples.

The differences between the population ρ^2 and the corresponding samples ρ^2 for the models based on correlated residuals were entered into a repeated measures ANOVA with Computation method (two levels, based on equations 4 and 5), Misspecification (correctly specified versus misspecified) and Number of factors (three levels) as within-subjects factors and Number of cases (three levels), Loading-size (two levels), and Obliqueness (orthogonal versus oblique) as between subjects factors. Misspecification was considered as within-subjects factor, because the same data sets were used for the correctly specified models and for the misspecified models. It was decided to consider Number of factors as within-subjects factor, because the four-factor models include the two factors of the two-factor models and the eight-factor models include the four factors of the four-factors model. A large main effect

Table 5: Mean population ρ^2 for the Two Different Calculation Methods

			According To Equation 4		According To Equation 5	
Model Type	Loading Size	Specification	Odd Factors	Even Factors	Odd Factors	Even Factors
With Correlated Residuals	.40	Correctly Specified	.738	.738	.751	.751
		Misspecified	.761	.761	.761	.761
	.60	Correctly Specified	.897	.897	.903	.903
		Misspecified	.906	.906	.906	.906
Without Correlated Residuals	.40	Correctly Specified	.751	.751	.751	.751
		Misspecified	.791	.697	.904	.623
	.60	Correctly Specified	.903	.903	.903	.903
		Misspecified	.922	.883	1.011	.823

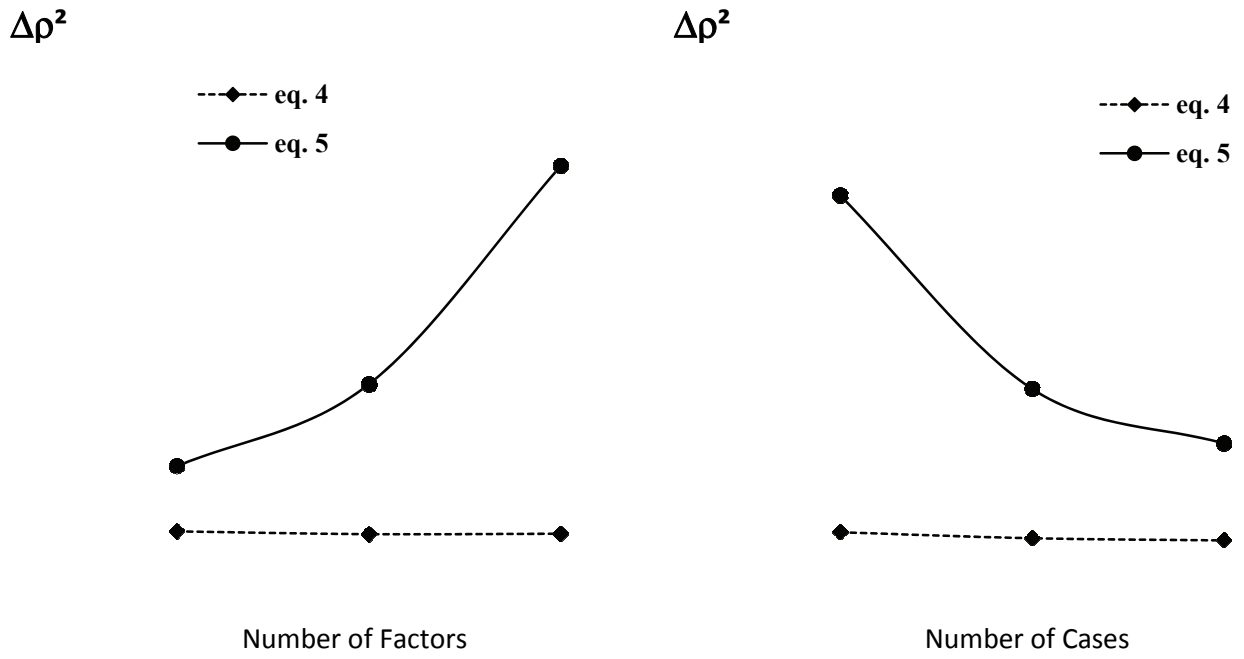
Notes: The column odd factors contains the mean ρ^2 for the factors with odd numbers, the column even factors contains the mean ρ^2 for the factors with even numbers.

occurred for Computation method (η^2= 0.94). The mean ρ^2-difference was 0.081 (SD= 0.068) when based on equation 4 and 0.171 (SD= 0.094) when based on equation 5. Thus, the mean difference between ρ^2 in the population and in the samples was about twice as large when it was based on equation 5. This indicates that the empirical covariance matrix (used in equation 5) introduces a substantial amount of sampling error into ρ^2.

A large effect size occurred for the interaction between computation method and number of factors (η^2= 0.94). This interaction is mainly due to a larger increase of the ρ^2-difference with number of factors when ρ^2 is computed according to equation 5 (see Figure 1a). Another large effect size occurs for the interaction of computation method and number of cases (η^2= 0.81). This interaction is mainly due to a larger increase of the ρ^2-difference with Number of cases when ρ^2 is computed according to equation 5 (see Figure 1b). Moreover, a large three-way interaction computation method x number of factors x number of cases occurred (η^2= 0.83). This three-way-interaction occurs because the size of the two-way interaction computation method x Number of factors is larger for the small samples (250 cases) than for the large samples (750 cases). In fact, the mean difference between the ρ^2-differences for the two computation methods is only 0.018 for the two-factor models based on 750 cases and it is 0.304 for the eight-factor models based on 250 cases.

Finally, the interaction of computation method with Obliqueness is of relevant size (η^2= 0.43). The difference between the computation methods is smaller for the orthogonal models than for the oblique models. Although there is a substantial main effect for misspecification (η^2= 0.47), the size of the interaction between computation method and misspecification is moderate (η^2= 0.16) and the interaction is extremely small in terms of mean differences: The difference between the ρ^2-differences for the two computation methods is 0.092 for the correctly specified models and it is 0.089 for the misspecified models; thus, misspecification had no relevant effect on the difference between the computation methods.

Figure 1: ρ^2-Differences for the Two Computation Methods Based on the Data Sets with Correlated Residuals: a) for 2-, 4-, and 8-factor models; b) for 250, 500, and 750 cases



The differences between the population ρ^2 and the corresponding samples ρ^2 based on the models without correlated residuals were entered into a repeated measures ANOVA with the same factors as the ρ^2-differences for the models based on correlated residuals. Again, a large main effect occurred for computation method ($\eta^2 = 0.97$), indicating that the mean ρ^2-difference between population and sample ρ^2 was considerably smaller when ρ^2 was computed according to equation 4.

The mean ρ^2-difference was only 0.01 (SD = 0.01) when ρ^2 was computed according to equation 4 and it was 0.14 (SD = 0.09) when ρ^2 was computed according to equation 5. A substantial interaction of computation method with number of factors occurred ($\eta^2 = 0.97$). An inspection of this interaction reveals that the computation methods had similar ρ^2-differences for the two-factor models, but that the computation method based on equation 5 yielded much larger ρ^2-differences in the eight-factor models (see Figure 2a). Another substantial interaction occurred for computation method and number of cases ($\eta^2= 0.77$), indicating that the ρ^2-differences increased more with decreasing sample size when ρ^2 was computed according to equation 5 (see Figure 2b).

The effect size of the three-way interaction Computation method x Number of factors x Number of cases was also substantial ($\eta^2= 0.83$). This relation of Number of factors and Number of cases with the Computation method can be described by the following result: The mean ρ^2-differences were rather similar for both Computation methods when based on the two-factor models with 750 cases (their difference was 0.033). The mean differences were, however, very different for the computation methods when based on the eight-factor models with 250 cases (their difference was 0.333). The ρ^2-differences based on equation 5 were larger than the ρ^2-differences based on equation 4 when the size of the loadings was larger (Computation method x Loading-size; $\eta^2 = 0.59$). The ρ^2-differences based on equation 5 were also larger than the ρ^2-differences based on equation 4 for orthogonal models than for oblique models (Computation method x Obliqueness; $\eta^2 = 0.92$). The effect of model misspecification on the ρ^2-differences for the two methods was, however, moderate ($\eta^2 = 0.17$). For the correctly specified models the difference between the computation methods was slightly larger (0.125) than for the misspecified models (0.122).

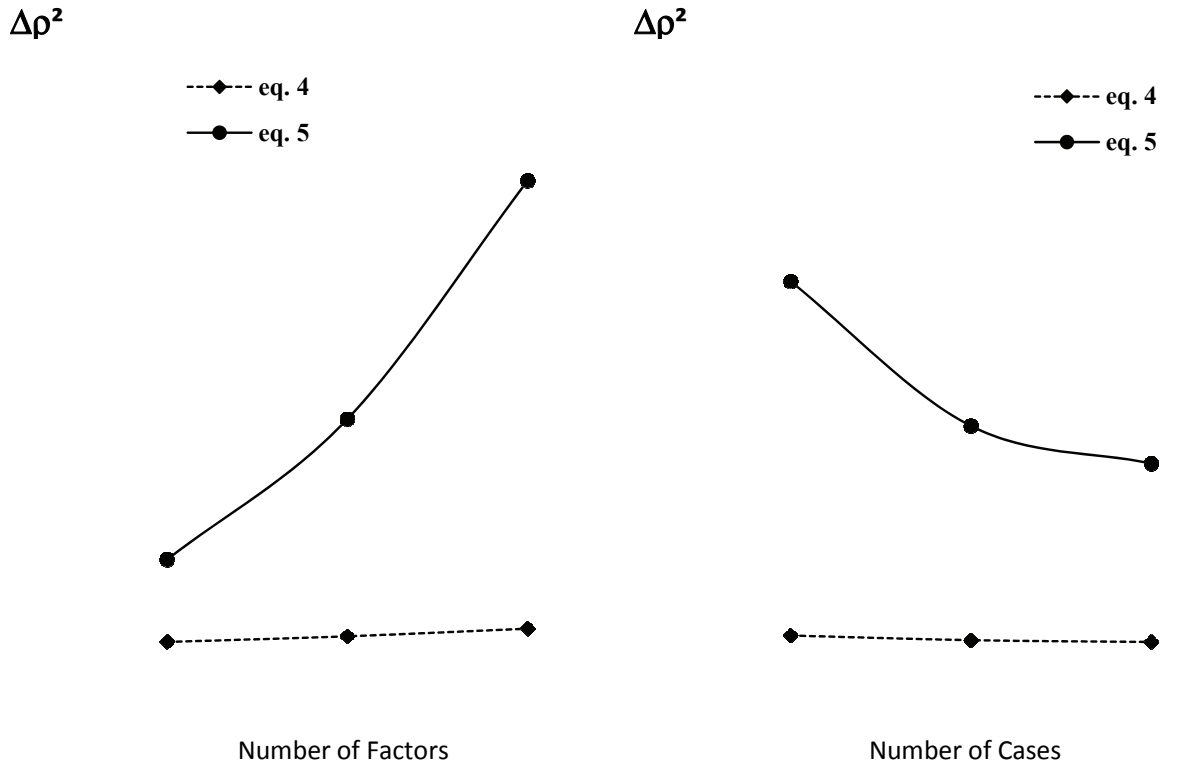
Conclusion

This study compared two calculation methods of the indeterminacy coefficient ρ^2 (or ρ) that allows for the evaluation of factor score estimates. Thereby it should be investigated which method should be preferred when a CFA model is slightly misspecified, as is often the case. Therefore, the two calculation methods for indeterminacy were compared in correctly and misspecified CFA models.

Correctly specified and misspecified models based on data sets with correlated residuals as well as on data sets without correlated residuals were investigated. For the models based on data sets with correlated residuals, the correlated residuals were not specified in order to generate misspecified models in addition to the correctly specified models. For the models based on data sets without correlated residuals misspecified models were generated by means of equality constraints imposed on unequal loadings.

Two computation methods for coefficients of indeterminacy were investigated: The first method is based on the correlations or covariances of the observed variables reproduced from the model (equation 4), the second method (equation 5) is based on the empirical correlations or covariances of the observed variables. Because both the computation of ρ^2 by means of the reproduced covariance matrix (McDonald, 1974; Mulaik & McDonald, 1978) and the computation of ρ^2 by means of the sample covariance matrix (Gorsuch, 1983; Grice, 2001; Heermann, 1963) have been proposed, an investigation of the differences between these methods was regarded as important. Moreover, in case of model misspecification, it is clear that the covariance matrix reproduced from the model ($\mathbf{\Sigma}$) contains some error. The errors due to model

Figure 2: ρ²-Differences for the Two Computation Methods Based on the Data Sets Without Correlated Residuals: a) for 2-, 4-, and 8-factor models; b) for 250, 500, and 750 cases



misspecification are not present in the empirical covariance matrix (**S**), so that the computation based on **S** might have been expected to work well for misspecified models. Therefore, the two computation methods were investigated both in correctly as well as in misspecified models. However, the model misspecifications were moderate in order to represent models that might be accepted according to conventional fit criteria (Hu & Bentler, 1999). The reason for the investigation of models with small amounts of misspecification was that this allows some insight into the effects of model misspecification on ρ² that might occur in empirical research with a given amount of accepted misfit. Sample size (250, 500, 750 cases), number of factors (2, 4, 8 factors), obliqueness (orthogonal versus correlated factors), and size of salient loadings (0.40/0.60 versus 0.60/0.80) were manipulated in the simulation study. The main limitations of the present simulation study are that only two types of model misspecification were explored and that the effects of severe model misspecification were not investigated. Nevertheless, the results of the simulation study shed some light on the effects of sampling error on ρ² for different types of correctly and misspecified CFA models.

The difference between ρ² computed from the population and the samples was substantially smaller when ρ² was computed according to equation 4 (as can be seen from the main effect of Computation method). This result can be interpreted as a larger effect of sampling error on ρ² when computed according to equation 5, as might be expected from using the sample covariance matrix **S** in equation 5 instead of the population covariance matrix **Σ**.

The interpretation that the use of **S** for the computation of ρ² introduces some sampling error into the coefficient is also supported by the interaction of computation method with sample size, indicating that the difference between the population ρ² and the sample ρ² was larger for smaller sample sizes, especially when ρ² was

computed according to equation 5 (based on **S**). Even in the misspecified models, when **Σ** suffers from the misspecification, due to its being reproduced from the (misspecified) model parameters, the mean differences between the populations ρ^2 and the samples ρ^2 was smaller when ρ^2 was computed on the basis of **Σ** (equation 4).

Although the model misspecifications used in the present study were not very large, it is still possible that advantages of using **S** for the computation of ρ^2 (equation 5) might occur for extreme amounts of model misspecification. On the other hand, it seems rather unlikely that severely misspecified models would generally be accepted according to fit indexes and it might be regarded as problematic to base the results of a simulation study on models that should not occur in empirical research. The results of the present study are therefore taken as support for a computation of ρ^2 by means of the reproduced correlation or covariance matrix (equation 4). Moreover, it was found for the population models that effects of misspecification can result in serious over-estimation of ρ^2, so that the validity of factor score predictors might be over-estimated, just because the respective models were incorrectly specified.

Nevertheless, the effect of sampling error and model misspecification on ρ^2 found in this study should not discourage researchers to report indeterminacy coefficients when factor score estimates are computed from CFA models. It is necessary to report indeterminacy coefficients – otherwise the validity of the factor score estimates remains unknown. Of course, indeterminacy coefficients might be even more biased than reported here when a model is more seriously misspecified; the case of extreme misspecification was not investigated in this study because factor score estimates should not at all be computed for seriously misspecified CFA models, thus the question of the validity of such scores is irrelevant.

References

Barrett, P. (2007). Structural equation modeling: Adjudging model fit. *Personality and Individual Differences*, *42*, 815-824.

Beauducel, A., & Rabe, S. (2009). Model-related factor score estimates for confirmatory factor analysis. *British Journal of Mathematical and Statistical Psychology*, *62*, 489-506.

Beauducel, A., & Wittmann, W. W. (2005). Simulation study on fit indices in confirmatory factor analysis based on data with slightly distorted simple structure. *Structural Equation Modeling*, *12*, 41-75.

Fan, X., & Sivo, S. A. (2007). Sensitivity of fit indices to model misspecification and model types. *Multivariate Behavioral Research*, *42*, 509-529.

Gorsuch, R. L. (1983). *Factor analysis* (*2nd Ed.*). Hillsdale, NJ: Lawrence Erlbaum Associates.

Green, B. F. (1976). On the factor score controversy. *Psychometrika*, *41*, 263-266.

Guttman, L. (1955). The determinacy of factor score matrices with applications for five other problems of common factor theory. *British Journal of Mathematical and Statistical Psychology*, *8,* 65-82.

Grice, J. W. (2001). Computing and evaluation of factor scores. *Psychological Methods*, *6*, 430-450.

Heermann, E. F. (1963). Univocal or orthogonal estimators of orthogonal factors. *Psychometrika*, *28*, 161-172.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*, 1-55.

Krijnen, W. P. (2006). Some results on mean square error for factor score prediction. *Psychometrika*, *71*, 395-409.

Krijnen, W. P., Wansbeek, T., & Ten Berge, J. M. F. (1996). Best linear predictors for factor scores. *Communications in Statistics: Theory and Methods*, *25*, 3013-3025.

Marsh, H. W., Hau, K.-T., & Wen, Z. (2004). In search of golden rules: comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Structural Equation Modeling*, *11*, 320–341.

McDonald, R. P., & Burr, E. J. (1967). A comparison of four methods of constructing factor scores. *Psychometrika*, *32*, 381-401.

McDonald, R. P. (1974). The measurement of factor indeterminacy. *Psychometrika*, *39*, 203-222.

Mulaik, S., & McDonald, R. P. (1978). The effect of additional variables on factor indeterminacy in models with a single common factor. *Psychometrika*, *43*, 177-192.

Muthén, L. K., & Muthén, B. O. (2004). *Mplus user's guide (Version 3.11).* Los Angeles: Author.

Schönemann, P. H. (1971). The minimum average correlation between equivalent sets of uncorrelated factors. *Psychometrika*, *36*, 21-30.

Schönemann, P. H., & Steiger, J. H. (1976). Regression component analysis. *British Journal of Mathematical and Statistical Psychology*, *29*, 175-189.

SPSS. (2005). *SPSS for Windows, Release 14.0.0.* Chicago: Author.

Thurstone, L. L. (1935). *The vectors of mind.* Chicago: University of Chicago Press.

Journal of Modern Applied Statistical Methods
November 2011, Vol. 10, No. 2, 599-606

1538 – 9472/11/$95.00

Maximum Log Likelihood Estimation using EM Algorithm and Partition Maximum Log Likelihood Estimation for Mixtures of Generalized Lambda Distributions

Steve Su
University of Western Australia,
Perth, Australia
Covance Pty Ltd, Sydney, Australia

Two mixture distribution fitting methods based on maximizing the likelihood using generalized lambda distributions are presented. The fitting algorithms are demonstrated on various data and the strengths and weakness of the algorithms which can influence their use under different mixture modeling situations are discussed. The procedures described are available in GLDEX package in R.

Key words: Fitting distributions, prior distributions, empirical data analysis, mixture distributions, generalized lambda distributions.

Introduction

Mixture distribution modeling is a substantial area of interest among statisticians; many works regarding fitting mixtures have appeared in the literature. Böhning and Seidel (2003) discussed the general strategy used in confronting various problems associated with mixture distribution modeling. Although there are generic works, such as finding initial values to ensure better optimization of the mixture fitting scheme (Karlis & Xekalaki, 2003) and finding the optimal number of components of mixtures (Miloslavsky & van der Laan, 2003), no work has been presented on using mixtures of the generalized Lambda distributions to fit multi-modal data. This is an important development because the use of generalized Lambda distributions has advantages over traditional distributions such as Normal, Weibull and Exponential in the sense that they have overwhelmingly rich shapes and can handle a wide range of different data sets (Freimer, et al., 1988; Karian & Dudewicz, 2000; Okur, 1988; Su, 2010a, 2010b, 2005, 2007a, 2007b). Fitting a mixture of generalized Lambda distributions can therefore be very beneficial because it is much more efficient to fit distributions to data using a smaller range of distributions rather than choosing and comparing across a wide range of different combination of distributions.

Though generalized Lambda distributions are flexible their uses are not as widespread; this may be due to the fact that these distributions are only explicitly defined by quantiles, thus, extensive numerical methods are required to perform standard calculations, such as finding the probability under the curve. As computing power continues to grow, maximum likelihood estimations conducted numerically may become more popular. This article discusses two different ways of fitting mixtures using generalized Lambda distributions (GλDs).

Steve Su is affiliated with School of Mathematics and Statistics at University of Western Australia and Covance Pty Ltd, Sydney, Australia. Email him at: allegro.su@gmail.com.

Methodology

The Ramberg-Schmeiser (1974) (RS) GλD is an extension of Tukey's Lambda distribution (Hastings, Mosteller, Tukey & Windsor 1947). It is defined by its inverse distribution function:

$$F^{-1}(u) = \lambda_1 + \frac{u^{\lambda_3} - (1-u)^{\lambda_4}}{\lambda_2} \qquad (1)$$

In (1), $0 \le u \le 1$, $\lambda_2 \neq 0$ and λ_1 ,λ_2, λ_3, λ_4 are respectively the location, inverse scale and shape parameters of the generalized Lambda distribution GλD(λ_1 ,λ_2, λ_3, λ_4). Karian, Dudewicz and MacDonald (1996) noted that GλD is defined if and only if:

$$\frac{\lambda_2}{\lambda_3 u^{\lambda_3 - 1} + \lambda_4 (1-u)^{\lambda_4 - 1}} \geq 0 \text{ for } u \in [0,1]. \tag{2}$$

Another distribution known as FKML GλD also exists (Freimer, Kollia, Mudholkar, & Lin, 1988). The FKML GλD can be written as:

$$F^{-1}(u) = \lambda_1 + \frac{\frac{u^{\lambda_3} - 1}{\lambda_3} - \frac{(1-u)^{\lambda_4} - 1}{\lambda_4}}{\lambda_2} \tag{3}$$

Under (2), $0 \le u \le 1$ and λ_1 ,λ_2, λ_3, λ_4 are consistent with the interpretations in RS GλD, namely λ_1 ,λ_2 are the location and inverse scale parameters and λ_3, λ_4 are the shape parameters.

The fundamental motivation for the development of FKML GλD is that the distribution is proper over all λ_3 and λ_4 (Freimer Mudholkar, Kollia & Lin, 1988). The only restriction on FKML GλD is that $\lambda_2 > 0$.

The most commonly used technique in mixture distributional fitting is maximum likelihood estimation. This is usually achieved by using the EM algorithm for explicitly defined probability functions such as the Normal, Gamma and Exponential. In the case of implicitly defined distributions such as the GλDs, it is possible to use two ways of estimating the parameters of the mixtures, the maximum likelihood estimation using the EM algorithm and the partitioned maximum likelihood method which utilizes the complete data log likelihood. Both methods are discussed below.

GλDs Fitting Mixture Algorithm

The fitting of mixture of two GλDs is completed using the following algorithm:

Step 1

Divide the data into two parts. This can be done using a variety of clustering methods. Practical experience has shown that clustering methods such as Clara and Fanny described in Kaufman and Rousseeuw (1990) worked well in a wide range of situations. However, any clustering method that gives a reasonable classification can be used. This step provides a starting value for p in the mixture distribution equation $pf_1+(1-p)f_2$, which will be optimized later. The Clara clustering method appears to work well for a wide variety of empirical data and all fitting results in this article uses this clustering method.

To maximize the partition log likelihood this is all that is required. In the case of maximizing the log likelihood using EM algorithm, each partition of the data set additionally contains the maximum and minimum values of the entire data set as well as 1% (it is often worthwhile to explore different percentages to obtain better initial values for the maximum likelihood fitting scheme) of randomly selected data from the other group.

For example, if data sets 1 and 2 both have 100 observations, data set 1 will contain 102 observations, including 1 observation randomly selected from data set 2 and 1 maximum value from data set 2 (if it was not selected already), assuming data set 1 already contains the minimum value of the original data set. This is to ensure that the partitioned data span the entire range of the data; a necessary step because the goal is to maximize the log likelihood for the mixture data

Step 2

For each part of the data, fit a statistical distribution using maximum likelihood estimation (Su 2007a, Su 2007b).

Step 3

After the distribution fits for both parts of the data are obtained, the final parameters are estimated by maximizing the appropriate formula in (4) (for partition maximum likelihood) or (5) (for the EM algorithm approach). The initial value of p comes from step 1 and the initial values for this stage of the optimization are from step 2. The maximization

is conducted numerically via the Nelder-Mead Simplex algorithm and only solutions that span the entire original data set are accepted. The formulae required in this maximization step are discussed below.

Let X, Z be the complete data, with X~ $f_1(x,\theta)$ if z = 0 and X ~ $f_2(x,\theta)$ if z=1, Then, the complete data log likelihood is given by:

$$l_c(\theta,p) = \sum_{i=1}^{n}(1-z)\begin{Bmatrix}\log(f_1(x_i,\boldsymbol{\theta_1}))+\\ \log(p)\end{Bmatrix}+z\begin{Bmatrix}\log(f_2(x_i,\boldsymbol{\theta_2}))+\\ \log(1-p)\end{Bmatrix} \quad (4)$$

Using standard statistical calculations, the conditional expectation of $l_c(\theta, p)$ given x is:

$$\sum_{i=1}^{n} T_i\begin{Bmatrix}\log(f_1(x_i,\boldsymbol{\theta_1}))\\ +\log(p)\end{Bmatrix}+S_i\begin{Bmatrix}\log(f_2(x_i,\boldsymbol{\theta_2}))\\ +\log(1-p)\end{Bmatrix} \quad (5)$$

and

$$S_i = \frac{f_2(x_i,\theta_2)(1-p)}{f_2(x_i,\theta_2)(1-p)+f_1(x_i,\theta_1)(p)} \quad (6)$$
$$1-S_i = T_i$$

where f_1 and f_2 are GλD distributions fitted to each partition of the data set and θ_1 and θ_2 representing the parameters associated with these distributions respectively. In the case of two RS GλDs mixture fits, for example, equation (4) becomes:

$$\left(\sum_{i=1}^{n_1}\log(p)+\log\left[\frac{\lambda_2}{\lambda_3 u_i^{\lambda_3-1}+\lambda_4(1-u_i)^{\lambda_4-1}}\right]\right)$$
$$+\left(\sum_{j=1}^{n_2}\log(1-p)+\log\left[\frac{\delta_2}{\delta_3 v_i^{\delta_3-1}+\delta_4(1-v_i)^{\delta_4-1}}\right]\right),$$

with $n_1 + n_2 = n$. Here the n_1 and n_2 are the number of observations in each partition of the data set and the δ_k for k = 1, 2, 3, 4 represents the parameters of the second GλD fit, similarly u_i and v_i represents the quantiles for each partition of the data set for the i^{th} observation.

All other combinations of different RS and FKML GλD fits for complete data log likelihood and maximum likelihood via EM algorithm can be found by substituting the required GλD into (4) or (5) and hence are not detailed herein.

Step 4

The parameters obtained in step 3 are then used to maximize (7). The results of this optimization process are the final parameters for the GλD mixture fits. This step was omitted in Su (2007a) but subsequent updates to the GLDEX package in R, by default, has added this optimization step for both partition and full maximum likelihood methods.

$$\sum_{i=1}^{n}\log(p(f_1(x_i,\boldsymbol{\theta_1}))+(1\text{-}p)(f_2(x_i,\boldsymbol{\theta_2}))) \quad (7)$$

Step 5

The final fitting result can be examined by plotting the result on the histogram with the fitted line, quantile plots as well as testing the goodness of fit using the Kolmogorov-Smirnov (KS) test. A two sample KS test is carried out by sampling 90% of the empirical data from the actual distributions and this is compared to equal number of data from the corresponding fitted distributions. This is repeated 1,000 times with the result of this test being the number of times the p-value exceeds 0.05 (or at a specified significance level) over 1,000 times. This will give the user an independent measure as to the adequacy of fits beyond a visual comparison.

Although this study is focused on fitting two mixtures of GλD, fitting three or more mixtures of GλD is a straightforward extension. In the case of three mixtures, it is possible to divide the data into three partitions, apply maximum likelihood estimation to each partition to find the initial values and maximize the following partition maximum likelihood or EM maximum likelihood formulae to find the parameters of the mixture distribution. To achieve this, let X, Z again be the complete data and X~$f_j(\mathbf{x},\boldsymbol{\theta})$ if z_j = 1, with j = 0, 1, 2. The proportion of the data in f_j are represented by p_j. The complete data likelihood or partition

maximum likelihood is given in (8) and the conditional expectation of complete data log likelihood given **x** is given in (9).

$$l_c(\boldsymbol{\theta},\mathbf{p}) = \sum_{i=1}^{n} z_0\{\log(f_0(x_i,\boldsymbol{\theta_0})) + \log(p_0)\} + z_1\{\log(f_1(x_i,\boldsymbol{\theta_1})) + \log(p_1)\} + z_2\{\log(f_2(x_i,\boldsymbol{\theta_2})) + \log(p_2)\} \quad (8)$$

$$\sum_{i=1}^{n} \frac{f_0(x_i,\boldsymbol{\theta_0})(p_0)}{w_i}\{\log(f_0(x_i,\boldsymbol{\theta_0})) + \log(p_0)\} + \frac{f_1(x_i,\boldsymbol{\theta_1})(p_1)}{w_i}\log(f_1(x_i,\boldsymbol{\theta_1})) + \log(p_1)\} + \frac{f_2(x_i,\boldsymbol{\theta_2})(p_2)}{w_i}\log(f_2(x_i,\boldsymbol{\theta_2})) + \log(p_2)\}$$
$$w_i = f_0(x_i,\boldsymbol{\theta_0})(p_0) + f_1(x_i,\boldsymbol{\theta_1})(p_1) + f_2(x_i,\boldsymbol{\theta_2})(p_2) \quad (9)$$

Based on the parameters obtained in maximizing (8) or (9), the last step of the optimization is to maximize (10), this gives the final parameters of the mixture distribution fit.

$$\sum_{i=1}^{n} \log(p_0(f_0(x_i,\boldsymbol{\theta_0})) + p_1(f_1(x_i,\boldsymbol{\theta_1})) + p_2(f_2(x_i,\boldsymbol{\theta_2}))) \quad (10)$$

The development of partition maximum likelihood method and maximum likelihood via EM algorithm is intended to cover two different types of modeling situations. The first situation is when two distributions are distinct and disjoint, in which partition maximum likelihood would be the method of choice. The second situation is where two distributions overlap with each other in which the full maximum likelihood would be more preferable. However, this does not preclude the use of either methods in any given situation and the choice of one method over the other could still be based on more objective measures such as KS test and QQ plots.

The method presented here and in Su (2007a, Su 2007b) optimizes the maximum likelihood directly rather than use the usual method of differentiation. This is a much more efficient and reliable method of achieving the maximum likelihood rather than differentiating and solving a system of linear equations because in many cases, GλD may be undefined for certain parameter values, rendering the technique of differentiation useless. Hence, it is usually preferable to use a general purpose optimization scheme such as the Nelder-Simplex algorithm to fit GλDs.

Results

The effectiveness of using the algorithm described earlier to fit mixture of two and three generalized lambda distributions to a range of simulated and empirical data are now illustrated. The graphical displays of resulting fits are shown in Figures 1 and 2, and the numerical goodness of fit assessments are shown in Tables 1 and 2. Partition maximum likelihood method and maximum likelihood method using the EM algorithm are abbreviated as PML and ML in the outputs respectively.

In Figure 1, data set 1 is generated by 70% of Normal (mean = 10, standard deviation = 3) and 30% of exponential distributions. Data set 4 is generated by 50% of double exponential and 50% of Normal (mean = 5, standard deviation = 2) distributions. Both data sets 1 and 4 consist of 1,000 observations. Data sets 2, 3 and 5 are various data collected from the internet by the author and consist of 72, 244 and 272 observations, respectively. The data illustrated in Figure 2 is a relatively well known galaxy of white dwarf stars and consists of 7,140 observations. Numerical summaries of these data are provided in Tables 1 and 2.

The QQ plots in Figure 1 indicates that the algorithm using either partition or full maximum likelihood are convincing fits to the empirical data, this is supported by the high values indicated by the KS tests and in many cases, the theoretical moments of the fitted GλDs are quite close to the empirical data. In particular, Figure 1b demonstrates the type of distributional fits expected from using partition maximum likelihood methods; there is a tendency for the method to make a sharper split between the two data. This is reinforced in the comparison between Figure 1d and 1e, where a

more abrupt separation of the two data sets can be observed in 1e using the partition maximum likelihood method. It is, however, not always true that the partition maximum likelihood will result in a jagged distributional shape; as Figure 1f shows, the resulting fit is smooth.

Overall, both methods of fitting mixtures provide a good fit to a range of data and it is recommended to examine both methods in most cases. For example, it may be preferable (due to closer match of to the moments of data and better KS test results) to use partition maximum likelihood with user defined setting for data in Figure 2, but the maximum likelihood using EM algorithm is preferred for data set 4. Clearly, no one fitting method will work the best in every case, so the choice of different methods is important to allow users to cope with different data with different tools. Sensitivity analysis using different distributional fits may also be carried out, to examine the robustness of a particular strategy under different representations of a probability distribution.

In many situations, the default setting of the GLDEX package works well. However, as known in mixture distribution modeling, the choice of initial values can have a large impact on the resulting fits. This is clearly demonstrated in Figure 2, where the default separation of the data into three parts using Clara classification scheme failed to give a very convincing fits as indicated in Figure 2a and 2b. The use of a user defined clustering regime in identifying the sub distributions (data < 100, data between 100 to 300, data > 300) leads to superior fits as shown in Figure 2c and 2d and the partition maximum likelihood with user defined data split is remarkably close to the first four moments of the empirical data.

Conclusion

This article demonstrates an algorithm to fit mixtures using the GλD distribution family. An important advantage of using GλD distribution is the elimination of the type of distributions that need to be used to model multi modal data. A critical improvement needed for all fitting methods of GλD is the search of suitable initial values. Although a fairly robust approach is provided here and in Su (2010b, 2007a, 2007b), it may be possible to directly find a set of good initial values from empirical data to speed up the optimization process and to increase the prospect of reaching a global maximum.

References

Böhning, D., & Seidel, W. (2003). Recent developments in mixture models. *Computational Statistics and Data Analysis*, *41*, 349-357.

Freimer, M., Kollia, G., Mudholkar, G., & Lin C. (1988). A study of the generalized Tukey lambda family. *Communications in Statistics- Theory and Methods*, *17*(*10*), 3547-3567.

Hastings, J. C., Mosteller, F., Tukey, J., & Windsor C. (1947). Low moments for small samples: A comparative study of order statistics. *The Annals of Statistics*, *18*, 413-426.

Karian, Z., & Dudewicz, E. (2000). *Fitting statistical distributions: The generalized lambda distribution and generalized bootstrap methods*. New York, NY: Chapman and Hall.

Karian, Z., Dudewicz, E., & McDonald, P. (1996). The extended generalized lambda distribution systems for fitting distributions to data: History, completion of theory, tables, applications, the final word on moment fits. *Communications in Statistics-Computation and Simulation*, *25*(*3*), 611-642.

Karlis, D., & Xekalaki E. (2003). Choosing initial values for the EM algorithm for finite mixtures. *Computational Statistics and Data Analysis*, *41*, 577-590.

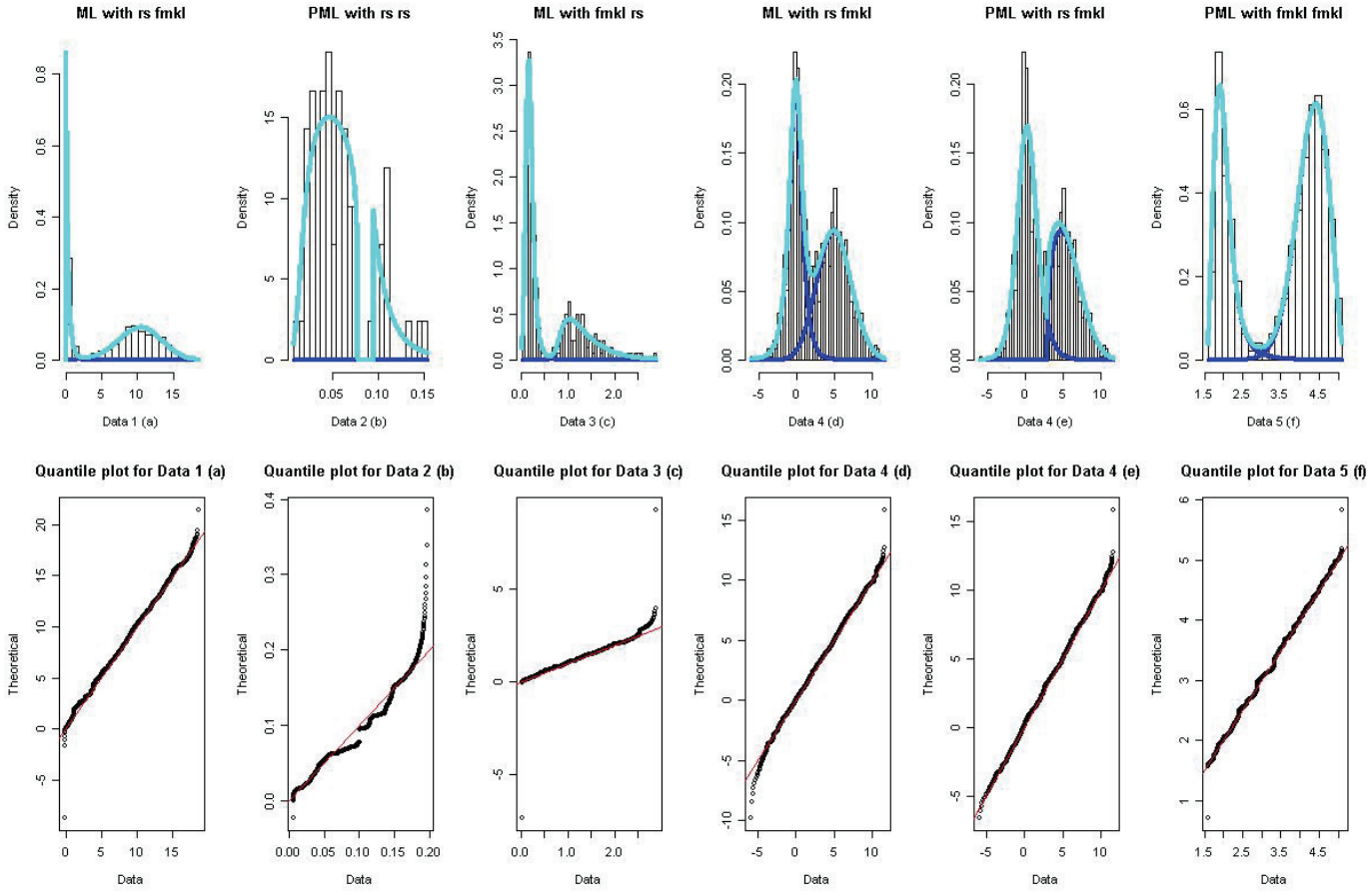
Kaufman, L., & Rousseeuw, P. J. (1990). *Finding groups in data: An introduction to cluster analysis*. New York, NY: Wiley.

Miloslavsky, M., & van der Laan, M. (2003). Fitting of mixtures with unspecified number of components using cross validation distance estimate. *Computational Statistics and Data Analysis*, *41*, 413-428.

Okur M., (1988). On fitting the generalised Lambda distribution to air pollution data. *Atmospheric Environment*, *22*, 2569-2572.

Ramberg, J., & Schmeiser, B. (1974). An approximate method for generating asymmetric random variables. *Communications of the Association for Computing Machinery*, *17*, 78-82.

Figure 1: Examples of Fitting Bimodal Data with a Mixture of Two Generalized Lambda Distributions



Su, S. (2010a). Fitting GLD to data via quantile matching method. In *Handbook of distribution fitting methods with R*, Z. Karian, & E. Dudewicz, Eds., 1171-1205. Boca Raton: CRC Press/Taylor & Francis.

Su, S. (2010b). Fitting gld to data using the GLDEX 1.0.4 in R. In *Handbook of distribution fitting methods with R*, Z. Karian, & E. Dudewicz, Eds., 585-608. Boca Raton: CRC Press/Taylor & Francis.

Su, S. (2005). A discretized approach to flexibly fit generalized lambda distributions to data. *Journal of Modern Applied Statistical Methods*, *4*, 408-424.

Su, S. (2007a). Fitting single and mixture of generalized lambda distributions to data via discretized and maximum likelihood methods: GLDEX in R. *Journal of Statistical Software*, *21*(*9*), 1-17.

Su, S. (2007b). Numerical maximum log likelihood estimation for generalized lambda distributions. *Computational Statistics and Data Analysis*, *51*(*8*), 3983-3998.

Table 1: Numerical Results Indicating Goodness of Fit In Terms of First Four Moments and Resample KS Tests for Figure 1

	Data 1	(a)	Data 2	(b)	Data 3	(c)	Data 4	(d)	(e)	Data 5	(f)
Mean	7.23	7.28	0.06	0.06	0.62	0.63	2.56	2.56	2.62	3.49	3.49
Variance	26.89	26.76	0.00	0.00	0.39	0.39	10.01	10.07	9.87	1.30	1.30
Skewness	-0.17	-0.20	1.09	1.76	1.19	1.21	0.36	0.33	0.28	-0.42	-0.41
Kurtosis	1.70	1.69	3.77	12.87	3.60	3.89	2.24	2.30	2054.78	1.50	2.11
Number of times KS test p value > 0.05 out of 1,000		912		949		948		985	833		943

Figure 2: Examples of Fitting Trimodal Data with a Mixture of Three Generalized Lambda Distributions (This example illustrates how splitting data manually can improve the fit beyond the default settings.)

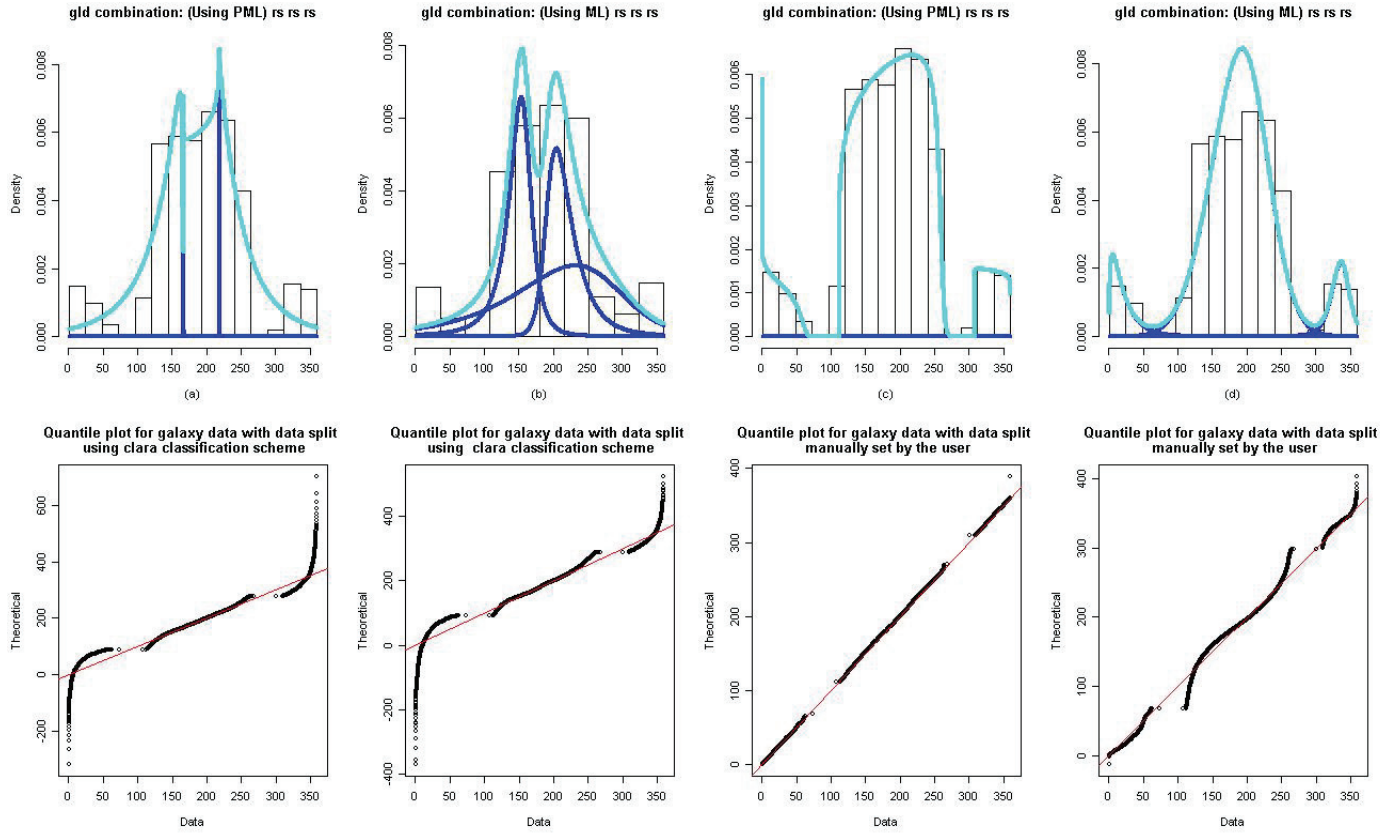


Table 2: Numerical Results Indicating Goodness of Fit In Terms of First Four Moments and Resample KS Tests for Figure 2

	Data	PML Using Clara Scheme	ML Using Clara Scheme	PML with Manual Setting	ML with Manual Setting
Mean	187.78	187.82	188.06	188.32	187.69
Variance	4870.03	5110.28	5665.51	4868.24	4946.95
Skewness	-0.18	-0.09	-4.02	-0.20	2.29
Kurtosis	3.85	7.32	NA	3.87	-1112094.77
Number of times KS test p value > 0.05 out of 1,000		850	769	938	317

Journal of Modern Applied Statistical Methods
November 2011, Vol. 10, No. 2, 607-617

1538 – 9472/11/$95.00

On Maximum Likelihood Estimators of the Parameters of a Modified Weibull Distribution Using Extreme Ranked Set Sampling

Amer Ibrahim Al-Omari
Al al-Bayt University,
Mafraq, Jordan

Said Ali Al-Hadhrami
College of Applied Sciences,
Nizwa, Oman

Extreme ranked set sampling (ERSS) is considered to estimate the three parameters and population mean of the modified Weibull distribution (MWD). The maximum likelihood estimator (MLE) is investigated and compared to the corresponding one based on simple random sampling (SRS). It is found that, the MLE based on ERSS is more efficient than MLE using SRS for estimating the three parameters of the MWD. The ERSS estimator of the population mean of the MWD is also found to be more efficient than the SRS based on the same number of measured units.

Key words: Modified Weibull distribution, extreme ranked set sampling, maximum likelihood estimator, simple random sampling, information number.

Introduction

The modified Weibull distribution (MWD) was suggested by Sarhan and Zaindin (2009). The probability density function (pdf) of the MWD is given by

$$f(x;\alpha,\beta,\gamma) = \left(\alpha+\beta\gamma x^{\gamma-1}\right)\exp\left(-\alpha x-\beta x^{\gamma}\right), \quad x>0, \tag{1}$$

and the corresponding distribution function (cdf) is

$$F(x;\alpha,\beta,\gamma) = 1-\exp\left(-\alpha x-\beta x^{\gamma}\right),\ x>0, \tag{2}$$

where $\gamma>0$ and $\alpha,\ \beta\geq 0$ such that $\alpha+\beta>0$. The MWD have two shape parameters γ and β, and a scale parameter α. The hazard function of the MWD is

$$h(x;\alpha,\beta,\gamma) = \alpha+\beta\gamma x^{\gamma-1}, \tag{3}$$

which increases for $\gamma>1$, decreases for $\gamma<1$ and remains constant for $\gamma=1$. Sarhan and Zaindin (2009) defined the k^{th} moment, μ_k, of the MWD random variable as

$$\mu_k = \begin{cases} \sum_{i=0}^{\infty}\frac{(-\beta)^i}{i!}\left(\frac{\Gamma(i\gamma+k+1)}{\alpha^{i\gamma+k}}+\frac{\beta\gamma\Gamma(i\gamma+\gamma+k)}{\alpha^{i\gamma+\gamma+k}}\right) & \text{if } \alpha,\beta>0, \\ \frac{\Gamma(k/\gamma+1)}{\beta^{k/\gamma}} & \text{if } \alpha=0,\beta>0, \\ \frac{\Gamma(k+1)}{\beta^{k}} & \text{if } \alpha>0,\beta=0. \end{cases} \tag{4}$$

The moment generating function of the MWD is given by

Amer Ibrahim Al-Omari is an Assistant Professor in the Department of Mathematics. Email him at: alomari_amer@yahoo.com. Said Ali Al-Hadhrami is an Assistant Professor in the Department of Mathematics, College of Applied Sciences. Email him at: abur1972@yahoo.co.uk.

$$M(t)=\begin{cases}\sum_{i=0}^{\infty}\frac{(-\beta)^i}{i!}\left(\frac{\alpha\Gamma(i\gamma+1)}{(\alpha-t)^{i\gamma+1}}+\frac{\beta\gamma\Gamma(i\gamma+\gamma)}{(\alpha-t)^{i\gamma+\gamma}}\right) & \text{if } \alpha,\beta>0,\alpha>t,\\ \sum_{i=0}^{\infty}\frac{t^i\Gamma(i/\gamma+1)}{\beta^{i/\gamma}} & \text{if } \alpha=0,\beta>0,\\ \frac{\alpha}{\alpha-t} & \text{if } \alpha>0,\beta=0,\alpha>t.\end{cases} \quad (5)$$

Some special cases of the MWD distribution are the exponential distribution, Raleigh distribution, linear failure rate distribution and Weibull distribution. For additional details about the MWD see: Sarhan & Zaindin (2009) and Zaindin & Sarhan (2009). The maximum likelihood estimator of the three parameters and the population mean of the modified Weibull distribution is examined, and compared to their counterparts based on simple random sampling. The MLE of the parameters based on ERSS is considered for two cases: when the set size is even and odd.

RSS and ERSS

Ranked set sampling (RSS) was proposed by McIntyre (1952) to improve the estimation of the population mean. The following steps are employed to obtain an RSS of size *m*:

Step 1: Randomly select m^2 units from the population; these units are randomly allocated into *m* sets, each of size *m*.

Step2: The *m* units of each set are ranked either visually or by any inexpensive method with respect to the variable of interest.

Step3: From the first set of *m* units, the smallest ranked unit is measured; from the second set of *m* units the second smallest ranked unit is measured. The process continued until the m^{th} smallest unit (largest) is measured from the last set.

Step 4: The procedure can be repeated *n* times if needed to increase the sample size to *nm* units.

It should be noted that the error in ranking reduces the efficiency of the method. Extreme ranked set sampling was proposed by Samawi, et al. (1996) as a useful modification of RSS. It requires identifying the extreme units only, as opposed to all ranks as in the usual RSS. The method gives an unbiased estimate of the population mean in the case of symmetric distributions and it is more efficient than SRS.

The extreme ranked set sampling (ERSS) method can be described as follows:

Step 1: Select *m* random samples each of size *m* units from the target population.

Step 2: Rank the units within each sample with respect to a variable of interest by visual inspection or any other inexpensive method.

Step 3: For actual measurement, if the sample size *m* is even, from the first $\frac{m}{2}$ sets select the lowest ranked unit of each set and from the other $\frac{m}{2}$ sets select the largest ranked unit. If the sample size is odd, from the first $\frac{m-1}{2}$ sets select the lowest ranked unit, from the other $\frac{m-1}{2}$ sets select the largest ranked unit, and from the remaining set the median ranked unit is selected.

Step 4: The procedure can be repeated *n* times if needed to increase the sample size to *nm* units.

Let $X_1,\ X_2\ \ldots, X_m$ be a simple random sample from the probability density function $f(x)$, with mean μ and variance σ^2. Let $X_{11},\ X_{12},\ldots,X_{1m};\ X_{21},X_{22},\ldots,X_{2m};\ldots;$

X_{m1}, X_{m2}, ..., X_{mm} be m independent SRS each of size m. Let $X_{i(1)}, X_{i(2)}$, ..., $X_{i(m)}$ be the order statistics of the sample X_{i1}, X_{i2}, ..., X_{im} for $(i=1,2,...,m)$. The pdf and cdf of the i^{th} order statistics, $X_{(i)}$, respectively are

$$f_{(i)}(x)=\frac{m!}{(i-1)!(m-i)!}\left[F(x)\right]^{i-1}\left[1-F(x)\right]^{m-i}f(x),$$

and

$$F_{(i)}(x)=\frac{m!}{(i-1)!(m-i)!}\int_0^{F(x)}v^{i-1}(1-v)^{m-i}dv.$$

The mean and the variance of $X_{(i)}$ are given by

$$\mu_{(i)}=\int_{-\infty}^{\infty}xf_{(i)}(x)dx$$

and

$$\sigma^2_{(i)}=\int_{-\infty}^{\infty}\left(x-\mu_{(i)}\right)^2 f_{(i)}(x)dx,$$

respectively (see David and Nagaraja, 2003). Takahasi and Wakimoto (1968) provided the mathematical properties of the RSS and gave the following identities

$$f(x)=\frac{1}{m}\sum_{i=1}^{m}f_{(i)}(x),\ \mu=\frac{1}{m}\sum_{i=1}^{m}\mu_{(i)},$$

and

$$\mathrm{Var}\left(\hat{\mu}_{RSS}\right)=\frac{\sigma^2}{m}-\frac{1}{m^2}\sum_{i=1}^{m}\left(\mu_{(i)}-\mu\right)^2.$$

They showed that the efficiency of RSS with respect to SRS is

$$1\le eff\left(\hat{\mu}_{RSS},\hat{\mu}_{SRS}\right)=\frac{\mathrm{Var}\left(\hat{\mu}_{SRS}\right)}{\mathrm{Var}\left(\hat{\mu}_{RSS}\right)}\le\frac{m+1}{2},$$

where $\hat{\mu}_{SRS}$ and $\hat{\mu}_{RSS}$ are unbiased estimators of the population mean μ using SRS and RSS, respectively.

When m is even, the ERSS estimator of the population mean is defined as

$$\hat{\mu}_{ERSS1}=\frac{1}{rm}\sum_{j=1}^{r}\left(\sum_{i=1}^{m/2}X_{(m)i,j}+\sum_{i=m/2+1}^{m}X_{(1)i,j}\right), \tag{6}$$

and when m is odd

$$\hat{\mu}_{ERSS2}=\frac{1}{rm}\sum_{j=1}^{r}\left(\begin{array}{l}\sum_{i=1}^{(m-1)/2}X_{(m)i,j}+\sum_{i=(m+1)/2}^{m-1}X_{(1)i,j}\\ +X_{((m+1)/2)i,j}\end{array}\right), \tag{7}$$

where $X_{(k)i,j}$ denotes the k^{th} ranked from the i^{th} set at the j_{th} cycle.

Samawi, et al. (1996) showed that the sample mean using ERSS is more efficient than that of SRS when the distribution is symmetric. Samawi and Al-Sagheer (2001) investigated the ERSS method to estimate the distribution function and Muttlak (2001) considered regression estimation using extreme and median ranked set samples methods. Samawi and Saeid (2004) studied the stratified ERSS and the ratio estimator based on ERSS. Al-Omari, et al. (2008) considered ratio type estimator based on ERSS. For more about RSS and its modifications see: Arnold, et al. (2009); Al-Omari & Jaber (2008); Bouza (2009); Shadid, et al. (2011); Al-Hadhrami & Al-Omari (2009); Islam, et al. (2009); Jemain & Al-Omari (2006); Sengupta & Mukhuti (2009).

Maximum Likelihood Estimation of the MWD: When m is Even

The maximum likelihood estimators (MLEs) of the three estimators α, β and γ when m is even are investigated based on the likelihood function L using ERSS as

$$L=\left(\begin{array}{l}h\prod_{i=1}^{r}\prod_{j=1}^{p}\left\{mf\left(x_{(m)i,j}\right)\left[F\left(x_{(m)i,j}\right)\right]^{m-1}\right\}\\ \times\prod_{i=1}^{r}\prod_{j=p+1}^{m}\left\{mf\left(x_{(1)i,j}\right)\left[1-F\left(x_{(1)i,j}\right)\right]^{m-1}\right\}\end{array}\right), \tag{8}$$

where $p = m/2$ and h is a constant. The variable $X_{(k)i,j}$ denotes the k^{th} ranked unit of the i^{th} sample at the j^{th} cycle. The log likelihood function of (8) is

$$\begin{aligned} L^* = C &+ \sum_{j=1}^{r}\sum_{i=1}^{p} \ln f\left(x_{(m)i,j}\right) \\ &+ (m-1)\sum_{j=1}^{r}\sum_{i=1}^{p} \ln F\left(x_{(m)i,j}\right) \\ &+ \sum_{j=1}^{r}\sum_{i=p+1}^{m} \ln f\left(x_{(1)i,j}\right) \\ &+ (m-1)\sum_{j=1}^{r}\sum_{i=p+1}^{m} \ln\left(1-F\left(x_{(1)i,j}\right)\right), \end{aligned} \tag{9}$$

where C is a constant. The first derivatives of L^* with respect to α, β and γ, respectively are

$$\begin{aligned} \frac{\partial L^*}{\partial \alpha} = &\sum_{j=1}^{r}\sum_{i=1}^{p}\left(\frac{1}{\alpha+\beta\gamma x_{(m)i,j}^{\gamma-1}} - x_{(m)i,j}\right) \\ &+ (m-1)\sum_{j=1}^{r}\sum_{i=1}^{p}\left(\frac{x_{(m)i,j}T_1}{1-T_1}\right) \\ &- (m-1)\sum_{j=1}^{r}\sum_{i=p+1}^{m} x_{(1)i,j} \\ &+ \sum_{j=1}^{r}\sum_{i=p+1}^{m}\left(\frac{1}{\alpha+\beta\gamma x_{(1)i,j}^{\gamma-1}} - x_{(1)i,j}\right), \end{aligned} \tag{10}$$

$$\begin{aligned} \frac{\partial L^*}{\partial \beta} = &\sum_{j=1}^{r}\sum_{i=1}^{p}\left(\frac{\gamma x_{(m)i,j}^{\gamma-1}}{\alpha+\beta\gamma x_{(m)i,j}^{\gamma-1}} - x_{(m)i,j}^{\gamma}\right) \\ &+ (m-1)\sum_{j=1}^{r}\sum_{i=1}^{p}\left(\frac{x_{(m)i,j}^{\gamma}T_1}{1-T_1}\right) \\ &- (m-1)\sum_{j=1}^{r}\sum_{i=p+1}^{m} x_{(1)i,j}^{\gamma} \\ &+ \sum_{j=1}^{r}\sum_{i=p+1}^{m}\left(\frac{\gamma x_{(1)i,j}^{\gamma-1}}{\alpha+\beta\gamma x_{(1)i,j}^{\gamma-1}} - x_{(1)i,j}^{\gamma}\right), \end{aligned} \tag{11}$$

$$\begin{aligned} \frac{\partial L^*}{\partial \gamma} = &\sum_{j=1}^{r}\sum_{i=1}^{p}\begin{pmatrix}\dfrac{\beta x_{(m)i,j}^{\gamma-1}\left(\gamma\ln\left(x_{(m)i,j}\right)+1\right)}{\alpha+\beta\gamma x_{(m)i,j}^{\gamma-1}} \\ -\beta\ln\left(x_{(m)i,j}\right)x_{(m)i,j}^{\gamma}\end{pmatrix} \\ &+ \sum_{j=1}^{r}\sum_{i=p+1}^{m}\begin{pmatrix}\dfrac{\beta x_{(1)i,j}^{\gamma-1}\left(\gamma\ln\left(x_{(1)i,j}\right)+1\right)}{\alpha+\beta\gamma x_{(1)i,j}^{\gamma-1}} \\ -\beta\ln\left(x_{(1)i,j}\right)x_{(1)i,j}^{\gamma}\end{pmatrix} \\ &+ (m-1)\sum_{j=1}^{r}\sum_{i=1}^{p}\begin{pmatrix}\dfrac{\beta\ln\left(x_{(m)i,j}\right)x_{(m)i,j}^{\gamma}T_1}{1-T_1} \\ -(m-1)\displaystyle\sum_{j=1}^{r}\sum_{i=p+1}^{m}\beta\ln\left(x_{(1)i,j}\right)x_{(1)i,j}^{\gamma}\end{pmatrix}, \end{aligned} \tag{12}$$

where

$$T_1 = \exp\left(-\alpha x_{(m)i,j} - \beta x_{(m)i,j}^{\gamma-1}\right)$$

and

$$T_2 = \exp\left(-\alpha x_{(1)i,j} - \beta x_{(1)i,j}^{\gamma}\right).$$

The MLE of the parameters $\alpha, \beta,$ and γ are the solution of equations (10), (11) and (12), respectively, when set them to zero. However, the solutions are not in closed forms, in order to obtain estimates for the parameters, the three equations may be solved numerically.

Fisher information (FI) numbers describe the amount of information that a sample provides about the parameters. The FI is defined as

$$I = -E\left(\frac{\partial^2 \log(L)}{\partial\theta^2}\right),$$

where θ is a parameter. The FI number from ERSS for estimating α, β and γ can be expressed as in equations, (13), (14) and (15), respectively as

$$I_{ERSS}(\alpha)=$$
$$-E\left\{\sum_{j=1}^{r}\sum_{i=1}^{p}\left[\frac{-1}{\left(\alpha+\beta\gamma x_{(m)i,j}^{\gamma-1}\right)^2}\right]+(m-1)\sum_{j=1}^{r}\sum_{i=1}^{p}\left(\frac{-x_{(m)i,j}^2T_1}{1-T_1}\right)\right.$$
$$\left.-(m-1)\sum_{j=1}^{r}\sum_{i=1}^{p}\left[\frac{\left(x_{(m)i,j}T_1\right)^2}{\left(1-T_1\right)^2}\right]+\sum_{j=1}^{r}\sum_{i=p+1}^{m}\left[\frac{-1}{\left(\alpha+\beta\gamma x_{(1)i,j}^{\gamma-1}\right)^2}\right]\right\},\tag{13}$$

$$I_{ERSS}(\beta)=-E\left\{\sum_{j=1}^{r}\sum_{i=1}^{p}-\left[\frac{\gamma x_{(m)i,j}^{\gamma-1}}{\alpha+\beta\gamma x_{(m)i,j}^{\gamma-1}}\right]^2\right.$$
$$-\sum_{j=1}^{r}\sum_{i=p+1}^{m}\left[\frac{\gamma x_{(1)i,j}^{\gamma-1}}{\alpha+\beta\gamma x_{(1)i,j}^{\gamma-1}}\right]^2$$
$$\left.-(m-1)\sum_{j=1}^{r}\sum_{i=1}^{p}\left[\frac{\left(x_{(m)i,j}^{\gamma}\right)^2T_1}{1-T_1}+\left(\frac{x_{(m)i,j}^{\gamma}T_1}{1-T_1}\right)^2\right]\right\},\tag{14}$$

and

$$I_{ERSS}(\beta)=$$
$$-E\left\{\sum_{j=1}^{r}\sum_{i=1}^{p}\left[\begin{array}{l}\dfrac{\partial^2}{\partial\gamma^2}\log\left(f\left(x_{(m)i,j}\right)\right)\\+(m-1)\dfrac{\partial^2}{\partial\gamma^2}\log\left(F\left(x_{(m)i,j}\right)\right)\end{array}\right]\right.$$
$$\left.-\sum_{j=1}^{r}\sum_{i=1}^{p}\left[\begin{array}{l}\dfrac{\partial^2}{\partial\gamma^2}\log\left(f\left(x_{(1)i,j}\right)\right)\\+(m-1)\dfrac{\partial^2}{\partial\gamma^2}\log\left[1-F\left(x_{(1)i,j}\right)\right]\end{array}\right]\right\},\tag{15}$$

where

$$\frac{\partial^2}{\partial\gamma^2}\log\left[f\left(x_{(.)i,j}\right)\right]=$$
$$\frac{\left[\beta x_{(m)i,j}^{\gamma-1}\ln\left(x_{(m)i,j}\right)\right]\left[\gamma\ln\left(x_{(m)i,j}\right)+2\right]}{\alpha+\beta\gamma x_{(m)i,j}^{\gamma-1}}$$
$$-\left\{\frac{\beta x_{(m)i,j}^{\gamma-1}\left[\gamma\ln\left(x_{(m)i,j}\right)+1\right]}{\alpha+\beta\gamma x_{(m)i,j}^{\gamma-1}}\right\}^2$$
$$-\beta x_{i,j}^{\gamma}\ln^2\left(x_{(m)i,j}\right),$$

$$\frac{\partial^2}{\partial\gamma^2}\log\left[F\left(x_{(.)i,j}\right)\right]=$$
$$\frac{\beta T_2x_{(.)i,j}^{\gamma}\ln^2\left(x_{(.)i,j}\right)\left(1-\beta x_{(.)i,j}^{\gamma}\right)}{1-T_2}$$
$$-\left[\frac{\beta\ln\left(x_{(.)i,j}\right)x_{(.)i,j}^{\gamma}T_1}{1-T_1}\right]^2$$

and

$$\frac{\partial^2}{\partial\gamma^2}\log\left[1-F\left(x_{(.)i,j}\right)\right]=-\beta x_{(.)i,j}^{\gamma}\ln^2\left(x_{(.)i,j}\right).$$

Maximum Likelihood Estimation of the MWD: When *m* is Odd

Based on ERSS, when *m* is odd, the likelihood function is

$$L=K\prod_{i=1}^{r}\prod_{j=1}^{q}\left(\begin{array}{l}\left\{mf\left(x_{(m)i,j}\right)\left[F\left(x_{(m)i,j}\right)\right]^{m-1}\right\}\\\prod_{i=1}^{r}\prod_{j=q+1}^{m-1}\left\{mf\left(x_{(1)i,j}\right)\left[1-F\left(x_{(1)i,j}\right)\right]^{m-1}\right\}\end{array}\right)$$
$$f\left(x_{((m+1)/2)j}\right)\left[F\left(x_{((m+1)/2)j}\right)\left(1-F\left(x_{((m+1)/2)j}\right)\right)\right]^{\frac{m-1}{2}}\tag{16}$$

where $q=(m-1)/2$ and K is a constant. The log likelihood function of (16) is

$$
\begin{aligned}
L^* &= Ln(L) \\
&= K^* + \sum_{j=1}^{r}\sum_{i=1}^{q} \ln f\left(x_{(m)i,j}\right) \\
&\quad + (m-1)\sum_{j=1}^{r}\sum_{i=1}^{q}\left[\ln F\left(x_{(m)i,j}\right)\right] \\
&\quad + \sum_{j=1}^{r}\sum_{i=q+1}^{m-1} \ln f\left(x_{(m)i,j}\right) \\
&\quad + (m-1)\sum_{j=1}^{r}\sum_{i=q+1}^{m-1}\left\{\ln\left[1-F\left(x_{(1)i,j}\right)\right]\right\} \\
&\quad + \log f\left(x_{(\frac{m+1}{2})j}\right) \\
&\quad + \frac{m-1}{2}\left[\log F\left(x_{(\frac{m+1}{2})j}\right) + \log\left(1-F(x_{(\frac{m+1}{2})j}\right)\right].
\end{aligned}
\tag{17}
$$

Taking the first derivative of L^* in (17) with respect to α, β and γ results in

$$
\begin{aligned}
\frac{\partial L^*}{\partial \alpha} &= \sum_{j=1}^{r}\sum_{i=1}^{q}\left(\frac{1}{\alpha+\beta\gamma x_{(m)i,j}^{\gamma-1}} - x_{(m)i,j}\right) \\
&\quad + \sum_{j=1}^{r}\sum_{i=q+1}^{m-1}\left(\frac{1}{\alpha+\beta\gamma x_{(m)i,j}^{\gamma-1}} - x_{(m)i,j}\right) \\
&\quad + (m-1)\sum_{j=1}^{r}\sum_{i=1}^{m-1}\frac{x_{(m)i,j}T_1}{1-T_1} \\
&\quad - (m-1)\sum_{j=1}^{r}\sum_{i=q+1}^{m-1} x_{(1)i,j} \\
&\quad + \sum_{j=1}^{r}\left(\frac{1}{\alpha+\beta\gamma x_{(\frac{m+1}{2})j}^{\gamma-1}} - x_{(\frac{m+1}{2})j}\right) \\
&\quad + \frac{m-1}{2}\sum_{j=1}^{r}\left(\frac{x_{(\frac{m+1}{2})j}T_3}{1-T_3} - x_{(\frac{m+1}{2})j}\right),
\end{aligned}
\tag{18}
$$

$$
\begin{aligned}
\frac{\partial L^*}{\partial \beta} &= \sum_{j=1}^{r}\sum_{i=1}^{q}\left(\frac{\gamma x_{(m)i,j}^{\gamma-1}}{\alpha+\beta\gamma x_{(m)i,j}^{\gamma-1}} - x_{(m)i,j}^{\gamma}\right) \\
&\quad + \sum_{j=1}^{r}\sum_{i=q+1}^{m-1}\left(\frac{\gamma x_{(1)i,j}^{\gamma-1}}{\alpha+\beta\gamma x_{(1)i,j}^{\gamma-1}} - x_{(1)i,j}^{\gamma}\right) \\
&\quad + (m-1)\sum_{j=1}^{r}\sum_{i=1}^{q}\left(\frac{x_{(m)i,j}^{\gamma}T_1}{1-T_1}\right) - (m-1)\sum_{j=1}^{r}\sum_{i=q+1}^{m-1} x_{(1)i,j}^{\gamma} \\
&\quad + \sum_{j=1}^{r}\left(\frac{\gamma x_{(\frac{m+1}{2})j}^{\gamma-1}}{\alpha+\beta\gamma x_{(\frac{m+1}{2})j}^{\gamma-1}} - x_{(\frac{m+1}{2})j}^{\gamma}\right) \\
&\quad + \frac{m-1}{2}\sum_{j=1}^{r}\left(\frac{x_{(\frac{m+1}{2})j}^{\gamma}T_3}{1-T_3} - x_{(\frac{m+1}{2})j}^{\gamma}\right),
\end{aligned}
\tag{19}
$$

and

$$
\begin{aligned}
\frac{\partial L^*}{\partial \gamma} &= \sum_{j=1}^{r}\sum_{i=1}^{q}\left[\frac{\beta x_{(m)i,j}^{\gamma-1}\left[\gamma\ln\left(x_{(m)i,j}\right)+1\right]}{\alpha+\beta\gamma x_{(m)i,j}^{\gamma-1}} - \beta\ln\left(x_{(m)i,j}\right)x_{(m)i,j}^{\gamma}\right] \\
&\quad + \sum_{j=1}^{r}\sum_{i=q+1}^{m-1}\left[\frac{\beta x_{(1)i,j}^{\gamma-1}\left[\gamma\ln\left(x_{(1)i,j}\right)+1\right]}{\alpha+\beta\gamma x_{(1)i,j}^{\gamma-1}} - \beta\ln\left(x_{(1)i,j}\right)x_{(1)i,j}^{\gamma}\right] \\
&\quad + (m-1)\sum_{j=1}^{r}\sum_{i=1}^{q}\left[\frac{\beta\ln\left(x_{(m)i,j}\right)x_{(m)i,j}^{\gamma}T_1}{1-T_1}\right] \\
&\quad - (m-1)\sum_{j=1}^{r}\sum_{i=1}^{q}\beta\ln\left(x_{(p+1)i,j}\right)x_{(p+1)i,j}^{\gamma} \\
&\quad + \sum_{j=1}^{r}\left[\frac{\beta x_{(\frac{m+1}{2})j}^{\gamma-1}\left[\gamma\ln\left(x_{(\frac{m+1}{2})j}\right)+1\right]}{\alpha+\beta\gamma x_{(\frac{m+1}{2})j}^{\gamma-1}} - \beta\ln\left(x_{(\frac{m+1}{2})j}\right)x_{(\frac{m+1}{2})j}^{\gamma}\right] \\
&\quad + \frac{m-1}{2}\sum_{j=1}^{r}\left[\frac{\beta\ln\left(x_{(\frac{m+1}{2})i,j}\right)x_{(\frac{m+1}{2})i,j}^{\gamma}T_3}{1-T_3} - \beta\ln\left(x_{(\frac{m+1}{2})i,j}\right)x_{(\frac{m+1}{2})i,j}^{\gamma}\right],
\end{aligned}
\tag{20}
$$

respectively, where

$$T_1 = \exp\left(-\alpha x_{(m)i,j} - \beta x_{(m)i,j}^{\gamma-1}\right),$$

$$T_2 = \exp\left(-\alpha x_{(1)i,j} - \beta x_{(1)i,j}^{\gamma}\right)$$

and

$$T_3 = \exp\left(-\alpha x_{((m+1)/2)i,j} - \beta x_{((m+1)/2)i,j}^{\gamma}\right).$$

The Fisher Information number of α, β and γ from the samples, respectively are

$$\begin{aligned} I_{ERSS}(\alpha) = -E\Bigg\{ & \sum_{j=1}^{r}\sum_{i=1}^{q}\frac{-1}{\left(\alpha+\beta\gamma x_{(m)i,j}^{\gamma-1}\right)^2} \\ & -(m-1)\sum_{j=1}^{r}\sum_{i=1}^{q}\left[\frac{x_{(m)i,j}^2 T_1}{1-T_1}+\frac{\left(x_{(m)i,j}T_1\right)^2}{\left(1-T_1\right)^2}\right] \\ & +\sum_{j=1}^{r}\sum_{i=q+1}^{m-1}\frac{-1}{\left(\alpha+\beta\gamma x_{(1)i,j}^{\gamma-1}\right)^2} \\ & +\sum_{j=1}^{r}\frac{-1}{\left(\alpha+\beta\gamma x_{(\frac{m+1}{2})j}^{\gamma-1}\right)^2} \\ & -\frac{m-1}{2}\sum_{j=1}^{r}\left[\frac{x_{(\frac{m+1}{2})j}^2 T_3}{1-T_3}+\left(\frac{x_{(\frac{m+1}{2})j}T_3}{1-T_3}\right)^2\right]\Bigg\}, \end{aligned} \tag{21}$$

$$\begin{aligned} I_{ERSS}(\beta) = -E\Bigg\{ & \sum_{j=1}^{r}\sum_{i=1}^{q}\left[\frac{-\left(\gamma x_{(m)i,j}^{\gamma-1}\right)^2}{\left(\alpha+\beta\gamma x_{(m)i,j}^{\gamma-1}\right)^2}\right] \\ & +\sum_{j=1}^{r}\sum_{i=q+1}^{m-1}\left[\frac{-\left(\gamma x_{(1)i,j}^{\gamma-1}\right)^2}{\left(\alpha+\beta\gamma x_{(1)i,j}^{\gamma-1}\right)^2}\right] \\ & -(m-1)\sum_{j=1}^{r}\sum_{i=1}^{q}\left[\frac{\left(x_{(m)i,j}^{\gamma}\right)^2 T_1}{1-T_1}+\left(\frac{x_{(m)i,j}^{\gamma}T_1}{1-T_1}\right)^2\right] \\ & -\sum_{j=1}^{r}\left(\frac{\gamma x_{(\frac{m+1}{2})j}^{\gamma-1}}{\alpha+\beta\gamma x_{(\frac{m+1}{2})j}^{\gamma-1}}\right)^2 \\ & -\frac{m-1}{2}\sum_{j=1}^{r}\left[\frac{\left(x_{(\frac{m+1}{2})j}^{\gamma}\right)^2 T_3}{1-T_3}+\left(\frac{x_{(\frac{m+1}{2})j}^{\gamma}T_3}{1-T_3}\right)^2\right]\Bigg\}, \end{aligned} \tag{22}$$

$$\begin{aligned} I_{ERSS}(\gamma) = & \\ & -E\Bigg\{\sum_{j=1}^{r}\sum_{i=1}^{q}\left[\begin{array}{l}\frac{\partial^2}{\partial\gamma^2}\log\left[f\left(x_{(m)i,j}\right)\right] \\ +(m-1)\frac{\partial^2}{\partial\gamma^2}\log\left[F\left(x_{(m)i,j}\right)\right]\end{array}\right] \\ & +\sum_{j=1}^{r}\sum_{i=1}^{q}\left[\begin{array}{l}\frac{\partial^2}{\partial\gamma^2}\log\left[f\left(x_{(1)i,j}\right)\right] \\ +(m-1)\frac{\partial^2}{\partial\gamma^2}\log\left[1-F\left(x_{(1)i,j}\right)\right]\end{array}\right] \\ & +\sum_{j=1}^{r}\frac{\partial^2}{\partial\gamma^2}\log\left[f\left(x_{(\frac{m+1}{2})i,j}\right)\right] \\ & +\frac{m-1}{2}\sum_{j=1}^{r}\frac{\partial^2}{\partial\gamma^2}\log\left[F\left(x_{(\frac{m+1}{2})i,j}\right)\right] \\ & +\frac{\partial^2}{\partial\gamma^2}\log\left[1-F\left(x_{(\frac{m+1}{2})i,j}\right)\right]\Bigg\}, \end{aligned} \tag{23}$$

where

$$\frac{\partial^2}{\partial\gamma^2}\log\left[f\left(x_{(.)i,j}\right)\right]=$$
$$\frac{\left[\beta x_{(.)i,j}^{\gamma-1}\ln\left(x_{(.)i,j}\right)\right]\left[\gamma\ln\left(x_{(.)i,j}\right)+2\right]}{\alpha+\beta\gamma x_{(.)i,j}^{\gamma-1}}$$
$$-\left[\frac{\beta x_{(.)i,j}^{\gamma-1}\left(\gamma\ln\left(x_{(.)i,j}\right)+1\right)}{\alpha+\beta\gamma x_{(.)i,j}^{\gamma-1}}\right]^2-\beta x_{i,j}^{\gamma}\ln^2\left(x_{(.)i,j}\right),$$

$$\frac{\partial^2}{\partial\gamma^2}\log\left[F\left(x_{(.)i,j}\right)\right]=$$
$$\frac{\beta T_2\left(x_{(.)i,j}^{\gamma}\right)\ln^2\left(x_{(.)i,j}\right)\left(1-\beta x_{(.)i,j}^{\gamma}\right)}{1-T_2}$$
$$-\left[\frac{\beta\ln\left(x_{(.)i,j}\right)x_{(.)i,j}^{\gamma}T_1}{1-T_1}\right]^2,$$

and

$$\frac{\partial^2\log}{\partial\gamma^2}\left[1-F\left(x_{(.)i,j}\right)\right]=-\beta x_{(.)i,j}^{\gamma}\ln^2\left(x_{(.)i,j}\right),$$

where

$$T_1=\exp\left(-\alpha x_{(m)i,j}-\beta x_{(m)i,j}^{\gamma-1}\right),$$

$$T_2=\exp\left(-\alpha x_{(1)i,j}-\beta x_{(1)i,j}^{\gamma}\right),$$

and

$$T_3=\exp\left(-\alpha x_{[(m+1)/2]i,j}-\beta x_{[(m+1)/2]i,j}^{\gamma}\right).$$

Methodology

Simulation Study

To investigate the properties of the MLEs of the three parameters of the MWD a simulation was conducted. The inverse transform method was used to generate samples from MWD (see Ros, 1997). The inverse transform algorithm can be described as: generate U from the uniform (0, 1), initiate X_1 and then find a new X_1 using $X_1=-\frac{\beta}{\alpha}X_1^{\gamma}-\frac{1}{\alpha}\ln\left(1-U\right)$; repeat until stability of X_1 is reached, which eventually represents a random number from MWD. The samples generated are then used to obtain the Fisher Information numbers, I_{ERSS} and I_{SRS}, when using ERSS and SRS. The asymptotic relative efficiency (RP) is found as the ratio I_{ERSS}/I_{SRS}.

Results

For $\alpha=3$, $\beta=1.2$ and $\gamma=1.3$, the results are presented in Tables 1, 2 and 3, respectively.

Table 1: Information Numbers and Asymptotic RP of the MLE of α Based on ERSS with respect to SRS

m	I_{ERSS}	I_{SRS}	Asymptotic RP
3	0.3613	0.1854	1.9490
4	0.6069	0.2392	2.5372
5	0.7933	0.2849	2.7845
6	0.9818	0.3478	2.8229
7	1.3030	0.4057	3.2119

Table (2): Information Numbers and Asymptotic RP of the MLE of β Based on ERSS with respect to SRS

m	I_{ERSS}	I_{SRS}	Asymptotic RP
3	0.1401	0.0542	2.5849
4	0.2894	0.1014	2.8554
5	0.4627	0.1551	2.9832
6	0.6335	0.1956	3.2382
7	0.8606	0.2314	3.7191

Table (3): Information Numbers and Asymptotic RP of the MLE of γ based on ERSS with respect to SRS

m	I_{ERSS}	I_{SRS}	Asymptotic RP
3	0.6451	0.5348	1.2062
4	1.1279	0.8684	1.2987
5	1.2494	0.7336	1.7032
6	1.7856	0.9847	1.8133
7	1.9196	0.8459	2.2693

For $\alpha = 2.3$, $\beta = 1.3$ and $\gamma = 1.6$, results are summarized in Tables 4, 5 and 6 respectively.

Table (4): Information Numbers and Asymptotic RP of the MLE of α Based on ERSS with respect to SRS

m	I_{ERSS}	I_{SRS}	Asymptotic RP
3	0.2523	0.1201	2.1010
4	0.3739	0.1563	2.3922
5	0.5461	0.1987	2.7483
6	0.6521	0.2254	2.8931
7	0.8796	0.2695	3.2632

Table (5): Information Numbers and Asymptotic RP of the MLE of β Based on ERSS with respect to SRS

m	I_{ERSS}	I_{SRS}	Asymptotic RP
3	0.1195	0.0481	2.4871
4	0.1681	0.0595	2.8263
5	0.2383	0.0766	3.1110
6	0.2913	0.0814	3.5787
7	0.3852	0.0919	4.1902

Table (6): Information Numbers and Asymptotic RP of the MLE of γ Based on ERSS with respect to SRS

m	I_{ERSS}	I_{SRS}	Asymptotic RP
3	1.2459	0.7885	1.5801
4	1.9567	1.0232	1.9123
5	2.9202	1.3510	2.1615
6	3.8283	1.5697	2.4388
7	5.0158	1.7128	2.9284

Tables 1-3 show that:

- The ERSS estimators dominate the estimators based on SRS.
- The information numbers from ERSS are greater than those of SRS.
- For odd and even sample sizes the Fisher information numbers are increasing when the sample size is increasing.
- The asymptotic relative precision values are increasing when sample size increasing.

Estimation of the Population Mean of the MWD

The problem of estimating the population mean of the MWD is now considered and compared with the SRS estimator of the population mean $\hat{\mu}_{SRS} = \sum_{i=1}^{m} X_i / m$, which has variance σ^2 / m. The efficiency of $\hat{\mu}_{ERSS1}$ and $\hat{\mu}_{ERSS2}$ respectively with respect to $\hat{\mu}_{SRS}$ are defined as

$$eff\left(\hat{\mu}_{ERSSi}, \hat{\mu}_{SRS}\right) = \frac{MSE\left(\hat{\mu}_{SRS}\right)}{MSE\left(\hat{\mu}_{ERSSi}\right)},\ i = 1, 2.$$

Simulation results are summarized in Tables 7-9 for some values of the population parameters.

From results shows in Tables 7-9, it may be concluded that the ERSS estimators are biased and more efficient than the SRS estimator for all cases considered in this study. However, as demonstrated by Samawi, et al. (1996) it is better to use ERSS with small sample size. Also note that the efficiency of the mean estimation depends on the values of α, β, γ, as well as the sample size.

Conclusion

Maximum likelihood estimators for the three parameters of the modified Weibull distribution were studied based on extreme ranked set sampling. These MLEs are not in closed forms, so numerical method is used. Results show that the Fisher information numbers obtained from ERSS are greater than that from SRS. Also, it was shown that ERSS is more efficient than SRS in estimating the population mean and it has a small bias. However, the ERSS estimators dominate the corresponding estimators based on SRS for estimating the population mean of the MWD.

Table 7: Efficiency and Bias Values of Estimating the Population Mean of the MWD Using ERSS with respect to SRS for $\alpha = 2$, $\beta = 1.2$ and $\gamma = 1.3$

m	*Bias(ERSS)*	*MSE(SRS)*	*MSE(ERSS)*	Efficiency
3	0.0015	0.0173	0.0097	1.7835
4	0.0284	0.0140	0.0075	1.8667
5	0.0256	0.0099	0.0053	1.8679
6	0.0544	0.0088	0.0064	1.3750
7	0.0486	0.0079	0.0049	1.6122

Table 8: Efficiency and Bias Values of Estimating the Population Mean of the MWD Using ERSS with respect to SRS for $\alpha = 4$, $\beta = 2$ and $\gamma = 3$

m	*Bias(ERSS)*	*MSE(SRS)*	*MSE(ERSS)*	Efficiency
3	0.0028	0.0088	0.0047	1.8723
4	0.0165	0.0071	0.0038	1.8684
5	0.0064	0.0051	0.0020	2.5500
6	0.0362	0.0047	0.0032	1.4688
7	0.0221	0.0041	0.0016	2.5625

Table 9: Efficiency and Bias Values of Estimating the Population Mean of the MWD Using ERSS with respect to SRS for $\alpha = 3.5$, $\beta = 2$ and $\gamma = 1.5$

m	*Bias(ERSS)*	*MSE(SRS)*	*MSE(ERSS)*	Efficiency
3	0.0285	0.0067	0.0036	1.8611
4	0.0029	0.0076	0.0041	1.8537
5	0.0071	0.0040	0.0014	2.8570
6	0.0122	0.0055	0.0027	1.8519
7	0.0025	0.0029	0.0008	3.6250

References

Al-Hadhrami, S., & Al-Omari, A. I. (2009). Bayesian inference on the variance of normal distribution using moving extremes ranked set sampling. *Journal of Modern Applied Statistical Methods*, *8*(*1*), 227-235.

Al-Omari, A. I., Jaber, K., & Al-Omari, A. (2008). Modified ratio type estimators of the mean using ranked set sampling. *Journal of Mathematics and Statistics*, *4*(*3*), 150-155.

Al-Omari, A. I., & Jaber, K. (2008). Percentile double ranked set sampling. *Journal of Mathematics and Statistics*, *4*(*1*), 60-64.

Arnold, B. C., Castillo, E., & Sarabia, J. M. (2009). On multivariate order statistics. Application to ranked set sampling. *Computational Statistics and Data Analysis*, *53*, 4555-4569.

Bouza, C. N. (2009). Ranked set sampling and randomized response procedures for estimating the mean of a sensitive quantitative character. *Metrika*, *70*, 267–277.

David, H. A., & Nagaraja, H. N. (2003). *Order statistics* (*3rd Ed.*). Hoboken, NJ: John Wiley & Sons, Inc.

Jemain, A. A., & Al-Omari, A. I. (2006). Double quartile ranked set samples. *Pakistan Journal of Statistics*, *22*(*3*), 217-228.

Islam, T., Shaibur, M. R., & Hossain, S. S. (2009). Effectivity of modified maximum likelihood estimators using selected ranked set sampling data. *Austrian Journal of Statistics*, *38*(*2*), 109-120.

Mahdizadeh, M., & Arghami, N. R. (2009). Efficiency of ranked set sampling in entropy estimation and goodness-of-fit testing for the inverse Gaussian law. *Journal of Statistical Computation and Simulation*, *80*(*7*), 761–774.

McIntyre, G. A. (1952). A method for unbiased selective sampling using ranked sets. *Australian Journal Agricultural Research*, *3*, 385-390.

Muttlak, H. A. (2001). Regression estimators in extreme and median ranked set samples. *Journal of Applied Statistics*, *28*(*8*), 1003-1017.

Ross, S. M. (1997). *Simulation* (*2nd Ed.*). New York: Academic Press, Inc.

Samawi, H., Abu-Dayyeh, W., & Ahmed, S. (1996). Extreme ranked set sampling. *The Biometrical Journal*, *30*, 577-586

Samawi, H. M., & Al-Sagheer, O. A. (2001). On the estimation of the distribution function using extreme and median ranked set sampling. *Biological Journal*, *43*(*3*), 357-373.

Samawi, H. M., & Tawalbeh, E. M. (2002). Double median ranked set sample: Comparison to other double ranked samples for mean and ratio estimators. *Journal of Modern Applied Statistical Methods*, *1*(*2*), 428-442.

Samawi, H. M., & Saeid, L. J. (2004). Stratified extreme ranked set sample with application to ratio estimators. *Journal of Modern Applied Statistics Methods*, *3*(*1*), 117-133.

Sarhan, A. M., & Zaindin, M. (2009). Modified Weibull distribution. *Applied Sciences*, *11*, 123-136.

Sengupta, S., & Mukhuti, S. (2009). Unbiased estimation of $P(X > Y)$ using ranked set sample data. *Statistics*, *42*(*3*), 223-230.

Shadid, M. R., Raqab, M., & Al-Omari, A. I. (2011). Modified BLUEs and BLIEs of the location and scale parameters and the population mean using ranked set sampling. *Journal of Statistical Computation and Simulation*, *81*(*3*), 261-274.

Zaindin, M., & Sarhan, A. M. (2009). Parameter estimation of the modified Weibull distribution. *Applied Mathematical Sciences*, *11*(*3*), 541-550.

Takahasi, K., & Wakimoto, K. (1968). On the unbiased estimates of the population mean based on the sample stratified by means of ordering. *Annals of the Institute of Statistical Mathematic*s, *20*, 1-31.

Journal of Modern Applied Statistical Methods
November 2011, Vol. 10, No. 2, 618-624

1538 – 9472/11/$95.00

Modeling Repairable System Failures with Interval Failure Data and Time Dependent Covariate

Jayanthi Arasan Samira Ehsani
University Putra Malaysia,
Malaysia

An application of a repairable system model for interval failure data with a time dependent covariate is examined. The performance of several models based on the NHPP when applied to real data on ball bearing failures is also explored. The best model for the data was selected based on results of the likelihood ratio test. The bootstrapping technique was applied to obtain the variance estimate for the estimated expected number of failures. Results demonstrate that the proposed model works well and is easy to implement, in addition the bootstrap variance estimate provides a simple substitute for the traditional estimate.

Key words: Interval, repairable, NHPP, covariate, bootstrap.

Introduction

A repairable system is a system that can be restored back to functionality after a failure has occurred. The period where the system is unable to function is referred to as repair time and is assumed to be negligible. Grouped data, also known as interval failure data occurs when a component's failure time falls within a certain interval (t_{i-i}, t_i) where t_{i-i} is the lower inspection time and t_i is the upper inspection time in the i^{th} interval. In reliability this phenomenon occurs when components are inspected periodically to carry out maintenance or repair actions. These types of data often arise in the medical field where patients are examined periodically, for example every 3 or 6 months, so the exact failure time is typically unknown.

Many stochastic models have been developed to describe the failure rate of a non-homogenous Poisson process (NHPP) such as the power law model proposed by Crow (1974) and based on the ideas of Duanne (1964). Other popular models are the log linear proposed by Cox and Lewis (1966) and linear models discussed by Vesely (1977) and Atwood (1992). Lawless and Thiagarajah (1996) introduced an important repairable system model that incorporates both time trends and renewal behavior, known as a proportional intensity model. Guo, et al. (2006) proposed a proportional intensity model that is based on the powerlaw model. Guo, et al. (2007) also developed a new general repair model based on the expected cumulative number of failures to capture the repair history. Samira and Arasan (2009) extended the model to include a time dependent covariate and applied it to pipe failures in water networks.

Other literature on repairable system models and recurrent events includes Brown (1975), Gasmi, et al. (2003), Kaminskiy and Krivtsov (1998), Kijima and Sumita (1986), Kijima (1989), Wang and Pham (1996) and Yanez, et al. (2002). Park, et al. (2008) presented an application of the log-linear and power law models for interval failure data in water distribution systems.

More details regarding recurrent event models for grouped and interval failure data can also be found in Meeker and Escobar (1998),

Jayanthi Arasan is a Senior Lecturer in the Department of Mathematics. Email: jayanthi@science.upm.edu.my. Samira Ehsani is a post graduate student in the Department of Mathematics.
Email: ehsani_samira@yahoo.com.

Lawless and Zhan (1998) and Cook and Lawless (2007).

The Model

Most recurrent event data, such as in the case of repairable systems, usually has recurrence times that are not be independent. The most widely used models for recurrence data are those based on the non-homogenous Poisson process, mainly the power law and log-linear models. This research extends the power law model to incorporate the analysis of grouped or interval failure data while accommodating the effect of covariates or other factors that may affect or contribute to system failure. Thus, the failure intensity or recurrence rate can be described as $\nu(t) = abt^{b-1}e^{gx(t)}$, where $x(t)$ is a time dependent covariate that may impact system failure.

Thus, the proposed model takes into account both the effect of time and a time dependent covariate on the recurrence rate of a system. Because it is dealing with interval failure data - and there can be more than one failure in any time interval - the number of intervals is always less or equal to number of failures observed.

Suppose d_i is the number of failures in the i^{th} interval and $x(t_i)$ is the value of covariate at time t_i. The expected number of recurrences

$$\mu(t_{i-1}, t_i) = E[N(t_{i-1}, t_i)] = \int_{t_{i-1}}^{t_i} \nu(u)du,$$

where $i = 1, 2, \cdots, n$.

If the intervals are contiguous, the Poisson process log-likelihood for a series of n time intervals is:

$$L(a,b,g) = \sum_{i=1}^{n} d_i g x\,(t_i) \ln\left[a\left({t_i}^b - t_{i-1}^b\right)\right] - \left[a e^{gx(t_i)}\left({t_i}^b - t_{i-1}^b\right)\right] - \ln\left(d_i!\right). \quad (1)$$

The first and second derivatives of the log-likelihood function are as follows:

$$\frac{\partial L(a,b,g)}{\partial a} = \sum_{i=1}^{n} \frac{d_i}{a} - \exp(gx(t_i))(t_i^b - t_{i-1}^b),$$

$$\frac{\partial L(a,b,g)}{\partial b} = \sum_{i=1}^{n} \frac{d_i(t_i^b ln(t_i) - t_{i-1}^b ln(t_{i-1}))}{(t_i^b - t_{i-1}^b)} - \exp(gx(t_i))a(t_i^b ln(t_i) - t_{i-1}^b ln(t_{i-1})),$$

$$\frac{\partial L(a,b,g)}{\partial g} = \sum_{i=1}^{n} d_i x\,(t_i) - x\,(t_i)\exp(gx\,(t_i)) \; a\left({t_i}^b - {t_{i-1}}^b\right),$$

$$\frac{\partial^2 L(a,b,g)}{\partial a^2} = \sum_{i=1}^{n} -\frac{d_i}{a^2},$$

$$\frac{\partial^2 L(a,b,g)}{\partial a \partial b} = \sum_{i=1}^{n} -\exp(gx(t_i))(t_i^b ln(t_i) - t_{i-1}^b ln(t_{i-1})),$$

$$\frac{\partial^2 L(a,b,g)}{\partial a \partial g} = \sum_{i=1}^{n} -x\,(t_i)\exp(gx\,(t_i))\left({t_i}^b - {t_{i-1}}^b\right),$$

$$\frac{\partial^2 L(a,b,g)}{\partial b^2} = \sum_{i=1}^{n} \frac{d_i(t_i^b ln(t_i)^2 - t_{i-1}^b ln(t_{i-1})^2)}{(t_i^b - t_{i-1}^b)} - \frac{d_i(t_i^b ln(t_i) - t_{i-1}^b ln(t_{i-1}))^2}{(t_i^b - t_{i-1}^b)^2} - \exp(gx(t_i))a(t_i^b ln(t_i)^2 - t_{i-1}^b ln(t_{i-1})^2),$$

$$\frac{\partial^2 L(a,b,g)}{\partial b \partial g} = \sum_{i=1}^{n} -x(t_i)\exp(gx(t_i))a(t_i^b ln(t_i) - t_{i-1}^b ln(t_{i-1})),$$

$$\frac{\partial^2 L(a,b,g)}{\partial g^2} = \sum_{i=1}^{n} -x(t_i)^2\exp(gx(t_i))a(t_i^b - t_{i-1}^b).$$

The extended power law model allows interval failure data to be analyzed by incorporating the effect of time and covariates simultaneously. Occasionally, the effect of covariates are insignificant, thus, the reduced form of the model may prove to be a better fit for the data; this can be obtained by setting $g = 0$. Another useful NHPP model is the log linear model, which has the failure intensity function $\lambda(t) = e^{a+bt}$, where a and b are the parameters of the model. The log linear model can also be extended to accommodate interval or grouped failure data. Let

$$v_i = e^{(a+bt_i)} - e^{(a+bt_{i-1})},$$
$$w_i = t_i e^{(a+bt_i)} - t_{i-1} e^{(a+bt_{i-1})},$$
$$z_i = t_i^2 e^{(a+bt_i)} - t_{i-1}^2 e^{(a+bt_{i-1})}.$$

The log-likelihood function for a series of n time intervals is:

$$L(a,b) = \sum_{i=1}^{n} d_i \ln\left(\frac{v_i}{b}\right) - \frac{v_i}{b} - \ln(d_i!). \tag{2}$$

The first and second derivatives of the log-likelihood function are:

$$\frac{\partial L(a,b)}{\partial a} = \sum_{i=1}^{n} \frac{d_i b - v_i}{b},$$
$$\frac{\partial L(a,b)}{\partial b} = \sum_{i=1}^{n} \frac{(w_i b - v_i)(d_i b - v_i)}{b^2 v_i},$$
$$\frac{\partial^2 L(a,b)}{\partial a^2} = \sum_{i=1}^{n} -\frac{v_i}{b},$$
$$\frac{\partial^2 L(a,b)}{\partial a \partial b} = \sum_{i=1}^{n} -\frac{(w_i b - v_i)}{b^2},$$
$$,$$
$$\frac{\partial^2 L(a,b)}{\partial b^2} = \sum_{i=1}^{n} \frac{d_i b^3 v_i z_i + d_i b v_i^2 - d_i b^3 w_i^2 - z_i b^2 v_i^2}{b^3 v_i^2} + \frac{2 w_i b v_i^2 - 2 v_i^3}{b^3 v_i^2}.$$

Application with Real Data

The real data used in this study consists of 25 time intervals to ball bearing failures in a conveyer belt in an automobile production. The failure occurrences are in intervals because the conveyer is only checked by the inspection team at certain times, referred to as inspection times (hours). There can be more than 1 failure in a certain time interval for which repair action is carried out. The time dependent covariate used is the number of maintenance actions taken throughout the study period.

Graphical methods are often used in modeling repairable systems to check trends in the data which then enables a reasonable model selection. Figure 1 displays the plot of the cumulative number of failures, $N(t_i)$ versus operating hours, t_i. Because data are failures within intervals, the graph was drawn using the upper interval point. The plot suggests that the use of a NHPP model might be appropriate because the failure rate appears to be inconsistent.

Figure 1: Cumulative Number of Failures vs. Time

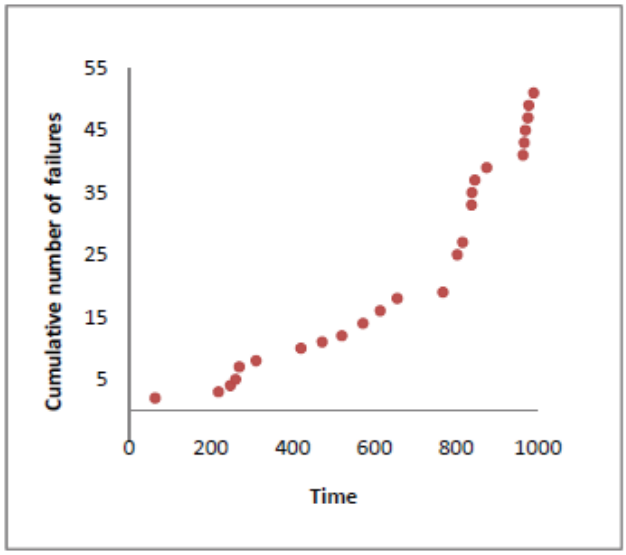


Table 1 shows the value of the parameter estimates and their standard errors when the data is fitted to the extended power law, power law, log-linear and HPP models. The table also shows the log likelihood value for each model at the estimated parameters. In the case of the extended power law model, the parameter estimate g has a positive value; this implies that the maintenance action could not prevent the system from deteriorating with time. In addition, the estimate of b shows a reliability improvement, but overall this fails to improve the system. All of the models show evidence of increasing failure intensity over time.

The extended power law model gives the highest log likelihood value, this implies that it fits the real data better than the other models. Figure 2 shows the estimates of the expected number of failures using the extended power law, power law, log linear and HPP models. The extended power law model shows the best fit for the real data, although the log linear appears to be a reasonable fit as well. The plot also shows

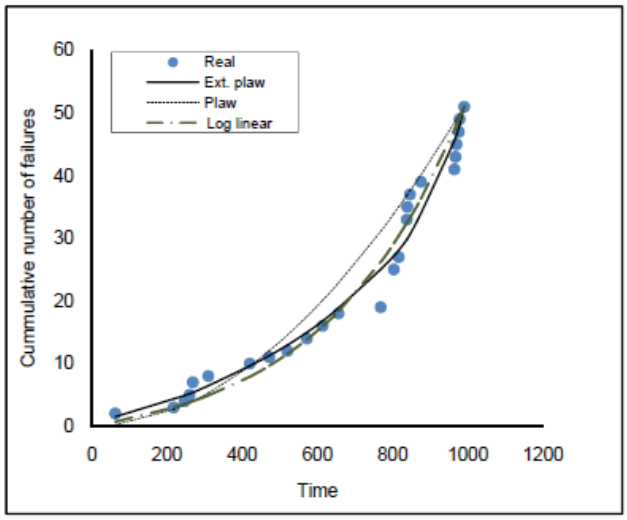
an obvious change in the slope towards the end on the process and certain data tend to form clusters, requiring further investigation.

Table 1: Parameter Estimates for Various Models

Ext. powerlaw	Powerlaw
$\hat{a}: 0.05542(0.11061)$	$\hat{a}: 0.00007(0.14160)$
$\hat{b}: 0.79426(0.31160)$	$\hat{b}: 1.95071(0.28005)$
$\hat{g}: 0.13092(0.03699)$	***
$L(\hat{a},\hat{b},\hat{g}) = -61.89$	$L(\hat{a},\hat{b}) = -66.60$

Log linear	HPP
$\hat{a}: -4.60097(0.42546)$	$\hat{a}: 0.05152(0.00721)$
$\hat{b}: 0.00272(0.00058)$	***
***	***
$L(\hat{a},\hat{b}) = -62.33$	$L(\hat{a}) = -75.40$

Figure 2: Real vs. Fitted for Several Models



Hypothesis Testing and Confidence Intervals

If parameters g and b are significant then there is evidence of both maintenance effect and time trend within the model. The significance of the parameters b and g can be tested using likelihood ratio (LR) test. The idea of a LR test is to compare the maximized likelihood of two nested models, the full model and the reduced model. The reduced model is restricted by certain conditions in H_0.

Let $\widehat{\theta}_r$ be the maximum likelihood estimator of the restricted model under H_0 and $\widehat{\theta}_f$ the maximum likelihood estimator of the full model. The maximized likelihood of the reduced model, $l(\widehat{\theta}_r)$ can never exceed the maximized likelihood of the full model, $l(\widehat{\theta}_f)$, because it is a subset of the full model. Thus, the ratio of the maximized likelihood of the reduced model to the full model is bounded between 0 and 1. A ratio close to 1 indicates that the reduced model is close to the full model whereas a ratio close to 0 indicates that the two models are very different and the reduced model is unacceptable. The likelihood ratio statistic for testing H_0 versus H_l is the given by:

$$\Psi = -2[L(\widehat{\theta}_r) - L(\widehat{\theta}_f)]. \qquad (3)$$

For a large sample size, Ψ is approximately $\chi^2_{(\nu)}$, where ν is the number of parameters in the full model minus the number of parameters in the reduced model. The test statistic for testing the significance of the parameter, g, is 9.41, which is higher than $\chi^2(0.05,1) = 3.841$, thus implying that the effect of g is significant at the 0.05 level. The test statistic for testing the significance of parameter b, is 27.014, thus implying that the effect of b is also significant at the 0.05 level. Thus, it may be concluded that the extended power law model is the most suitable model for the data.

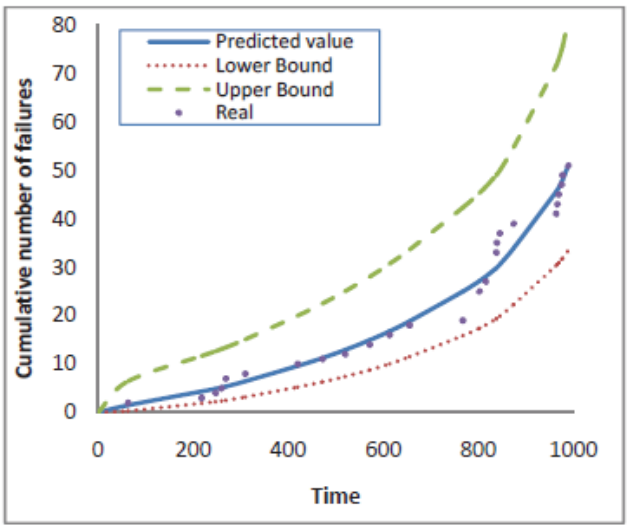
Confidence intervals for the expected number of failures over interval (a,b), $E[N(a,b)] = \mu(a,b)$ can be obtained by using the log normal distribution. The variance of an estimator can be calculated using the Delta method. The Delta method uses the 2^{nd} order Taylor expansion to approximate the variance of a function of random variables. Thus,

$$Var(\hat{\mu}(a,b)) = \left(\frac{\partial \mu(a,b)}{\partial a}\right)^2 Var(\hat{a}) + \left(\frac{\partial \mu(a,b)}{\partial b}\right)^2 Var(\hat{b}) + \left(\frac{\partial \mu(a,b)}{\partial g}\right)^2 Var(\hat{g}) + 2\left(\frac{\partial \mu(a,b)}{\partial a}\right)\left(\frac{\partial \mu(a,b)}{\partial b}\right) Cov(\hat{a},\hat{b}) + 2\left(\frac{\partial \mu(a,b)}{\partial a}\right)\left(\frac{\partial \mu(a,b)}{\partial g}\right) Cov(\hat{a},\hat{g}) + 2\left(\frac{\partial \mu(a,b)}{\partial b}\right)\left(\frac{\partial \mu(a,b)}{\partial g}\right) Cov(\hat{b},\hat{g}).$$

Following this, the confidence interval for $\mu(a,b)$ is

$$\left[\hat{\mu}(a,b)e^{-\frac{Z_{\frac{\alpha}{2}}\sqrt{Var(\hat{\mu}(a,b))}}{\hat{\mu}(a,b)}}, \hat{\mu}(a,b)e^{\frac{Z_{\frac{\alpha}{2}}\sqrt{Var(\hat{\mu}(a,b))}}{\hat{\mu}(a,b)}}\right].$$

Figure 3: Confidence Interval for $\mu(a,b)$



Another way to obtain the variance of $\hat{\mu}(a,b)$ is to use the bootstrap technique. Recently, alternative techniques requiring only minimal assumption have become popular. The bootstrapping technique was proposed by Efron (1993) and the procedure depends on how the bootstrap sampling is done. Efron (1993) showed that, in certain cases, the bootstrap estimate of variance or standard error can be used as an alternative for numerically estimating the traditional variance or standard error estimate.

Several different methods for generating bootstrap samples exist, namely parametric and nonparametric sampling procedures. This study utilizes the parametric bootstrap sampling procedure where B bootstrap samples of size n are generated from an assumed parametric distribution. The number of failures over interval (a,b) follows a Poisson distribution with mean $\mu(a,b)$. Thus, random samples can be generated from the Poisson distribution and bootstrap estimates of the mean, $\widehat{\theta}^{\,b}$, can be calculated where $b = 1, 2, \cdots, B$ are estimates calculated from each of the bootstrap samples of size n.

The bootstrap estimate of the variance of $\mu(a,b)$ is

$$\widehat{Var}_B = \frac{1}{B-1}\sum_{b=1}^{B}\left(\widehat{\theta}^{\,b} - \widehat{\theta}_{(.)}\right)^2,$$

where

$$\widehat{\theta}_{(.)} = \sum_{b=1}^{B}\frac{\widehat{\theta}^{\,b}}{B}.$$

Following this, the confidence interval for $\mu(a,b)$ can be obtained in the similar way as

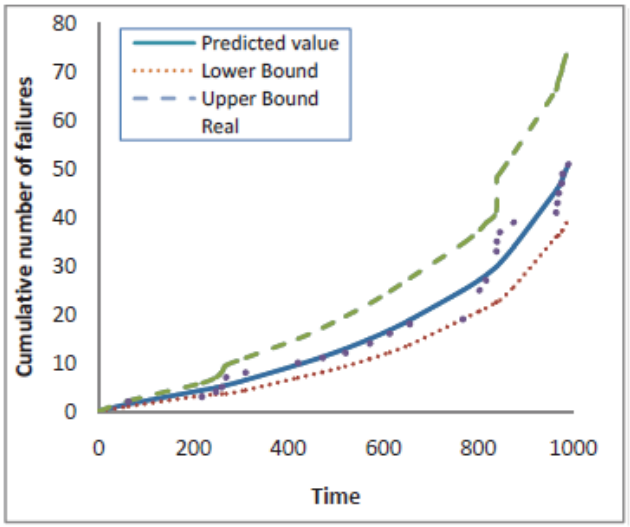
$$\left[\hat{\mu}(a,b)e^{-\frac{Z_{\frac{\alpha}{2}}\sqrt{\widehat{Var}_B}}{\hat{\mu}(a,b)}}, \hat{\mu}(a,b)e^{\frac{Z_{\frac{\alpha}{2}}\sqrt{\widehat{Var}_B}}{\hat{\mu}(a,b)}}\right].$$

Figure 4 shows the 95 % confidence interval for the expected cumulative number of failures using the bootstrap standard error estimate. This shows that the interval estimation using the bootstrap standard error estimate provides a good alternative and is slightly narrower than the traditional method.

Conclusion

This article proposed the use of the extended power law model for repairable systems with interval or grouped failure data and a time dependent covariate. The model reduces to the power law and HPP as a special case, thus it is convenient and useful. The model also allows incorporation and analysis of both time trend

Figure 4: Confidence Interval for $\mu(a,b)$ Using Bootstrap Standard Error



and covariate effects simultaneously. More research may be done by implementing the methods discussed herein to other repairable system models to determine if similar results are obtained. The use of other types of log-linear and linear models that can incorporate interval failure data with covariates should also be investigated.

The parametric bootstrap computer based technique was also employed to obtain the variance estimate for the estimated expected number of failures Alternative computer intensive techniques are simpler to implement and - in many cases - provide better estimates than traditional methods. Bootstrapping techniques are useful particularly when traditional methods become unreliable and certain assumptions are not satisfied. The high capability of modern day computers makes these methods practical.

Other parametric bootstrapping techniques and block jackknifing techniques for confidence interval estimation could also be explored. There may also be chances of applying other bootstrap confidence interval estimates such as percentile bootstrap, bootstrap-t and BCa. These intervals are usually known to be more reliable and give better coverage probabilities and, as noted by Arasan (2008), are more symmetrical. However, their use with repairable system data should be done with caution; some modifications are also likely necessary to avoid violating the basic assumptions.

References

Arasan, & Lunn. (2008). Alternative interval estimation for parameters of bivariate exponential model with time varying covariate. *Computational Statistics*, *23*, 605-622.

Atwood, C. L. (1992). Parametric estimation of time-dependent failure rates for probabilistic risk assessment. *Reliability Engineering and System Safety*, *37*, 181-194.

Brown, C. (1975). On the use of indicator variable for studying the time dependence of parameters in a response-time model. *Biometrics*, *31*, 863-872.

Cook, R. J., & Lawless, J. F. (2007). *The statistical analysis of recurrent events*. New York: Springer.

Cox, D. R., & Lewis, P. A. (1966). *Statistical analysis of series of events*. London: Methuen.

Crow, L. H. (1974). Reliability analysis for complex, repairable systems. In *Reliability and biometry*, F. Proschan & R. J. Serfling, Eds, 379-410. Philadelphia, PA: *SIAM*.

Duanne, J. T. (1964). Learning curve approach to reliability monitoring. *IEEE Transactions on Aerospace*, *2*, 563-566.

Efron, B., & Tibshirani, R. J. (1993). An *introduction to the bootstrap*. New York, NY: Chapman and Hall.

Gasmi, S., Love, C. E., & Kahle, W. A. (2003). General repair, proportional hazards, framework to model complex repairable systems. *IEEE Transactions on Reliability*, *52*(*1*), 26-32.

Guo, H. R., Liao, H., Zhao, W., & Mettas, A. (2007). A new stochastic model for systems under general repairs. *IEEE Transactions on Reliability*, *56*(*1*), 40-49.

Guo, H. R., Zhao, W., & Mettas, A. (2006). Practical methods for modeling repairable systems with time trends and repair effects. *Annual Reliability and Maintainability Symposium*, 182-188.

Jeng, S. L., & Meeker, W. Q. (2000) Comparison for approximate confidence interval procedure for type I censored data. *Technometrics*, *42*(*2*), 135-148.

Kaminskiy, M., & Krivtsov, V. (1998). *A Monte Carlo approach to repairable system repairable analysis, probabilistic safety assessment and management*. New York, NY: Springer.

Kijima, M. (1989). Some results for repairable systems with general repair. *Journal of Applied Probability*, *26*, 89-102.

Kijima, M., & Sumita, N. (1986). A useful generalization of renewal theory: Counting process governed by non-negative Markovian increments. *Journal of Applied Probability*, *23*, 71-88.

Lawless, J. F., & Thiagarajah, K. (1996). A point-process model incorporating renewals and time trends, with application to repairable systems. *Technometrics*, *38*(*2*), 131-138.

Lawless, J. F., & Zhan, M. (1998). Analysis of interval-grouped recurrent-event data using piecewise constant rate functions. *The Canadian Journal of Statistics*, *26*, 549-565.

Meeker, W. Q., & Escobar, L. A. (1998). *Statistical methods for reliability data*. New York, NY: Wiley.

Park, S., Jun, H., Kim, B. J., & Im, G. C. (2008). Modeling of water main failure rates using the log-linear rocof and the power law process. *Water Resource Management*, *22*, 1311-1324.

Samira, E. F. & Arasan, J. (2009). A counting process for repairable system under general repair with repair history and time varying covariate. *Proceedings of 17th National Symposium on Mathematical Sciences*, *2*, 823-828.

Vesely, W. E. (1977). Estimating common cause failure probabilities in reliability and risk analysis: Marshall-Olkin specialization. In *Nuclear System Reliability Engineering and Risk Assessment*, J. B. Fussell & G. R. Burdick, Eds., 314-341. Philadelphia, PA: SIAM.

Wang, H., & Pham, N. (1996). Optimal age dependent preventive maintanance policies with imperfect maintanance. *International Journal of Reliability, Quality and Safety Engineering*, *3*, 119-135.

Yanez, M., Joglar, F., & Modarres, M. (2002). Generalized renewal process for analysis of repairable systems with limited failure experience. *Reliability Engineering and System Safety*, *77*, 167-180.

Journal of Modern Applied Statistical Methods
November 2011, Vol. 10, No. 2, 625-631

1538 – 9472/11/$95.00

Explicit Equations for ACF in Autoregressive Processes In the Presence of Heteroscedasticity Disturbances

Samir Safi
The Islamic University of Gaza,
Gaza

The autocorrelation function, ACF, is an important guide to the properties of a time series. Explicit equations are derived for ACF in the presence of heteroscedasticity disturbances in p^{th} order autoregressive, *AR(p)*, processes. Two cases are presented: (1) when the disturbance term follows the general covariance matrix, Σ, and (2) when the diagonal elements of Σ are not all identical but $\sigma_{i,j} = 0 \ \forall i \neq j$.

Key words: Heteroscedasticity, homoscedasticity, autocorrelation, autoregressive, covariance, disturbance, time series.

Introduction

When disturbance terms are identically distributed, it implies that they have the same variance for all observations: this is known as homoscedasticity. If they are not, it causes serious problems for estimates and must be corrected in order to obtain reliable estimates. A sequence, or a vector, of random variables is heteroscedastic if the random variables have different variances. Heteroscedastic means differing variance and is derived from the Greek *hetero*, meaning different, and *skedasis*, meaning dispersion. The word heteroscedasticity indicates a time-varying variance and is a deviation from the identically distributed assumption because the variances are not the same for each value.

Heteroscedasticity occurs when observations are based on average data and in a number of random coefficient models. It has two forms, conditional and unconditional. Conditional heteroscedasticity identifies non-constant volatility when future periods of high and low volatility cannot be identified.

Unconditional heteroscedasticity is when future periods of high and low volatility can be identified. For example, periods of low and high volatility for the prices of stocks and bonds cannot be predicted over any period of time, and therefore would be described as conditional heteroscedasticity. By contrast, unconditional heteroscedasticity can be used discussing variables that have identifiable seasonal variability, such as electricity usage.

The consequences of heteroscedasticity are problematic in general, and it is well known that the consequences of heteroscedasticity for ordinary least squares (OLS) estimation are very serious. Although parameter estimates remain unbiased, they are no longer efficient, meaning they are no longer best linear unbiased estimators (BLUE) among the class of all the linear unbiased estimators. The standard errors typically computed for the least squares estimators are no longer appropriate, hence, confidence intervals and hypothesis tests that use these standard errors are invalid. Because the estimated error's variance-covariance is not efficient, it invalidates the t-statistic, sometimes making insignificant variables appear to be statistically significant. Heteroscedasticity causes the OLS estimates of the standard error to be biased, leading to unreliable hypothesis testing. The most serious implication of heteroscedasticity is a misleading inference when the standard tests are used such as t and F tests.

Samir Safi is an Associate Professor of Statistics. Email: samirsafi@gmail.com.

The disturbance term in time series data is modeled under an assumption of constant variance and the assumption of heteroscedastic disturbances has traditionally been considered in the context of cross-sectional data. With time series data the disturbance term is modeled with some kind of stochastic process, and most of the conventional stochastic processes assume homoscedasticity (Judge, et al., 1985). The econometrician Robert Engle won the 2003 Nobel Memorial Prize for Economics for his studies on regression analysis in the presence of heteroscedasticity, which led to his formulation of the AutoRegressive Conditional Heteroscedasticity (ARCH) modeling technique.

Background

Heteroscedasticity is a problem often faced by statisticians and econometricians. A wealth of literature related to estimating and testing heteroscedasticity exists, see for example, Wallentin and Agren (2002), Kalirajan (1989), Evans and King (1988) and Farebrother (1987).

Safi (2009) derived explicit equations for ACF in the presence of heteroscedasticity disturbances in first-order autoregressive, *AR(1)*, process. He showed two cases: (1) when the disturbance follows the general covariance matrix, Σ, and (2) when the diagonal elements of Σ are not all identical but $\sigma_{i,j} = 0 \ \forall i \neq j$, that is, $\Sigma = \mathrm{diag}\left(\sigma_{11}, \sigma_{22}, \ldots, \sigma_{tt}\right)$. This article extends the Safi (2009) results for the general autoregressive, *AR(p)*, process.

Praetz (2008) discussed the effect of auto-correlated disturbances when they are not modeled on statistics used in drawing inferences in the multiple linear regression model. He derived biases for the F and R^2 statistics and evaluated them numerically. He discussed the reflections for empirical research on the causes, detection and treatment of autocorrelation.

Bera, et al. (2005) investigated conditional and unconditional heteroscedasticities as well as normality in the market model. They showed that conditional heteroscedasticity is more widespread than unconditional heteroscedasticity, suggesting the necessity of model refinements that take conditional heteroscedasticity into account. They provided an alternative estimation of betas of individual securities and portfolios based on the autoregressive conditional heteroscedastic (ARCH) model introduced by Engle. The efficiency of the market model coefficients is markedly improved across all firms in the sample through the ARCH technique. Demos (2000) derived expressions for the autocovariance of the observed series and the squared errors as a function of the parameters, something which facilitates the comparison of the observed properties of the data with the theoretical properties of the models, and consequently may play an important part in model identification.

Studies of many econometric time series models for financial markets reveal that it is unreasonable to assume that conditional variance of the disturbance term is constant as it for many stochastic processes. Two exceptions are the heteroscedastic stochastic processes proposed by Engle (1982) and Cragg (1982). Engle (1982), showed that, for many economic models, it is unreasonable to assume that the conditional forecast variance $\mathrm{var}\left(y_t \mid y_{t-1}\right)$ is constant, and that is more realistic to assume that $\mathrm{var}\left(y_t \mid y_{t-1}\right)$ depends on y_{t-1}.

Bumb, and Kelejian (1983) studied the auto-correlated and heteroscedastic disturbances in linear regression analysis. They discussed various procedures to test for the possibility that the disturbance terms of a linear regression model are auto-correlated in a first order process with a constant autoregressive coefficient.

Autocorrelation Function (ACF)

The autocorrelation function (ACF), is an important guide to the properties of a time series. It measures the correlation between observations at different distances apart. This behavior is a powerful tool to identify a preliminary time series model. The ACF provides a better understanding of correlation structure of the data and, within the Box Jenkins framework, a rough idea of the order of the components to be used in any autoregressive model. The estimate of ACF may suggest which of the many possible stationary time series models is a suitable candidate for representing

the dependence in the data, Brockwell and Davis (2002). The forms of the explicit equations depend on the autoregressive coefficients.

General Heteroscedastic Autocorrelation Function (GHACF)

Autoregressive processes are regressions on themselves. In other words, in autoregressive processes, the current value of the process Z_t is expressed as a finite linear combination of the *p* most recent past values of itself plus an innovation term e_t which incorporates everything new in the series at time *t* that is not explained by past values. Thus, for every *t*, it is assumed that e_t is independent of $Z_{t-1}, Z_{t-2}, \ldots$. If the values of a process at equally spaced times *t*, *t-1*, *t-2*,…, denoted by $Z_t, Z_{t-1}, Z_{t-2}, \ldots$, then $Z_t = \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + \cdots + \phi_p Z_{t-p} + e_t$ is called a p^{th} order autoregressive process, abbreviated *AR(p)*.

The p^{th} order autoregressive model may be written in terms of backward shift operator B as:

$$\left(1 - \phi_1 B - \cdots - \phi_p B^p\right) Z_t = \phi(B) Z_t = e_t. \tag{1}$$

A special notation used to simplify the representation of lag values, with $B^j Z_t = Z_{t-j}$. Z_t is the time series under investigation and e_t is the white noise series normally distributed with mean zero and variance σ_e^2. For the general *AR(p)* process, $Z_t = \phi^{-1}(B) e_t$, results in

$$\phi(B) = (1 - G_1 B)(1 - G_2 B) \cdots (1 - G_p B)$$

where $G_1^{-1}, \ldots, G_p^{-1}$ are the roots of $\phi(B) = 0$, and expanding $\phi^{-1}(B)$ in partial fractions yields

$$Z_t = \phi^{-1}(B) e_t = \sum_{i=1}^{p} \frac{K_i}{1 - G_i B} e_t.$$

(See for example Box, et al., 1994.) Thus, if $\psi(B) = \phi^{-1}(B)$ is to be a convergent series for $|B| \le 1$, then the weights $\psi_j = \sum_{i=1}^{p} K_i G_i^j$ must be absolutely summable so that the *AR(p)* will represent a stationary process, $|G_i| < 1$ for $i = 1, 2, \ldots, p$. Equivalently, the roots of $\phi(B) = 0$ must lie outside the unit circle. From the relation $\phi(B)\psi(B) = 1$ it follows that the weights ψ_j for the *AR(p)* process satisfy the difference equation:

$$\psi_j = \phi_1 \psi_{j-1} + \phi_2 \psi_{j-2} + \cdots + \phi_p \psi_{j-p}, \; j > 0 \tag{2}$$

with $\psi_0 = 1$ and $\psi_j = 0$ for $j < 0$, from which the weights ψ_j can easily be computed recursively in terms of the ϕ_i.

The *AR(p)* autoregressive process $Z_t = \phi^{-1}(B) e_t$ may be written as:

$$Z_t = \sum_{j=0}^{\infty} \psi_j e_{t-j}, \quad t = 0, \pm 1, \pm 2, \ldots. \tag{3}$$

It is assumed that the disturbance term has mean zero, E (**e**) = **0**, and the covariance matrix $\mathrm{Cov}(e_i, e_j) = \Sigma$ where:

$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1t} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2t} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{t1} & \sigma_{t2} & \cdots & \sigma_{tt} \end{bmatrix}. \tag{4}$$

Definition 1

The covariance between Z_t and Z_{t+k}, separated by k intervals of time (which under the stationary assumption must be the same for all t) is called the autocovariance function at lag k (ACVF) and is defined by

$$\gamma_k = \mathrm{Cov}(Z_t, Z_{t+k}) = \mathrm{E}\left[(Z_t - \mu)(Z_{t+k} - \mu)\right], \tag{5}$$

assuming that Z_t has zero mean. A nonzero mean can be introduced by replacing Z_t by $Z_{t-\mu}$ throughout the equations.

Definition 2

The autocorrelation function at lag k, that is the correlation between Z_t and Z_{t+k}, is defined by

$$\rho_k = \frac{\gamma_k}{\gamma_0} \tag{6}$$

where $\gamma_0 = \sigma_Z^2$ is the same at time t+k as at time t.

Lemma 1

Consider the general *AR(p)* process, $Z_t = \sum_{j=0}^{\infty} \psi_j e_{t-j}$, with E ($\mathbf{e}_t$) = **0**, and $\mathrm{Cov}(e_i, e_j) = \Sigma$, where Σ is given in (4). The autocovariance function at lag k is given by

$$\gamma_k = \sum_{j=0}^{t-k-1} \sum_{i=0}^{t-1} \psi_i \psi_j \sigma_{t-i,t-k-j}. \tag{7}$$

Proof

Using (3),

$$\begin{aligned} Z_t Z_{t-k} &= \left(\sum_{i=0}^{t-1} \psi_i e_{t-i} \right) \left(\sum_{j=0}^{t-k-1} \psi_j e_{t-k-j} \right) \\ &= \sum_{j=0}^{t-k-1} \sum_{i=0}^{t-1} \psi_i \psi_j e_{t-i} e_{t-k-j}. \end{aligned}$$

and using (5), the ACVF at lag k is

$$\begin{aligned} \gamma_k &= \mathrm{E}(Z_t Z_{t-k}) \\ &= \mathrm{E}\left[\sum_{j=0}^{t-k-1} \sum_{i=0}^{t-1} \psi_i \psi_j e_{t-i} e_{t-k-j} \right] \\ &= \sum_{j=0}^{t-k-1} \sum_{i=0}^{t-1} \psi_i \psi_j \mathrm{E}(e_{t-i} e_{t-k-j}) \\ &= \sum_{j=0}^{t-k-1} \sum_{i=0}^{t-1} \psi_i \psi_j \sigma_{t-i} \sigma_{t-k-j}. \end{aligned}$$

Theorem 1: Deriving the GHACF at Lag k when $\sigma_{i,j} \neq 0$ for all $i \neq j$ In an *AR(p)* Process

Consider the general AR(p) process $Z_t = \sum_{j=0}^{\infty} \psi_j e_{t-j}$, ψ_j is given in (2), with E($\mathbf{e}_t$) = **0**, and $\mathrm{Cov}(e_i, e_j) = \Sigma$, where Σ is given in (4), with $\sigma_{i,j} \neq 0 \ \forall i \neq j$, then the GHACF at lag k is given by:

$$\rho_k = \frac{\sum_{j=0}^{t-k-1} \sum_{i=0}^{t-1} \psi_j \psi_i \sigma_{t-i,t-j-k}}{\sum_{j=0}^{t-1} \sum_{i=0}^{t-1} \psi_j \psi_i \sigma_{t-j,t-i}}. \tag{8}$$

Proof

Using (7), the ACVF at lag 0 is

$$\begin{aligned} \gamma_0 = {} & \psi_0^2 \sigma_{t,t} + \psi_1 \sigma_{t,t-1} + \psi_2 \sigma_{t,t-2} + \psi_3 \sigma_{t,t-3} + \cdots + \psi_{t-1} \sigma_{t,1} + \\ & \psi_1 \sigma_{t-1,t} + \psi_1^2 \sigma_{t-1,t-1} + \psi_1 \psi_2 \sigma_{t-1,t-2} + \cdots + \psi_1 \psi_{t-1} \sigma_{t-1,1} + \\ & \psi_2 \sigma_{t-2,t} + \psi_2 \psi_1 \sigma_{t-2,t-1} + \psi_2^2 \sigma_{t-2,t-2} + \cdots + \psi_2 \psi_{t-1} \sigma_{t-2,1} + \\ & \vdots \\ & \psi_{t-1} \sigma_{1,t} + \psi_{t-1} \psi_1 \sigma_{1,t-1} + \psi_{t-1} \psi_2 \sigma_{1,t-2} + \cdots + \psi_{t-1} \psi_{t-2} \sigma_{1,2} + \psi_{t-1}^2 \sigma_{1,1}. \end{aligned} \tag{9}$$

Collecting terms, the ACVF at lag 0, that is, the variance of the process is:

$$\gamma_0 = \sum_{j=0}^{t-1} \sum_{i=0}^{t-1} \psi_j \psi_i \sigma_{t-j,t-i}. \tag{10}$$

Using (7), the ACVF at lag 1 is

$$\begin{aligned}\gamma_1 = &\\ &\sigma_{t,t-1} + \psi_1\sigma_{t,t-2} + \psi_2\sigma_{t,t-3} + \psi_3\sigma_{t,t-4} + \cdots + \psi_{t-2}\sigma_{t,1} +\\ &\psi_1\sigma_{t-1,t-1} + \psi_1^2\sigma_{t-1,t-2} + \psi_1\psi_2\sigma_{t-1,t-3} + \cdots + \psi_1\psi_{t-2}\sigma_{t-1,1} +\\ &\psi_2\sigma_{t-2,t-1} + \psi_2\psi_1\sigma_{t-2,t-2} + \psi_2^2\sigma_{t-2,t-3} + \cdots + \psi_2\psi_{t-2}\sigma_{t-2,1} +\\ &\vdots\\ &\psi_{t-1}\sigma_{1,t-1} + \psi_{t-1}\psi_1\sigma_{1,t-2} + \psi_{t-1}\psi_2\sigma_{1,t-3} + \cdots + \psi_{t-1}\psi_{t-2}\sigma_{1,1}.\end{aligned} \tag{11}$$

Collecting terms, the ACVF at lag 1 is

$$\gamma_1 = \sum_{j=0}^{t-2}\sum_{i=0}^{t-1}\psi_j\psi_i\sigma_{t-i,t-j-1} \tag{12}$$

similarly, the ACVF at lag k is

$$\gamma_k = \sum_{j=0}^{t-k-1}\sum_{i=0}^{t-1}\psi_j\psi_i\sigma_{t-i,t-j-k}. \tag{13}$$

Dividing (13) by (10), results in (8), which completes the proof.

Corollary 1: GHACF at Lag k for an *AR(1)* Process

Consider an *AR(1)* process $Z_t = \sum_{j=0}^{\infty}\psi_j e_{t-j}$, $\psi_j = \phi\psi_{j-1}$ with E ($\mathbf{e_t}$) = $\mathbf{0}$, and $\mathrm{Cov}(e_i, e_j) = \Sigma$, where Σ is given in (4), with $\sigma_{i,j} \neq 0 \ \forall i \neq j$. The GHACF at lag k is given by

$$\rho_k = \frac{\sum_{j=0}^{t-k-1}\sum_{i=0}^{t-1}\phi^{j+i}\sigma_{t-i,t-j-k}}{\sum_{j=0}^{t-1}\sum_{i=0}^{t-1}\phi^{j+i}\sigma_{t-j,t-i}}. \tag{14}$$

Proof

For an *AR(1)* process , because $\psi_j = \phi\psi_{j-1}$, it follows that $\psi_j = \phi^j$, for $j \geq 0$. From equations (10) and (13),

$$\gamma_0 = \sum_{j=0}^{t-1}\sum_{i=0}^{t-1}\phi^{j+i}\sigma_{t-j,t-i}$$

and

$$\gamma_k = \sum_{j=0}^{t-k-1}\sum_{i=0}^{t-1}\phi^{j+i}\sigma_{t-i,t-j-k}$$

are obtained, thus completing the proof.

Heteroscedastic Autocorrelation Function (HACF)

Heteroscedasticity exists if the diagonal elements of Σ in (4) are not all identical and the disturbance term is free from autocorrelation, meaning, the disturbances are pairwise uncorrelated. This assumption is likely to be realistic one when using cross-sectional data. In this case Σ can be written as a diagonal matrix with the i^{th} diagonal element given by σ_{ii}. Assume E($\mathbf{e_t}$) = $\mathbf{0}$, and $\mathrm{Cov}(e_i, e_j) = \Sigma$, where $\Sigma = \mathrm{diag}(\sigma_{11}, \sigma_{22}, \ldots, \sigma_{tt})$. Thus,

$$\Sigma = \begin{bmatrix} \sigma_{11} & 0 & \cdots & 0 \\ 0 & \sigma_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_{tt} \end{bmatrix}. \tag{15}$$

Theorem 2: HACF, at Lag k when $\sigma_{i,j} = 0$ for all $i \neq j$, i.e. $\Sigma = \mathrm{diag}(\sigma_{11}, \sigma_{22}, \ldots, \sigma_{tt})$ In an *AR(p)* Process

Consider the general *AR(p)* process $Z_t = \sum_{j=0}^{\infty}\psi_j e_{t-j}$, ψ_j as given in (2), with E ($\mathbf{e_t}$) = $\mathbf{0}$ and $\mathrm{Cov}(e_i, e_j) = \Sigma$, with $\sigma_{i,j} = 0 \ \forall i \neq j$, that is, $\Sigma = \mathrm{diag}(\sigma_{11}, \sigma_{22}, \ldots, \sigma_{tt})$ as given in (15). The HACF at lag k is then given by

$$\rho_k = \frac{\sum_{i=k}^{t-1}\psi_i\psi_{i-k}\sigma_{t-i,t-i}}{\sum_{i=0}^{t-1}\psi_i^2\sigma_{t-i,t-i}}. \tag{16}$$

Proof

Using (9) with $\sigma_{i,j} = 0 \ \forall i \neq j$, the ACVF at lag 0 is

$$\gamma_0 = \psi_0^2\sigma_{t,t} + \psi_1^2\sigma_{t-1,t-1} + \psi_2^2\sigma_{t-2,t-2} + \psi_3^2\sigma_{t-3,t-3} + \cdots + \psi_{t-1}^2\sigma_{1,1}$$

and the ACVF at lag 0, that is, the variance of the general *AR(p)* process is

$$\gamma_0 = \sum_{i=0}^{t-1}\psi_i^2\sigma_{t-i,t-i}. \qquad (17)$$

Using (11) with $\sigma_{i,j} = 0 \ \forall i \neq j$, the ACVF at lag 1 is

$$\gamma_1 = \psi_1\sigma_{t-1,t-1} + \psi_2\psi_1\sigma_{t-2,t-2} + \cdots + \psi_{t-1}\psi_{t-2}\sigma_{1,1}$$

so that the ACVF at lag 1 is

$$\gamma_1 = \sum_{i=1}^{t-1}\psi_i\psi_{i-1}\sigma_{t-i,t-i}. \qquad (18)$$

Similarly, the ACVF at lag k is

$$\gamma_k = \sum_{i=k}^{t-1}\psi_i\psi_{i-k}\sigma_{t-i,t-i}. \qquad (19)$$

Dividing (19) by (17), results in (16), which completes the proof.

Corollary 2: HACF at Lag k for an *AR(1)* Process

Consider an *AR(1)* process, $Z_t = \sum_{j=0}^{\infty}\psi_j e_{t-j}$, $\psi_j = \phi\psi_{j-1}$, with E (**e**$_t$) = **0**, and $\mathrm{Cov}(e_i, e_j) = \Sigma$, with $\sigma_{i,j} = 0 \ \forall i \neq j$, that is, $\Sigma = \mathrm{diag}(\sigma_{11}, \sigma_{22}, \ldots, \sigma_{tt})$ as given in (15). Then the HACF at lag k is given by

$$\rho_k = \frac{\sum_{i=k}^{t-1}\phi^{2i-k}\sigma_{t-i,t-i}}{\sum_{i=0}^{t-1}\phi^{2i}\sigma_{t-i,t-i}}. \qquad (20)$$

Proof

For an *AR(1)* process , because $\psi_j = \phi\psi_{j-1}$, it follows that $\psi_j = \phi^j$, for $j \geq 0$. From equations (17) and (19), $\gamma_0 = \sum_{i=0}^{t-1}\phi^{2i}\sigma_{t-i,t-i}$ and $\gamma_k = \sum_{i=k}^{t-1}\phi^{2i-k}\sigma_{t-i,t-i}$ are obtained and the proof is complete.

Special Case

Homoscedasticity exists if the diagonal elements of Σ in (4) are all identical and the disturbance term, e, is free from autocorrelation, that is, $\sigma_{ij} = 0 \forall i \neq j$. In this case, the disturbance term is a sequence of independent, identically distributed random variables.

Corollary 3.3: ACF at Lag k for an *AR(1)* Process Using Theorem (3.2)

Consider an *AR(1)* process, $Z_t = \sum_{j=0}^{\infty}\psi_j e_{t-j}$, $\psi_j = \phi\psi_{j-1}$, with E (**e**$_t$) = **0**, $\sigma_{i,j} = 0 \ \forall i \neq j$, and $\mathrm{Var}(e_t) = \sigma^2 \ \forall t$. For an *AR(1)*, $\psi_j = \phi^j$ for $j \geq 0$, taking $t \to \infty$ in equations (17) through (19), results in

$$\gamma_0 = \sigma^2\sum_{i=0}^{\infty}\phi^{2i} = \frac{\sigma^2}{1-\phi^2},$$

$$\gamma_1 = \sigma^2\sum_{i=1}^{\infty}\phi^{2i-1} = \phi\frac{\sigma^2}{1-\phi^2},$$

and

$$\gamma_k = \sigma^2\sum_{i=k}^{\infty}\phi^{2i-k} = \phi^k\frac{\sigma^2}{1-\phi^2},$$

respectively. The ACF at lag k is then given by $\rho_k = \phi^k, k \geq 0$, which is the well-known ACF for an *AR(1)* process.

Conclusion

This study investigated an important statistical problem concerning the autocorrelation function (ACF) in the presence of heteroscedasticity disturbances in p^{th} order autoregressive (*AR(p)*) processes. Explicit equations were derived for ACF when the disturbance follows the general covariance matrix, Σ, and when the diagonal elements of Σ are not all identical but $\sigma_{ij} = 0 \forall i \neq j$, i.e., $\Sigma = \text{diag}(\sigma_{11}, \sigma_{22}, \ldots, \sigma_{tt})$. Future research is needed to extend the explicit equations derived in this article for ACF in the presence of heteroscedasticity disturbances in the general form of the moving average models with order *q*, *MA(q)*.

References

Bera, A., Bubnys, E., & Park, H. (2005). Conditional heteroscedasticity in the market model and efficient estimates of betas. *Financial Review*, *23*(*2*), 201-214.

Box, G. E., Jenkins, G. M., & Reinsel, G. C. (1994). *Time series analysis* (3rd Ed.). New Jersey: Prentice Hall.

Brockwell, P. J., & Davis, R. A. (2002). *Introduction to time series and forecasting*. New York: Springer.

Bumb, B., & Kelejian, H. (1983). Autocorrelated and heteroscedastic disturbances in linear regression analysis: A Monte Carlo study. *The Indian Journal of Statistics*, *45*, *Series B*(*2*), 257-270.

Cragg, J. G. (1982). Estimation and testing in time series regression models with heteroscedastic disturbances. *Journal of Econometrics*, *20*, 135-157.

Demos, A. (2000). The autocorrelation function of conditionally heteroskedastic in mean models. *Athens University of Economics and Business, Department of International and European Economic Studies*.

Engle, R. F. (1982). Autoregressive Conditional Heteroscedasticity with Estimation of the variance of United Kingdom Inflation. *Econometrica*, *50*, 987-1008.

Evans, M. A., & King, M. L. (1988). A further class of tests for heteroscedasticity. *Journal of Econometrics*, *37*, 265-276.

Farebrother, R. W. (1987). The statistical foundations of a class of parametric tests for heteroscedasticity. *Journal of Econometrics*, *36*, 359-368.

Judge, G. G., Griffiths, W. E., Hill, R. C., Lütkepohl, H., & Lee. T. (1985). *The theory and practice of econometrics*. New York: John Wiley and Sons.

Kalirajan, K. P. (1989). A test for heteroscedasticity and non-normality of regression residuals. *Economics Letters*, *30*, 133-136.

Praetz, P. (2008). A note on the effect of autocorrelation on multiple regression statistics. *Australian & New Zealand Journal of Statistics*, *23*, 309-313.

Safi, S. (2009). Explicit equations for ACF in the presence of heteroscedasticity disturbances in first-order autoregressive models, AR(1). *The Journal of the Islamic University of Gaza*, *17*(*2*), 97-107.

Wallentin, B., & Agren, A. (2002). Test of heteroscedasticity in a regression model in the presence of measurement errors. *Economics Letters*, *76*, 205-211.

Journal of Modern Applied Statistical Methods
November 2011, Vol. 10, No. 2, 632-638

1538 – 9472/11/$95.00

Control Balanced Designs Involving Sequences of Treatments

Cini Varghese Seema Jaggi
Indian Agricultural Statistics Research Institute,
New Delhi, India

Designs involving sequences of treatments for test vs. control comparisons are suitable for research in which each experimental unit receives treatments over time in order to compare several test treatments to one (or more) control treatment(s). These designs can be advantageously used in screening experiments and bioequivalence trials. Three series of such designs are constructed in incomplete sequences wherein the first class of designs is variance balanced while the other two classes of designs are partially variance balanced for test versus test comparisons of both direct and residual effects of treatments.

Key words: Change over designs, direct effects, residual effects, control balance, variance balance, partial balance, bioequivalence trials.

Introduction

Change over designs (COD) are designs in which each experimental unit receives one or more treatments, one at a time, in successive periods. These designs also known as repeated measurement designs, crossover trials and designs involving sequences of treatments; they have been widely used in several fields of research, notably in nutrition experiments with dairy cattle, clinical trials, educational/ learning experiments, long-term agricultural field experiments and bioequivalence trials. A COD is one of the most suitable designs for experiments with animals as experimental units (different treatments) are often applied to the same animal in different periods. The distinguishing feature of such an experiment is that any treatment applied to a unit in a certain period influences the responses of the unit not only in the period of its application but also leaves residual effects in the succeeding periods.

In some experimental situations involving treatment sequences, researchers are interested in comparing several new (test) treatments to one (or more) established (standard or control) treatment(s) rather than in all pair-wise comparisons. That is, the researcher is interested in drawing inferences based on a subset of comparisons among treatments; special designs giving more importance to test versus control comparisons must be developed to meet requirements in these cases. Using such a design would allow a researcher to screen out best test treatments as compared to existing control treatment(s). This type of design is also useful in bioequivalence trials (such as veterinary medicinal trials) where a set of test formulations are to be compared to established reference formulations before sanctioning the marketing patent for a newly produced formulation.

Usage of CODs for test versus control comparisons began with the introduction of control balanced CODs by Pigeon and Raghavarao (1987), who derived a set of necessary conditions for the existence of control balanced CODs (CODs balanced for test vs. control comparisons). They provided construction methods using existing balanced

Cini Varghese is a Senior Scientist in Agricultural Statistics. Her research interests include: construction and analysis of experimental designs, characterization properties of experimental designs, web generations of designs and software development. Email her at: cini_v@iasri.res.in. Seema Jaggi is a Senior Scientist in Agricultural Statistics. Her research interests include: design of experiments, statistical computing, statistical techniques in agricultural research, web solutions to experimental designs and analysis. Email her at: seema@iasri.res.in.

CODs, pairwise balanced designs and also the method of differences. Majumdar (1988) obtained some optimal control balanced designs involving sequences of treatments when number of treatments is less than the number of periods and showed that the designs can be constructed from existing strongly balanced uniform circular/non-circular CODs in test treatments by changing some test treatment labels into control. Koch, et al. (1989) studied a two-period COD for the comparison of two active treatments and placebo. Hedayat and Zhao (1990) investigated two classes of efficient CODs for the purpose of comparing several test treatments to a control treatment when the number of periods is two.

Ting (2002) constructed optimal designs for the estimation of control-test treatment contrasts in a COD set up. Aggarwal, et al. (2004) developed families of CODs for test versus control comparisons by juxtaposing Williams (1949) Latin square(s) by using block contents of various classes of balanced incomplete block designs and an orthogonal array of type 1 and strength 2. Aggarwal, et al. (2004) showed that these designs are optimal. Hedayat and Yang (2005) provided some construction methods for obtaining control balanced CODs. Most of these designs are balanced for carryover effects, but require a large number of experimental periods as well as subjects. Hedayat and Yang (2005) also characterized a class of designs that are optimal for comparing several test treatments with a control. Yang and Park (2007) obtained efficient CODs for comparing test treatments with a control treatment with three periods. Aggarwal and Jha (2009) suggested methods for constructing CODs to compare v test treatments with a control treatment when the number of periods is no larger than v+1.

This study constructed a series of control balanced designs involving sequences of treatments in three periods that are variance balanced. Another class of partially balanced designs involving incomplete sequences based on mutually orthogonal Latin squares was also obtained. In addition, a third series of control balanced designs in incomplete sequences of two distinct sets of treatments was obtained to compare one set of test treatments with two control treatments. Some definitions are given below that would be used in the subsequent sections.

Definitions

The following designs relate to studies involving treatment sequences.

Control Balanced Design

A control balanced COD for t + c (= t test + c control) treatments in p periods and n experimental units for test versus control comparisons is said to be balanced in the presence of residual effects, if:

(a) Each test treatment occurs ω_t times and each control treatment occurs ω_c times in each period;

(b) Each test treatment is immediately preceded by every other test treatment equally often, for example, $\upsilon_{tt'}$ $(t \neq t')$;

(c) Each control treatment is immediately preceded by every other control treatment equally often, for example, $\upsilon_{cc'}$ $(c \neq c')$; and

(d) Each control treatment is immediately preceded by every test treatment and vice versa equally often, for example, υ_{tc}.

It may be noted that when $\omega_t = \omega_c$ and $\upsilon_{tt'} = \upsilon_{cc'} = \upsilon_{tc}$, these designs reduce to conventional CODs balanced for first order residual effects.

Variance Balanced Design

A control balanced COD for t + c (= t test + c control) treatments in p periods and n experimental units for test versus control comparisons is said to be variance balanced in the presence of residual effects, if all elementary contrasts pertaining to:

(a) Direct (residual) effects among test treatments are estimated with the same variance, $V_{tt'd}$ ($V_{tt'r}$) $(t \neq t')$; and

(b) Direct (residual) effects among test versus control treatment are estimated with the same variance, V_{tcd} (V_{tcr}).

Partially Balanced Design

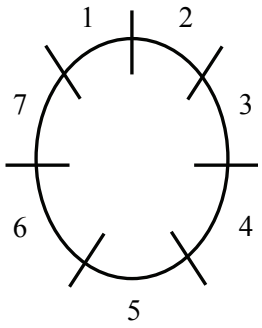
A control balanced COD for t + c (= t test + c control) treatments in p periods and n experimental units for test versus control comparisons is said to be partially variance balanced with an underlying m-class association scheme in the presence of residual effects, if all elementary contrasts pertaining to:

(a) Direct (residual) effects among test treatments that are i^{th} associates to each other, are estimated with the same variance $V_{tt'id}$ ($V_{tt'ir}$) ($t \neq t'$; i = 1,2,…m); and

(b) Direct (residual) effects of test and control treatment are estimated with the same variance V_{tcd} (V_{tcr}).

Circular Association Scheme

Let there be t test treatments arranged on the circumference of a circle. For a given treatment, the treatments that appear at i^{th} positions on its either side are i^{th} associates [i = 1, 2, …, (t–1)/2 if t is odd, or t/2 if t is even]. For odd t , there are always two i^{th} associates of each treatment for i = 1, 2, …, (t–1)/2, and for an even t there are two i^{th} associates of every treatment for i = 1, 2, …,(t-1)/2 and one associate for i = t/2. The arrangement of 7 treatments in a circular association scheme could be:



The first, second and third associates of the 7 treatments are:

Treatment	First Associates	Second Associates	Third Associates
1	2, 7	3, 6	4, 5
2	1, 3	4, 7	5, 6
3	2, 4	1, 5	6, 7
4	3, 5	2, 6	1, 7
5	4, 6	3, 7	1, 2
6	5, 7	1, 4	2, 3
7	6, 1	2, 5	3, 4

Experimental Design 1: Control Balanced Designs Involving Treatment Sequences in Three Periods

Arrange all possible distinct pairs from t test treatments ($2\,^{t}C_{2}$ pairs) excluding the identical pairs in $2\,^{t}C_{2}$ rows of size 2 each; repeat the $2\,^{t}C_{2}$ pairs 3 times. In the first set, append a column containing all elements as control treatment (t+1), $2\,^{t}C_{2}$ times as the first column; in the second set append a column of control treatment (t+1), $2\,^{t}C_{2}$ times as the second column and in the third set append a column of control treatment (t+1), $2\,^{t}C_{2}$ times as the third column. Juxtapose the three sets, side by side, so that the resulting arrangement has 3 columns and $6\,^{t}C_{2}$ rows. Treating columns as periods and rows as experimental units, this arrangement yields a control balanced COD balanced for first residual effects for testing v treatments with a control in 3 periods and $6\,^{t}C_{2}$ units.

Example 1

A control balanced three-period COD balanced for first residual effects for comparing 3 test treatments (denoted by 1, 2, 3) with one control treatment (denoted by 0) in 18 experimental units is:

Experimental Unit	Period		
	i	ii	iii
i	0	1	2
ii	0	1	3
iii	0	2	1
iv	0	2	3
v	0	3	1
vi	0	3	2
vii	1	0	2
viii	1	0	3
ix	2	0	1
x	2	0	3
xi	3	0	1
xii	3	0	2
xiii	1	2	0
xiv	1	3	0
xv	2	1	0
xvi	2	3	0
xvii	3	1	0
xviii	3	2	0

A program was developed using SAS software PROC IML for calculating the variance estimates of contrasts among test treatments and the variances estimates of contrasts pertaining to test versus control treatments for direct and residual treatment effects. Table 1 shows a list of designs for comparing t (≤10) test treatments with c (=1) control treatment in p (≤10) periods, n (≤100) units, along with variances.

Table 1 also shows that the designs are variance balanced. It also shows that estimate variances of the contrasts between test versus control treatment of direct effects is less than those of residual effects. Further, variances of the estimates of contrasts between test versus control treatment is less compared to those of test versus test treatments in the case of both direct and residual treatment effects.

Experimental Design 2: Control Balanced Designs Involving Incomplete Treatment Sequences Using MOLS

Append a complete set of (t−1) mutually orthogonal Latin squares (MOLS) for prime number t of treatment symbols (Fisher & Yates, 1963) one after another. This arrangement has t columns and (t−1)×t rows. Delete the last column of the array resulting in (t−1) columns and (t−1)×t rows. Replace the first set of t elements in the first column, second set of t elements in the second column, …, $(t-1)^{th}$ set of t elements in the last column, by the control treatment (t+1). Treating columns as periods and rows as experimental units, the final arrangement results into a control balanced COD for t tests treatments and 1 control treatment in p (= t−1) periods and (t−1)×t units.

Table 1: List of Control Balanced Designs Involving Treatment Sequences in Three Periods

S. No.	t	p	n	$\sigma^{-2} V_{tt'd}$	$\sigma^{-2} V_{tcd}$	$\sigma^{-2} V_{tt'r}$	$\sigma^{-2} V_{tcr}$
1	3	3	18	0.2455	0.1860	0.4091	0.3239
2	4	3	36	0.1741	0.1174	0.2813	0.1992
3	5	3	60	0.1349	0.0852	0.2143	0.1420
4	6	3	90	0.1101	0.0667	0.1731	0.1096

Example 2

A control balanced COD for comparing 5 test treatments (denoted by 1, 2, 3, 4, 5) with one control treatment (denoted by 0) in 4 periods and 20 units is:

Experimental Unit	Period			
	i	ii	iii	iv
i	0	2	3	4
ii	0	3	4	5
iii	0	4	5	1
iv	0	5	1	2
v	0	1	2	3
vi	1	0	5	2
vii	2	0	1	3
viii	3	0	2	4
ix	4	0	3	5
x	5	0	4	1
xi	1	4	0	5
xii	2	5	0	1
xiii	3	1	0	2
xiv	4	2	0	3
xv	5	3	0	4
xvi	1	5	4	0
xvii	2	1	5	0
xviii	3	2	1	0
xix	4	3	2	0
xx	5	4	3	0

Table 2 shows a list of designs has been prepared for t test treatments and c (=1) control treatment, where t is a prime number less than 15. As shown, the designs are partially variance balanced with an underlying varying circular association scheme for test versus test comparisons. Hence, average variance was computed for such comparisons for both the cases of direct ($\sigma^{-2}\overline{V}_{tt'd}$) as well as residual ($\sigma^{-2}\overline{V}_{tt'r}$) effects. Variances of the estimates of contrasts between test versus control treatment of direct effects is less than those of residual effects (see Table 2). Variances of the estimates of contrasts between test versus control treatment is less compared to those of test versus test treatments in both cases of direct effects as well as residual treatment effects.

Experimental Design 3: Control Balanced Designs Involving Incomplete Sequences of Two Distinct Sets of Treatments

In the (t–1) columns and (t–1)×t rows arrangement previously obtained with the MOLS method, replace the first set of t elements in the first column by the first control and first set of the last column by the second control, second set of t elements in the second column by the first control and second set of last but one column by the second control and so on. Thus in each set of t rows, t treatments is replaced by the first control in a staircase descending fashion and t treatments are replaced by the second control in a staircase fashion circularly until each column is replaced by both controls. Treating columns as periods and rows as

Table 2: List Control Balanced Designs Involving Incomplete Sequences of Two Distinct Sets of Treatments

S. No.	t	p	n	$\sigma^{-2}\overline{V}_{tt'd}$	$\sigma^{-2}V_{tcd}$	$\sigma^{-2}\overline{V}_{tt'r}$	$\sigma^{-2}V_{tcr}$
1	5	4	20	0.2122	0.1582	0.2954	0.2248
2	7	6	42	0.0733	0.0610	0.0902	0.0754
3	9	8	72	0.0375	0.0329	0.0434	0.0382
4	11	10	110	0.0229	0.0206	0.0257	0.0232
5	13	12	156	0.0155	0.0142	0.0170	0.0156

experimental units, the final arrangement results in a control balanced design involving sequences of treatments for t test treatments and 2 control treatments in p (= t−1) periods and (t−1)×t units.

Example 3

A control balanced design involving sequences of treatments for comparing 5 test treatments (denoted by 1, 2, 3, 4, 5) with 2 control treatments (denoted by 0_1 and 0_2) in 4 periods and 20 units is:

Experimental Unit	Period			
	i	ii	iii	iv
i	0_1	2	3	0_2
ii	0_1	3	4	0_2
iii	0_1	4	5	0_2
iv	0_1	5	1	0_2
v	0_1	1	2	0_2
vi	1	0_1	0_2	2
vii	2	0_1	0_2	3
viii	3	0_1	0_2	4
ix	4	0_1	0_2	5
x	5	0_1	0_2	1
xi	1	0_2	0_1	5
xii	2	0_2	0_1	1
xiii	3	0_2	0_1	2
xiv	4	0_2	0_1	3
xv	5	0_2	0_1	4
xvi	0_2	5	4	0_1
xvii	0_2	1	5	0_1
xviii	0_2	2	1	0_1
xix	0_2	3	2	0_1
xx	0_2	4	3	0_1

Table 3 shows a list of designs prepared for comparing t test treatments with c (=2) control treatments, where t is a prime number less than 15. These designs are partially balanced based on varying circular association scheme for test versus test comparisons pertaining to direct as well as residual effects of treatments. Hence average variance was calculated for these comparisons in case of direct ($\sigma^{-2}\,\bar{V}_{tt'd}$) as well as residual ($\sigma^{-2}\,\bar{V}_{tt'r}$) effects.

Table 3 shows that the variances of estimates of contrasts between test versus control treatment of direct effects is less than those of residual effects. Also, that variances of estimates of the contrasts between test versus control treatment is less as compared to those of test versus test treatments in both the cases.

References

Aggarwal, M. L., Deng, L-Y, & Jha, M. K. (2004). Some new residual treatment effects designs for comparing test treatments a control. *Journal of Applied Statistics*, *31*(*9*), 1065-1081.

Aggarwal, M. L., & Jha, M. K. (2009). Constructions of residual treatment effects designs for comparing test treatments with a control. *Communications in Statistical Theory and Methods*, *38*(*15*), 2567-2577.

Fisher, R. A., & Yates, F. (1963). Statistical tables for Biological. *Agricultural and Medical Research*, *6*, 88-89.

Hedayat, A. S., & Zhao, W. (1990). Optimal two-period repeated measurements designs. *Annals of Statistics*, *18*(*4*), 1805-1816.

Hedayat, A. S., & Yang, M. (2005). Optimal and efficient crossover designs for comparing test treatments with a control treatment. *Annals of Statistics*, *33*(*2*), 915-943.

Koch, G. G., Amara, I. A., Brown, Jr, B. W., Colton, T., & Gillings, D. B. (1989). A two-period cross-over design for the comparison of two active treatments and placebo. *Statistics in Medicine*, *8*, 487-504.

Majumdar, D. (1988). Optimal repeated measurement designs for comparing test treatments with a control. *Communications in Statistical Theory and Methods*, *17*(*11*), 3687-3703.

Pigeon, J. G., & Raghavarao, D. (1987). Cross-over designs for comparing treatments with a control. *Biometrika*, *72*(*2*), 321-328.

Ting, P.C. (2002). Optimal and efficient repeated measurements designs for comparing test treatments with a control. *Metrika*, *56*(*3*), 229-238.

Table 3: List of Control Balanced Designs Involving Incomplete Sequences of Two Distinct Sets of Treatments

S. No.	t	p	n	$\sigma^{-2}\bar{V}_{tt'd}$	$\sigma^{-2}V_{tcd}$	$\sigma^{-2}\bar{V}_{tt'r}$	$\sigma^{-2}V_{tcr}$
1	5	4	20	0.3204	0.2210	0.4663	0.3065
2	7	6	42	0.0960	0.0719	0.1176	0.0887
3	9	8	72	0.0446	0.0363	0.0516	0.0421
4	11	10	110	0.0261	0.0222	0.0292	0.0249
5	13	12	156	0.0171	0.0150	0.0188	0.0165

Williams, E. J. (1949). Experimental designs balanced for the estimation of residual effects of treatments. *Australian Journal of Scientific Resources*, *A2*, 149-168.

Yang, M., & Park, M. (2007). Efficient crossover designs for comparing test treatments with a control treatment when p = 3. *Journal of Statistical Planning and Inference*, *137*(*6*), 2056-2067.

Journal of Modern Applied Statistical Methods
November 2011, Vol. 10, No. 2, 639-645

1538 – 9472/11/$95.00

Construction of Control Charts Based On Six Sigma Initiatives for the Number of Defects and Average Number of Defects per Unit

R. Radhakrishnan
P.S.G. College of Arts and Science,
Coimbatore, India

P. Balamurugan
The Kavery Engineering College,
Tamilnadu, India

A control chart is a statistical device used for the study and control of a repetitive process. In 1931, Shewart suggested control charts based on 3 sigma limits. Today manufacturing companies around the world apply Six Sigma initiatives, with a result offewer product defects. Companies practicing Six Sigma initiatives are expected to produce 3.4 or less number of defects per million opportunities, a concept suggested by Motorola in 1980. If companies practicing Six Sigma initiatives use control limits suggested by Shewhart, then no points will fall outside the control limits due to the improvement in the quality of the process. ASix Sigma based control chart is constructed for the number of defects and average number of defects per unit. Tables are providedto aid engineers in decision making.

Key words: Six Sigma quality level, control chart, process control, Six Sigma.

Introduction

The concept of Six Sigma was introduced in 1980 by engineer M. Harry at Motorola. Harry analyzed variations in outcomes of the company's internal procedures and realized that by measuring variations it was possible to improve the working of the system. The procedure was designed to improve overall performance. Companies practicing Six Sigma are expected to produce 3.4 or less number of defects per million opportunities. Radhakrishnan and Sivakumaran (2008a, 2008b, 2008c, 2009a, 2009b, 2010) used the concept of Six Sigma in the construction of sampling plans, such as single, double and repetitive group sampling plans indexed through Six Sigma Quality Levels (SSQLs) with the Poisson distribution as the base line distribution. Radhakrishnan (2009) suggested a single sampling plan indexed through SSQLs based on Intervened Random Effect Poisson Distribution and the Weighted Poisson Distribution as the base line distributions. Radhakrishnan and Balamurugan (2010) constructed Six Sigma based control charts for the number of defectives. The control charts originated by W. A. Shewhart (1931) were based on 3 sigma control limits; if these same charts are used for the products of companies adopting Six Sigma initiatives in their processes, then no points will fall outside the control limits due to the improvement in quality. Thus, a separate control chart is required to monitor the outcomes of the companies that adopt Six Sigma initiatives.

R. Radhakrishnan holds a Bachelor and post graduate degrees in Statistics, M.Phil., Ph.D and a post graduate degree in Business Administration. He has 31 years of experience teaching theoretical and applied statistics, has presented more than 150 papers at national and international conferences and has published more than 100 articles. He is a quality auditor for ISO certification and a certified Six Sigma Black Belt. Email him at: rkrishnan_cbe@yahoo.com. P. Balamurugan is a Lecturer in Statistics. He holds a Bachelor degree, a post graduate degree and M.Phil. in statistics. He is a Research Scholar under the guidance of R. Radhakrishnan.

Definitions

- Upper specification limit (USL): The greatest amount specified by the producer for a process or product to have acceptable performance.

- Lower specification limit (LSL): The smallest amount specified by the producer for a process or product to have acceptable performance.

- Tolerance level (TL): The difference between USL and LSL, TL = USL–LSL.

- Process capability (C_p): The ratio of tolerance level to six times standard deviation of the process.

$$c_p = \frac{T_l}{6\sigma} = \frac{USL - LSL}{6\sigma}$$

- Subgroup size (N):The total number of samples.

- Subgroup size (n): The choice of the sample size n and the frequency of sampling.

- Quality control constants ($L_{6\sigma}$ & $R_{6\sigma}$): The constants introduced in this article, $L_{6\sigma}$ and $R_{6\sigma}$, determine the control limits based on Six Sigma initiatives for the number of defects and average number of defects per unit respectively.

Conditions for Application

1. Human involvement should be less in the manufacturing process; and

2. The company adopts Six Sigma quality initiatives in its processes.

Construction of Control Charts Based On Six Sigma Initiatives for the Number of Defects

Fix the tolerance level (TL) and process capability (C_P) to determine the process standard deviation ($\sigma_{6\sigma}$). Apply the value of $\sigma_{6\sigma}$ in the control limits $\overline{c} \pm L_{6\sigma}\sigma_{6\sigma}$, to find the control limits for the Six Sigma based control chart for the number of defects. The value of $L_{6\sigma}$ = 4.831 isobtained using

$$p(z \le z_{ss}) = 1 - \alpha_1, \alpha_1 = 3.4 \text{ x } 10^{-6},$$

where z is a standard normal variate. For a specified TL and C_P of the process, the values of σ (termed as $\sigma_{6\sigma}$) are calculated from $c_p = \frac{T_L}{6\sigma}$ using a C program and are presented in Table 3 for various combinations of TL and C_P. The control limits based on Six Sigma initiatives for the number of defects are:

$$\text{UCL}_{6\sigma} = \overline{c} + L_{6\sigma}\sigma_{6\sigma}$$
$$\text{Central Line CL} = \overline{c}$$
$$\text{LCL}_{6\sigma} = \overline{c} - L_{6\sigma}\sigma_{6\sigma}.$$

Example 1

Consider an example from Mahajan (2005).Table 1 shows the numbers of missing rivets noted at aircraft final inspection.

Table 1: Missing Rivets Noted for Aircraft

Airplane No.	No. of Missing Rivets
1	8
2	16
3	14
4	19
5	11
6	15
7	8
8	11
9	21
10	12
11	23
12	16
13	9
14	25
15	15
16	9
17	9
18	14
19	11
20	9
21	10
22	22
23	7
24	28
25	9

Where

$$\bar{c} = \frac{\text{Number of defects in all samples}}{\text{Total number of samples}}$$

and

$$\bar{c} = \frac{\sum c}{N} = \frac{351}{25} = 14.04.$$

Three Sigma Control Limits for the Number of Defects

The 3σ control limits suggested by Shewhart (1931) are:

$$\begin{aligned} UCL_{3\sigma} &= \bar{c} + 3\sqrt{\bar{c}} \\ &= 14.04 + 3\sqrt{14.04} = 25.28 \\ CL_{3\sigma} &= \bar{c} = 14.04 \\ LCL_{3\sigma} &= \bar{c} - 3\sqrt{\bar{c}} \\ &= 14.04 - 3\sqrt{14.04} = 2.80 \end{aligned}$$

Figure 1 shows that airplane number 24falls above the upper control limit; therefore the process does not exhibit statistical control.

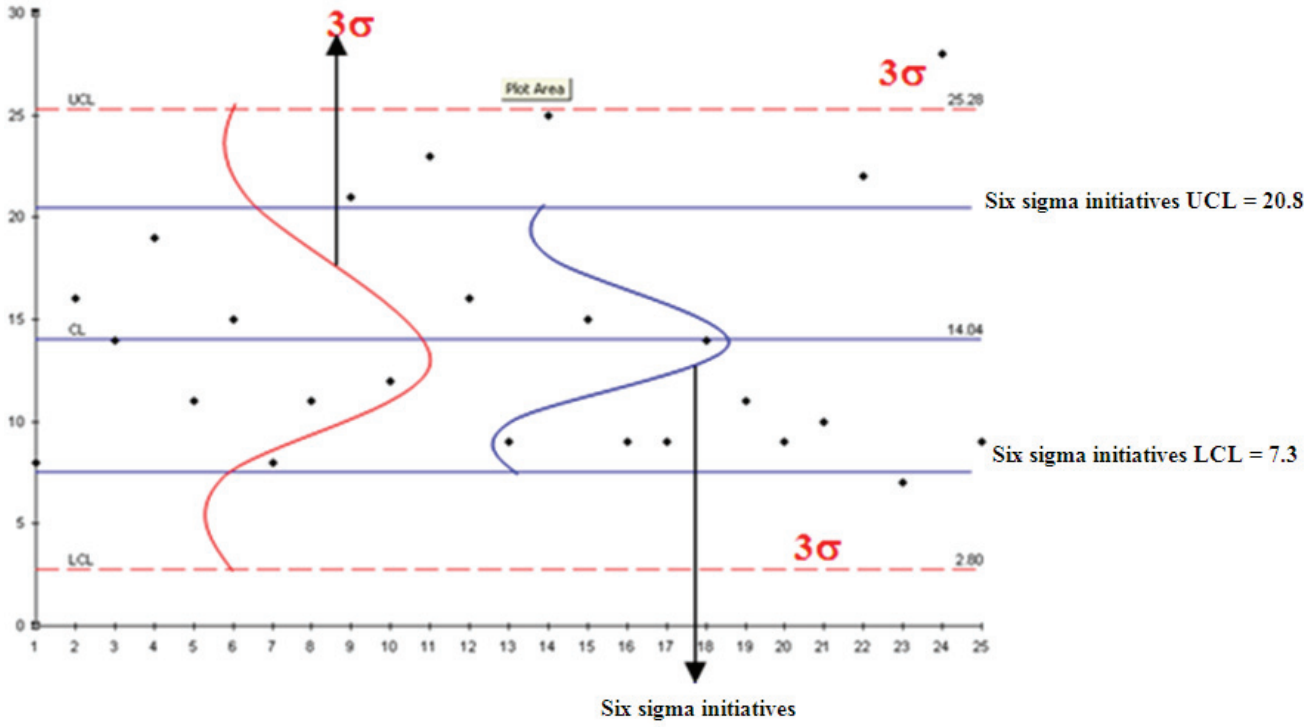
Control Limits Based on Six Sigma Initiatives for the Number of Defects

For a given TL = 21 (USL-LSL =28-7) & C_P = 2.5, Table 3 shows that the value of $\sigma_{6\sigma}$ is 1.4. The control limits based on Six Sigma initiatives for the number of defects for a specified TL and $L_{6\sigma}$ are $\bar{c} \pm 4.831\sigma_{6\sigma}$ with

$$\begin{aligned} UCL_{6\sigma} &= \bar{c} + L_{6\sigma}\sigma_{6\sigma} \\ &= 14.04 + (4.831 \times 1.4) = 20.8 \\ CL_{6\sigma} &= \bar{c} = 14.04 \\ LCL_{6\sigma} &= \bar{c} - L_{6\sigma}\sigma_{6\sigma} \\ &= 14.04 - (4.831 \times 1.4) = 7.3 \end{aligned}$$

Figure 1 shows that airplane numbers 9, 11, 14, 22 and 24 are above the upper control limit and airplane number 23falls below the lower control limit; therefore the process does not exhibit statistical control.

Figure 1: Process Comparison for 3σLimits and Control Limits Using Six Sigma Initiatives



Construction of Control Chart Based On Six Sigma Initiatives for Average Number Defects per Unit

Fix the tolerance level (TL) and process capability (C_P) to determine the process standard deviation ($\sigma_{6\sigma}$). Apply the value of $\sigma_{6\sigma}$ in the control limits $\bar{u} \pm R_{6\sigma}\sigma_{6\sigma}$, to obtain the control limits for the control chart based on Six Sigma initiatives for average number of defects per unit. The value of $R_{6\sigma}$ is obtained using

$$p(z \le z_{ss}) = 1 - \alpha_1, \alpha_1 = 3.4 \text{ x } 10^{-6}$$

where z is a standard normal variate. For a specified TL and C_P of the process, the value of σ (termed as $\sigma_{6\sigma}$) is calculated from $c_p = \frac{T_L}{6\sigma}$ using a C program. Table 4 presents calculated 6σ values for various combinations of TL and C_P. Further, the value of $R_{6\sigma}$ is obtained using the procedure given above and presented in Table 5 for various sample sizes. The control limits based on six sigma initiatives for average number of defects per unit are

$$UCL_{6\sigma} = \bar{u} + R_{6\sigma}\sigma_{6\sigma}$$
$$\text{Central Line, } CL_{6\sigma} = \bar{u}$$
$$LCL_{6\sigma} = \bar{u} - R_{6\sigma}\sigma_{6\sigma}.$$

Example 2

Consider an example provided by Mahajan (2005). Table 2 shows the average number of outlet leaks per radiator for 10 lots (n) of 100 radiators (N) each.

The mean number of defects per unit in the lot, based on all the n samples is given by

$$\bar{u} = \frac{1}{n}\sum_{i=1}^{n} u_i = \frac{1.23}{10} = 0.123.$$

Table 2: Average Number of Outlet Leaks per Radiator

Lot No.	No. of Leaks (c)	Leaks per Radiator (c/N)
1	15	0.15
2	17	0.17
3	12	0.12
4	16	0.16
5	14	0.14
6	5	0.05
7	14	0.14
8	11	0.11
9	9	0.09
10	10	0.10
Total		$\sum$1.23

Three Sigma Control Limits for Average Number of Defects per Unit

The 3σ control limits suggested by Shewhart (1931) are

$$UCL_{3\sigma} = \bar{u} + 3\sqrt{\bar{u}/n}$$
$$= 0.123 + 3\sqrt{0.123/100} = 0.228$$
$$CL_{3\sigma} = \bar{u} = 0.123$$
$$LCL_{3\sigma} = \bar{u} - 3\sqrt{\bar{u}/n}$$
$$= 0.123 - 3\sqrt{0.123/100} = 0.018$$

Figure 1 shows that the process is in control because all the samples lie within the control limits.

Control Limits Based on Six Sigma Initiatives for Average Number of Defects per Unit

For a given TL = 0.12 (USL-LSL =0.17-0.05) and C_P = 2.5, Table4 shows that the value of $\sigma_{6\sigma}$ is 0.008. The control limits based on Six Sigma initiatives for the average number of defects per unit chart for a specified TL and $\sigma_{6\sigma}$ are $\bar{u} \pm R_{6\sigma}\sigma_{6\sigma}$ with

$$UCL_{6\sigma} = \bar{u} + R_{6\sigma}\sigma_{6\sigma}$$
$$= 0.123 + (0.4831 \times 0.008) = 0.127$$
$$CL_{6\sigma} = \bar{u} = 0.123$$
$$LCL_{6\sigma} = \bar{u} - R_{6\sigma}\sigma_{6\sigma}$$
$$= 0.123\text{-}(0.4831 \times 0.008) = 0.12.$$

Figure 2 illustrates that the process is out of control because only one airplane number lies inside the control limits; thus, the process does not exhibit statistical control.

Conclusion

This article provided a procedure to construct control charts based on Six Sigma initiatives for the number of defects and average number of defects per unit. Using examples, itwas found that the examined processeswere not in control even when Six Sigma initiatives were adopted. It is clear from the comparison that when the process is centered with reduced variation many points fall outside the control limits, thus indicating that the processes are not at expected levels; thus, a correction in the process is required to reduce variations. The charts suggested herein may be useful for companies practicing Six Sigma initiatives in their process. These charts can be used to replace existing Shewhart (1931) control charts implemented when companies first started implementing Six Sigma Initiatives.

References

Mahajan, M. (2005).*Statistical quality control* (*Revised Ed.*).Delhi, India: Dhanpat Rai & Co.,Ltd.

Radhakrishnan, R. (2009). *Construction of Six Sigma based sampling plans*. Unpublished D.Sc. thesis, Bharathiar University, Coimbatore, India.

Radhakrishnan, R., & Sivakumaran, P. K. (2008a). Construction and selection of Six Sigma sampling plan indexed through Six Sigma quality level.*International Journal of Statistics and Systems*, *3*(*2*), 153-159.

Radhakrishnan, R., & Sivakumaran, P. K. (2008b). Construction of Six Sigma repetitive group sampling plan.*International Journal of Mathematics and Computation*, *1*(*8*), 75-83.

Figure 2: Process Comparison of 3σLimits and Control Limits Using Six Sigma Initiatives

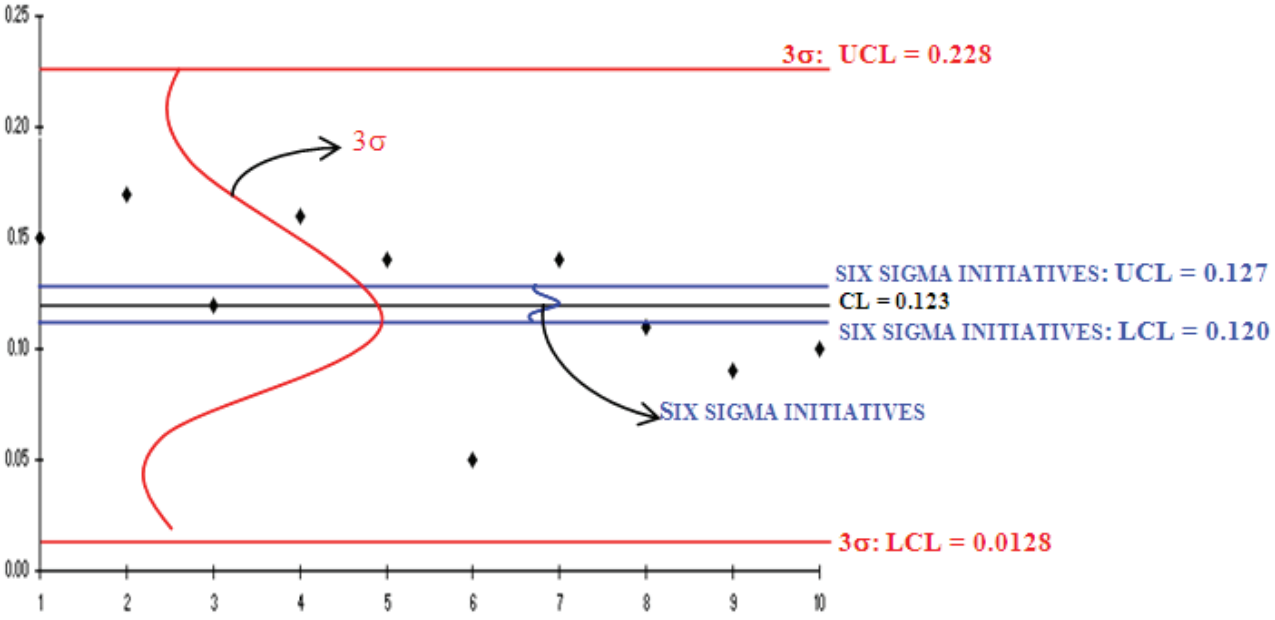


Table 3: $\sigma_{6\sigma}$ Values for Specified C_p and TLfor the Number of Defects

C_p	TL					
	20	21	22	23	24	25
1.0	3.3	3.5	3.7	3.8	4.0	4.2
1.1	3.0	3.2	3.3	3.5	3.6	3.8
1.2	2.8	3.0	3.1	3.2	3.3	3.5
1.3	2.6	2.7	2.8	2.9	3.1	3.2
1.4	2.4	2.5	2.6	2.7	2.9	3.0
1.5	2.2	2.3	2.4	2.6	2.7	2.8
1.6	2.1	2.2	2.3	2.4	2.5	2.6
1.7	2.0	2.1	2.2	2.3	2.4	2.5
1.8	1.9	1.9	2.0	2.1	2.2	2.3
1.9	1.8	1.8	1.9	2.0	2.1	2.2
2.0	1.7	1.8	1.8	1.9	2.0	2.1
2.1	1.6	1.7	1.7	1.8	1.9	2.0
2.2	1.5	1.6	1.7	1.7	1.8	1.9
2.3	1.4	1.5	1.6	1.7	1.7	1.8
2.4	1.4	1.5	1.5	1.6	1.7	1.7
2.5	1.3	1.4	1.5	1.5	1.6	1.7

Table 4: $\sigma_{6\sigma}$ Values for Specified C_P and TLfor the Average Number of Defects per Unit

C_p	TL					
	0.10	0.11	0.12	0.13	0.14	0.15
1.0	0.017	0.018	0.020	0.022	0.023	0.025
1.1	0.015	0.017	0.018	0.020	0.021	0.023
1.2	0.014	0.015	0.017	0.018	0.020	0.021
1.3	0.013	0.014	0.015	0.017	0.018	0.019
1.4	0.012	0.013	0.014	0.015	0.017	0.018
1.5	0.011	0.012	0.013	0.014	0.016	0.017
1.6	0.010	0.011	0.013	0.014	0.015	0.016
1.7	0.010	0.011	0.012	0.013	0.014	0.015
1.8	0.009	0.010	0.011	0.012	0.013	0.014
1.9	0.009	0.010	0.010	0.011	0.012	0.013
2.0	0.008	0.009	0.010	0.010	0.012	0.013
2.1	0.008	0.009	0.010	0.010	0.011	0.012
2.2	0.008	0.008	0.009	0.010	0.011	0.011
2.3	0.007	0.008	0.009	0.009	0.010	0.011
2.4	0.007	0.008	0.008	0.009	0.010	0.010
2.5	0.007	0.007	0.008	0.009	0.009	0.010

Table 5: $R_{6\sigma}$ Values for a Specified Subgroup Size (n) for Average Number of Defects per Unit

Subgroup Size (n)	$R_{6\sigma}$
100	0.4831
101	0.4807
102	0.4783
103	0.4760
104	0.4737
105	0.4715
106	0.4692
107	0.4670
108	0.4649
109	0.4627
110	0.4606

Radhakrishnan, R., & Sivakumaran, P. K. (2008c). Construction and selection of conditional double sampling plan indexed through Six Sigma quality levels. *Sri Lankan Journal of Applied Statistics*, *9*, 74-83.

Radhakrishnan, R., & Sivakumaran, P. K. (2009a). Construction of Six Sigma double sampling plans. *Proceedings of the National Seminar IT and Business Intelligence*, Nagpur, India.

Radhakrishnan, R., & Sivakumaran, P. K. (2009b). Construction of double sampling plans through Six Sigma quality levels. *Proceedings of the IEEE Second International Joint Conference on Computational Sciences and Optimization*, Sanya, Hainan, China, 1027-1030.

Radhakrishnan, R., & Sivakumaran, P. K. (2010). Construction and selection of tightened-normal-tightened schemes of type tnt-(n_1,n_2; c) indexed through Six Sigma quality levels. *Proceedings of the 2010 International Conference onIndustrial Engineering and Operations Management*, Dhaka, Bangladesh, 93.

Radhakrishnan, R., & Balamurugan, P. (2010). Six Sigma based control charts for number of defectives. *Proceedings of the 2010 International Conference on Industrial Engineering and Operations Management*, Dhaka, Bangladesh, 92.

Shewhart, W. A. (1931). *Economic control of quality of manufactured product*. New York, NY: Van Nostrand.

Journal of Modern Applied Statistical Methods
November 2011, Vol. 10, No. 2, 646-655

1538 – 9472/11/$95.00

Non-homogenous Poisson Process for Evaluating Stage I & II Ductal Breast Cancer Treatment

Chris P. Tsokos
University of South Florida,
Tampa, FL

Yong Xu
Radford University,
Radford, VA

Non-Homogenous Poisson Process (NHPP), also known as the Power Law process (PLP) or the Weibull Process, is used to evaluate the effectiveness of a given treatment for Stage I & II ductal breast cancer patients. The behavior of the shape parameter of the intensity function is examined to evaluate the response of a given treatment with respect to its effectiveness for a cancer subject.

Key words: Statistical modeling, power law process, Weibull process, non-homogenous Poisson process, intensity function, cancer analysis.

Introduction

Breast cancer (malignant breast neoplasm) is cancer originating from breast tissue, most commonly from the inner lining of milk ducts or the lobules that supply the ducts with milk (Sariego, 2010). This study uses the Non-Homogenous Poisson Process (NHPP), also known as the Power Law Process (PLP) or the Weibull Process, to evaluate the effectiveness of a given treatment for Stage I & II ductal breast cancer patients. The behavior of the shape parameter of the intensity function is examined to evaluate the response of a given treatment with respect to its effectiveness for the cancer subject. Data from the Surveillance Epidemiology and End Results (SEER) Program is used to test the proposed model. This data is collected by the U.S. National Institutes of Health (NIH) (2010) and includes information on incidence, survival and prevalence from specific geographic areas representing 26% of the U.S. population; the NIH also compiles reports on several types of cancer and includes mortality rates in the SEER database.

Chris Tsokos is a Distinguished University Professor in mathematics and Statistics at the University of South Florida. His research interests are in modeling Global Warming, analysis and modeling of cancer data, parametric, Bayesian and nonparametric reliability, and stochastic systems, among others. He is a fellow of both ASA and ISI. Email him at: profcpt@cas.usf.edu. Yong Xu is an Assistant Professor in the Department of Mathematics and Statistics at Radford University. Email him at: yxu10@radford.edu.

Historical Review

Many authors have contributed to the literature on point processes. Billingsley (1961) proposed a statistical inference method for Markov processes. Duane (1964) suggested a learning curve approach to reliability monitoring. Cox and Lewis (1966) studied statistical inference problems in point processes and their applications. Cox and Isham (1980) discussed random collection of point processes, and Basawa and Parkasa Rao (1980) studied different stochastic processes with the applications. Dharmadhikari, et al. (1989) estimated the scale parameter of a power law process using power law counts. Bain and Enelhardt (1991) presented a statistical analysis of reliability and compared several life testing models. Kingman (1993) discussed methods of Poisson Sampling. Tsokos (1997) presented the parameter estimation of Power Law Process. Rigdon and Basu (2000) proposed several statistical methods for the reliability of repairable systems using a power law process.

Methodology

The schematic diagram presented in Figure 1 provides a picture of the database used in this study. A randomized data set was generated to reduce random errors by performing simple random sampling procedures. From a total 578,134 cancer patients in the SEER database, 500,000 breast cancer patients' information was randomly selected. Out of these 500,000 breast cancer patients, 496,783 are female and 3,217 are male. The female patients are categorized into three different racial groups: Caucasian, African-American and Asian (which includes others). Within these groups, there are 426,302 Caucasian, 39,681 African-American, 29,015 Asian and 1,785 unspecified patients. Within each patient group there are four types of breast cancer: ductal, medullary, lobular and other (unspecified). For each type of breast cancer, patients are further divided according to the American Joint Committee on Cancer (AJCC) Cancer Staging, such as, stage I, II, III, IV and others. Breast cancer, particularly the ductal form, is a common occurrence among Caucasian females; thus, this study focuses on ductal breast cancer among Caucasian females.

Caucasian Ductal Cancer Patients in Stage I

WD stage I stands for Caucasian ductal cancer patients in AJCC stage I. Similarly, WD stage II, III and IV stand for Caucasian ductal cancer patients in AJCC stages II, III and IV. WD patients in stage I were divided into two groups: (1) patients who are still living, and (2) patients who are deceased (see Figure 2). Deceased patients were grouped into (1) patients who are deceased due to breast cancer and, (2) patients who are deceased due to other reasons. For those patients who are deceased due to breast cancer, different treatment information is available. A NHPP was constructed with respect to WD stage I patients in order to compare the effects of the four different treatments.

Caucasian Ductal Cancer Patients in Stage II

Caucasian ductal patients in stage II were divided into two groups, patients who are still living and patients who are deceased (see Figure 3). Deceased patients were further divided into groups of patients who (1) are deceased due to breast cancer, and (2) patients who are deceased due to other reasons. For those patients who are deceased due to breast cancer, different treatment information is available. A NHPP was constructed with respect to WD stage II patients in order to compare the effects of the four different treatments.

The most common stages to classify breast cancer patients are stages I and II. Thus, these are the stages considered herein using the NHPP to determine the effectiveness of the four different treatments (see Figures 2 and 3).

Non-Homogeneous Poisson Process Analysis

According to Tsokos (1997), the non-homogeneous Poisson process (NHPP) is also known as the Power Law Process (PLP) or the Weibull process (WP), in addition, the NHPP is also considered a counting process. Let $\{N(t), t \geq 0\}$ be a counting process with the following three properties:

1. $N(t) \geq 0$.

2. $N(t)$ is an integer.

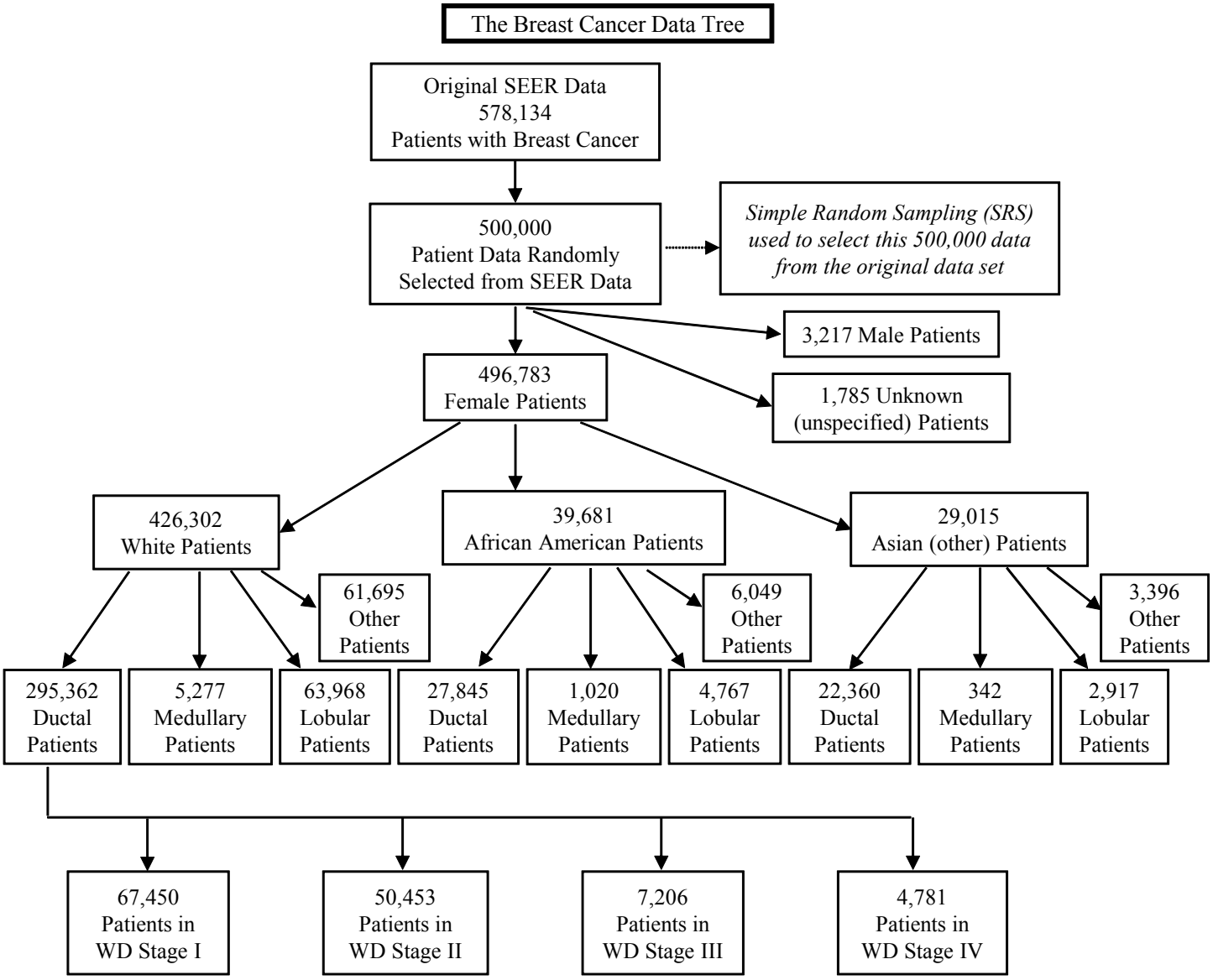
3. If $s \leq t$, then $N(s) \leq N(t)$.

If $s \prec t$, then $N(t) - N(s)$ is the number of events occurring during the interval $(s,t]$.

A Poisson process is a stochastic process in which events occur continuously and independently of one another. The Poisson process is a collection {N(t): t ≥ 0} of random variables, where N(t) is the number of events that have occurred up to time t (starting from time 0). The number of events between times a and b is given as N(b) − N(a) and has a Poisson distribution. Each realization of the process {N(t)} is a non-negative integer-valued step function that is non-decreasing.

For NHPP, the rate parameter may change over time. In this case, the generalized

Figure 1: Breast Cancer Data Tree Diagram
(WD stage I stands for White Ductal cancer patients in AJCC Stage I)



rate function is given as λ(t), thus, the expected number of events between time a and time b is:

$$\lambda_{a,b} = \int_a^b \lambda(t)dt. \quad (1)$$

Therefore, the number of arrivals in the time interval (a, b], given as N(b) − N(a), follows a Poisson distribution with associated parameter λ, a, b as:

$$P[(N(b)) - N(a)) = k] = \frac{e^{-\lambda_{a,b}}(\lambda_{a,b})^k}{k!}, \quad k = 0,1,... \quad (2)$$

A homogeneous Poisson process may be viewed as a special case when λ(t) = λ, a constant rate.

Figure 2: Breast Cancer Data Diagram White Ductal Stage I Patients

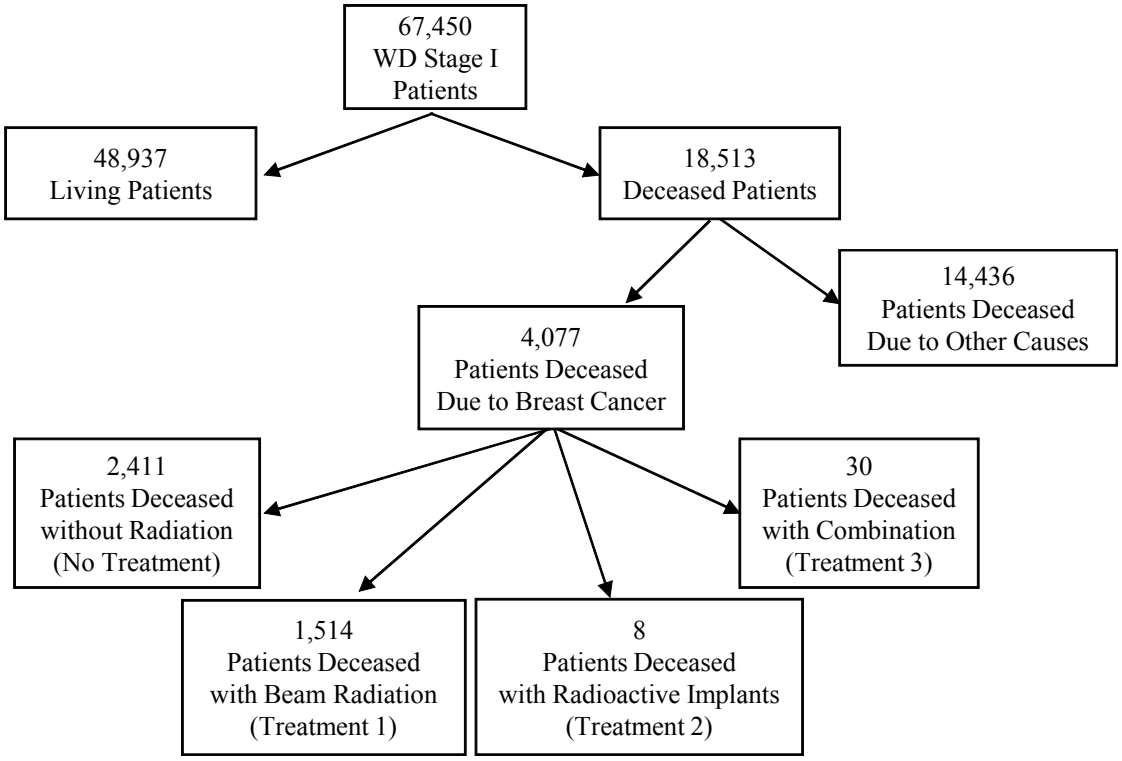
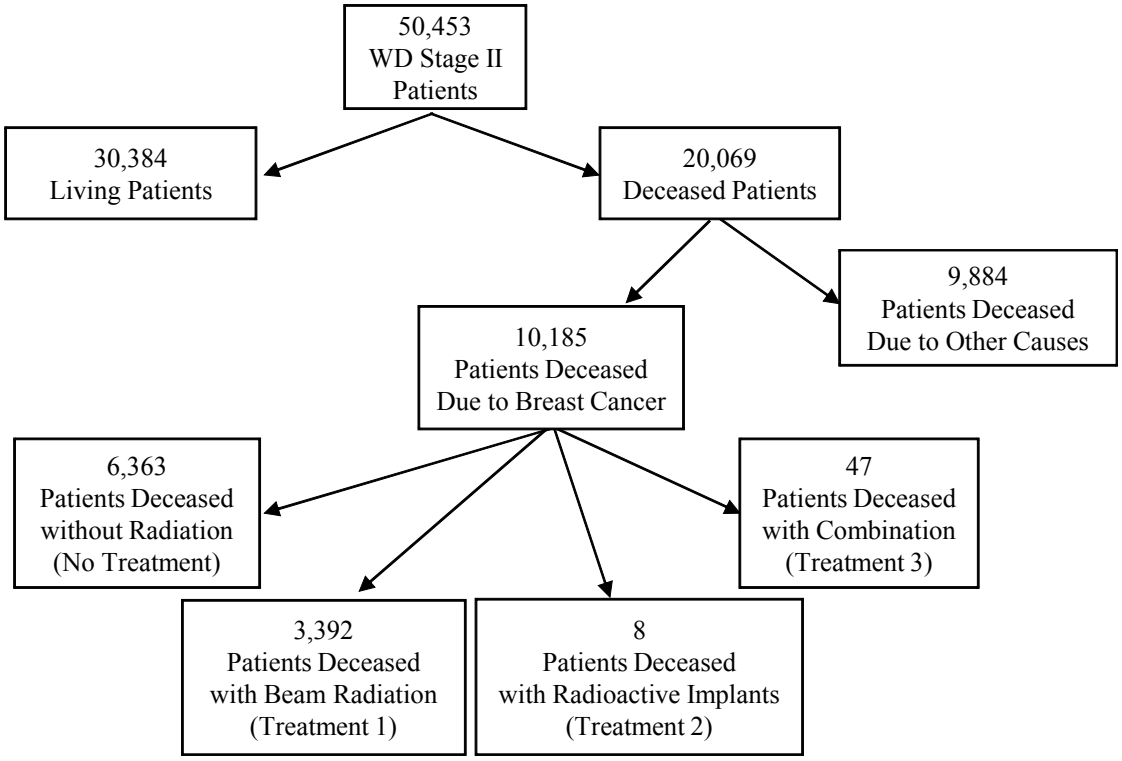


Figure 3: Breast Cancer Data Diagram White Ductal Stage II Patients



The mean value function $\lambda(t)$ of the process is:

$$\begin{aligned}\lambda(t) &= E(N(t)) \\ &= \int_0^t v(s)ds \\ &= \int_0^t \left(\frac{\beta}{\alpha}\right)\left(\frac{s}{\alpha}\right)^{\beta-1} ds \\ &= \left(\frac{t}{\alpha}\right)^{\beta}.\end{aligned} \quad (4)$$

It is known that, if the parameter beta is greater than one in survival analysis, then the failure time increases; this indicates a decrease in survival rate. If beta is less than one in the survival analysis, then the failure time decreases, meaning the survival rate increases. If beta equals one then the failure time is constant and the NHPP will become a homogenous Poisson process (HPP) (Rigdon & Basu, 2010).

The NHPP has the intensity function

$$v(t) = \left(\frac{\beta}{\alpha}\right)\left(\frac{t}{\alpha}\right)^{\beta-1}, \text{ for } \alpha \succ 0, \beta \succ 0, t \succ 0. \quad (3)$$

The unbiased estimator of beta is (Bain & Enelhardt, 1991):

$$\begin{aligned}\hat{\beta}_U &= \frac{n-1-\gamma}{n} \times \beta_{MLE} \\ &= \frac{n-1-\gamma}{\sum_{i=1}^{n} \log\left(\frac{t_n}{t_i}\right)}.\end{aligned} \quad (5)$$

where γ is an indicator function. If $\gamma = 1$ the system will be failure time truncated, meaning the system is restricted by a number of tails and testing will stop when that number of tails is reached. If $\gamma = 0$ then the system will be time truncated, which means the system is restricted by a final failure time and will stop when that time is reached.

The other parameter alpha can be calculated by equation 6, below.

$$\hat{\alpha} = \frac{t_n}{n^{\frac{1}{\hat{\beta}}}} \quad (6)$$

This study belongs to the first case; that is, the time of cases has been fixed. Patients were divided into four groups according to their cancer stage and, within each stage, it is known what kind of treatment the patient received, including if the patient did not receive any radiation treatment at all. Therefore, within each stage patients are divided into four groups with respect to treatment they received, namely, without treatment, treatment 1, 2 or 3. Treatment 1 refers to beam radiation, treatment 2 refers to radioactive implants and treatment 3 is a combination treatment. Few patients in the data source had treatments 2 or3, thus, those are the smallest groups.

Results

After calculating alpha and beta values for the NHPP for each treatment, results were compared and emerging patterns observed. Because the Caucasian race is the major population and ductal patients are the dominate type, this study focused on Caucasian ductal breast cancer patients. The estimation of the parameter is shown in Table 1.

Figure 4 shows the pattern for the key parameter beta. For example, β_{11} is 1.11 which means if a patient does not receive any treatment, the patient's condition will likely become worse because this indicates tumor growth which will lead to the progression of cancer. It may lead the patient to move from stage I to stage II or higher. Examining β_{31} and β_{32}, it is possible to determine whether a patient who receives treatment 3 in stage I will have a better result than a patient who receives the same treatment as a patient in stage II.

It was found that, for cases when beta are less than one, a decreased tumor size is indicated, meaning the treatment for breast cancer works. Results show that patients in early stages (for example, I and II) without treatment will experience increased tumor size and shorter time until death (see Table 1). Beam radiation

Figure 4: Evaluation Chain for NHPP

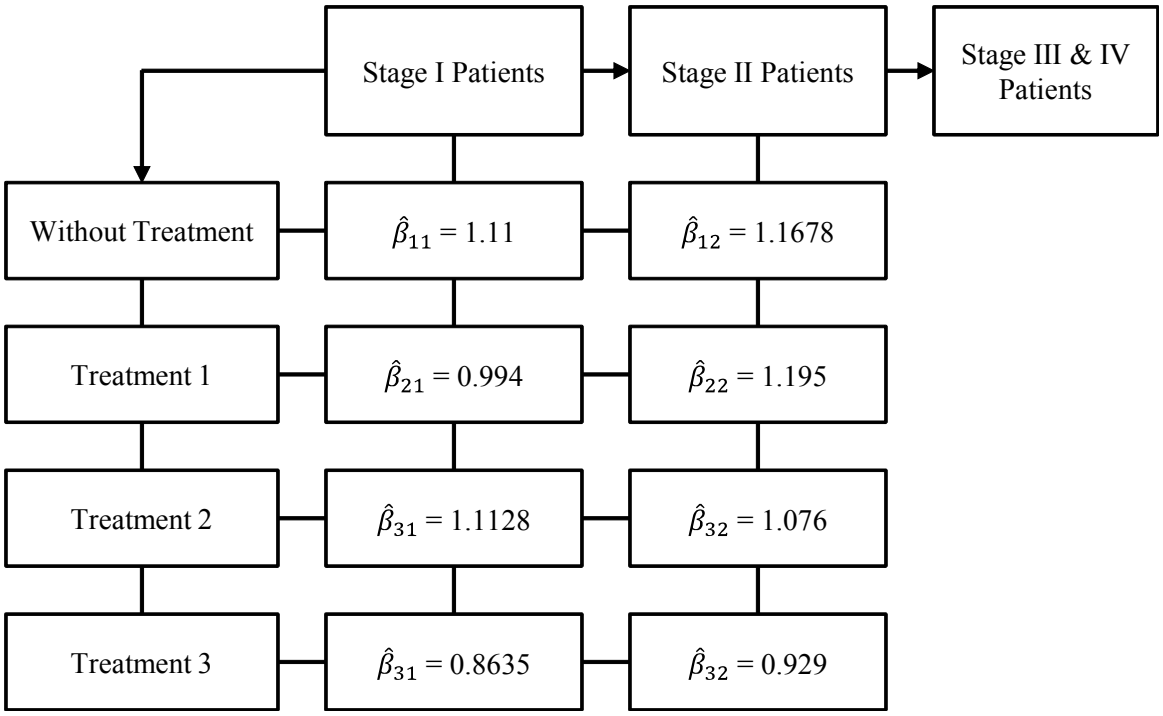


Table 1: Parameter Estimation for NHPP

		Stage I	Stage II
Alpha	Without Treatment	94.5112	113.8267
	With Treatment 1	56.17724	92.982
	With Treatment 2	76.03755	66.60
	With Treatment 3	33.8427	41.35
Beta	Without Treatment	1.110023	1.167756
	With Treatment 1	0.9943948	1.1195
	With Treatment 2	1.112772	1.076
	With Treatment 3	0.8635	0.929

(treatment 1) works for stage I but not for stage II. Radioactive implants (treatment 2) do not work well for either stage I or II. Treatment 3, a combination of the treatments, works well in stages I and II. (There is not enough data to conduct the NHPP for stages III and IV.) Intensity function plots are shown in Figures 5 - 12.

Figure 5 shows that, as the cumulative time of a patient increases, the intensity function also increases: this indicates, as expected, that tumor size is increasing and cancer is progressing. This result verifies the result obtained from parameter estimate β_{11}. Figure 6 shows that, as the cumulative time of a patient increases, the intensity function also decreases; this indicates, as expected, that the cancer will decrease with treatment 1 for stage I ductal Caucasian patients. This result leads to the same result obtained from parameter estimate β_{12}.

Figure 7 shows that, as the cumulative time of a patient increases, the intensity function also increases; this indicates, as expected, that the cancer progresses without treatment. This result verifies the result obtained from parameter estimate β_{13}. Figure 8 shows that, as the cumulative time of a patient increases, the intensity function decreases; this indicates that the cancer will improve with treatment 1 for stage 1 ductal Caucasian patients. This result leads to the same result obtained from the parameter estimate β_{14}.

Following a similar method, Figures 9, 10 and 11 show that, as the cumulative time of a patient increases, the intensity function also increases. This indicates that the cancer progresses without treatment or with treatment 1 or 2 for stage II ductal Caucasian patients. This result leads to the same result obtained from the parameter estimates β_{21}, β_{22} and β_{23}.

Figure 12 shows that, as the cumulative time of a patient increases, the intensity function decreases, this indicates - as expected - that the cancer will improve with treatment 3 for stage II ductal Caucasian patients. This result attests to the estimation obtained from parameter estimate β_{24} (see Table 1).

In summary, results indicate that, for Caucasian ductal breast cancer patients, it would be recommended to provide either a combination or a beam radiation treatment when they are in early stages I and II.

Conclusion

Based on breast cancer patients from the SEER database, adequate data exists to apply the NHPP analysis to Caucasian ductal cancer female patients in two early stages. Based on the results obtained from applying the proposed model, the following conclusions are put forth:

- With no treatment, the intensity function in stage I and stage II increases exponentially, implying that the tumor size of the patients increases at the same rate.
- With treatment 1 (beam radiation) in stage I the intensity function decreases, implying that the tumor size decreases. However, the same treatment in stage II shows the opposite result.
- With treatment 2 (radioactive implants) the intensity function in stage I increases and similar behavior is observed for the same treatment in stage II, this implies that the tumor size of the patients increases at the same rate.
- With treatment 3 (combination treatment) the intensity function in stages I and II decreases exponentially, this implies that the tumor size of the patients decreases at the same rate.

The study reported here is part of a larger, ongoing study. We will continue to obtain data and, eventually to construct a NHPP for each stage and each tumor size available for all treatments and compare the results. With more data and a broader range of patients and cancer stages, it will be possible to make suggestions for the particular treatment that will be best for patients with a particular tumor size. NHPP may also be applied to Bayesian survival analysis to compare and improve results.

Figure 5: Stage I Breast Cancer Intensity Function without Treatment

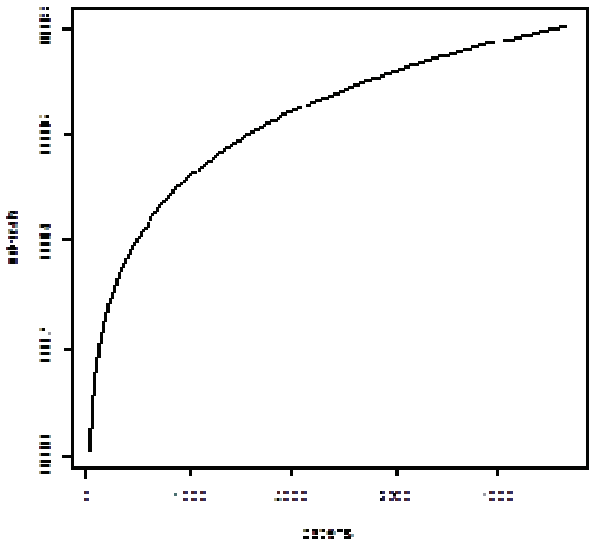


Figure 6: Stage I Breast Cancer Intensity Function with Treatment 1

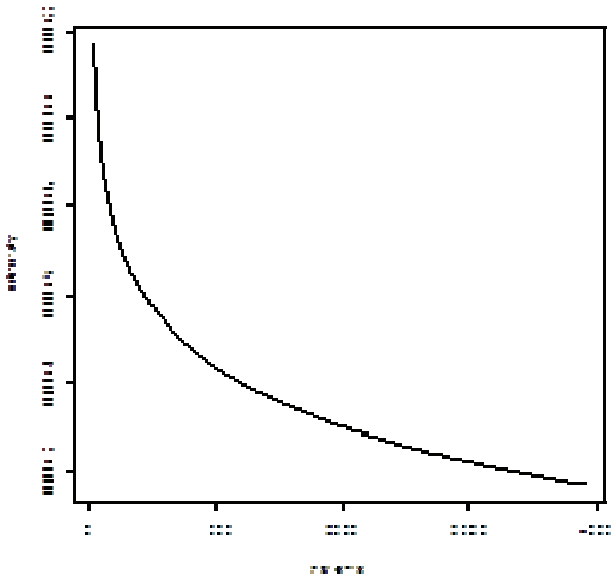


Figure 7: Stage I Breast Cancer Intensity Function with Treatment 2

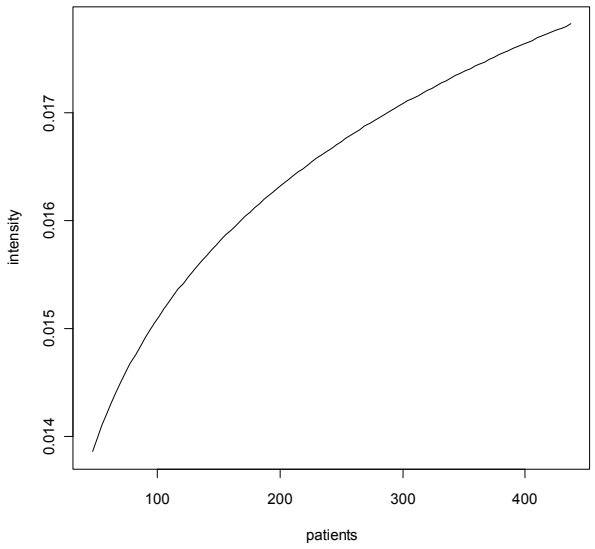


Figure 8: Stage I Breast Cancer Intensity Function with Treatment 3

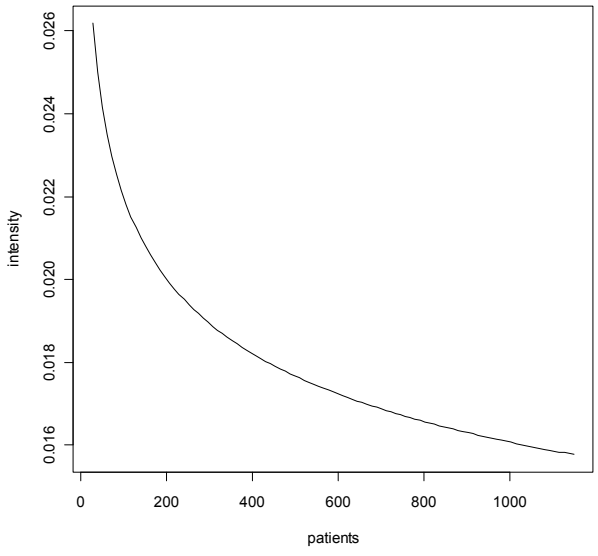


Figure 9: Stage II Breast Cancer Intensity Function without Treatment

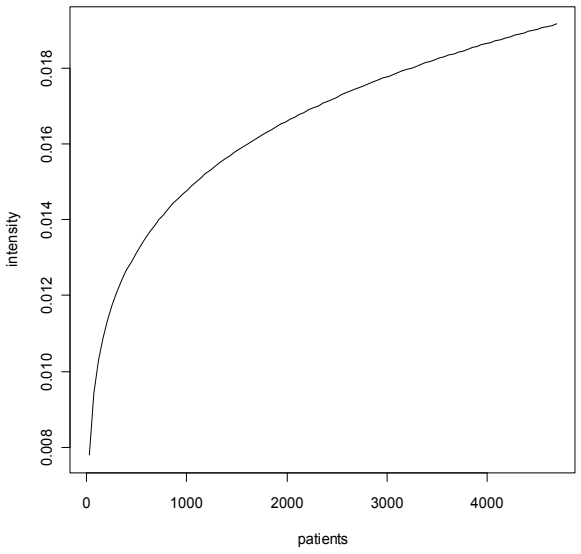


Figure 10: Stage II Breast Cancer Intensity Function with Treatment 1

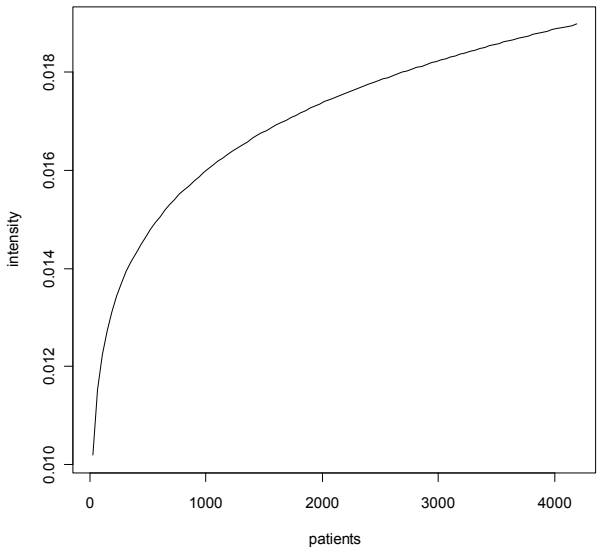


Figure 11: Stage II Breast Cancer Intensity Function with Treatment 2

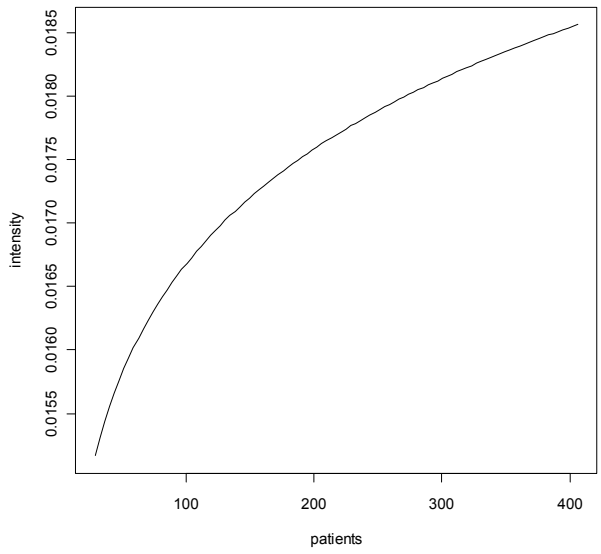
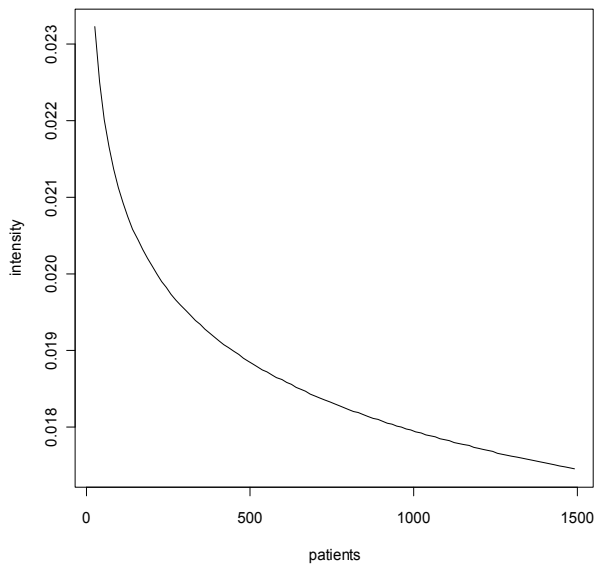


Figure 12: Stage II Breast Cancer Intensity Function with Treatment 3



References

Bain, L. J., & Enelhardt, M. (1991). *Statistical analysis of reliability and life testing models*, *2nd Ed*. New York, NY: Marcel Dekker.

Basawa, I. V., & Rao, B. L. S. P. (1980). *Statistical inference for stochastic processes*. London, England: Academic Press.

Billingsley, P. (1961). *Statistical inference for Markov processes*. Chicago, IL: University of Chicago Press.

Cox, D. R., & Lewis, P. A. W. (1966). The statistical analysis of series of events. *Annals of Mathematical Statistics*, *37*(*6*), 1852-1853.

Cox, D. R., & Isham, V. (1980). *Point processes*. London, England: Chapman & Hall.

Dharmadhikari, A. D., et al. (1989). Estimation of the scale parameter of a power law process using power law counts. *Annals of the Institute of Statistical Mathematics*, *41*(*1*), 139-148.

Duane, J. T. (1964). Learning curve approach to reliability monitoring. *IEEE Transactions on Aerospace*, *2*, 563-566.

Kingman, J. F. C. (1993). *Poisson processes*. Oxford, England: Oxford University Press.

Tsokos, C. P. (1997). Parameter estimation of power law process. In *Nonlinear Problems in Aviation and Aerospace*, 575-86. United States, CRC Press.

Rigdon, S. E., & Basu, A. P. (2000). *Statistical methods for the reliability of repairable systems*. New York, NY: Wiley.

Sariego, J. (2010). Breast cancer in the young patient. *The American Surgeon*, *76*(*12*), 1397-1400.

U. S. National Institutes of Health (2010). Access at: http://seer.cancer.gov.

Journal of Modern Applied Statistical Methods
November 2011, Vol. 10, No. 2, 656-668

1538 – 9472/11/$95.00

Salary Equity Studies: An Analysis of Using the Blinder-Oaxaca Decomposition to Estimate Differences in Faculty Salaries by Gender

Sally A. Lesik Carolyn R. Fallahi
Central Connecticut State University,
New Britain, CT

Parameter estimates for equity studies tested for stability are described. Bootstrap simulation can test whether parameter estimates remain stable given changes in the sample data; fractional polynomials can be used to access functional form specification; and variance inflation factors can be used to test for multicollinearity.

Keywords: Equity studies; Blinder-Oaxaca decomposition, stability, bootstrap simulation, fractional polynomial analysis, variance inflation factors.

Introduction

Significant progress has been made in gender and racial equality over the last several decades since the introduction of the Equal Pay Act of 1963 and the Civil Rights Act of 1964 (Baker, Wendt, & Slonaker, 2002). However, many researchers believe that inequities continue to exist in higher education in the areas of hiring practices, salary, promotion and tenure (Perna, 2005; Hampton, et al, 2000; Sampson & Moore, 2008). Although many national studies continue to address gender and racial equity in academia, it is necessary and prudent to conduct studies within individualized institutions to address all of the variables within these institutions that could affect equity (McLaughlin & McLaughlin, 2003).

Gender and Race Equity

Study after study has concluded that a society where men and women are treated equitably in higher education - or where the gap between white and minority professionals is being bridged - does not currently exist.

Regarding the status of higher education the National Center for Education Statistics (2009) reported that, in the fall of 2007, 55% of those tenured were male as compared to 41% females. Furthermore, four out of five faculty tenured during that same semester were reportedly white (Caucasian). Women in academia also fall significantly below their male counterparts in academic rank, salary and full-time status (Jacob, 2004). Throughout the public sector internationally, the wage differential is significantly lower for women (Fransson & Thörnqvist, 2006; Kjeldal, Rindfleish, & Sheridan, 2005; Lips, 2003); women are also significantly underrepresented within government systems as well as in high-ranking business positions (Connell, 2006).

Although there are a plethora of equity studies involving gender at the local and national level, few examine these issues considering race/ethnicity equity (Barbezat, 2002). This is due in part to the fact that there are not many minority faculty. For example, Barbezat (2002) found that no minority groups constitute more than 5% of faculty involved in teaching and research at the university/college level. Hearn (in Barbezat, 2002) concludes that trends in salary equity for minorities cannot be studied due to the low numbers of minorities in academia. Compensation for minorities in academia, as compared to Caucasian faculty, has not been investigated in relationship to how being a male or female faculty of color affects outcomes.

Sally A. Lesik is a Professor in the Department of Mathematics. Email: lesiks@ccsu.edu. Carolyn R. Fallahi is an Associate Professor in the Department of Psychology. Email: fallahic@ccsu.edu.

The Study of Equity

One of the most famous gender equity studies was the Massachusetts Institute of Technology (MIT, 1999). Gender issues were brought to the forefront due to international media attention. Of interest was the notion that despite diversity incentives at MIT, women faculty were not considered to be equal with their male counterparts (Bailyn, 2003). Bailyn pointed out that, although there have been many equity studies conducted within academia, there had not been any noticeable effect on the policies or practices at such universities. Fewer studies results quantified the experiences of race or ethnicity as compared with Caucasians in academia or the workforce, and when researchers did take race into account, they frequently lacked statistical power as the sample size is often too small to find a reasonably sized effect (Toutkoushian, 1998).

Authors of several studies sought to explain the lack of advancement for women and minorities in academia and other disciplines. For example, Ash et al. (2004) conducted a cross-sectional study of women in academic medicine and found that female physicians earned less in both academia and private practice, but also did not advance to higher ranks as compared to their male counterparts. Some of these differences were explained by other factors, such as the fact that women have significantly less productivity with publishing (Cooperstein, 2008; Friedman, 2004) and that women's careers are more affected by family responsibilities (Friedman, 2004). Probert (2005) found that high rates of separation and divorce and family needs accounted for some of the disparity in academic rank. Peterson et al. (2004) concluded, on the basis of a self-reported questionnaire, that minorities in academic medicine are promoted at a slower rate and failed to attain more senior academic ranks as compared to their white counterparts.

Equity in academia and the workforce continues to be a hotly debated topic. Multiple studies conclude that disparities exist for both women and minorities, particularly in terms of salary and senior positions, but many argued that these differences may in fact be due to unexplained factors (Green & Ferber, 2005; Ferber & Loeb, 2002). Others argued that such salary disparities were due to continued discrimination (Gibelman, 2003). Historically, salary equity studies were divided into two different types, (1) total wage gap studies that examine the differences in the average salary for different groups of employees, and (2) unexplained wage gap studies where employee characteristics are considered in order to try and account for these differences (Toutkoushian, 1998).

Green and Ferber (2005) attempted to introduce many variables that are often not included in equity studies in order to evaluate whether they help to explain the gap in earnings. Many researchers have argued that when comparing salary and other equity data, if there is a difference, it is assumed that the difference implies discrimination. However, such differences may in fact be due to unexplained variables that are not included in the study (O'Neill, 2003). Some of the variables that helped explain the reduction in salary for women have included controlling for factors such as experience, educational history, field of study and scholarly productivity (Toutkoushian, 1998, Creamer, 1998).

McLaughlin and McLaughlin (2003) argued that scholarly productivity has been operationally defined by multiple methods in the history of equity studies. For example, researchers have examined the number of publications, the number of times a researcher's work is cited, internal and external grant dollars received, and the quality of publications as markers to indicate scholarly productivity. These studies argue that, without measures of scholarly productivity, only the magnitude of the salary differences can be estimated, not which employees need a review of their salaries in order to correct the inequities.

Additional variables studied in salary equity studies have included age (differences in pay disparity for younger faculty appears be less as compared to more senior faculty) (Toutkoushian, 1998), and seniority. Although McLaughlin and McLaughlin (2003) argued that rewarding seniority does not make sense and is probably not an appropriate variable to include because most faculty are rewarded for productivity as opposed to how many years they have been a faculty member. Another

controversial variable in the study of salary equity involved part-time status. Women engaged in significantly more hours in part-time work as compared with male faculty (Thornley, 2007; Jacobs, 2003), although many researchers did not include part-time faculty or contingent faculty despite the fact that in academia there is a trend towards hiring these contingent faculty (Curtiss, 2005).

Marital status and children (Jacobs & Winslow, 2004), as well as discipline specialty, have been extensively studied. Umbach (2006) argued that labor market conditions may affect salary; he argues that disciplines with a high concentration of women and heavy teaching loads were valued less in the academy and therefore more inequities existed. Gibelman (2003) expanded on this idea to include differential patterns of salaries associated with fields that are primarily female, e.g. nursing and social work, and concluded that gender is a better predictor of salary than any of the characteristics or variables that are typically studied within an equity analysis.

Further, Becker & Toutkoushian (2003) noted that many studies include factors such as academic experience, seniority, academic attainment and - most controversial of all - academic rank. They argued that salary and rank go hand and hand; if a woman is not promoted despite the necessary qualifications, this leads to salary regression and qualifies as rank discrimination. Despite the importance of rank in salary equity, they reviewed a number of studies that did not include academic rank as a factor in predicting salaries. They also argued that because faculty tend not to be terminated when they are tenured, yet if a faculty member is not promoted, it does not appear to look like discrimination.

Methods for studying equity remain an important topic because estimating wage gap differences based on gender and minority status have important and far-reaching consequences. Recent legislation such as the Lilly Ledbetter Fair Pay Act of 2009 and the Paycheck Fairness Act, brought equity discrimination to the forefront by allowing employees to file lawsuits for current and past equity discrimination in their place of employment (Deere, 2010). Furthermore, company officers fear that when inequities do exist, not only will they be at risk for litigation, but this also affects employee's morale and work performance (Romanoff, Boehm & Benson, 1986).

Given the vast body of research on equity studies, it is clear that many studies relied on statistical methods and techniques to make an inference to a larger population of interest. However, one limitation of most of the previous research was that many studies did not assess whether parameter estimates obtained for a gender or race salary inequity remain stable given small changes in the underlying data. This is an important consideration that often is ignored because methods and techniques are often not easily available to access model stability. Clearly, if small changes in the sample data produce parameter estimates that vary greatly, then any inferences would be suspect. Also, if a statistical model is considered, then the functional form of the model needs to be correct. Various functional forms can often give different and contradictory parameter estimates. Given that claims of discrimination are often based on the findings of such analyses, accessing the stability of any findings is crucial for making a valid inference.

The purpose of this study is three-fold. First, a study on salary equity is described that uses the Blinder-Oaxaca decomposition to partition a wage difference as both a portion that can be explained as well as a portion that is left unexplained. Second, a series of simulation analyses is presented that can be used to assess the stability the parameter estimates that are found using the Blinder-Oaxaca decomposition. Third, fractional polynomial modeling is introduced as a way to determine the appropriate functional form of a regression model and variance inflation factors are calculated to assess model stability.

Methodology

The Blinder-Oaxaca decomposition (Blinder, 1973; Oaxaca, 1973) is a fairly simple extension of multiple regression modeling that is often used to describe wage differences between two different groups. The basic idea behind the Blinder-Oaxaca decomposition is to partition the estimated effect of a binary predictor variable into two portions: one portion that represents the

explained difference between the two groups, and the other portion that describes the unexplained difference between the two groups. For example, a binary predictor variable could be used to describe gender (i.e., male is assigned the value of 0; female is assigned the value of 1). Many studies have used the Blinder-Oaxaca technique to decompose wage differences into explained and unexplained portions, and often the unexplained portion is used to infer discrimination (Neumark, 1988).

Data

A sample of $n = 110$ newly hired tenure-track faculty were considered for this study. The sample represented all newly hired tenure-track faculty members who joined the institution during a four-year period between the years 2004 and 2008. Variables considered for this study are described in detail below.

Predictor Variables

- Year of hire: This is a series of five separate binary variables that represent the beginning of the academic year of hire (YR04, YR05, YR06, YR07, YR08). For the YR04 variable, if a faculty member was hired during the academic year 2004-2005, then they are assigned the value 1. If they were not hired during the 2004-2005 academic year, they are assigned the value 0. Similar assignments are made for the faculty hires for the years 2005-2006, 2006-2007, 2007-2008 and 2008-2009.

- Rank at hire: This is a series of three separate binary variables that represent the rank at hire (ASST, ASSOC, PROF). For the ASSOC variable, if a faculty member was hired as an Associate Professor, they are assigned the value 1. If they were not hired as an Associate Professor, they are assigned the value 0. Similar assignments were made for Assistant (ASST) and Full Professor (PROF).

- Age at hire: This is a continuous predictor variable representing a new faculty member's age in years at the time of hire.

- School of hire: This is a series of five binary variables representing the new hire's school (Arts and Sciences, Education, Business, Engineering and Technology, Other).

- Female: This is a binary variable representing new faculty's self-identified gender (Female = 0 if the new hire identifies as Male, and Female = 1 if the new hire identifies as Female).

- Minority: This is a binary variable representing new faculty's self-identified minority status (Minority = 0 if the new hire identifies as White/Caucasian, and Minority = 1 if the new hire identifies as Non-White/Caucasian).

Means and standard deviations for the continuous predictor variables are presented in Table 1, percentages for the binary control variables are presented in Table 2.

Response Variable

- Ln(Wages): This variable represents the natural logarithm of yearly wages (in dollars). As with many wage studies, the natural logarithm of the yearly wages was used in order to estimate a constant percentage effect (Wooldridge, 2002, 2003).

Table 1: Mean and Standard Deviation for Continuous Variables Yearly Wages and Age at Hire for Newly Hired Faculty ($n = 111$)

Continuous Variable	Mean	Standard Deviation
Yearly Wages	60127.52	11002.19
Age at Hire	41.41	9.42

Table 2: Percentages of Binary Variables for Tenured and Tenure-Track New Faculty Hires

Binary Variable	Percentage
Year of Hire 04	21.62
Year of Hire 05	18.92
Year of Hire 06	18.92
Year of Hire 07	20.72
Year of Hire 08	19.82
Assistant	80.91
Associate	15.45
Full Professor	3.64
Arts & Science	49.55
Business	20.72
Engineering & Technology	6.30
Education	18.02
Other	5.41
Female	45.05
Male	54.95
Minority*	19.44
White/Caucasian	80.56

*Three observations did not self-report

The results from the following generalized ln-wage equation for the model that includes males and females pooled together are presented in Table 3.

$$\ln(wage) = \beta_0 + \beta_1 YEAR + \beta_2 RANK + \beta_3 AGE + \beta_4 SCHOOL + \beta_5 GENDER + \varepsilon \quad (1)$$

Initial Blinder-Oaxaca Results

Version 10 of STATA® was used to conduct the Blinder-Oaxaca decomposition technique to estimate the wage difference between males and females and to partition the wage difference into two components (Jann, 2008). The explained component is determined based on observed characteristics, and the unexplained component is based on unobserved characteristics (Jann, 2008). The results from these analyses are summarized in Table 4.

Notice in Table 4 that the mean of the ln(wages) for the generalized ln-wage equation is estimated to be approximately 11.02 for males and 10.95 for females. This suggests that there is a total wage difference of 0.069 as represented on the logarithmic scale. The exponentiated results from the last column in Table 4 (which express the estimate on the dollar scale) indicate that the (geometric) mean yearly wages for males is estimated to be approximately $61,160.46 as compared to approximately $57,057.39 for females. This indicates that there is an estimated total wage difference of approximately 7.19% between male and female new faculty hires. The decomposition portion of Table 4 suggests that if females were hired with the same characteristics as males (for example if females had the same year at hire, age at hire, rank at hire, and school of hire), then the total wage gap observed between males and females would be decreased by approximately 4.78%. This leaves a wage gap of approximately 2.30% that cannot be accounted for by the given observed characteristics between male and female new faculty hires.

Model Instability

Many different scenarios can generate different and often contradictory parameter estimates. Such differences can often be attributed to the model not being stable given changes in the underlying data, the functional form of the model being not being specified correctly, or some of the predictor variables being highly correlated with each other. Model instability can occur if small changes in the data generate vastly different parameter estimates (Royston & Sauerbrei, 2009). Also, if the functional form of the model is not specified correctly, then differences from different model specifications can also generate vastly different parameter estimates (Griffin, Montgomery & Rister, 1987; Royston & Sauerbrei, 2008, 2009). Furthermore, including predictor variables that are highly correlated with each other can also cause the estimated parameters to be unstable (Graham, 2003; Lesik, 2010).

Assessing Model Instability Due to Changes in the Data: Bootstrapping

One of the more common techniques for assessing model instability due to small changes in the underlying data is to use bootstrap resampling (Sauerbrei & Schumacher, 1992). Bootstrap resampling entails drawing repeated samples (with replacement) from the sample of interest, estimating the parameter of interest, empirically estimating the distribution for the parameter of interest, and finally determining if the parameter of interest is significant in the model.

A bootstrap simulation program was written for version 10 of STATA® (see Appendix). This program draws a bootstrap sample from the initial 110 new faculty hires and then conducts the Blinder-Oaxaca decomposition. Line 5 of the bootstrap program [generate nsamp = cond(sex, 49, 61)] ensures that the bootstrap sample was drawn to represent the underlying percentages of males and females at the institution (of the 110 new faculty hires, 49 were females and 61 were males). The mean exponentiated percent unexplained difference for the simulation analysis run with 10,000 replicates was 2.2260% with a standard deviation of 1.3173%. The distribution of the mean exponentiated unexplained difference is shown in Figure 1. It was also found that for all of the bootstrap resamples, 58.86% had significant unexplained differences ($p < 0.10$).

Also calculated from the bootstrap simulation analysis were descriptive statistics of the unexplained differences being negative (this would indicate that males made less than females). Of the 10,000 simulation analyses, only 444 (only 4.44%) indicated that the unexplained percent difference was negative. Of these 444 bootstrap samples, only 13 were significant at the 10% level, thus suggesting that only 0.13% of the 10,000 bootstrap simulations showed that males made less than females (significant at the 10% level). Given these results of the bootstrap simulation, it appears that the estimated unexplained percent difference stable, even given small changes in the underlying data set.

Assessing Model Stability from Functional Form Misspecification: Fractional Polynomial Modeling

Because the Blinder-Oaxaca decomposition used in this study is a simple extension of ordinary least squares regression, it relies on some basic model assumptions. One such assumption is that the functional form of the model is specified correctly with respect to the relationship between the continuous predictor variables and the response variable. Different functional forms can often yield different and even contradictory parameter estimates.

The generalized ln(wage) model given in equation (1) is specified such that the continuous predictor variable which corresponds to the age at hire is linear. Fractional polynomial modeling was used to see if changes in the functional form of the generalized ln(wage) model would present different parameter estimates. Fractional polynomial modeling can be used to determine if a linear model is appropriate for virtually any type of regression modeling, even logistic regression (i.e. Hosmer & Lemeshow, 2000).

The basic idea underlying fractional polynomial modeling is to include powers of continuous predictor variables to determine if this improves the fit of the model (Royston & Sauerbrei, 2008, 2009). Royston and Altman (1994) suggest that a restricted set of fractional polynomial powers is sufficient in transforming continuous predictor variables for better model fit.

Given a single continuous predictor variable (as is the case with this study), the general form of a population linear regression model is:

$$y = \beta_0 + \beta_1 x_1 + \varepsilon$$

Powers of the continuous variable, $f_k(x_1)$ can be included into the regression model as follows:

$$y = \beta_0 + \sum_{i=1}^{k} \beta_i \cdot f_k(x_1) + \varepsilon$$

Table 3: Parameter Estimates, Standard Errors and 95% Confidence Intervals for the Predictor Variables of the Generalized ln(wage) Equation (1) for all New Full-Time Tenure-Track Faculty Who were Hired During the Academic Years 2004-2008 (n = 110).

Variable	Parameter Estimate [Standard Error]	95% Confidence Interval
Year 04	-0.1443*** [0.0180]	-0.1801, -0.1085
Year 05	-0.0827*** [0.0187]	-0.1198, -0.0457
Year 06	-0.0603** [0.0187]	-0.0974, -0.0232
Year 07	-0.0335~ [0.0183]	-0.0699, 0.0030
Assistant	-0.3403*** [0.0365]	-0.4127, -0.2679
Associate	-0.0904* [0.0348]	-0.1594, -0.0214
Age at Hire	0.0012 [0.0008]	-0.0003, 0.0027
Arts & Sciences	-0.0409 [0.0255]	-0.0915, 0.0098
Business	0.0727* [0.0299]	0.0134, 0.1320
Engineering & Technology	0.0725* [0.0338]	0.0053, 0.1397
Education	0.0030 [0.0283]	-0.0531, 0.0592
Gender	-0.0227~ [0.0121]	-0.0468, 0.0013
Constant	11.3074*** [0.0599]	11.1884, 11.4263
R-squared	0.8900	
Adjusted R-Squared	0.8764	
Sample Size	110	

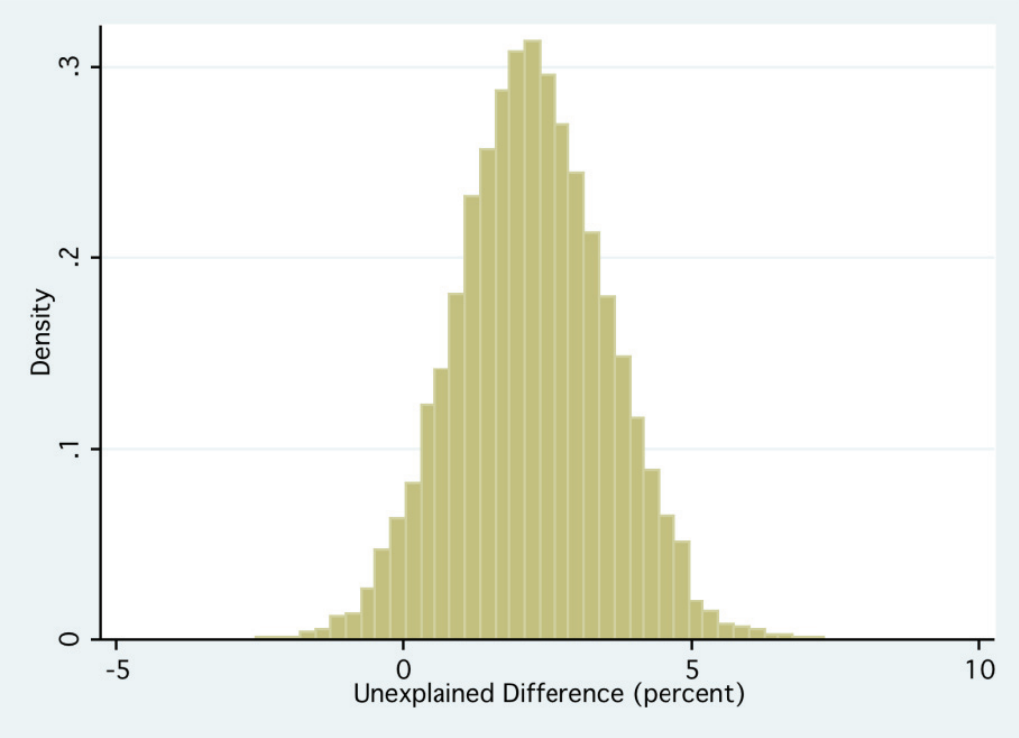
~ $p < 0.10$; * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

Table 4: Ln-Scale Parameter Estimates and Exponentiated Estimates (in Dollars), and Standard Errors for the Blinder-Oaxaca Decomposition for Initial Faculty Salaries Based on Gender

Differential Category	Ln-Scale Parameter Estimate [Standard Error]	Exponentiated Parameter Estimate [Standard Error]
Males	11.0213*** [0.0220]	61160.46*** [1348.526]
Females	10.9518*** [0.0224]	57057.39*** [1275.899]
Total Difference	0.0694* [0.0314]	1.0719* [0.0337]
Decomposition		
Explained Difference	0.0467 [0.0298]	1.0478 [0.0312]
Unexplained Difference	0.0227* [0.0116]	1.0230* [0.0118]

~ $p < 0.10$; * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

Figure 1: Distribution of the Unexplained Wage Difference for the 10,000 Bootstrap Samples Using the Blinder-Oaxaca Decomposition



$\{-2, -1, -0.5, 0, 0.5, 1, 2, 3\}$. The powers of the continuous variable x_1 can then be included in the model:

$$f_k(x_1) = \begin{cases} x^p & \text{if } p_k \neq p_{k-1} \\ f_{k-1}(x_1) \cdot \ln(x_1) & \text{if } p_k = p_{k-1} \end{cases},$$

where $k = 1, 2, 3, \ldots$. For example if $k = 2$, with powers 0.5 and 0.5, then $f_1(x_1) = x_1^{0.5}$ and $f_2(x_1) = x_1^{0.5} \cdot \ln(x_1)$. Therefore, $y = \beta_0 + \beta_1 x_1^{0.5} + \beta_2 x_1^{0.5} \cdot \ln(x_1) + \varepsilon$. For another example if it is supposed that $k = 4$ with powers -2, 2, 3 and 3, then $f_1(x_1) = x_1^{-2}$, $f_2(x_1) = x_1^2$, $f_3(x_1) = x^3$, and $f_4(x_1) = f_3(x_1)\ln(x_1) = x_1^3 \cdot \ln(x_1)$. Thus,

$$y = \beta_0 + \beta_1 x_1^{-2} + \beta_2 x_1^2 + \beta_3 x_1^3 + \beta_4 x_1^3 \cdot \ln(x_1) + \varepsilon.$$

Version 10 of STATA® was used to find the best fractional model that has a maximum of $k = 4$ (STATA Corporation, 2005). The STATA routine fracpoly finds the best fractional polynomial models for each of the values. For example, the best model for $k = 2$ has the powers -2 and -2. The table also provides deviance statistics and *p*-values for comparing the improvement in fit for each successive pairs of models (Royston & Altman, 1994). The deviance statistic is calculated as follows:

$$D = n\left[1 - \overline{w} + \ln\left(\frac{2\pi}{SSR}\right)\right],$$

where n is the sample size, $\overline{w}$ is the mean of the normalized weights, and *SSR* is the residual sum of squares. Although somewhat conservative, these *p*-values indicate whether the fit of the model improved by including the predictor variable with the additional powers (see Table 5).

Based on the *p*-values presented in Table 5, no improvement is observed in model fit for including the predictor variable that represents the age at hire, as well as any fractional powers of the variable. Thus, the age at hire is not significant in predicting starting salaries for new faculty hires.

Highly Correlated Predictor Variables: Variance Inflation Factors

One common technique to determine if the predictor variables are highly correlated with each other is to calculate the variance inflation factor for each predictor variable in the generalized ln(wage) model. Variance inflation factors (*VIF*) for each predictor variable can be found by assigning each predictor variable as the response variable and running a regression analysis with all the other predictor variables. The *VIF* for each variable can then be calculated as follows:

$$VIF_j = \frac{1}{1 - R_j^2},$$

where $j = 1, 2, \ldots p - 1$, where p is the total number of beta parameters being estimated in the model (including the constant parameter), and R_j^2 is the coefficient of determination for the model in which variable x_i is represented as the response and all the other variables are included as predictor variables (Lesik, 2010). None of the variance inflation factors were above 10, thus suggesting that the individual predictor variables do not appear to be highly correlated with each other (the minimum *VIF* was 1.143 and the maximum was 6.453).

Conclusion

Concern over methods related to estimating the wage gaps in equity studies prompted our interest in determining the stability of wage gap estimates that are found in equity studies. As employers and employees are increasingly sensitive to gender and race equity for salary, an increasing number of studies are being done in both the public and private sector internationally (Fransson & Thornqvist, 2006). Authors of many equity studies, as well as studies on related topics, note concern over the stability of the estimate of the wage gap between males and females; yet to date, these concerns have not been addressed (Graham, 2003; Griffin, et al., 1987; Royston & Sauerbrei, 2008).

Table 5: Results of Fractional Polynomial Model Comparisons for the Generalized ln(wage) Equation

Age at Hire	Degrees of Freedom	Deviance	Residual Standard Deviation	Difference in Deviance	*p*-value	Powers
Not in Model	0	-322.579	0.059468	2.300	0.987	
Linear	1	-324.516	0.059253	0.363	1.000	1
k = 1	2	-324.663	0.059213	0.216	1.000	-2
k = 2	4	-324.675	0.059521	0.203	0.996	-2 -2
k = 3	6	-324.871	0.059783	0.008	0.997	0 0 0
k = 4	8	-324.879	0.060102	---	---	-2 -2 -2 -2

This study shows that the estimate of the wage gap between males and females remained stable given small changes in the underlying data as well as for various fractional powers of the continuous predictor variable that represents the age at hire. Also, none of the predictor variables were highly correlated with each other, thus there was no concern that highly correlated predictor variables could be influencing the estimated parameters. Given more powerful statistical software for bootstrap simulations and fractional polynomial analysis, as well as calculating variance inflation factors, these tools can be used to ensure that the estimates provided herein are not only accurate, but are stable given small changes in the data as well as the functional form of the regression model at hand.

Although this study was conducted in order to address some of the concerns that can generate unstable parameter estimates, there are still some limitations to note. One limitation of the Blinder-Oaxaca decomposition is that it can only decompose a regression model based on only two groups. Even though two groups are adequate to quantify gender, the decomposition cannot be used to compare more than two groups, such as would be the case with various classifications of race.

Limitations to fractional polynomial modeling include loss of power and sensitivity to outliers (Royston & Sauerbrei, 2008). Furthermore, because fractional polynomial modeling can identify the powers of a continuous predictor variable that suggest the best model fit, including continuous predictor variables with such powers can greatly increase the complexity of a regression model, thus making interpretation more difficult.

Acknowledgements

This research was supported in part by Central Connecticut State University. Portions of this article may represent material from a study about race/gender equity conducted by the authors and commissioned by Central Connecticut State University. The full study is online at http://www.ccsu.edu/page.cfm?p=4595. Lisa L. Leishman, a graduate student in Psychology at Central Connecticut State University and Law student at Western New England College School of Law, Springfield, MA provided help coding the data and with the maintenance of the database.

References

Ash, A. S., Carr, P. L., Goldstein, R., & Friedman, R. (2004). Compensation and advancement of women in academic medicine: Is there equity? *American College of Physicians*, *141*, 205-212. doi: 10.1016.j.ajo.2004.09.012

Bailyn, L. (2003). Academic careers and gender equality: Lessons learned from MIT. *Gender, Work, and Organization*, *10*(*2*), 137-153.

Baker, B., Wendt, A., Slonaker, W. (2002). An Analysis of gender equity in the federal labor relations career field. *Public Personal Management*, *31*(*4*), 559-567.

Barbezat, D. A. (2002). History of pay equity studies. *New Directions for Institutional Research*, *115,* 9-39.

Barbezat, D. A., & Hughes, J. W. (2005). Salary structure effects and the gender pay gap in academia. *Research in Higher Education*, *46*(*6*), 621-640. DOI: 10.1007/s11162-004-4137-1.

Becker, W. E., & Toutkoushian, R K. (2003). Measuring gender bias in the salaries of tenured faculty members. *New Directions for Institutional Research*, *117,* 5-20.

Blinder, A. S. (1973). Reduced form and structural estimates. *Journal of Human Resources*, *8*, 436-455.

Cooperstein, D. (2008). The long road to pay equity for women at Adelphi. *Academe*, *94*(*1*), 34-36.

Creamer, E. G. (1998). *Assessing faculty publication productivity: Issues of equity*. Retrieved from ERIC, ED420242.

Connell, R. (2006). Glass ceiling or gendered institutions? Mapping the gender regimes of public sector worksites. *Public Administration Review*, 837-849. DOI: 10.1111/j.1540.6210.2006.00652.x

Curtis, J. W. (2005). Inequities persist for women and non-tenure-track faculty: The annual report on the economic status of the profession. *Academe*, *91*(*2*), 20-44.

Deere, D. (2010). *Conduct a pay equity study to mitigate litigation risks.* Retrieved from: http://www.shrm.org/hrdisciplines/compensation/Articles/Pages/PayEquityStudy.aspx.

Ferber, M. A., & Loeb, J. W. (2002). Issues in conducting an institutional salary-equity study. *New Directions for Institutional Research*, *115,* 41-69.

Ferree, M. M., & McQuillan, J. (1998). Gender-based pay gaps: Methodological and policy issues in university salary studies. *Gender and Society*, *12*(*1*), 7-39.

Fransson, S., & Thörnqvist, C. (2006). Some notes on workplace equality renewal in Swedish labour market. *Gender, Work and Organization*, *13*(*6*), 606-620. DOI: 10.1111/j.1468-0432.2006.00324.x.

Gibelman, M. (2003). So how far have we come? Pestilent and persistent gender gap in pay. *National Association of Social Workers*. *48*(*1*), 22-32.

Graham, M. H. (2003). Confronting multicollinearity in ecological multiple regression. *Ecology*, *84*(*11*), 2809-2815.

Green, C., & Ferber, M. (2005). Do detailed work histories help to explain gender and race ethnic wage differentials? *Review of Social Economy*, *13*(1), 55-85. DOI: 10.1080/00346760500047982

Haignere, L. (2002). *Paychecks. A guide to conducting salary-equity studies for higher education faculty*. Washington, DC: American Association of University Professors.

Griffin, R. C., Montgomery, J. M., & Rister, M. E. (1987). Selecting functional form in production function analysis. *Western Journal of Agricultural Economics*, *12*(*2*), 216-227.

Hampton, M., Oyster, C., Pena, L., & Rodgers, P. (2000). Gender inequality in faculty pay. *Compensation & Benefits Review*, 54-59.

Hosmer, D., & Lemeshow, S. (2000). *Applied Logistic Regression* (*2nd Ed.*). New York, NY: Wiley.

Kjeldal, S., Rindfleish, J., & Sheridan, A. (2005). Deal-making and rule breaking: Behind the facade of equality in academia. *Gender and Education*, *17*(*4*), 431-447. DOI: 10.1080/09540250500145/30.

Jacobs, J. A. (2004). The faculty time divide. *Sociological Forum*, *19*(*1*), 3-27.

Jacobs, J. A., & Winslow, S. E. (2004). The academic life course, time pressures, and gender inequality. *Community, Work & Family*, *7*(*2*), 143-161.

Jann, B. (2008). The Blinder-Oaxaca decomposition for linear regression models. *Stata Journal*, *8*(*4*), 453-479.

Lesik, S. (2010). *Applied statistical inference with MINITAB*. Boca Raton, FL: CRC Press.

Lips, H. M. (2003). The gender pay gap: Concrete indicator of women's progress toward equality. *Analyses of Social Issues and Public Policy*, *3*(*1*), 87-109.

Massachusetts Institute of Technology. (1999). *A study on the status of women faculty in science at MIT*. Cambridge, MA: Massachusetts Institute of Technology.

National Center for Education Statistics. (2009). *Digest of Education Statistics (NCES 2009-020)*. Washington, DC: US Department of Education, Office of Educational Research and Improvement.

Neumark, D. (1988). Employers' discriminatory behavior and the estimation of wage discrimination. *Journal of Human Resources*, *23*(*3*), 279-295.

O'Neill, J. (2003). The gender gap in wages, circa 2000. *American Economic Review*, *93*(*2*), 309-314.

Oaxaca, R. L. (1973). Male-female wage differentials in urban labor markets. *International Economic Review*, *14*, 693-709.

Oaxaca, R. L., & Ransom, M. R. (1994). On discrimination and the decomposition of wage differentials. *Journal of Econometrics*, *61*(*1*), 5-21.

Oaxaca, R. L., & Ransom, M. R. (2002). Regression methods for correcting salary inequities between groups of academic employees. *New Directions for Institutional Research*, *115*, 91-103.

Peterson, N. B., Friedman, R. H., Ash, A. S., Franco, S., &Carr, P. L., (2004). Faculty self-reported experience with racial and ethnic discrimination in academic medicine. *Journal of General Internal Medicine*, *19*, 259-265. DOI: 10.1111/j.1525-1497.2004.20409.x.

Probert, B. (2005). "I just couldn't fit in": Gender and unequal outcomes in academic careers. *Gender, Work and Organization*, *12*(*1*), 50-72. DOI: 10.1111/j.1468-0432.2005.00262.x.

Romanoff, K., Boehm, K., & Benson, E. (1986). Pay equity: Internal and external considerations. *Compensation and Benefit Review*, *18*, 17-25.

Royston, P., & Altman, D. G. (1994). Regression using fractional polynomials of continuous covariates: Parsimonious parametric modelling (with discussion). *Applied Statistics*, *43*, 429-467.

Royston, P., & Sauerbrei, W. (2008). *Multivariable model building: A pragmatic approach to regression analysis based on fractional polynomals for modelling continuous Variables*. Chichester, UK: Wiley.

Royston, P., & Sauerbrei, W. (2009). Bootstrap assessment of the stability of multivariable models. *The STATA Journal*, *9*(*4*), 547-570.

Sampson, S. D., & Moore, L. L. (2008). Is there a glass ceiling for women in development? *Nonprofit Management & Leadership*, *18*(*3*), 321-339.

Sauerbrei, W., & Schumacher, M. (1992). A bootstrap resampling procedure for model building: Application to the Cox regression model. *Statistics in Medicine*, *11*, 2093-2109.

STATA Corporation. (2005). *STATA base reference manual* (*Vol. 1*). College Station, TX: STATA Press.

Thornley, C. (2007). Working part-time for the state: Gender, class and the public sector pay gap. *Gender, Work and Organization*, *14*(*5*), 454-474. DOI: 10.1111/j.1468-0432.2007.00360.x.

Toutkoushian, R. K., & Hoffman, E. P. (2002). Alternatives for measuring the unexplained wage gap. *New Directions for Institutional Research*, *115*, 71-89.

Umbach, P. D. (2006*). Gender equality in the academic labor market: An analysis of academic disciplines*. Paper presented at the 46th Annual Association for Institutional Research Forum, Chicago, IL.

Wooldridge, J. (2002). *Econometric analysis of cross section and panel data*. Cambridge, MA: The MIT Press.

Wooldridge, J. (2003). *Introductory econometrics: A modern approach* (*2nd Ed.*). Mason, Ohio: South-Western.

Appendix:

STATA program for bootstrap resampling.

```
program BlinderSim, rclass
version 10.1
        drop _all
        use "BlinderOaxaca.dta"
        generate nsamp = cond(sex, 49, 61)
        bsample nsamp, strata(sex)
        oaxaca lnwage yr04 yr05 yr06 yr07 asst
assoc ageathire as business engrtech educ,
by(sex) pooled
        matrix list e(b)
        matrix list e(V)
        matrix define C = e(b)
        matrix define S = e(V)
        local undiff = el(C,1,5)
        local seundiff = sqrt(el(S,5,5))
        local zstat = `undiff'/`seundiff'
        local pvalue = 2*normal(-abs(`zstat'))
                if `pvalue' <= 0.10 {
                        local inmodel = 1
                        }
                        else {
                        local inmodel = 0
                        }
        local expundiff = 100*(exp(`undiff')-1)
        local checkval = 0
                if `expundiff' < 0 {
                        local checkval = 1
                        }
                        else {
                        local checkval = 0
                        }
        return scalar undiff = `undiff'
        return scalar seundiff = `seundiff'
        return scalar zstat = `zstat'
        return scalar pvalue = `pvalue'
        return scalar inmodel = `inmodel'
        return scalar expundiff = `expundiff'
        return scalar checkval = `checkval'
end
```

Journal of Modern Applied Statistical Methods
November 2011, Vol. 10, No. 2, 669-675

1538 – 9472/11/$95.00

A Sequential Monte Carlo Approach for Online Stock Market Prediction Using Hidden Markov Models

Ahani E. Bridget O. Abass
University of Lagos,
Nigeria, Africa

A sequential Monte Carlo (SMC) algorithm prediction approach is developed based on joint probability distribution in hidden Markov Models (HMM). SMC methods, a general class of Monte Carlo methods, are typically used for sampling from sequences of distributions and simple examples of these algorithms are found extensively throughout the tracking and signal processing literature. Recent developments indicate that these techniques have much more general applicability and can be applied very effectively to statistical inference problems. Due to the problem involved in estimating the parameter of HMM, the HMM is represented in a state space model and the sequential Monte Carlo (SMC) method is used. Predictions are made using the SMC method in HMM and the corresponding on-line algorithm is developed. Daily stock price data from the banking sector of the Nigerian Stock Exchange (NSE) (price index between the years 1 January 2005 to 31 December 2008) are analyzed; experimental results reveal that the method proposed is effective.

Key words: Sequential Monte Carlo, hidden Markov model, state-space model, stock market.

Introduction

State space, or hidden Markov models (HMM), are convenient means to statistically model a process that varies over time. The state space model (Doucet& Johansen, 2008) of a hidden Markov model is represented by the following two equations:

the state equation,

$$X_t \mid (X_{t-1} = x_{t-1}) \sim f(x_t \mid x_{t-1}) \qquad (1)$$

and the observation equation,

$$Y_t \mid (X_t = x_t) \sim g(y_t \mid x_t). \qquad (2)$$

Ahani Bridget is a Lecturer in the Department of Mathematics. Email her at: briggetk2002ng@yahoo.com. O. Abass is a Professor in the Department of Computer Science. Email him at: olabass@unilag.edu.ng.

The state variables x_t and observations y_t may be continuous-valued, discrete-valued or a combination of the two, $f(x_t \mid x_{t-1})$, which indicates the probability density associated with moving from x_{t-1} to x_t, and $g(y_t \mid x_t)$ are the state (transition) and observation densities. Practically, the *x*'s are the unseen true signals in signal processing (Liu & Chen 1995), the actual words in speech recognition (Rabiner 1989), the target features in a multitarget tracking problem (Avitzour 1995; Gordon, et al 1993; Gordon, et al 1995), the image characteristics in computer vision (Isard& Blake 1996), the gene indicator in a DNA sequence analysis (Churchill 1989), or the underlying volatility in an economical time series (Pitt &Shephard 1997). Hidden Markov Models represent the applications of dynamic state space model in DNA and protein sequence analysis (Krogh, et al 1994; Liu, et al 1997).

Using the functions provided by C++ to expand an on-line algorithm for predicting a hidden Markov model, this articleutilizes Johansen (2009) SMCTC: Sequential Monte Carlo in C++. Further supports were derived from results on predicted and actual data of

monthly national air passengers in America (Zhang, et al., 2007). Cheng, et al. (2003) applied SMC methodology to the problems of optimal filtering and smoothing in hidden Markov models and SMC have also stirred great interest in the engineering and statistical literature (see Doucet, et al., 2000, for a summary). SMC methods have been applied for resolving a marginal Maximum Likelihood problem (Johansen, 2008) and Gordon, et al. (1993) applied SMC to optimal filtering. Herein the SMC method is developed for prediction of state by estimating the probability $p(x_t|y_{1-t-1})$.

Hidden Markov Models (HMM)

Initially introduced and studied as far back as 1957 and into the early 1970's, HMM statistical methods have enjoyed more recent popularity.An HMM is a bivariate discrete-time process $\{X_k, Y_k\}_{k\geq 0}$ where $\{X_k\}_{k\geq 0}$ is a homogeneous Markov chain that is not directly observed, it can only be observed through $\{Y_k\}_{k\geq 0}$ that produces the observation. $\{Y_k\}_{k\geq 0}$, which is a sequence of independent random variables such that the conditional distribution of Y_k only depends on X_k. The underlying Markov chain $\{X_k\}_{k\geq 0}$ is called the state. In general, the random variables X_k and Y_k can be of any dimension and of any domain, such as discrete, real or complex. K elements of X_k and Y_k for $k = 1, 2, \cdots, K$ are collected to construct the vectors X_k and Y_k, respectively. Due to the Markov assumption, the probability of the current true state given the immediately previous one is conditionally independent of the other earlier states:

$$p(x_k|x_{k-1}, x_{k-2}, \cdots, x_0) = p(x_k|x_{k-1}).$$

Similarly, the measurement at thek^{th} time step is dependent only upon the current state, thus it is conditionally independent of all other states given the current state:

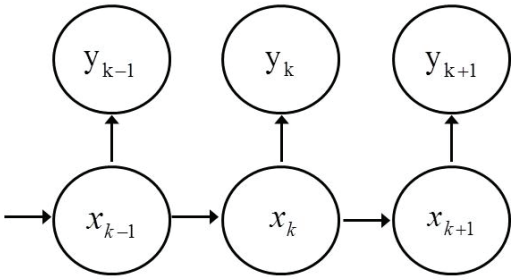
$$p(y_k|x_k, x_{k-1}, \cdots, x_0) = p(y_k|x_k).$$

Using these assumptions the probability distribution over all states of the HMM can be written simply as:

$$p(x_0, \cdots, x_k, y_1, \cdots, y_k) = p(x_1)p(y_1|x_1)\prod_{k=2}^{K} p(x_k|x_{k-1})p(y_k|x_k)$$

which is reflected graphically as:

$$y_k \sim g(y_k \mid x_k)$$



$$x_k \sim f(x_k \mid x_{k-1})$$

Given $p(x_{k-1} \mid y_{k-1})$, $p(x_k \mid y_k)$ can be found using the following prediction and update steps:

Prediction

$$p(X_k \mid Y_{1:k-1}) = \int p(X_k \mid X_{k-1})p(X_{k-1} \mid Y_{1:k-1})dx_{-1}$$

Update

$$p(X_k \mid Y_{1:k}) = \frac{p(Y_k \mid X_k)p(X_k \mid Y_{1:k-1})}{\int p(Y_k \mid X_k)p(X_k \mid Y_{1:k-1})dx_k}$$

In this case numerical integration is used, which becomes computationally complex when the number of states of x_k are large: one particular Monte Carlo based approach to solve this for the HMM is the Sequential Monte Carlo Method (SMC).

Sequential Monte Carlo Methods (SMC)

Since their pioneering contribution in 1993 (Gordon, et al., 1993), SMC have become a well-known class of numerical methods for the

solution of optimal estimation problems in non-linear non-Gaussian scenarios.The main idea of the SMC method is to represent the posterior density function $p(x_{0:k-1} \mid y_{0:k-1})$ at time $k-1$ by samples and associated weights, $\{x_{0:k-1}^{(i)}, w_{0:k-1}^{(i)} i = 1, \cdots, N\}$ and to compute estimates based on these samples and weights.As the number of samples becomes very large, this Monte Carlo characterization develops into an equivalent representation to the functional description of the posterior probability density function (Sanjeev, et al., 2002).

If $\{x_{0:k-1}^{(i)}, w_{0:k-1}^{(i)} i = 1, \cdots, N\}$ are samples and associated weights approximating the density function, then $p(x_{0:k-1} \mid y_{0:k-1})$, $\{x_{0:k-1}^{(i)}\}_{i=1}^{N}$ is a set of particles with associated weights $\{w_{0:k-1}^{(i)}\}_{i=1}^{N}$ with $\sum_{i=1:N} w^{(i)}_{\ k-1} = 1$, and the density function are approximated by:

$$p(x_{0:k-1} \mid y_{0:k-1}) \approx \sum_{i=1}^{N} w_{k-1}^{(i)} \delta(x_{k-1} - x_{k-1}^{(i)})$$

where $\delta(x)$ signifies the Dirac delta role, y_k becomes available when a new observation arrives, and the density function $p(x_k \mid y_k)$ is obtained recursively in two stages:

1. Drawing samples $x_k^i \sim p(x_k x_{k-1})$,

and

2. Updating the weight with the principle of importance sampling. (For details on SMC, see Doucel, et al., 2000; Sanjeev, 2002).

The particles are proliferated over time by Monte Carlo simulation to obtain new particles and weights (usually as new information are received), hence forming a series of PDF approximations over time. The reason that it works can be understood from the theory of (recursive) importance sampling.

Methodology

Procedural Functions

Consider a particular algorithm for the SMC, known also as the Sampling Importance Resampling (SIR) (Gordon, 1993; Carpenter, et al., 1999; Johansen, 2009). The algorithm can be summarized as follows: The algorithm is initiated by setting $k = 1$, for which $p(x_k \mid x_{k-1}) = p(x_k)$ is defined.

Prediction for Step k:

Draw N samples from the distribution $p\left(x_k \mid x_{k-1} = s_{k-1}^{(i)}\right)$ $\forall_i$ to form the particles $\{\hat{s}_k^{(i)}, \tilde{w}_k^{(i)}\}_{i=1:N}$. The weight is $\tilde{w}_k^{(i)} = \dfrac{\hat{w}_k^{(i)}}{\sum_i \hat{w}_k^{(i)}}$ where $\hat{w}_k^{(i)}$ is calculated from the conditional PDF $p\left(y_k \mid x_k = \hat{s}_k^{(i)}\right)$, given observation Y_k.

Resample for Step k:

Resample the random measure $\{\hat{s}_k^{(i)}, \tilde{w}_k^{(i)}\}_{i=1:N}$ obtained in the prediction procedure to obtain $\left\{s_k^{(i)}, \dfrac{1}{N}\right\}_{i=1:N}$ which has uniform weights.

The importance of the prediction step is clear by establishing the following results.Using a importance function $q(x_k \mid y_k)$ satisfying the property

$$q(x_k \mid x_{k-1}, y_k) = q\left(x_k \mid x_{k-1}, Y_i\right),$$

$\{\hat{s}_k^{(i)}, \tilde{w}_k^{(i)}\}_{i=1:N}$ is the random measure for estimating $p(x_k \mid y_k)$, where $\hat{s}_i = \left[\hat{s}_1^{(i)}, \cdots, \hat{s}_k\right]$ is the trajectoryfor particle i and where $\tilde{w}_k^{(i)} = \hat{w}_k\left(\hat{s}_k^{(i)}\right)$ is the normalized weights of particle i at time k which can be calculated recursively.

Let $\hat{w}_k^{(i)} = \hat{w}_k\left(\hat{s}_k^{(i)}\right)$, according to the argument at the k^{th} step, the density function estimate for $p(x_k \mid y_k)$ is

$$p(\hat{x}_k \mid y_k) = \sum_{i=1}^{N} \tilde{w}_k^{(i)} \delta(x_k - \hat{s}_k^{(i)}).$$

After the density function $\hat{p}(x_k \mid y_k)$ has been estimated, the observation prediction $\hat{y}_k$ with some samples with associated weights can be made. Accordingly, $p(\hat{y}_k \mid y_{k-1})$ are approximated by a new set of samples $\{\hat{y}_k^1, w_{k-1}^{(i)}\}_{i=1:N}$ and the observation prediction equation is:

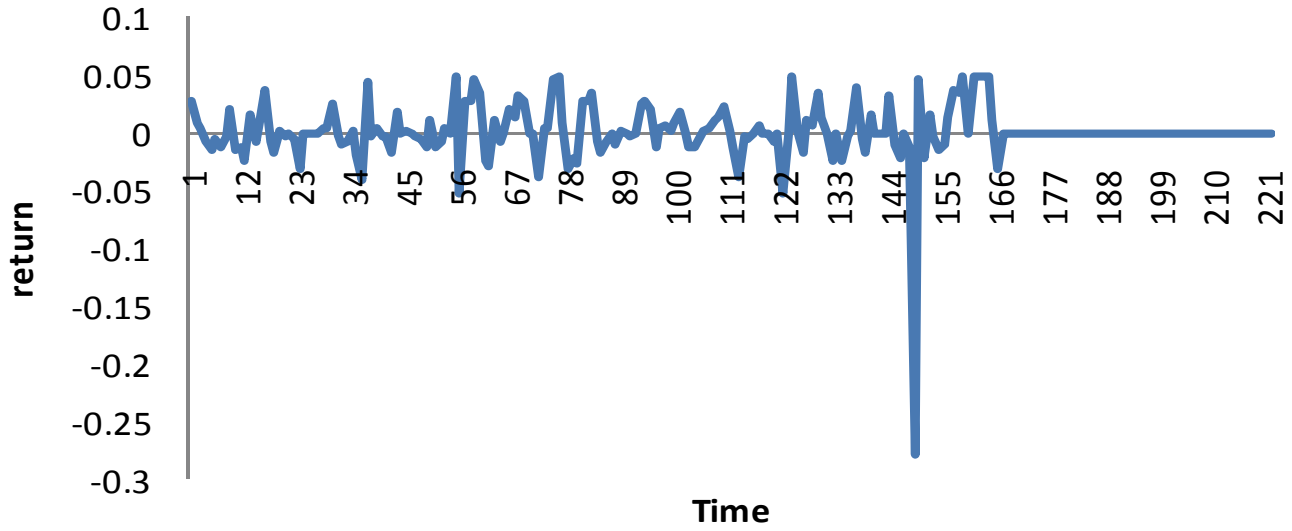
$$\hat{p}(\hat{y}_k \mid y_k) = \sum_{i=1}^{N} \tilde{w}_k^{(i)} \delta(y_k - y_k^{(i)}).$$

Data Description

The above method is applied to the data sets of daily stock prices in the banking sector of the Nigerian Stock Exchange for price indices between the years 1 January 2005 to 31 December 2008 (see www.cashcraft.com/pricemovement.asp and Figure 1). Three hidden states are studied: bull, bear and even. These hiddenstates along with the observable sequences of large rise, small rise, no change, large drop and small drop were used to develop the hidden Markov model.The sequence of observation is obtained by subtracting the prior price from the current price, the percentage change then gives the classification of the sequence of observation, where P_t is the price of an asset at time t, and the daily price relative/log returnis calculated as $r_t = \log p_t / p_{t-1}$.

Stock prices regularly alter in stock markets as observed in the price index on Tuesday, 5 February 2006; it fell by more than 100% (see Figure 2). No infallible system exists that indicates the precise movement of stock price. Instead, stock price is subjective to the influence of various factors, such as company fundamentals, external factors, and market behavior. These decide the state of the market which may be in bull or bear state. It grows along time through different market states, which are hidden states. The state of the market can be a Markovian process and is modeled in HMM.

Figure 1: Daily Stock Prices in the Banking Sector of the Nigerian Stock Exchange (Price Index between the Years 1 January 2005 to 31 December 2008)



Results

Using the functions provided by C++, this study develops an on-line algorithm of predicting hidden Markov model (Johansen, 2009). The on-line prediction using SMC begins with states producing signals that follow the normal distribution.The numbers of hidden states in the Markov chain are defined as bull (state 1), even (state 2) and bear (state 3). Figure2 shows the predicted and actual daily stock prices and Table 1 shows predicted representational prices of the NSE and predicted errors.

The stock price is modeled in HMM and prediction is made based on available observations.Due to the strong statistical foundation of the HMM and SMC methods, the model can predict similar patterns proficiently (see Figure 2). Table 1 shows that the mean absolute percentage error (MAPE) is 0.068, hence, the predictive exactness is high.

Conclusion

An online sequential Monte Carlo method is used to predict ahidden Markov model. A C++ (Sequential Monte Carlo in C++) template class library (Johansen, 2009) enabled the development of an online, sequential Monte Carlo for prediction. HMM and SMC method were introduced and the density function with a set of random samples with associated weights was approximated.Lastly, the data sets of daily stock prices in the banking sector of the Nigerian Stock Exchange were analyzed and experimental results revealed that the online algorithm is effective.

References

Avitzour, D. (1995). A Stochastic Simulation Bayesian Approach to Multitarget Tracking. *Proceedings of the IEEE on Radar, Sonar and Navigation*, *142*, 41-44.

Carpenter, J. Clifford, P., & Fearnhead, P. (1999). Improved Particle Filter for Nonlinear problems. *Proceedings of the IEEE on Radar, Sonar and Navigation*, *146*, (*1*), 2-7.

Doucet, A., & Johansen, A. M. (2008). A Tutorial on Particle Filtering and Smoothing. In Oxford Handbook of Nonlinear Filtering, D. Crisan & B. Rozovsky, Eds. Oxford University Press.

Figure 2: Daily Stock Prices in the Banking Sector of the Nigerian Stock Exchange
(Red line represents predicted stock price, Blue line represents actual stock price)

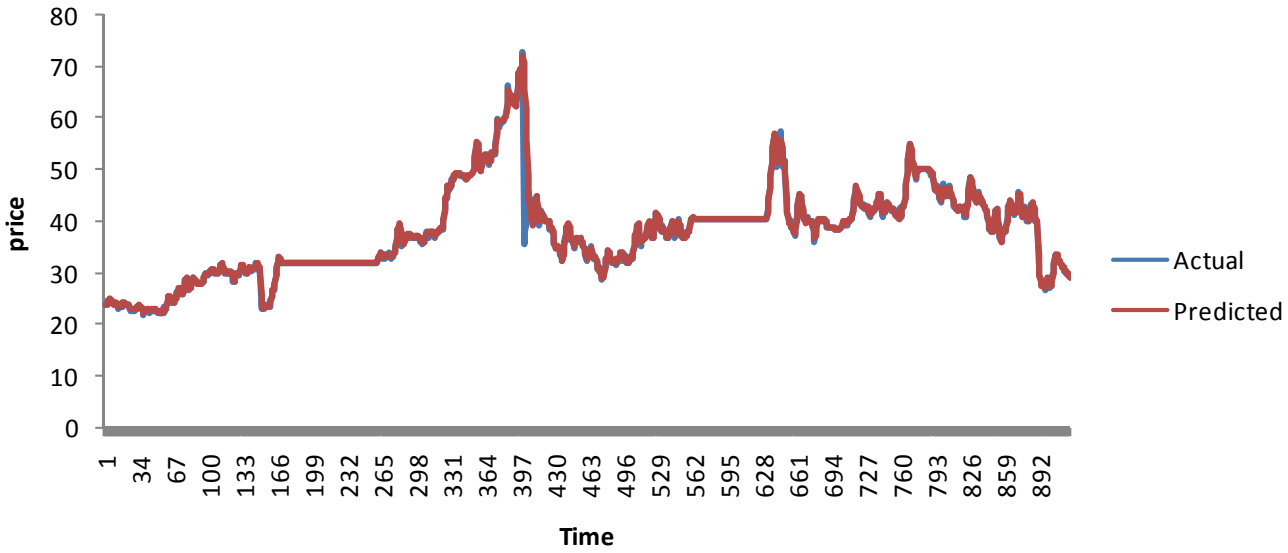


Table 1: Predicted Daily Stock Price in the Banking Sector of the NSE

Actual	Predicted	R.E(%)	MAPE(%)	Actual	Predicted	R.E(%)
24	23.8489	0.629583	0.068285	22.75	22.6411	0.478681
24.7	24.0614	2.585425		22.5	22.5232	-0.10311
24.9	24.4768	1.699598		22.35	22.373	-0.10291
25	24.941	0.236		22.45	22.3671	0.369265
24.8	24.9793	-0.72298		22.46	22.4187	0.183882
24.45	24.688	-0.97342		23.58	23.1687	1.744275
24.3	24.3934	-0.38436		22.41	22.7752	-1.62963
23.99	24.0885	-0.41059		23.06	22.9608	0.430182
23.95	23.933	0.070981		23.7	23.5019	0.835865
24.47	24.2088	1.06743		24.8	24.4987	1.214919
24.09	24.1513	-0.25446		25.68	25.5147	0.643692
23.8	23.922	-0.51261		25.08	25.5347	-1.813
23.22	23.4166	-0.84668		24.4	24.9159	-2.11434
23.6	23.4176	0.772881		24.7	24.7253	-0.10243
23.42	23.377	0.183604		24.49	24.4938	-0.01552
23.6	23.4982	0.431356		24.5	24.4089	0.371837
24.49	24.1671	1.318497		25.03	24.763	1.06672
24.3	24.3828	-0.34074		25.4	25.2465	0.604331
23.88	24.1404	-1.09045		26.24	26.0237	0.824314
23.94	24.018	-0.32581		27	26.8721	0.473704
23.85	23.89	-0.16771		27	27.2044	-0.75704
23.86	23.8301	0.125314		26.98	27.2338	-0.9407
23.73	23.7339	-0.01643		26	26.5007	-1.92577
23	23.1971	-0.85696		26.09	26.1648	-0.2867
22.98	22.9523	0.12054		26.17	26.0937	0.291555
22.99	22.8886	0.441061		27.39	26.8896	1.826944
23	22.9326	0.293043		28.75	28.2272	1.818435
23	22.955	0.195652		28.98	29.0147	-0.11974
23.1	23.055	0.194805		28.07	28.6229	-1.96972
23.2	23.1768	0.1		27.5	27.8895	-1.41636
23.78	23.6018	0.749369		26.77	27.0194	-0.93164
23.7	23.7578	-0.24388		27.5	27.1466	1.285091
23.45	23.6338	-0.7838		28.24	27.8034	1.546034
23.3	23.4173	-0.50343		29.22	28.843	1.290212
23.35	23.344	0.025696		28.99	29.1623	-0.59434
22.89	23.0174	-0.55657		28.5	28.8644	-1.2786
22	22.2651	-1.205		28.31	28.5203	-0.74285
22.97	22.5771	1.710492		28.3	28.3238	-0.0841
22.9	22.7748	0.546725		28.02	28.0612	-0.14704
23	22.9519	0.20913		28.08	27.9971	0.295228
22.95	22.9895	-0.17211		28.05	27.9861	0.227807
22.91	22.9678	-0.25229		27.95	27.9407	0.033274
22.55	22.6986	-0.65898		27.91	27.9132	-0.01147
22.95	22.826	0.540305		28.6	28.3646	0.823077
22.94	22.8994	0.176983		29.4	29.1204	0.95102
23	22.9894	0.046087		29.99	29.8659	0.413805
22.98	23.0266	-0.20279		29.65	29.9393	-0.97572
22.94	23.0066	-0.29032		29.75	29.9012	-0.50824
22.8	22.8641	-0.28114		29.96	29.9926	-0.10881
22.51	22.6008	-0.40338		29.99	30.0266	-0.12204

Churchill, G. A. (1989). Stochastic Models for Heterogeneous DNA Sequences. *Bulletin of Mathematical Biology*, *51*, 79-94.

Gordon, N. J., Salmond, D. J., & Smith, A. (1993). Novel Approach to Nonlinear/Non-Gaussian Bayesian State Estimation. *IEEE Proceedings on Radar Signal Process*, *140*(*2*), 107-113.

Gordon, N. J., Salmon, D. J., & Ewing, C. M. (1995). Bayesian State Estimation for Tracking and Guidance Using the Bootsrap Filter. *Journal of Guidance, Control and Dynamics*, *18*, 1434-1443.

Isard, M., & Blake, A. (1996). Contour Tracking by Stochastic Propagation of Conditional Density. In *Computer Vision*, Buxton & R. Cipolla, Eds. New York: Springer.

Doucet, A., de Freitas, J. F. G., & Gordon, N. J. (2000). *Sequential Monte Carlo Methods in Practice*. New York: Springer-Verlag.

Johansen, A. M., Doucet, A., & Davy, M. (2008). Particle methods for Maximum Likelihood Parameter Estimation in Latent Variable Models. *Statistics and Computing*, *18*(*1*):47-57.

Johansen, A. M. (2009). Sequential Monte Carlo in C++. *Journal of Statistical Software*, *30*(*6*), http://www.jstatsoft.org/.

Krogh, A., Brown, M., Mian, S., Sjolander, K., & Haussler, D. (1994).Proetein Modeling Using Hidden Markov Models. *Journal of Molecular Biology*, *235*, 1501-1531.

Liu, J. S., & Chen, R. (1995). Blind Deconvolution via Sequential Imputations. *Journal of the American Statistical Association*, *90*, 567-576

Liu, J. S., Neuwald, A. F., & Lawrence, C. E. (1997). *Markov Structures in Biological Sequence Alignment*. Technical Report, Stanford University.

Pitt, M. K., &Shephard, N. (1997). *Filtering via simulation: Auxiliar particle filters*. www.nuff.ox.ac.uk/users/shephard.

Rabiner, L. R. (1989). A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceedings of the IEEE*, *77*, 257-286.

Sanjeev, A., Maskell, S., Gordon, N., & Clapp, T. (2002). A tutorial on particle filter for on-line non-linear/non-Guassian Bayesian tracking. *IEEE Transactions on Signal Processing*, *50*, 174-188.

Journal of Modern Applied Statistical Methods
November 2011, Vol. 10, No. 2, 676-685

1538 – 9472/11/$95.00

Height-Diameter Relationship in Tree Modeling Using Simultaneous Equation Techniques in Correlated Normal Deviates

S. O. Oyamakin
Forestry Research Institute of Nigeria
Ibadan, Nigeria

In other to study the complex simultaneous relationships existing in forest/tree growth modeling, six estimation methods of a simultaneous equation model are examined to determine how they cope with varying degrees of correlation between pairs of random deviates using average parameter estimates. A two-equation simultaneous system assumed covariance matrix was considered. The model was structured to have a mutual correlation between pairs of random deviates: a violation of the assumption of mutual independence between pairs of such random deviates. The correlation between the pairs of normal deviates were generated using three scenarios r = 0.0, 0.3 and 0.5. The performances of estimators considered were examined at various sample sizes (N = 20, 25, 30) and correlation levels with 50 replications for each. Using the average of parameter estimates criterion, 2 3SLIML were the best estimators followed by FIML and OLS for the three cases studied. Also, as sample size increases from 20 to 25 to 30, 2-3SLIML performed best and was most consistent.

Key words: Growth models, Monte Carlo, random deviates, mutual correlation, average of parameter estimates, simultaneous equation models.

Introduction

Growth models assist forest researchers and managers in many ways. Some important uses include the ability to predict future yields and to explore silvicultural options. Models provide an efficient way to prepare resource forecasts, but a more important role may be their ability to explore management options and silvicultural alternatives. For example, foresters may wish to know the long-term effect on both the forest and on future harvests, of a particular silvicultural decision, such as changing the cutting limits for harvesting. With a growth model, they can examine the likely outcomes; both with the intended and alternative cutting limits and can make their decision objectively. The process of developing a growth model may also offer interesting new insights into stand dynamics. Forest growth models are very useful to forest managers and forestry researchers in many respects. A forest growth model aims to describe the dynamics of the forest closely and precisely enough to meet the needs of the forester or forestry researcher (dynamics includes all the change processes throughout the forest or tree's lifetime). The primary changes in the forestry field are related to the incorporation, growth and death of trees, a forest's key asset.

There are many forest growth models. Forest models the individual tree. The most common uses of these models for managers are to forecast timber production or, less often, other forestry products (cones, cork, etc.) and to simulate different forestry management alternatives with a view to decision-making. The models help to forecast what long-term effects a forestry management intervention is likely to have on both timber production and the future conditions of the actual forest, as well as the impact of interventions on other forest values. For forestry researchers, models are most useful as tools for researching forest dynamics.

S. O. Oyamakin is a Research Officer in the Department of Planning, Research, Statistics & Biometrics. Email him at: fm_oyamakin@yahoo.com.

Monte Carlo simulation is a method of analysis based on recreating a chance process (usually with a computer), running it many times, and directly observing the results. The term Monte Carlo method was coined by physicists working on nuclear weapons projects at the Los Alamos National Laboratory. Monte Carlo methods are extensively used in many fields such as operational research, nuclear physics and econometrics, where there are a variety and complexity of problems beyond the available resources of the theoretician (Adepoju, 2009a, c). Many modern investigations have employed Monte Carlo Methods, notable examples include: Wagner (1958); Nagar (1960); Johnston (1972); Anderson & Sawa (1979); Basmann (1963); Cragg (1966); Anderson (1990); Metropolis (1987); Fomby, Hill & Johnson (1988); and Smith (1973).

In Monte Carlo studies, data sets are generated with stochastic terms that are free of the problems of multicollinearity, non-spherical disturbances, measurement error and specification error. In the context of a simultaneous equation system, the design of Monte Carlo experiments requires the generation of orthogonal normal deviates or mutually independent sequences distributed as $N(0,1)$. These normal deviates are then transformed to ensure that the disturbance terms are distributed as $N(0,\Sigma)$, which are not serially correlated, where Σ is the assumed variance-covariance matrix of the disturbances: However, in real life situations, the errors are not completely correlation free (Adepoju, 2009b; Johnston & DiNardo, 1984; Anderson & Sawa, 1973). This study examined the performance of estimators of a two-equation simultaneous model to varying degrees of correlation between pairs of normal deviates.

General Study Framework

Simultaneous equation models (SEM) are at the heart of a class of models in a data generation process that depends on more than one equation interacting together to produce observed data. Unlike a single-equation model, in which a dependent (*y*) variable is a function of independent (*x*) variables, other *y* variables are among the independent variables in each SEM equation. The *y* variables in the system are jointly (or simultaneously) determined by the equations in the system.

The following two structural equations are assumed:

$$Y_{t1} = \beta_{21}Y_{t2} + \gamma_{11}X_{t1} + \gamma_{21}X_{t2} + U_{t1}$$

and

$$Y_{t2} = \beta_{12}Y_{t1} + \gamma_{12}X_{t1} + \gamma_{32}X_{t3} + U_{t2}.$$

These equations can be rewritten as:

$$-Y_{t1} = \beta_{21}Y_{t2} + \gamma_{11}X_{t1} + \gamma_{21}X_{t2} + U_{t1}$$

and

$$\beta_{12}Y_{t1} = Y_{t2} + \gamma_{12}X_{t1} + \gamma_{32}X_{t3} + U_{t2}.$$

These equations are exactly identified.

The reduced form model is derived as

$$\beta Y = \Gamma X + U$$
$$\Rightarrow Y = \beta^{-1}\Gamma X + \beta^{-1}U \text{ i.e } \pi X + V$$

where, $\pi = \beta^{-1}\Gamma$, and by extension, the following endogenous equations are obtained:

$$Y_{t1} = \frac{1}{1-\beta_{21}\beta_{12}}\begin{pmatrix}\gamma_{11}X_{t1} + \gamma_{21}X_{t2} + \beta_{21}\gamma_{12}X_{t1} \\ +\beta_{21}\gamma_{32}X_{t3} + U_{t1} + \beta_{21}U_{t2}\end{pmatrix}$$

$$Y_{t2} = \frac{1}{1-\beta_{21}\beta_{12}}\begin{pmatrix}\gamma_{12}X_{t1} + \beta_{12}\gamma_{11}X_{t1} + \beta_{12}\gamma_{21}X_{t2} \\ +\gamma_{32}X_{t3} + \beta_{12}U_{t1} + U_{t2}\end{pmatrix}$$

$$Y_{t1} = \left(\frac{\gamma_{11}+\beta_{21}\gamma_{12}}{1-\beta_{21}\beta_{12}}\right)X_{t1} + \left(\frac{\gamma_{21}}{1-\beta_{21}\beta_{12}}\right)X_{t2} + \left(\frac{\beta_{21}\gamma_{32}}{1-\beta_{21}\beta_{12}}\right)X_{t3} + \left(\frac{U_{t1}+\beta_{21}U_{t2}}{1-\beta_{21}\beta_{12}}\right)$$

and

$$Y_{t2} = \left(\frac{\beta_{12}\gamma_{11} + \gamma_{12}}{1 - \beta_{21}\beta_{12}} \right) X_{t1} + \left(\frac{\beta_{12}\gamma_{21}}{1 - \beta_{21}\beta_{12}} \right) X_{t2} + \left(\frac{\gamma_{32}}{1 - \beta_{21}\beta_{12}} \right) X_{t3} + \left(\frac{\beta_{12}U_{t1} + U_{t2}}{1 - \beta_{21}\beta_{12}} \right).$$

Monte Carlo Data Generation

Monte Carlo simulation was used to examine the properties of different statistics computed from sample data. In other words, test-drive estimators were tested, to determine how different recipes perform under different circumstances. The procedure was as follows: In each case an artificial environment was created in which the values of important parameters and the nature of the chance process were specified; then the computer simulated the chance process repeatedly and displayed the results of the experiment.

The main task was the generation of stochastic dependent (endogenous) variables $Y_{it}(i = 1,2; t = 1,...,T)$, which are subsequently used in estimating the parameters of the model. To achieve this, the following assumptions were necessary:

(i) Values of the predetermined variables X_{1t}, X_{2t}, and $X_{3t}(t = 1,...,T)$;

(ii) Values of the parameters: β_{12}, β_{21}, γ_{11}, γ_{12}, γ_{32}; and

(iii) Values of elements Ω.

The simulation of the error term $U_{it}(i = 1,2,...,T)$ is the most complex step in generating stochastic dependent variables. To conduct the Monte Carlo experiment, first, the sample size N was specified as N= 20, 25, 30. After specifying the sample size, numerical values were arbitrarily assigned to each structural parameter as follows: $\beta_{12} = 1.5$, $\beta_{21} = 1.8$, $\gamma_{11} = 1.5$, $\gamma_{11} = 1.5$ $\gamma_{12} = 0.5$, $\gamma_{32} = 2.0$ for all cases. The covariance matrix of the disturbances was specified arbitrarily as:

$$\Omega = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix} = \begin{pmatrix} 5.0 & 2.5 \\ 2.5 & 3.0 \end{pmatrix}.$$

The standard random number generator with values obtained from the uniform distribution with mean 0 and standard deviation 1 (Kmenta, 1971) was used to generate the values of the exogenous variables, $X_{it}(i = 1,2,3; t = 1,...,T)$.

Generation of Random Disturbance Term, U

A 3-stage process was employed to generate random disturbance terms. In the first stage, independent series of normal deviates of required length (N = 20, 25, 30) were generated. At the second stage, these series were standardized to a normal distribution with mean zero and variance 1. Lastly, the random disturbance terms were generated assuming three degrees of correlation between pairs of random deviates:

(i) Case I: no correlation between the random deviates ($r_{\varepsilon_1,\varepsilon_2} = 0$);

(ii) Case II: 0.3 correlation level between the random deviates ($r_{\varepsilon_1,\varepsilon_2} = 0.3$); and

(iii) Case III: 0.5 correlation level between the random deviates ($r_{\varepsilon_1,\varepsilon_2} = 0.5$).

The samples sizes considered for each scenario were N = 20, 25 and 30. The pairs of random normal deviates based on these sample sizes were generated and each was replicated 50 times. The deviates were then standardized and appropriately transformed to have a specific variance-covariance matrix Σ assumed in the model. Numerical values were generated for exogenous variables of the model as described. Next, selected $(\varepsilon_{1t}\varepsilon_{2t})$ were transformed to be distributed as $N(0,\Sigma)$ where Σ was $Cov(U_t U_t') = \Omega \otimes I_T$ and elements of Ω were

decomposed by a non- singular matrix ρ such that $\rho\rho' = \Omega$.

Recall, $V = \beta^{-1}U$

$$\begin{pmatrix} V_{t1} \\ V_{t2} \end{pmatrix} = \begin{pmatrix} \beta^* & \beta^*\beta_{21} \\ \beta^*\beta_{12} & \beta^* \end{pmatrix} \begin{pmatrix} U_{t1} \\ U_{t2} \end{pmatrix}$$

According to Nagar (1960), M independent terms of standard normal deviates of length N can be transformed into M series of random normal variables with mean 0 and a predetermined covariance matrix. In this model, $M = 2$, i.e. U_{1t}, U_{2t}, if the covariance matrix is

$$\Omega = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix}$$

where $\text{var}(U_1) = \sigma_{11}$, $\text{var}(U_2) = \sigma_{22}$ and $\text{cov}(U_1U_2) = \sigma_{12}$, considering both upper and lower triangular matrices. If the upper triangular matrix is

$$P_1 = \begin{pmatrix} \eta_{11} & \eta_{12} \\ 0 & \eta_{22} \end{pmatrix},$$

and the lower triangular matrix is

$$P_2 = \begin{pmatrix} \eta_{11} & 0 \\ \eta_{21} & \eta_{22} \end{pmatrix},$$

then

$$\Omega = P_1P_1' = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix}.$$

The pair of standard deviates can be transformed into a pair of random normal variables with mean Z^n variance σ_{11}, σ_{22} and covariance σ_{12} using

$$\begin{bmatrix} U_{1t} \\ U_{2t} \end{bmatrix} = U_t = \eta_1\varepsilon_t = \begin{pmatrix} \eta_{11} & \eta_{12} \\ 0 & \eta_{22} \end{pmatrix} \begin{bmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{bmatrix}$$

to obtain a pair of random disturbances for the upper triangular matrix:

$$\begin{aligned} U_{1t} &= \eta_{11}\varepsilon_{1t} + \eta_{12}\varepsilon_{2t} \\ &= 1.707825128\varepsilon_{1t} + 1.4043 \end{aligned}$$

and

$$\begin{aligned} U_{2t} &= \eta_{22}\varepsilon_{2t} \\ &= 1.732050808\varepsilon_{2t}. \end{aligned}$$

where $t = 1, 2, ..., T$. Similarly, an alternative solution can be obtained for the lower triangular matrix:

$$\begin{aligned} U'_{1t} &= \eta'_{11}\varepsilon_{1t} \\ &= 2.236067978\varepsilon_{1t} \end{aligned}$$

and

$$\begin{aligned} U'_{2t} &= \eta'_{12}\varepsilon_{1t} + \eta'_{22}\varepsilon_{2t} \\ &= 1.118033989\varepsilon_{1t} + 1.322875656\varepsilon_{2t}. \end{aligned}$$

Generation of Endogenous Variables

Assigning numerical values to the structural parameters provided all values required to generate the endogenous variables. Considering the upper and lower triangular matrix U_{t1}, U_{t2} defined as

$$\begin{bmatrix} U_{1t} \\ U_{2t} \end{bmatrix} = \begin{pmatrix} 1.707825128 & 1.443375673 \\ 0 & 1.732050808 \end{pmatrix} \begin{bmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{bmatrix},$$

and the lower triangular matrix U'_{1t}, U'_{2t} defined as

$$\begin{bmatrix} U'_{1t} \\ U'_{2t} \end{bmatrix} = \begin{pmatrix} 1.707825128 & 0 \\ 1.443375673 & 1.732050808 \end{pmatrix} \begin{bmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{bmatrix},$$

then, solving Y_{t1} and Y_{t2} using upper triangular matrix results in:

$$\begin{aligned} Y_{1t} = &-1.411764706X_{t1} - 0.588235294X_{t2} \\ &-2.117647059X_{t3} - 0.588235294U_{t1} \\ &-0.88235294U_{t2} \end{aligned}$$

and

$$Y_{2t} = -1.411764706X_{t1} - 0.588235294X_{t2} - 2.117647059X_{t3} - 0.88235294U_{t1} - 0.588235294U_{t2}.$$

Solving Y_{t1} and Y_{t2} using lower triangular matrix results in:

$$Y_{1t} = -1.411764706X_{t1} - 0.588235294X_{t2} - 2.117647059X_{t3} - 0.588235294U'_{t1} - 0.88235294U'_{t2}$$

and

$$Y_{2t} = -1.411764706X_{t1} - 0.588235294X_{t2} - 2.117647059X_{t3} - 0.88235294U'_{t1} - 0.588235294U'_{t2}.$$

Results

In theory, and as confirmed by Johnson (1991), when an equation is just identified, estimates of the parameter obtained by 2SLS, 3SLS and LIML should be identical. The results obtained in this study show that 2SLS, 3SLS and LIML estimators yielded virtually identical results, but the OLS, ILS and FIML yielded clearly different results from those estimators. Because 2SLS, 3SLS and LIML have the same results; the estimators shall be denoted as 2-3SLIML.

Analysis of results show that, in case I, 2-3SLIML performed best; it had the closet values to the assumed values in most cases (22) followed by FIML (8 cases) and OLS (5 cases); ILS did not perform at all. Also, as the sample size increased from 20 to 25 to 30, the value of the estimates moved closer to the true estimates of the parameters in about 72% of the cases across the upper and lower triangular matrices. For Equation I, the estimates improve from the lower triangular matrices to the upper triangular matrices.

Case II revealed that as the sample size increased, the estimates obtained by 2 3 SLIML were - in most cases - better than the remaining estimators, which did not show any clear pattern. For both P1 and P2 comparing cases I, II and III across the lower and upper triangular matrices, the performance of estimators under case I was better than those for case II and case III.

Case III revealed that, as the sample size increased from 20 to 25 to 30, the value of the estimates moved closer to the true estimates of the parameters across the upper and lower triangular matrices. For Equation I, the estimates improve from the lower triangular matrices to the upper triangular matrices.

As an illustration, for OLS over the three magnitudes of the correlation coefficient the estimates of β_{21} fell consistently for sample sizes N = 20, 25 and 30, that is, column wise comparison for the six estimates:

	N = 20	N = 25	N = 30
Case 1	0.92455	0.9256	0.9286
Case 2	0.9105	0.9098	0.9108
Case 3	0.9024	0.9045	0.9052

A comparison of the three entries in each row shows that estimates rose and fell in CASE 2, and rose consistently in both CASE 1 and CASE 3. Also, along the columns the estimates fell consistently at the three cases of the correlation coefficient at sample sizes N=20, 25 and 30.

The best OLS estimates for β_{21} , γ_{11} and γ_{21} of Equation 1 respectively are: 0.92455 (CASE 1), 0.9256 (CASE 1), 0.9286 (CASE 1) for β_{21}, 0.0077 (CASE 2), 0.0487 (CASE 2), 0.0323 (CASE 1), for γ_{11v} and 0.0065 (CASE 2), 0.0594 (CASE 3), 0.0022 (CASE 3) for γ_{21}. Thus, entries 3 (r = 0.0), 0 (r = 0.3) and 0 (r = 0.5) under β_{21}, 1 (r = 0.0), 2 (r = 0.3), 0 (r = 0.5) under γ_{11} and 0 (r = 0.0), 1 (r = 0.3), 2 (r = 0.5) under γ_{21} (See Table 1).

Similarly, for equation 2, the best OLS estimates for γ_{12} are observed for case 1. Hence, 3(r = 0.0), 0 (r = 0.3) and 0 (r = 0.5). For β_{12} they are 0 (r = 0.0), 1 (r = 0.3; 1.0757) and 2 (r = 0.5; 1.0944, 1.0914) and finally, 1 (r = 0.0; 0.06858), 1 (r = 0.3; 0.0272) and 1 (r = 0.5; 0.0955) for γ_{32}. This is repeated for the other three estimators. Results are displayed in Tables

1 and 2 for P1 and P2 respectively. Hence Tables 1 and 2 reflect the sensitivity of distribution of best estimates to varying correlation coefficients.

Tables 3 and 4 are derived from Tables 1 and 2. Each table contains the correlation-based distribution of estimators which yielded best estimates of not less than 50 percent of the parameters for each equation. Tables 3 and 4 show that CASE 2, where the error term has 0.3 level of correlation, has the least proportion of best estimates and hence fewest so-called best estimators. The most frequent estimator in this interval is the ILS and 2-3SLS.

As shown in Table 5 under P1, when error terms are not correlated (r = 0.0), OLS, 2-3SLS and FIML are best for estimating equation 1, OLS and ILS are good at CASE 2 (r = 0.3), and 2-3SLS is best at CASE 3 (r = 0.5). For equation 2, 2-3SLS is best at CASE 1, ILS is best at CASE 2 and FIML performed best at CASE 3. Under P2, the parameters of the first equation are poorly estimated at CASE 2 of the correlation coefficient (r = 0.3), ILS is best at CASE 1 followed by OLS at CASE 3. Results show that 2-3SLS performed equally well for this equation when the error term is positively correlated as in CASE 3. For equation 2, OLS and ILS are best at CASE 1, 2-3SLS is best at CASE 2 and FIML is best at CASE 3. There is a greater scope of estimating equation 2 at the three cases of correlation coefficient by several estimators.

The scope of estimating the parameter of the first equation is more sensitive to the varying correlation between the error terms than for the equation 2 and this observation is more obvious for P2 than for P1. The ranking of the estimators as displayed in Tables 6 and 8 shows that the estimators rank differently depending on whether the upper (P1) or lower (P2) triangular matrices were used. The ranking also shows that, although ILS ranks highly as the best estimator for the error term with r = 0.0, OLS is best for the error term with r = 0.3 and FIML is best for the error term with r = 0.5. The estimator rankings shown in Table 10, in which P1 and P2 are combined, is dominated in part by the ranking obtained under P2. In that table, ILS ranks high in case 1, 2-3SLS in case 2 and FIML ranks high in case 3 where the error terms are positively correlated.

Conclusion

The finite sampling property of estimators used in this work was the average of parameter estimate. Using the average of parameter estimates criterion, 2 3SLIML are the best estimators, followed by FIML and OLS, respectively, for the three cases studied. Also, as the sample size increased from 20 to 25 to 30, 2-3SLIML continued to perform best (that is, 2-3 SLIML is consistent); as the sample size increased, the estimates moved closer to the true parameter estimate in most cases. The result of this study will be used to determine the parameter estimation of simultaneous relationships of tree growth models with independent variables like Temperature, rainfall and relative humidity.

References

Adepoju, A. A. (2009a). Comparative assessment of simultaneous equation techniques to correlated random deviates. *European Journal of Scientific Research*, *28*(*2*), 253-265.

Adepoju, A. A. (2009b). Performances of the full information estimators in a two-equation structural model with correlated disturbances. *Global Journal of Pure and Applied Sciences*, *15*(*1*), 101-107.

Adepoju, A. A. (2009c). Comparative performance of the limited information technique in a two-equation structural model. *European Journal of Scientific Research*, *28*(*2*), 253-265.

Anderson, G. (1980). The structure of simultaneous estimation: A comment. *Journal of Econometrics*, *14*, 271-276.

Anderson, T., & Sawa, T. (1973). Distributions of estimates of coefficients of a single equation in a simultaneous system and their asymptotic expansions. *Econometrica*, *41*, 683-714.

Anderson, T., & Sawa, T. (1979). Evaluation of the distribution function of the two- stage least squares estimate. *Econometrica*, *47*, 163-182.

Table 1: Sensitivity of Estimators Using Average N= 20, 25, 30, R= 50 (P1)

Estimators		Equation 1			Equation 2		
		$\beta_{21}=1.8$	$\gamma_{11}=1.5$	$\gamma_{21}=1.0$	$\beta_{12}=1.5$	$\gamma_{12}=0.5$	$\gamma_{32}=2.0$
OLS	C1	3	1	0	0	3	1
	C2	0	2	1	1	0	1
	C3	0	0	2	2	0	1
ILS	C1	1	0	2	1	1	2
	C2	1	2	1	1	2	1
	C3	1	1	0	1	0	0
2-3SLS	C1	2	2	0	2	1	2
	C2	1	0	0	0	0	1
	C3	0	1	3	1	2	0
FIML	C1	1	2	1	1	1	0
	C2	0	1	1	0	1	2
	C3	2	0	1	2	1	1

Table 2: Performance of Estimators Using Average of Parameter Estimate N= 30, R= 50 (P2)

Estimators		Equation 1			Equation 2		
		$\beta_{21}=1.8$	$\gamma_{11}=1.5$	$\gamma_{21}=1.0$	$\beta_{12}=1.5$	$\gamma_{12}=0.5$	$\gamma_{32}=2.0$
OLS	C1	0	2	0	3	1	0
	C2	0	1	1	0	1	2
	C3	3	0	2	0	1	1
ILS	C1	1	2	2	1	2	1
	C2	0	0	1	0	0	2
	C3	2	1	0	2	1	0
2-3SLS	C1	0	1	0	1	1	1
	C2	1	1	1	2	1	1
	C3	2	1	2	0	1	1
FIML	C1	1	0	1	1	1	0
	C2	0	1	1	0	1	2
	C3	2	2	1	2	1	1

Table 3: Correlation-Based Sample Size-Free Distribution of Best Estimators
N = 20, 25, 30. R = 50, (P1)

Level of Correlation	Equation 1	Equation 2
CASE 1	OLS/2-3SLS/FIML	2-3SLS/OLS/ILS
CASE 2	-	ILS
CASE 3	2-3SLS	FIML

Source: Table 1

Table 4: Correlation-Based Sample Size-Free Distribution of Best Estimators
N = 20, 25, 30. R = 50, (P2)

Level of Correlation	Equation 1	Equation 2
CASE 1	ILS	OLS/ILS
CASE 2	-	2-3SLS
CASE 3	OLS/2-3SLS/FIML	FIML

Source: Table 2

Table 5: Sample and Replication-Free Distribution of Best Estimates of P1

Equation 1			Equation 2		
Case 1	Case 2	Case3	Case 1	Case 2	Case 3
OLS(4)	OLS(3)	2-3SLS(4)	2-3SLS(5)	ILS(4)	FIML(4)
2-3SLS(4)	ILS(3)	FIML(3)	OLS(4)	FIML(3)	OLS(3)
FIML(4)	FIML(2)	OLS(2)	ILS(4)	OLS(2)	2-3SLS(3)
ILS(3)	2-3SLS(1)	ILS(2)	FIML(2)	2-3SLS(1)	ILS(1)

Table 6: Rank of Estimators Using Level of Correlation (P1) for Eq1 and Eq2

Case 1	Case 2	Case 3
2-3 SLS(9)	ILS(7)	2-3SLS(7)
OLS(8)	OLS(5)	FIML(7)
ILS(7)	FIML(5)	OLS(5)
FIML(6)	2-3SLS(2)	ILS(3)

Table 7: Sample and Replication-Free Distribution of Best Estimates of P2

Equation 1			Equation 2		
Case 1	Case 2	Case3	Case 1	Case 2	Case 3
ILS(5)	2-3SLS(3)	OLS(5)	OLS(4)	2-3SLS(4)	FIML(5)
OLS(2)	OLS(2)	2-3SLS(5)	ILS(4)	OLS(3)	ILS(3)
FIML(2)	FIML(2)	FIML(5)	2-3SLS(3)	FIML(3)	OLS(2)
2-3SLS(1)	ILS(1)	ILS(3)	FIML(2)	ILS(2)	2-3SLS(2)

Table 8: Rank of Estimators Using Level of Correlation (P2) For Eq1 and Eq2

Case 1	Case 2	Case 3
ILS(9)	2-3SLS(7)	FIML(10)
OLS(6)	OLS(5)	OLS(7)
2-3SLS(4)	FIML(5)	2-3SLS(7)
FIML(4)	ILS(3)	ILS(6)

Table 9: Sample and Replication – Free Distribution of Best Estimates of P1 and P2

Equation 1			Equation 2		
Case 1	Case 2	Case3	Case 1	Case 2	Case 3
ILS(8)	OLS(5)	2-3SLS(9)	OLS(8)	ILS(6)	FIML(9)
OLS(6)	1LS(4)	FIML(8)	ILS(8)	FIML(6)	OLS(5)
FIML(6)	2-3SLS(4)	OLS(7)	2-3SLS(8)	OLS(5)	2-3SLS(5)
2-3SLS(5)	FIML(4)	ILS(5)	FIML(4)	2-3SLS(5)	1LS(4)

Table 10: Rank of Estimators Using Level of Correlation (P1 and P2 Combined)

Case 1	Case 2	Case 3
ILS(16)	OLS(10)	FIML(17)
OLS(14)	ILS(10)	2-3SLS(14)
2-3SLS(13)	FIML(10)	OLS(12)
FIML(10)	2-3SLS(9)	ILS(9)

Basmann, R. L. (1963). A note on the exact finite sample frequency functions of generalized classical linear estimators in a leading three-equation case. *Journal of the American Statistical Association*, *58*, 161-171.

Cragg, J. G. (1966). On the sensitivity of simultaneous-equations estimators to the stochastic assumptions of the models. *Journal of the American Statistical Association*, *61*, 136-151.

Fomby, T. B., Hill, R. C., & Johnson, S. R. (1988). *Advanced Econometrics methods*. New York, NY: Springer-Verlag.

Johnston, J. (1972). *Econometric methods*, *2nd Edition*. New York, NY: McGraw Hill.

Johnston, J., & DiNardo, J. (1984). *Econometric methods*, *4th Edition*. New York, NY: McGraw-Hill International.

Kmenta, J. (1971). *Elements of econometrics*. New York, NY: MacMillian.

Nagar, A. (1959). The bias and moment matrix of the general k-class estimators of the parameters in simultaneous equations. *Econometrica*, *27*, 575-595.

Nagar, A. (1960). A Monte Carlo study of alternative simultaneous equation estimators. *Econometrica*, *28*, 573-590.

Metropolis, N. (1987). The beginning of Monte Carlo method. *Los Alamos Science Special Issue Dedicated to Stanislaw Ulam*, 125-130.

Smith, V. (1973). *Monte Carlo methods*. Lexington, MA: D.C. Health.

Wagner, H. (1958). A Monte Carlo study of estimates of simultaneous linear structural equations. *Econometrica*, *26*, 117-133.

Journal of Modern Applied Statistical Methods
November 2011, Vol. 10, No. 2, 686-691

1538 – 9472/11/$95.00

Brief Report

Higher Order C(t, p, s) Crossover Designs

James F. Reed III
Christiana Care Hospital System,
Newark, Delaware

A crossover study is a repeated measures design in which each subject is randomly assigned to a sequence of treatments, including at least two treatments. The most damning characteristic of a crossover study is the potential of a carryover effect of one treatment to the next period. To solve the first-order crossover problem characteristic in the classic AB|BA design, the design must be extended. One alternative uses additional treatment sequences in two periods; a second option is to add a third period and repeat one of the treatments. Assuming a traditional model that specifies a first-order carryover effect, this study investigates the following alternative crossover trial designs: (1) two-treatment two-period four-sequence design (Balaam, 1968) design, (2) two treatments-three period-four sequence design (Ebbutt, 1984), and (3) three treatment-two period-six sequence design (Koch, 1983). Each design has attractive properties and, when properly applied, allows both treatment and carryover effects to be estimated.

Key words: Crossover design, Balaam's crossover design, Ebbutt's crossover design, Koch crossover design.

Introduction

The most damning characteristic of a crossover study is the potential for a carryover effect of one treatment to the next period. To manage this, researchers typically include washout periods in study designs. These washout periods are thought to be of sufficient length to negate any lingering effect of one treatment into the next period. In this article, and in most of the literature on crossover designs, the persistence of a carryover effect is assumed to (1) last for only a single period (a first-order carryover effect), and (2) a carryover effect is different for different treatments. If a carryover effect is suspected in any crossover trial, then a term for this effect must be included in the model and accounted for in subsequent analysis.

This study assumes a traditional model that specifies a first-order carryover effect and outlines three higher-order crossover designs: (1) a two-treatment two-period four-sequence design (Balaam, 1968), (2) a two treatments-three period-four sequence design (Ebbutt, 1984), and (3) a three treatment-two period-six sequence (Koch, 1983) design. Each design has appealing properties and - when properly applied - estimate both treatment and carryover effects.

James F. Reed III, Ph.D., is the Director, Clinical Business Intelligence Biostatistician, Adventist Health. Email him at: ReedJF@ah.org.

The Traditional Crossover Design Model with Continuous Data

The traditional crossover design with t-treatments, p-periods, and s-sequences, C(t, p, s), assumes that each treatment has a simple first-order carryover effect that does not interact with the direct effect of the treatment in the subsequent period, and that subject effects are either fixed or random. Though a variety of models are considered in the literature, virtually all work in crossover designs uses the same underlying statistical model. This model assumes the following for the response of patient y_{ij}: If y_{ij} denotes the observed response of subject j (j = 1, …, n) in period i (i = 1, …, p), then

$$y_{ij} = \mu + \pi_i + \tau_{d(i,j)} + \lambda_{d(i-1,j)} + \beta_j + \varepsilon_{ij}.$$

Where π_i is the effect of period i, $\tau_{d(i,j)}$ is the direct effect of treatment D, $\lambda_{d(i-1,j)}$ is the simple first-order carryover effect of treatment D, d(i, j) is the treatment allocated to patient j in period i, and $\lambda_{d(0,j)} = 0$ for all j. It is assumed that all effects are fixed effects. β_j is the effect of patient j and ε_{ij} is the error term. The random subject effect, β_j, and the experimental error, ε_{ij}, are assumed to be mutually independently distributed as N $(0, \sigma^2_\beta)$ and N $(0, \sigma^2_\varepsilon)$.

The primary purpose of a crossover design comparing treatments A and B is to estimate the treatment contrast $\tau_A - \tau_B$. The period effects (π_1 and π_2), the first order carryover effects (λ_A and λ_B) and μ are typically regarded as nuisance parameters that are desirable to eliminate from any estimate. To solve the first-order crossover problem in the two-treatment two-period crossover design, one possible solution is to extend the design to four sequences. Balaam's C(2, 2, 4) design (Balaam, 1968), AA|AB|BA|BB, is generally accepted as optimal for estimating treatment effects and is also more efficient than the classic C(2, 2, 2) design (Laska, Meisner & Kushner, 1983). If the carryover effect is absent, this design is inefficient because many subjects likely will not contribute any information to the estimate of treatment differences in the two sequences AA and BB. Using Balaam's design, unbiased estimates of the treatment differences and carryover effects are easily derived (see Table 1).

The second design strategy is to extend the classic design by adding a third period and repeating one of the two treatments. The treatment sequences will ensure that the first two trial periods constitute a conventional two-period crossover trial if the third treatment period leads to excessive subject drop-outs. Ebbutt's efficient C(2, 3, 4) design, the ABB| BAA|ABA|BAB (Ebbutt, 1984) illustrates this second strategy. This design, with equal number of subjects per sequence, is able to estimate all parameters in the traditional model and provide an unbiased estimate of the treatment contrast (Ebbutt, 1984; Heydat & Stufken, 2003; Liang & Carriere, 2010) (see Table 2). The expected values for each of the sequences are: $E[c_1] = E[(2y_{11} - y_{21} - y_{31})]$, $E[c_2] = E[(2y_{21} - y_{22} - y_{32})]$, $E[c_3] = E[(2y_{31} - y_{32} - y_{33})]$, and $E[c_4] = E[(2y_{41} - y_{42} - y_{43})]$. The linear contrast of $\frac{1}{2}(c_1 - c_2 + c_3 - c_4)$ forms an unbiased estimate of $\tau_A - \tau_B$. In testing for carryover effect, let c_i, i = 5, …, $8 = E[y_{1i} + y_{2i} - y_{3i}]$. The contrast $c_5 - c_6 + c_7 - c_8$ forms an unbiased estimate of $\lambda_A - \lambda_B$.

Koch's crossover design comparing two treatments A and B to a placebo P, uses six sequences AB, BA, AP, BP, PA, and PB (see Table 3). These six sequences enable the estimation of period effects, treatment effects and carryover effects from within-subject information. The four hypotheses of interest are: (1) $\tau_A - \tau_B$, (2) $\tau_A - \tau_P$, (3) $\tau_{\mathbf{B}} - \tau_P$, and (4) $\lambda_B - \lambda_A$. The linear contrast $(c_5 - c_6)$ forms an unbiased estimate of $\tau_A - \tau_B$; the linear contrast $(c_4 - c_2)$ forms an unbiased estimate of $\tau_A - \tau_P$; the linear contrast $(c_1 - c_3)$ forms an unbiased estimate of $\tau_{\mathbf{B}} - \tau_P$; and the linear contrast $(c_2 - c_1)$ forms an unbiased estimate of $\lambda_B - \lambda_A$.

Koch's C(3, 2, 6) design has six sequences, AB, BA, AC, CA, BC and CB (see Table 4). In this design, the hypotheses of interest are: (1) $\tau_A - \tau_B$, (2) $\tau_A - \tau_C$, (3) $\tau_{\mathbf{B}} - \tau_C$, (4) $\lambda_A - \lambda_B$, (5) $\lambda_A - \lambda_C$, and (6) $\lambda_B - \lambda_C$. The linear contrast $(c_1 - c_3)$ forms an unbiased estimate of $\tau_B - \tau_C$; the linear contrast $(c_2 - c_5)$ forms an unbiased estimate of $\tau_A - \tau_C$; and the linear contrast $(c_4 - c_6)$ forms an unbiased estimate of $\tau_{\mathbf{B}} - \tau_A$. For the three carryover hypotheses the linear contrast $(c_{1'} - c_{2'})$ forms an unbiased estimate of $\lambda_A - \lambda_B$; the linear contrast $(c_{3'} - c_{4'})$ forms an unbiased estimate of $\lambda_A - \lambda_C$; and the linear contrast $(c_{5'} - c_{6'})$ forms an unbiased estimate of $\lambda_B - \lambda_C$.

Conclusion

Optimal crossover designs are statistically efficient and require fewer subjects for the same number of observations than do non-crossover designs. Because variability is typically less within a subject than between different subjects, there is a corresponding increase in the precision of observations. The result: fewer subjects are required to detect a treatment difference. For example, if $N_{parallel}$ is the total number of subjects required for a two-way parallel trial to detect a treatment effect (δ) with 5% significance and 80% power, the total number of subjects $N_{crossover}$ required for a 2 x 2 crossover trial to detect the same effect is approximately

$N_{crossover} = (1 - r)N_{parallel}/2$, where r is a correlation coefficient among the repeated measurements of the primary endpoint.

The major concern - and subject of countless discussions - in a crossover study is the presence of a carryover effect. The standard way to avoid the carryover effect is to include a rest period between successive periods, hoping that the carryover effect will wash out. The inclusion of a rest period between each pair of successive periods increases the total duration of the experiment and there is no guarantee that any carryover effect will be eliminated.

To address the potential of first-order carryover effects, the classic AB|BA crossover design could easily be extended to one of the designs outlined herein. In effect, either the added sequence(s) or added treatment period permits direct estimates of treatment effect and examination of any carryover effects.

References

Balaam, L. N. (1968). A two-period design with t^2 experimental units. *Biometrics*, *24*, 61-73.

Ebbutt, A. F. (1984). Three-period crossover designs for two treatments. *Biometrics*, *40*, 219-24.

Hedayat, A. S., & Stufken, J. (2003). Optimal and efficient crossover designs under different assumptions about the carryover effects. *Journal of Biopharmaceutical Statistics*, *13*, 519-28.

Koch, G. C., Amara, K. A., Brown, Jr., B. W., Colton, T., & Gillings, D. B. (1983). A two-period crossover design for the comparison of two active treatments and placebo. *Statistics in Medicine*, *8*, 487-504.

Laska, E., Meisner, M., & Kushner, H. B. (1983). Optimal crossover designs in the presence of carryover effects. *Biometrics*, *39*, 1087-1091.

Liang, Y., & Carriere, K. C. (2010). On the role of baseline measurements for crossover designs under the self and mixed carryover effects model. *Biometrics*, *66*, 140-148.

Table 1: Balaam's Design (AB|BA|AA|BB)

AB|BA Design	Period 1 (k = 1)	Period 2 (k = 2)
Sequence AB (i = 1)	$\mu + \pi_1 + \tau_A$	$\mu + \pi_2 + \tau_B + \lambda_A$
Sequence BA (i = 2)	$\mu + \pi_1 + \tau_B$	$\mu + \pi_2 + \tau_A + \lambda_B$
Sequence AA (i = 3)	$\mu + \pi_1 + \tau_A$	$\mu + \pi_2 + \tau_A + \lambda_A$
Sequence BB (i = 4)	$\mu + \pi_1 + \tau_B$	$\mu + \pi_2 + \tau_B + \lambda_B$

Table 1 Notes:

Sequence AB (i = 1): $E(y_{AB,1}) = \mu_{AB,1} = \mu + \pi_1 + \tau_A$, $E(y_{AB,2}) = \mu_{AB,2} = \mu + \pi_2 + \tau_B + \lambda_A$
Sequence BA (i = 2): $E(y_{BA,1}) = \mu_{BA,1} = \mu + \pi_1 + \tau_B$, $E(y_{AB,2}) = \mu_{BA,2} = \mu + \pi_2 + \tau_A + \lambda_B$
Sequence AA (i = 3): $E(y_{BA,1}) = \mu_{AA,1} = \mu + \pi_1 + \tau_A$, $E(y_{AB,2}) = \mu_{BA,2} = \mu + \pi_2 + \tau_A + \lambda_A$
Sequence BB (i = 4): $E(y_{BA,1}) = \mu_{BA,1} = \mu + \pi_1 + \tau_B$, $E(y_{AB,2}) = \mu_{BA,2} = \mu + \pi_2 + \tau_B + \lambda_B$

In sequence AB, contrast c_1 has expected value: $E[c_1] = E[y_{11} - y_{21}] = (\pi_1 - \pi_2) + (\tau_A - \tau_B) - \lambda_A$
In sequence BA, contrast c_2 has expected value: $E[c_2] = E[y_{21} - y_{22}] = (\pi_1 - \pi_2) - (\tau_A - \tau_B) - \lambda_B$
In sequence AA, contrast c_3 has expected value: $E[c_3] = E[y_{31} - y_{32}] = (\pi_1 - \pi_2) - \lambda_A$
In sequence BB, contrast c_4 has expected value: $E[c_4] = E[y_{41} - y_{42}] = (\pi_1 - \pi_2) - \lambda_B$

In sequence AB, contrast c_5 has expected value: $E[c_5] = E[y_{11} + y_{21}] = 2\mu + (\pi_1 + \pi_2) + (\tau_A + \tau_B) + \lambda_A$
In sequence BA, contrast c_6 has expected value: $E[c_6] = E[y_{21} + y_{22}] = 2\mu + (\pi_1 + \pi_2) + (\tau_A + \tau_B) + \lambda_B$

Table 2: Ebbutt AAB|BAA|ABA|BAB Design

AB|BA Design	Period 1 (k = 1)	Period 2 (k = 2)	Period 3 (k = 3)
ABB (i = 1)	$\mu + \pi_1 + \tau_A$	$\mu + \pi_2 + \tau_B + \lambda_A$	$\mu + \pi_3 + \tau_B + \lambda_B$
BAA (i=2)	$\mu + \pi_1 + \tau_B$	$\mu + \pi_2 + \tau_A + \lambda_B$	$\mu + \pi_3 + \tau_A + \lambda_A$
ABA (i = 3)	$\mu + \pi_1 + \tau_A$	$\mu + \pi_2 + \tau_B + \lambda_A$	$\mu + \pi_3 + \tau_A + \lambda_B$
BAB (i = 4)	$\mu + \pi_1 + \tau_B$	$\mu + \pi_2 + \tau_A + \lambda_B$	$\mu + \pi_3 + \tau_B + \lambda_A$

Table 2 Notes:

ABB (i = 1): $E(y_{ABB,1}) = \mu + \pi_1 + \tau_A$, $E(y_{ABB,2}) = \mu + \pi_2 + \tau_A + \lambda_B$, $E(y_{ABB,3}) = \mu + \pi_3 + \tau_A + \lambda_B$
BAA (i = 2): $E(y_{BAB,1}) = \mu + \pi_1 + \tau_B$, $E(y_{BAB,2}) = \mu + \pi_2 + \tau_A + \lambda_B$, $E(y_{BAB,2}) = \mu + \pi_3 + \tau_A + \lambda_A$
ABA (i = 3): $E(y_{ABA,1}) = \mu + \pi_1 + \tau_A$, $E(y_{ABA,2}) = \mu + \pi_2 + \tau_A + \lambda_A$, $E(y_{ABA,3}) = \mu + \pi_3 + \tau_A + \lambda_B$
BAB (i = 4): $E(y_{AAB,1}) = \mu + \pi_1 + \tau_B$, $E(y_{AAB,2}) = \mu + \pi_2 + \tau_A + \lambda_B$, $E(y_{AAB,3}) = \mu + \pi_3 + \tau_B + \lambda_A$

In sequence ABB, the expected value $E[c_1]=E[(2y_{11} - y_{21} - y_{31})]=\{(2\pi_1 - \pi_2 - \pi_3) + 2(\tau_A - \tau_B) - \lambda_A - \lambda_B\}$
In sequence BAA, the expected value $E[c_2]=E[(2y_{21} - y_{22} - y_{32})]=\{(2\pi_1 - \pi_2 - \pi_3) + 2(\tau_A - \tau_B) - \lambda_A - \lambda_B\}$
In sequence ABA, the expected value $E[c_3]=E[(2y_{31} - y_{32} - y_{33})]=\{(2\pi_1 - \pi_2 - \pi_3) + (\tau_A - \tau_B) - \lambda_A - \lambda_B\}$
In sequence BAB, the expected value $E[c_4]=E[(2y_{41} - y_{42} - y_{43})]=\{(2\pi_1 - \pi_2 - \pi_3) - (\tau_A - \tau_B) - \lambda_A - \lambda_B\}$

In sequence ABB, the expected value $E[c_5]=E[(y_{11} + y_{21} - y_{31})]=\{2\mu + (\pi_1 + \pi_2 - \pi_3) + \tau_A + (\lambda_A - \lambda_B)\}$
In sequence BAA, the expected value $E[c_6]=E[(y_{21} + y_{22} - y_{32})]=\{2\mu + (\pi_1 + \pi_2 - \pi_3) + \tau_B - (\lambda_A - \lambda_B)\}$
In sequence ABA, the expected value $E[c_7]=E[(y_{31} + y_{32} - y_{33})]=\{2\mu + (\pi_1 + \pi_2 - \pi_3) + \tau_B + (\lambda_A - \lambda_B)\}$
In sequence BAB, the expected value $E[c_8]=E[(y_{41} + y_{42} - y_{43})]=\{2\mu + (\pi_1 + \pi_2 - \pi_3) + \tau_A - (\lambda_A - \lambda_B)\}$

Table 3: Koch Design (Treatments A, B and Placebo P)

Sequence	Period 1 (k = 1)	Period 2 (k = 2)
AB (i = 1)	$\mu + \pi_1 + \tau_A$	$\mu + \pi_2 + \tau_B + \lambda_A$
BA (i = 2)	$\mu + \pi_1 + \tau_B$	$\mu + \pi_2 + \tau_A + \lambda_B$
AP (i = 3)	$\mu + \pi_1 + \tau_A$	$\mu + \pi_2 + \tau_P + \lambda_A$
BP (i = 4)	$\mu + \pi_1 + \tau_B$	$\mu + \pi_2 + \tau_P + \lambda_B$
PA (i = 5)	$\mu + \pi_1 + \tau_P$	$\mu + \pi_2 + \tau_A + \lambda_P$
PB (i = 6)	$\mu + \pi_1 + \tau_P$	$\mu + \pi_2 + \tau_B + \lambda_P$

Table 3 Notes:

Sequence AB (i = 1): $E(y_{AB,1}) = \mu + \pi_1 + \tau_A$, $E(y_{AB,1}) = \mu + \pi_2 + \tau_B + \lambda_A$
Sequence BA (i = 2): $E(y_{AB,1}) = \mu + \pi_1 + \tau_B$, $E(y_{AB,1}) = \mu + \pi_2 + \tau_A + \lambda_B$
Sequence AP (i = 3): $E(y_{AB,1}) = \mu + \pi_1 + \tau_A$, $E(y_{AB,1}) = \mu + \pi_2 + \tau_P + \lambda_A$
Sequence BP (i = 4): $E(y_{AB,1}) = \mu + \pi_1 + \tau_B$, $E(y_{AB,1}) = \mu + \pi_2 + \tau_P + \lambda_B$
Sequence PA (i = 5): $E(y_{AB,1}) = \mu + \pi_1 + \tau_P$, $E(y_{AB,1}) = \mu + \pi_2 + \tau_A + \lambda_P$
Sequence PB (i = 6): $E(y_{AB,1}) = \mu + \pi_1 + \tau_P$, $E(y_{AB,1}) = \mu + \pi_2 + \tau_B + \lambda_P$

In sequence AB, contrast c_1 has expected value: $E[c_1] = E[(y_{11} - y_{12})] = (\pi_1 - \pi_2) + (\tau_A - \tau_B) - \lambda_A$
In sequence BA, contrast c_2 has expected value: $E[c_2] = E[(y_{21} - y_{22})] = (\pi_1 - \pi_2) - (\tau_A - \tau_B) - \lambda_B$
In sequence AP, contrast c_3 has expected value: $E[c_3] = E[(y_{31} - y_{32})] = (\pi_1 - \pi_2) + (\tau_A - \tau_P) - \lambda_A$
In sequence BP, contrast c_4 has expected value: $E[c_4] = E[(y_{41} - y_{42})] = (\pi_1 - \pi_2) + (\tau_B - \tau_P) - \lambda_B$
In sequence PA, contrast c_4 has expected value: $E[c_5] = E[(y_{51} - y_{52})] = (\pi_1 - \pi_2) - (\tau_A - \tau_P) - \lambda_P$
In sequence PB, contrast c_6 has expected value: $E[c_6] = E[(y_{61} - y_{62})] = (\pi_1 - \pi_2) - (\tau_B - \tau_P) - \lambda_P$

Table 4: Koch Design (Three Treatments, Two Periods)

Sequence	Period 1 (k = 1)	Period 2 (k = 2)
AB (i = 1)	$\mu + \pi_1 + \tau_A$	$\mu + \pi_2 + \tau_B + \lambda_A$
BA (i = 2)	$\mu + \pi_1 + \tau_B$	$\mu + \pi_2 + \tau_A + \lambda_B$
AC (i = 3)	$\mu + \pi_1 + \tau_A$	$\mu + \pi_2 + \tau_C + \lambda_A$
CA (i = 4)	$\mu + \pi_1 + \tau_C$	$\mu + \pi_2 + \tau_A + \lambda_C$
BC (i = 5)	$\mu + \pi_1 + \tau_B$	$\mu + \pi_2 + \tau_C + \lambda_B$
CB (i = 6)	$\mu + \pi_1 + \tau_C$	$\mu + \pi_2 + \tau_B + \lambda_C$

Table 4 Notes:

Sequence AB (i = 1): $E(y_{AB,1}) = \mu + \pi_1 + \tau_A$, $E(y_{AB,2}) = \mu + \pi_2 + \tau_B + \lambda_A$
Sequence BA (i = 2): $E(y_{AB,1}) = \mu + \pi_1 + \tau_B$, $E(y_{AB,2}) = \mu + \pi_2 + \tau_A + \lambda_B$
Sequence AC (i = 3): $E(y_{AB,1}) = \mu + \pi_1 + \tau_A$, $E(y_{AB,2}) = \mu + \pi_2 + \tau_C + \lambda_A$
Sequence CA (i = 4): $E(y_{AB,1}) = \mu + \pi_1 + \tau_C$, $E(y_{AB,2}) = \mu + \pi_2 + \tau_A + \lambda_C$
Sequence BC (i = 5): $E(y_{AB,1}) = \mu + \pi_1 + \tau_B$, $E(y_{AB,2}) = \mu + \pi_2 + \tau_C + \lambda_B$
Sequence CB (i = 6): $E(y_{AB,1}) = \mu + \pi_1 + \tau_C$, $E(y_{AB,2}) = \mu + \pi_2 + \tau_B + \lambda_C$

In sequence AB, contrast c_1 has expected value: $E[c_1] = E[(y_{11} - y_{12})] = (\pi_1 - \pi_2) + (\tau_A - \tau_B) - \lambda_A$
In sequence BA, contrast c_2 has expected value: $E[c_2] = E[(y_{21} - y_{21})] = (\pi_1 - \pi_2) - (\tau_A - \tau_B) - \lambda_B$
In sequence AC, contrast c_3 has expected value: $E[c_3] = E[(y_{31} - y_{21})] = (\pi_1 - \pi_2) + (\tau_A - \tau_C) - \lambda_A$
In sequence CA, contrast c_4 has expected value: $E[c_4] = E[(y_{41} - y_{21})] = (\pi_1 - \pi_2) - (\tau_A - \tau_C) - \lambda_C$
In sequence BC, contrast c_5 has expected value: $E[c_5] = E[(y_{51} - y_{21})] = (\pi_1 - \pi_2) + (\tau_B - \tau_C) - \lambda_B$
In sequence CB, contrast c_6 has expected value: $E[c_6] = E[(y_{61} - y_{21})] = (\pi_1 - \pi_2) - (\tau_B - \tau_C) - \lambda_C$

In sequence AB, contrast $c_{1'}$ has expected value: $E[c_{1'}] = E[(y_{11} + y_{12})] = 2\mu + (\pi_1 + \pi_2) + (\tau_A + \tau_B) + \lambda_A$
In sequence BA, contrast $c_{2'}$ has expected value: $E[c_{2'}] = E[(y_{21} + y_{22})] = 2\mu + (\pi_1 + \pi_2) + (\tau_A + \tau_B) + \lambda_B$
In sequence AC, contrast $c_{3'}$ has expected value: $E[c_{3'}] = E[(y_{31} + y_{32})] = 2\mu + (\pi_1 + \pi_2) + (\tau_A + \tau_C) + \lambda_A$
In sequence CA, contrast $c_{4'}$ has expected value: $E[c_{4'}] = E[(y_{41} + y_{42})] = 2\mu + (\pi_1 + \pi_2) + (\tau_A + \tau_C) + \lambda_C$
In sequence BC, contrast $c_{5'}$ has expected value: $E[c_{5'}] = E[(y_{51} + y_{52})] = 2\mu + (\pi_1 + \pi_2) + (\tau_B + \tau_C) + \lambda_B$
In sequence CB, contrast $c_{6'}$ has expected value: $E[c_{6'}] = E[(y_{61} + y_{62})] = 2\mu + (\pi_1 + \pi_2) + (\tau_B + \tau_C) + \lambda_C$

Journal of Modern Applied Statistical Methods
November 2011, Vol. 10, No. 2, 692-698

1538 – 9472/11/$95.00

Emerging Scholars

A Pooled Two-Sample Median Test Based on Density Estimation

Vadim Y. Bichutskiy
George Mason University
Fairfax, Virginia

A new method based on density estimation is proposed for medians of two independent samples. The test controls the probability of Type I error and is at least as powerful as methods widely used in statistical practice. The method can be implemented using existing libraries in R.

Key words: Sample median, two-sample hypothesis test, adaptive kernel density estimation.

Introduction

Let $X_1, X_2, \ldots, X_n$ be iid having cdf F and pdf f with $F(\eta) = 1/2$ so that η is the population median. Suppose f is continuous at η with $f(\eta) > 0$. Denote the sample median by H. It is known that H is asymptotically normal with mean η and variance $1/4nf^2(\eta)$. Estimating the asymptotic standard error of the sample median requires an estimate of the population density at the median. Besides being a challenging problem, density estimation was difficult to apply in practice prior to the computer revolution; due to this, several alternative methods for estimating the standard error of the sample median have been developed (Maritz & Jarrett, 1978; McKean & Schrader, 1984; Price & Bonett, 2001; Sheather & Maritz, 1983; Sheather, 1986).

Comparing medians based on two independent samples is a well-studied problem (see Wilcox & Charlin, 1986; Wilcox, 2005; Wilcox, 2006; Wilcox, 2010 also has a good discussion). The methods fall into two main categories. The first uses the bootstrap (Efron, 1979), and the second assumes the sample median or some other estimator of the population median is approximately normal and uses one of several methods for estimating the standard error of the sample median. Virtually all methods are very conservative, particularly for heavy-tailed populations.

A new two-sample test is proposed for comparing medians. When population shapes can be assumed to be the same, a pooled test statistic, analogous to a pooled two-sample Student's t statistic for comparing means, is derived. Computer-intensive Monte Carlo simulations in R (R Development Core Team, 2009) are used to study the properties of the test and compare it to other methods. The method offers several additional benefits to practitioners: (1) a parameter that controls the trade-off between making the test conservative and liberal with a suitable value of the parameter producing a test with a nominal significance level; (2) the test is easy to implement in R using the QUANTREG (Koenker, 2009) library.

Methodology

Two-Sample Test Statistic for Difference in Medians

Let $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_m$ be two independent random samples of sizes n and m from populations with densities f_x, f_y that are continuous at the medians η_x, η_y with $f_x(\eta_x) > 0$, $f_y(\eta_y) > 0$, respectively. Denote sample medians by H_x, H_y. The test hypotheses are:

Vadim Y. Bichutskiy is a Ph.D. student in the Department of Statistics. This work was completed when he was a M.S. student in the Department of Statistics and Biostatistics at California State University, East Bay (Hayward). Email him at: vbichuts@masonlive.gmu.edu.

$$H_0 : \eta_x - \eta_y = \Delta$$
$$vs.$$
$$H_1 : \eta_x - \eta_y \neq \Delta,$$

where Δ is a specified difference in medians, and is often 0.

For sufficiently large n and m:

$$H_x \dot{\sim} N\left(\eta_x \ , \ 1/4nf_x^2(\eta_x)\right),$$

$$H_y \dot{\sim} N\left(\eta_y \ , \ 1/4mf_y^2(\eta_y)\right),$$

$$H_x - H_y \dot{\sim} N\left(\eta_x - \eta_y \ , \ \frac{1}{4}\left\{\frac{1}{nf_x^2(\eta_x)} + \frac{1}{mf_y^2(\eta_y)}\right\}\right),$$

$$\frac{H_x - H_y - (\eta_x - \eta_y)}{\frac{1}{2}\sqrt{\frac{1}{nf_x^2(\eta_x)} + \frac{1}{mf_y^2(\eta_y)}}} \dot{\sim} N(0,1).$$

Assuming the normal approximation holds when the standard error of the difference in medians is estimated, then under the null hypothesis, the V statistic is:

$$V = \frac{(H_x - H_y) - \Delta}{\frac{1}{2}\sqrt{\frac{1}{n\hat{f}_x^2(H_x)} + \frac{1}{m\hat{f}_y^2(H_y)}}} \dot{\sim} N(0,1)$$

where $\hat{f}_x(H_x)$ and $\hat{f}_y(H_y)$ are respective population density estimates at the median.

Further, if it is assumed that the two populations have the same shape, possibly with a difference in location, then $f_x(\eta_x) = f_y(\eta_y)$, and the density estimates can be pooled to obtain a pooled test statistic:

$$V_p = \frac{(H_x - H_y) - \Delta}{\frac{1}{2\hat{f}_p(H)}\sqrt{\frac{1}{n} + \frac{1}{m}}} \dot{\sim} N(0,1) \qquad (1)$$

where

$$\hat{f}_p(H) = \sqrt{\frac{n\hat{f}_x^2(H_x) + m\hat{f}_y^2(H_y)}{n+m}}$$

is the pooled estimate of the population density at the median.

Simulations

The software R was used to simulate the power of the pooled test statistic (1). Two cases were considered: (i) population shapes are assumed to be known, and (ii) population shapes are unknown. The assumption of known population shapes is analogous to the assumption of known population variances in the z-test for comparing the means of two normal populations since the variance determines the shape of the normal distribution. The goal was to see how the test would perform for samples of moderate size from symmetric heavy-tailed populations. Parent populations investigated were Cauchy, Laplace and Student's t distributions with 2 and 3 degrees of freedom. In all settings, the parent populations were of the same shape, shifted under the alternative, and a two-sided test H_0: $\eta_x = \eta_y$ versus H_1: $\eta_x \neq \eta_y$ was performed.

Adaptive Kernel Density Estimation

When population shapes are unknown, $f_x(\eta_x)$ and $f_y(\eta_y)$ are estimated with $\hat{f}_x(H_x)$ and $\hat{f}_y(H_y)$, respectively, using adaptive kernel density estimation (AKDE).

Let $X_1, X_2, \ldots, X_n \in \mathbb{R}^d$ be a sample from unknown density f. The AKDE is a three step procedure:

1. Find a pilot estimate $\tilde{f}(X)$ that satisfies $\tilde{f}(X_i) > 0$, $i = 1, 2, \ldots, n$.
2. Define local bandwidth factors $\lambda_i = \{\tilde{f}(X_i)/g\}^{-\gamma}$ where g is the geometric mean of the $\tilde{f}(X_i)$ and $0 \leq \gamma \leq 1$ is the sensitivity parameter.
3. The adaptive kernel estimate is defined by

$$\hat{f}(X) = n^{-1}\sum_{i=1}^{n} h^{-d}\lambda_i^{-d}K\{h^{-1}\lambda_i^{-1}(X - X_i)\}$$

where $K(.)$ is a kernel function and h is the bandwidth.

The AKDE method varies the bandwidth among data points and is better suited for heavy-tailed populations than ordinary KDE (Silverman, 1998, pp. 100-110). Intuitively, the AKDE is based on the idea that for heavy-tailed populations a larger bandwidth is needed for data points in the tails of the distribution (i.e., for outliers). In R, function AKJ in library QUANTREG implements AKDE. Obtaining the pilot estimate requires the use of another density estimation method, such as ordinary KDE. The general view in the literature is that AKDE is fairly robust to the method used for the pilot estimate (Silverman, 1998) and that the choice of the sensitivity parameter γ is more critical. When using AKDE with Gaussian kernel, if the parent population has tails close to normal then $\gamma < .5$ should be used, however, if the parent population is heavy-tailed then $\gamma > .5$ should be used. Thus, $\gamma = .5$ is a good choice and has been shown to reduce bias (Abramson, 1982).

Results

Case 1: Known Population Shapes

Figure 1 shows the power curves for the pooled test when population shapes are assumed to be known at the 5% level of significance. Each point on the curves is based on 10,000 simulated samples. The Type I error rate is controlled very well.

Case 2: Unknown Population Shapes

Figure 2 shows the power curves for the pooled test when population shapes are unknown at the 5% level of significance and using AKDE with $\gamma = .5$. Each point on the curves is based on 10,000 simulated samples. The Type I error rate is controlled very well.

Comparisons with Other Methods

The test was compared to the following methods: (i) Student's t-test; (ii) Mann-Whitney-Wilcoxon (MWW) rank sum test; (iii) bootstrap (Efron & Tibshirani, 1993. p. 221); and (iv) permutation test. Figure 3 shows the receiver operating characteristic (ROC) curves for a balanced design with n = m = 30. The parent populations were of the same shape in each case and the difference in population medians was set to 1. For the bootstrap and the permutation test, the difference in medians was used as the metric. Each point on the curves is based on 10,000 simulated samples.

Conclusion

Tests for comparing medians tend to be very conservative. The proposed test is able to control the probability of Type I error. It is as powerful as the permutation test and the bootstrap and is more powerful than the MWW test for heavy-tailed populations. The more heavy-tailed the parent population, the greater the power advantage of the proposed test over the MWW test; when the parent population is light-tailed, the MWW test is more powerful than the proposed test.

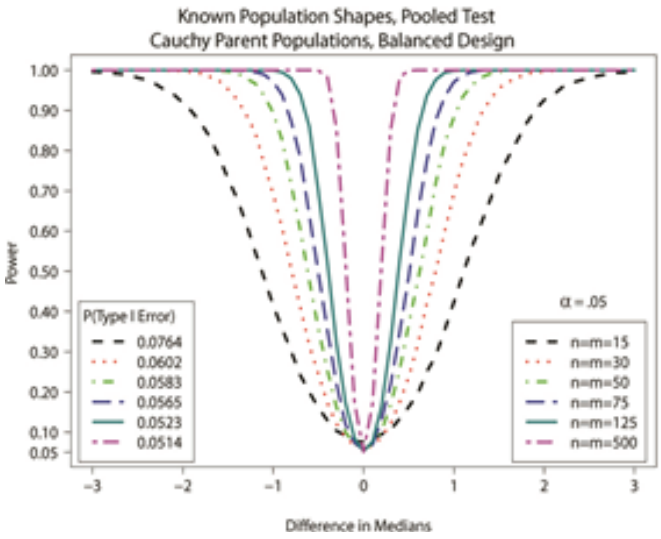
A key precept of the method is that AKDE provides a better estimate of the population density at the median, especially for heavy-tailed populations, than ordinary KDE. As expected, using ordinary KDE makes the test very conservative where the Type I error rate can be as low as 0.02 at the 5% significance level.

These experiments show that the sensitivity parameter γ in AKDE controls the trade-off between making the test conservative and liberal, with a suitable value of γ producing a test with a nominal significance level. The Type I error rate of the test can be increased (decreased) by increasing (decreasing) γ.

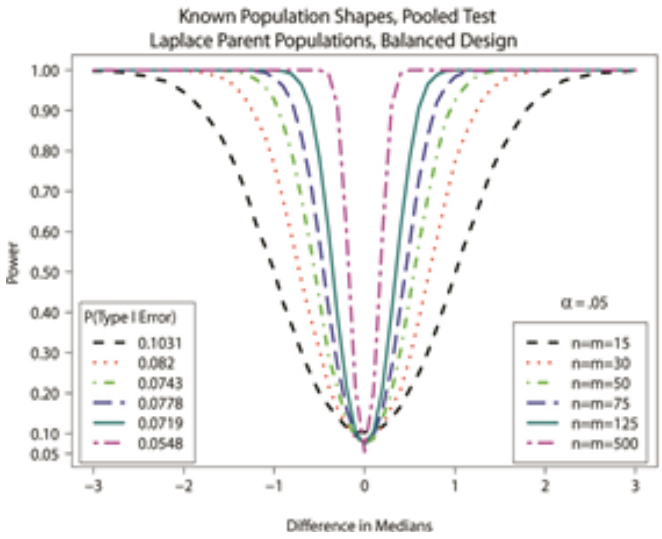
The asymptotic distribution of the sample median has been known for over 50 years (Chu, 1955; Chu & Hotelling, 1955), but it is only now with the improvement in computing power that this theory can be practically employed to derive useful statistical methodology, illustrating the interplay between theory, methodology and computation in the 21st century.

Figure 1: Power Curves for Known Population Shapes (10,000 Simulated Samples)

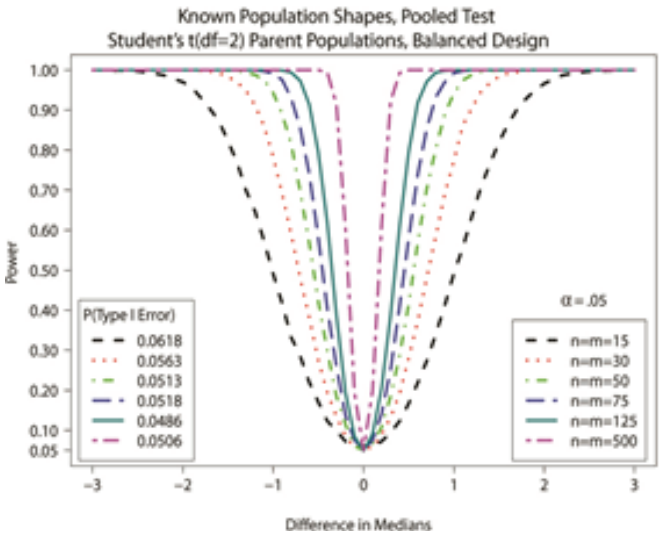
Cauchy



Laplace



Student's t (df = 2)



Student's t (df = 3)

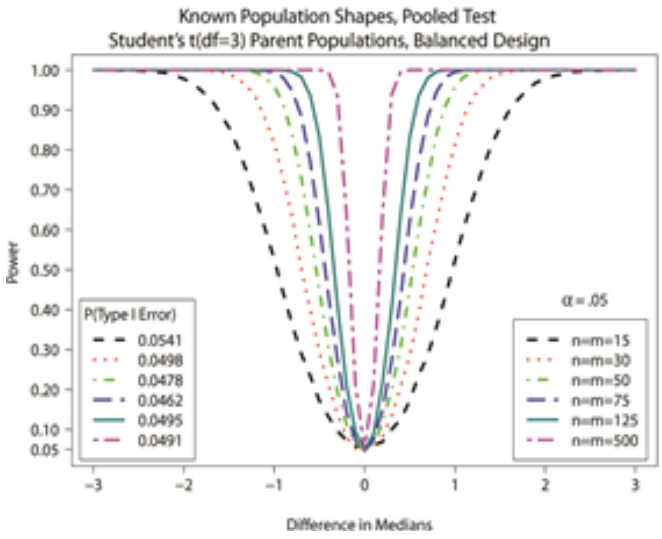
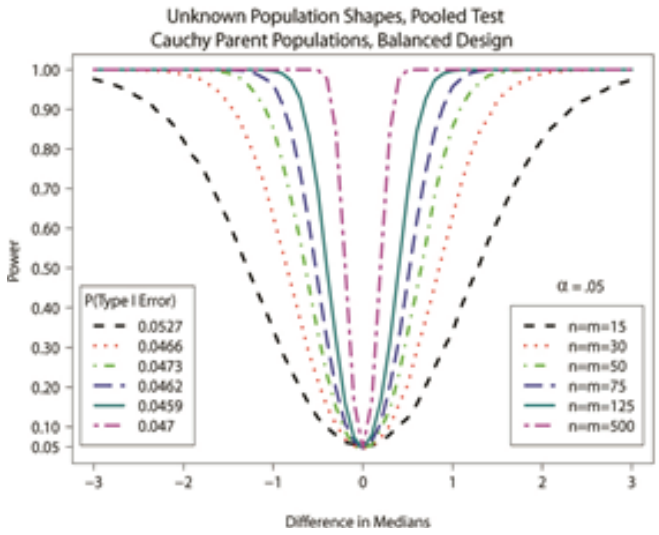
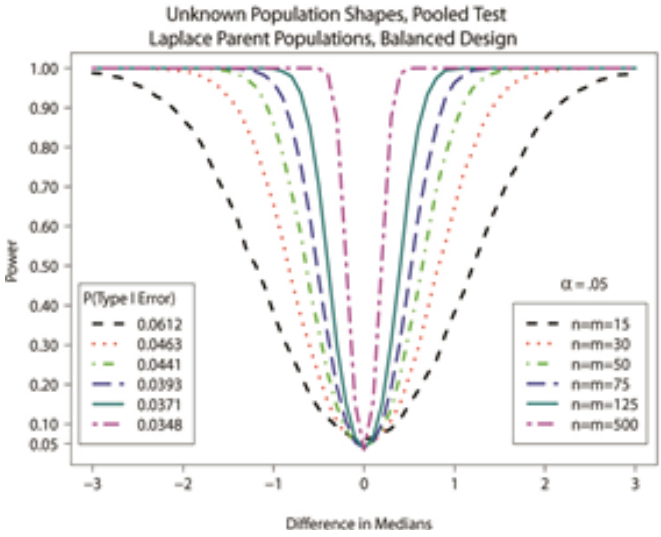


Figure 2: Power Curves for Unknown Population Shapes
(10,000 Simulated Samples, AKDE with $\gamma = .5$)

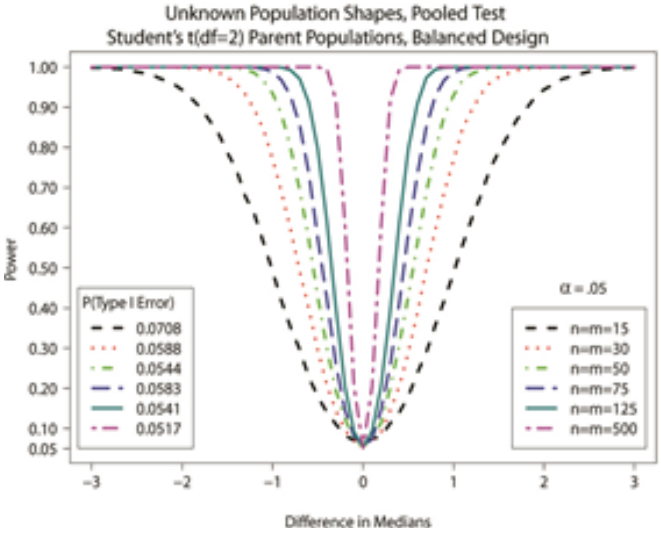
Cauchy



Laplace



Student's t (df = 2)



Student's t (df = 3)

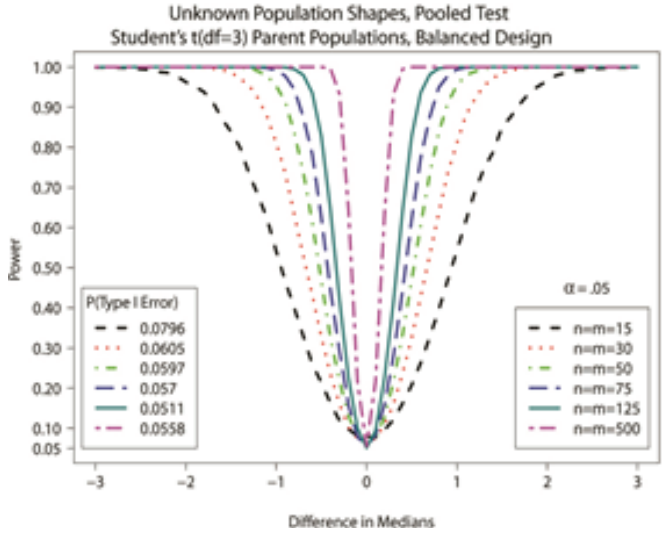
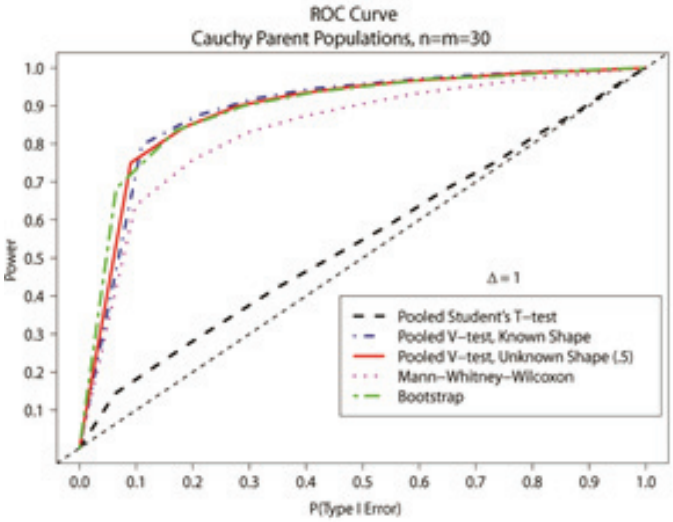


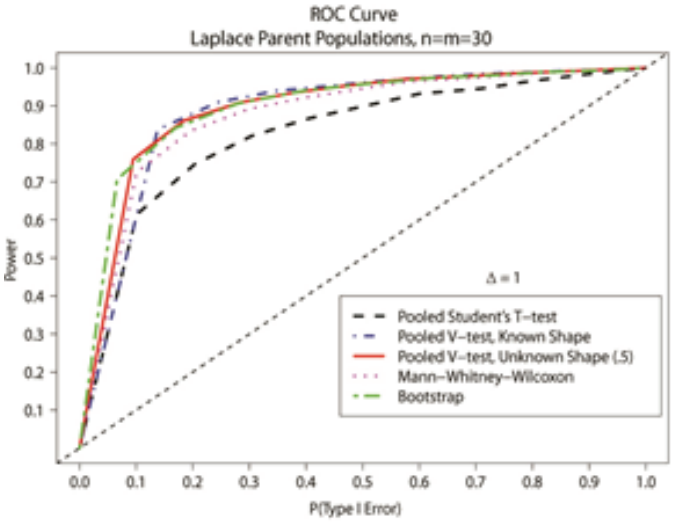
Figure 3: ROC Curves. Balanced Design with n = m = 30 (10,000 Simulated Samples)

(The curves for the permutation test coincide closely with the curves for the proposed test and have been omitted for clarity.)

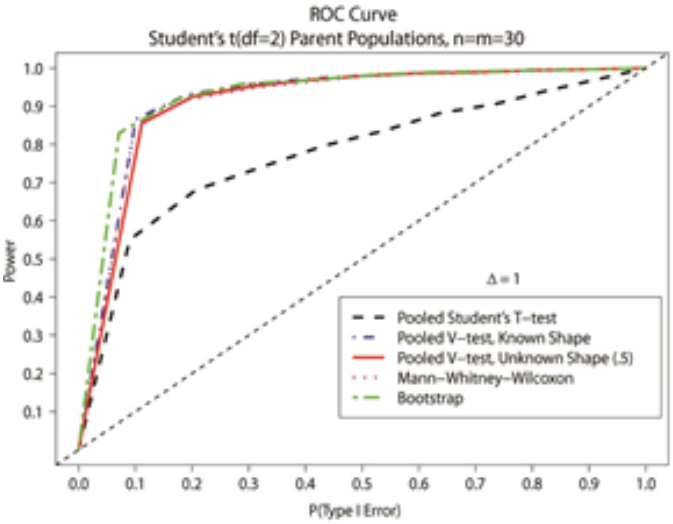
Cauchy



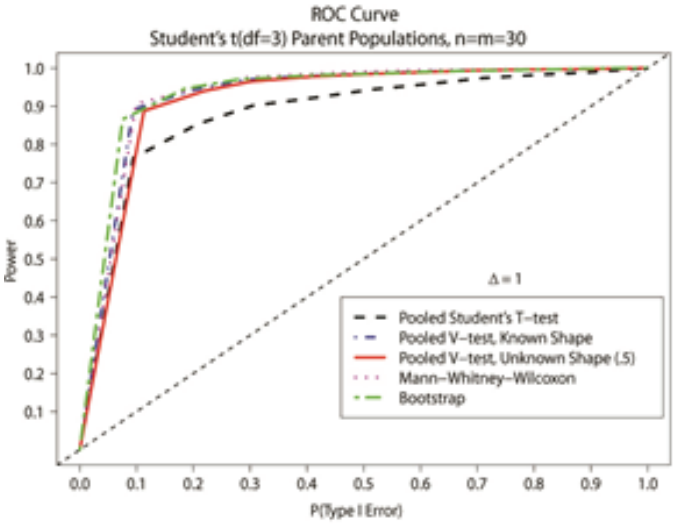
Laplace



Student's t (df = 2)



Student's t (df = 3)



Acknowledgements

The author thanks Professor Emeritus Bruce E. Trumbo, Professor Eric A. Suess and Professor Joshua D. Kerr at California State University, East Bay (Hayward), for helpful discussions and suggestions. The journal staff improved the prose. Earlier versions of this work were contributed at Joint Statistical Meetings (Bichutskiy, Kerr & Trumbo, 2009; Bichutskiy, et al., 2010).

References

Abramson, I. S. (1982). On bandwidth variation in kernel estimates: A square root law. *The Annals of Statistics*, *10*, 1217-1223.

Bichutskiy, V. Y., Kerr, J., & Trumbo, B. E. (2009). Classroom simulation: Investigation of the asymptotic distribution of the sample median. In *JSM Proceedings*, Statistical Education Section, Alexandria, VA: American Statistical Association, 3715-3728.

Bichutskiy, V. Y., Kerr, J. D., Suess, E. A., & Trumbo, B. E. (2010). Classroom derivation and simulation: An asymptotic two-sample test for comparing population medians. In *JSM Proceedings*, Statistical Education Section, Alexandria, VA: American Statistical Association, 4531-4545.

Chu, J. T. (1955). On the distribution of the sample median. *The Annals of Mathematical Statistics*, *26*, 112-116.

Chu, J. T., & Hotelling, H. (1955). The moments of the sample median. *The Annals of Mathematical Statistics*, *26*, 593–606.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, *7*(*1*), 1-26.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. Boca Raton, FL: Chapman & Hall/CRC.

Koenker, R. (2009). *quantreg: Quantile regression*, R package version 4.44. http://CRAN.R-project.org/package=quantreg.

Maritz, J. S., & Jarrett, R. G. (1978). A note on estimating the variance of the sample median. *Journal of the American Statistical Association*, *73*, 194-196.

McKean, J. W., & Schrader, R. M. (1984). A comparison of methods for Studentizing the sample median. *Communications in Statistics – Simulation and Computation*, *13*, 751-773.

Price, R. M., & Bonett, D. G. (2001). Estimating the variance of the sample median. *Journal of Statistical Computation and Simulation*, *68*, 295-305.

R Development Core Team. (2009). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Sheather, S. J., & Maritz, J. S. (1983). An estimate of the asymptotic standard error of the sample median. *Australian Journal of Statistics*, *25(1)*, 109-122.

Sheather, S. J. (1986). A finite sample estimate of the variance of the sample median. *Statistics and Probability Letters*, *4*, 337-342.

Silverman, B. W. (1998). *Density estimation for statistics and data Analysis*. Boca Raton, FL: Chapman & Hall/CRC.

Wilcox, R. R., & Charlin, V. L. (1986). Comparing medians: A Monte Carlo study. *Journal of Educational Statistics*, *11*(*4*), 263-274.

Wilcox, R. R. (2005). Comparing medians: An overview plus new results on dealing with heavy-tailed distributions. *The Journal of Experimental Education*, *73*(*3*), 249-263.

Wilcox, R. R. (2006). Comparing medians. *Computational Statistics & Data Analysis*, *51*, 1934-1943.

Wilcox, R. R. (2010). *Fundamentals of modern statistical methods: Substantially improving power and accuracy, 2nd Edition*. New York, NY: Springer.

Journal of Modern Applied Statistical Methods
November 2011, Vol. 10, No. 2, 699-709

1538 – 9472/11/$95.00

Tests for Correlation on Bivariate Non-Normal Data

L. Beversdorf
North Carolina State University,
Raleigh, NC

Ping Sa
University of North Florida,
Jacksonville, FL

Two statistics are considered to test the population correlation for non-normally distributed bivariate data. A simulation study shows that both statistics control type I error rates well for left-tailed tests and have reasonable power performance.

Key words: Saddlepoint approximation, Fisher's transformation, tests for correlation, bivariate non-normal distribution.

Introduction

Bivariate data are data in which two variables are measured on an individual. If the variables are quantitative, a researcher may be interested in describing the relationship between them. One measure used to describe the strength of linear relation between two quantitative variables is the linear correlation coefficient, denoted by ρ.

The true relationship between two variables of interest is always unknown. Different estimators have been proposed for ρ and two of them are used frequently: (1) the Spearman Rank Order Correlation, which is used for ordinal data, and (2) the Pearson Product Moment Correlation, which is applied to interval and ratio data. The maximum likelihood estimator of ρ is the Pearson product-moment correlation coefficient. When the data is not bivariate normal and the sample size exceeds 10, the nonparametric Spearman rank correlation is useful. Little work has been done for cases when the distribution of the data is unknown and the sample size is relatively small.

The most popular ρ estimator is the Pearson Product Moment Correlation Coefficient, *r*, which is a biased point estimator for ρ, however, the bias is small when *n* (sample size) is large. Given two variables Y_1 and Y_2, the statistic is:

$$r = \frac{\sum_{i=1}^{n}\left(Y_{i1}-\bar{Y}_1\right)\left(Y_{i2}-\bar{Y}_2\right)}{\left[\sum_{i=1}^{n}\left(Y_{i1}-\bar{Y}_1\right)^2\left(Y_{i2}-\bar{Y}_2\right)^2\right]^{1/2}},$$

where (Y_{i1}, Y_{i2}) is the i^{th} observation of the bivariate data $(Y_{11}, Y_{12}),\ldots,(Y_{n1}, Y_{n2})$, $\bar{Y}_1$ is the sample mean of Y_1 and $\bar{Y}_2$ is the sample mean of Y_2.

Researchers have done intensive work on the distribution of *r* when the population is bivariate normal (Fisher, 1915; Stuart & Ord, 1994). It has been found that, when $n = 2$, the distribution of *r* can be regarded as an extreme case of a U-shaped distribution, for $n = 3$ the density is still U-shaped, but if $n = 4$ the distribution is uniform when $\rho = 0$ and J-shaped otherwise. For $n > 4$ the density function is unimodal and has increased skew as $|\rho|$ increases, this follows from the fact that the

Ping Sa is a Professor of the Department of Mathematics and Statistics. She received her Ph.D. in Statistics from the University of South Carolina in 1990. She has published more than 20 papers. Her recent scholarly activities have involved research in multiple comparisons, quality control and statistical inference for non-normal data. Email: psa@unf.edu. Ms. Beversdorf received her Master's degree in Statistics from the University of North Florida in 2008. She is currently a Ph. D. candidate at North Carolina State University. Email: louanneb@gmail.com.

mode moves with ρ and *r*. For any ρ, the distribution of *r* slowly tends to normality as $n \rightarrow \infty$ (Stuart & Ord, 1994).

When the population is bivariate normal and has equal variance, a test statistic

$$t_r^* = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \qquad (1)$$

can be derived to test H_0: ρ = 0. Under H_0, t_r^* follows the Student's *t*-distribution with (*n*–2) degrees of freedom, denoted $t_{(n-2)}$. Disadvantages of this test include the need for a relatively large sample or bivariate normal data and the ability to test only for ρ = 0.

When the population is not bivariate normal and the sample size exceeds 10, a non-parametric statistic, the Spearman Rank Correlation Coefficient (Spearman), is typically used to measure the association between two variables when no transformation for the data can be found to approximate a bivariate normal distribution. Spearman, denoted by r_s, is then defined as the ordinary Pearson product-moment correlation coefficient based on data ranking:

$$r_s = \frac{\sum\left(R_{i1}-\overline{R}_1\right)\left(R_{i2}-\overline{R}_2\right)}{\left[\sum\left(R_{i1}-\overline{R}_1\right)^2\left(R_{i2}-\overline{R}_2\right)^2\right]^{1/2}},$$

where (R_{i1}, R_{i2}) are the ranks of (Y_{i1}, Y_{i2}) respectively; and $\overline{R}_1$ is the mean of the ranks of R_{i1}, i = 1,2,… n, and $\overline{R}_2$ is the mean of the ranks of R_{i2}, i = 1,2,… n.

Spearman can also be used to test the association between the two variables with the null hypothesis, H_0, stating: there is no association between Y_1 and Y_2. When sample size *n,* exceeds 10, the test statistic:

$$t_{rs}^* = \frac{r_s\sqrt{n-2}}{\sqrt{1-r_s^2}}, \qquad (2)$$

can be used. t_{rs}^* is approximately a *t*-distribution with *n*–2 degrees of freedom under H_0. This is a nonparametric test and thus may result in lower power performance. Again this test can only be used for testing whether an association exists. The purpose of this study is to test H_0: $\rho = \rho_0$, where ρ_0 can be values other than zero, for bivariate non-normal data. Fisher's Z-transformation and a saddlepoint transformation are investigated and tested.

Methodology

Two statistics for testing the correlation coefficient of bivariate non-normal populations are investigated: (1) Fisher's z-transformation, denoted r_F, and (2) the saddlepoint approximation, denoted r_L. These methods are used on bivariate non-normal data sets with small sample sizes. The goal is to determine if either of the two methods is appropriate for hypothesis testing about the population correlation coefficient, specifically for bivariate non-normal data sets with a small sample size.

Fisher's Z-Transformation

The sampling distribution of *r* is complicated when ρ ≠ 0 even when the population is bivariate normal. Fisher (1921) derived an approximation procedure based on a transformation of *r*, $z' = \frac{1}{2}\log\frac{1+r}{1-r}$ and it tends to normality much faster than *r*. After standardizing, the statistic for Fisher's classical transformation is given by:

$$r_F = \left(\frac{1}{2}\log\frac{1+r}{1-r} - \frac{1}{2}\log\frac{1+\rho}{1-\rho} - \frac{\rho}{2(n-1)}\right)\sqrt{n-3}. \qquad (3)$$

Saddlepoint Approximation

Saddlepoint approximations were introduced by Daniels (1954). However, computations of these approximations only recently became feasible with the availability of inexpensive computing power. In practice, statistical inference often involves test statistics with normal distributions, which are valid as sample sizes increase. For small sample size problems, these distributions tend to provide inaccurate results. Saddlepoint methods offer approximations that are accurate to a higher order than first-order approximations and their

accuracy holds for extremely small sample sizes (Huzurbazar, 1999). Saddlepoint approximations also provide good estimates to very small tail probabilities or to the density in the tails of the distributions.

Jensen (1995) transforms the Pearson correlation coefficient using Laplace transformations to derive a function of *r* that can be normalized and he claims that r_L is normally distributed to a high accuracy. Assuming a bivariate normal data set with correlation ρ, the saddlepoint approximation, denoted r_L, provided by Jensen (1995) is:

$$r_L = v + \frac{1}{v}\log\frac{u}{v}, \tag{4}$$

where

$$v = \operatorname{sgn}(r-\rho)\left\{2(n-4)\log\left(\frac{1-\rho r}{\sqrt{1-\rho^2}\sqrt{1-r^2}}\right)\right\}^{\frac{1}{2}},$$

$$u = \sqrt{n-4}\left(\frac{1-\rho r}{1-\rho^2}\right)^{\frac{3}{2}}\frac{r-\rho}{1-r^2},$$

and sgn(.) is the sign of $(r-\rho)$.

Proposed Test

A new test is required to investigate the hypothesis $H_0: \rho = \rho_0$ versus three possible alternative hypotheses, $H_a: \rho \neq \rho_0$, $H_a: \rho > \rho_0$ and $H_a: \rho < \rho_0$, when a data set is bivariate non-normal and sample size is relatively small to moderate. Although both the Fisher and saddlepoint transformations are derived for bivariate normal data, little work has been done to investigate if they can also be used for non-normal bivariate data; thus, the two approximations, r_F in (3) and r_L in (4), are used as the test statistics for the hypothesis $H_0: \rho = \rho_0$. Note that ρ_0 should be used in both equations whenever ρ is present. The decision rule to reject the null hypothesis for the two-tailed, upper-tailed and lower-tailed tests is $|r_F| > z_{\alpha/2}$ or $|r_L| > z_{\alpha/2}$, $r_F, r_L > z_\alpha$, and $r_F, r_L < -z_\alpha$, respectively.

Simulation Study: Generating Bivariate Non-Normal Data

Fleishman (1978) derived a method for generating univariate non-normal random variables. Fleishman's method is based on the variable Y defined as

$$Y = a + bZ + cZ^2 + dZ^3 \tag{5}$$

where Z is a standard normal random variable, and *a*, *b*, *c* and *d* are constants chosen in such a way that Y has the desired coefficients of skewness and kurtosis, γ_1 and γ_2, respectively.

Fleishman showed that $a = -c$ and the constants *b*, *c* and *d* are determined by simultaneously solving the following three equations:

$$\begin{aligned} &b^2 + 6bd + 2c^2 + 15d^2 - 1 = 0 \\ &2c\left(b^2 + 24bd + 105d^2 + 2\right) - \gamma_1 = 0 \\ &24\left\{\begin{matrix} bd + c^2\left(1 + b^2 + 28bd\right) \\ +d^2\left(12 + 48bd + 141c^2 + 225d^2\right)\end{matrix}\right\} - \gamma_2 = 0 \end{aligned} \tag{6}$$

Using these equations, a non-normal random variable Y can be obtained by generating a standard normal variable Z and using the equation (5).

Vale and Maurelli (1983) proposed generating multivariate non-normal random variables with a specified correlation structure based on Fleishman's method. For bivariate non-normal random data, (Y_1, Y_2) with desired coefficients of skewness and kurtosis, (γ_{11} and γ_{21}) for Y_1 and (γ_{12} and γ_{22}) for Y_2, solutions to the system of equations (6) given in Fleishman's method must be found. Let Z_1, Z_2 be two standard normal correlated variables. Y_1 and Y_2 can be calculated with the following equations:

$$\begin{aligned} Y_1 &= a_1 + b_1 Z_1 + c_1 Z_1^2 + d_1 Z_1^3, \\ Y_2 &= a_2 + b_2 Z_2 + c_2 Z_2^2 + d_2 Z_2^3 \end{aligned} \tag{7}$$

Vale and Maurelli (1983) found that the correlation coefficient between Y_1 and Y_2 is:

$$\rho_{Y_1,Y_2} = \rho_{Z_1,Z_2}\left(b_1b_2 + 3b_1d_2 + 3b_2d_1 + 9d_1d_2\right) + \rho^2_{Z_1,Z_2}\,2c_1c_2 + \rho^3_{Z_1,Z_2}\,6d_1d_2 \quad (8)$$

For a desired correlation, ρ_{y_1,y_2}, the intermediate correlation, ρ_{Z_1,Z_2}, can be determined by solving the above cubic equation. The bivariate non-normal random variate (Y_1, Y_2) can then be obtained by first generating a set of bivariate standard normal random variate with correlation ρ_{Z_1,Z_2}, and then using equation (7).

Simulation Description

Different values of skewness and kurtosis were chosen for the simulation study in order to reflect different population distributions. Four values of skewness, −3, −1, 1, 3 and three values of kurtosis, 3, 7, 25 were used, resulting in 78 possible pairs of populations. A relatively small sample size of 10 and a moderate sample size of 20 were used in the study and the test statistics r_L and r_F were investigated for type I error rates with left-tailed, right-tailed and two-tailed tests with the nominal levels of 0.01 and 0.05 for each sample.

Comparisons in the simulation study use r_L and r_F against three critical values, z_α, $t_{(n-2,\ \alpha),}$ and $(z_\alpha+t_{(n-2,\ \alpha)})/2$, to draw conclusions. Four ρ_0 values 0, 0.5, 0.7 and 0.9 were evaluated as the hypothesized values for H_0: $\rho = \rho_0$. When $\rho_0 = 0$, the t_r^* in (1) and t_{rs}^* in (2) are also included in the study for comparison purposes. The simulation study has two parts: the type I error rate comparisons and the power study. The steps of the simulation are:

Data Generation: Steps (1) – (5)

1) Input the five population parameters: skewness and kurtosis for each of the two populations and the desired population correlation;
2) Solve the system of equations (6) to calculate coefficients *a*, *b*, *c* and *d* for the two populations;
3) Solve $\rho_{z_1z_2}$ by equation (8);
4) Generate n bivariate standard normal variables (Z_1, Z_2) with correlation $\rho_{z_1z_2}$;
5) Apply the transformation in (7) to obtain the non-normal sample data Y_1 and Y_2;

Evaluation: Steps (6) – (8)

6) Evaluate r_L and r_F and compare to critical values z_α, $t_{(n-2,\ \alpha),}$ and $(z_\alpha+t_{(n-2,\ \alpha)})/2$; if $\rho_0 = 0$, t_r^* and t_{rs}^* are evaluated and compared to $t_{(n-2)}$ critical value;
7) Repeat steps (4) – (6) 99,999 times;
8) Calculate type I error rate for each method by finding the proportion of rejection in the 100,000 samples.

In the power study, an extra parameter ρ_a (which is different from ρ_0) is input in step (1) and used to generate the data as the true population correlation, however, all test statistics in step (6) are evaluated under ρ_0. All other steps in the power study are identical to the type I error rate study. All the simulations were run with Fortran 77 for Windows on a Toshiba Satellite-A105 Laptop Computer.

Results

Type I Error Rate Comparison

Tables 1-4 provide comparisons of type I error rates with sample size $n = 10$. The set of population parameters for skewness and kurtosis are in the first column with the first population's parameters in the first row and the second in the second row. Comparisons were made between the tests for saddlepoint and Fisher's transformation, given in the table as the two adjacent numbers within a given correlation column, r_L and r_F, respectively. Three critical

points $t_{n-2,\alpha}$, $\frac{z_\alpha + t_{n-2,\alpha}}{2}$ and z_α were used for the two proposed methods. The results are the first, second and third numbers in the respective column. Pearson and Spearman are evaluated with a critical value $t_{n-2,\alpha}$ for ρ = 0 only, and the type I error rates are reported in the first column with Pearson first and Spearman underneath.

Due to similar results in the study, only 12 pairs of populations and the small sample size n = 10 are reported in the tables. Also, although all the tests are done with levels of significance 0.05 and 0.01, both levels are reported here only for the left-tailed tests. (For complete simulation results, please contact the first author.)

Left-Tailed Type I Error Rates

Left-tailed type I error rates are given in Tables 1 and 2. Table 1 uses a significance level of 0.05 and Table 2 uses a significance level of 0.01. It can be observed that only slight differences in type I error rates are present between the results for the saddlepoint and Fisher's transformations. This same result was observed throughout the simulation study.

Results using the *t* critical value achieve very good type I error rates for all of the distributions. The *z* critical value results in a few slightly inflated type I error rates and only by the saddlepoint approximation. The worst case found in the study, produced by the saddlepoint approximation, is for the pair populations with the same (skewness, kurtosis) = (3, 25) under ρ = 0.9 using z_α as the critical point. The type I error rate for this case is 0.0688. However, after the critical point was changed to $\frac{z_\alpha + t_{n-1,\alpha}}{2}$, the type I error rate decreased to 0.0564 and it further decreased to 0.0458 when $t_{n-2,\alpha}$ is used. The Fisher's transformation, by contrast, controls the type I error rates properly for nearly all cases considered.

For the important case when ρ = 0, results show that both the r_L and r_F statistics control type I error rates using any of the three critical values at the 0.05 significance level. When the significance level is lowered to 0.01, some of the type I error rates using the *z* critical value are slightly inflated but within acceptable range. Surprisingly, Pearson controls the type I error rates better than the Spearman method. It performed very well for the 0.05 significance level; however, those involving a population with larger kurtosis are slightly inflated when the significance level is lowered to 0.01. Spearman has some slightly inflated type I error rates at both significance levels. Overall, it is fair to say that essentially all cases studied produced controlled type I error rates for the left-tailed test.

Right-Tailed & Two-Tailed Type I Error Rates

Right-tailed type I error rates are shown in Tables 3 with significance level of 0.05. (Although the 0.01 level of significance is also studied, the table is omitted due to the similar results.) With the right-tailed test, most type I error rates are inflated, the only values that stand out are for tests where the *t* critical values were used and both the skewness and kurtosis were relatively small. A great result is found for the *t* critical values when ρ = 0, type I error rates are controlled for both the r_L and r_F. As opposed to the left-tailed test, the Spearman *t*-test works better than the Pearson; however, results are still not as good as the corresponding results by r_L and r_F.

Overall, both Saddlepoint and Fisher's statistics are better candidates for testing $H_a: \rho > 0$. The *t* critical value produces more stable results then the *z* critical value, although the two statistics can also be used for other ρ_0 values if the populations have small kurtosis with *t* critical points, in general the two statistics are not recommended for a right-tailed test.

Two-tailed type I error rates are shown in Table 4. As expected, the results of the two-tailed tests are more controlled than that of the right-tailed test. However, because the methods essentially failed for the right-tailed tests, they are not recommended to be used to perform a two-tailed test.

Power Results

Table 5 summarizes the results of the power study for left-tailed tests with $H_o: \rho = 0.7$ versus various ρ_a values such that

$\rho_a < 0.7$. Five different ρ_a values and two levels of significance were investigated, but only three ρ_a and $\alpha = 0.05$ results are reported here. Power results for both methods show reasonable rate of convergence to probability 1. As expected, the *z* critical values have higher power than the other two tests. (For complete simulation results, please contact the first author.)

Conclusion

This study proposed and examined two statistics, the saddlepoint transformation, r_L, and Fisher's transformation, r_F, for testing a correlation which may or may not be zero for any bivariate non-normal population. The simulation study indicates that the two statistics perform similarly. They both have very good robust performance for all the distributions studied when testing a left-tailed test; they maintain the type I error rates close to the nominal level and show reasonably good power.

The two statistics are not recommended for testing a right-tailed test or a two-tailed test unless the practitioner knows for certain that the populations have both small skewness and kurtosis. In these cases, the two test statistics with a *t* critical point can properly control the type I error rates.

The two statistics can also be used for testing H_o: $\rho = 0$ versus any of the three possible alternative hypotheses. They control type I error rates better than the existing Pearson and Spearman *t*-tests. Because the two statistics are derived based on bivariate normal population, a sample size of at least 10 is recommended.

References

Daniels, H. E. (1954). Saddlepoint approximations in statistics. *Annals of Mathematical Statistics*, *25*, 631-650.

Fisher, R. A. (1915). Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika*, *10*(*4*), 507-521.

Fisher, R. A. (1921). On the probable error of a coefficient of correlation deduced from a small sample. *Metron*, *1*(*4*), 1.

Fleishman, A. (1978). A method for simulating non-normal distribution. *Psychometrika*, *43*, 521-532.

Huzurbazar, S. (1999). Practical saddlepoint approximation. *The American Statistician*, *53*(*3*), 225-232.

Jensen, J. (1995). *Saddlepoint approximation*. New York: Oxford University Press, Inc.

Stuart, A., & Ord, J. K. (1994). *Kendall's advanced theory of Statistics*, *Vol. 1*, *6th Ed*. New York: Halsted Press.

Vale, C., & Maurelli, V. (1983). Simulating multivariate non-normal distributions. *Psychometrika*, *48*, 465-471.

Table 1: Type I Error Rates for Left-Tailed Test, 0.05 Level of Significance

Skewness	Kurtosis	RHO = 0			RHO = 0.5		RHO = 0.7		RHO = 0.9	
		Pearson Spearman	r_L	r_F	r_L	r_F	r_L	r_F	r_L	r_F
3	25	0.0416	0.0284	0.0281	0.021	0.0179	0.0323	0.0263	0.0458	0.0356
3	25	0.0522	0.0348	0.0339	0.029	0.0248	0.0421	0.0344	0.0564	0.0447
			0.0424	0.0408	0.0395	0.0334	0.0537	0.0444	0.0688	0.055
-3	25	0.0428	0.0292	0.0289	0.0218	0.0188	0.0325	0.0261	0.0445	0.0342
-3	25	0.0538	0.0357	0.0348	0.0298	0.0255	0.0429	0.0348	0.0553	0.0433
			0.0436	0.042	0.0397	0.0337	0.0549	0.0451	0.0674	0.054
-1	7	0.0475	0.0302	0.0298	0.0285	0.0256	0.0289	0.0241	0.0311	0.0235
-1	7	0.0516	0.0381	0.037	0.0368	0.0324	0.0369	0.0307	0.0395	0.0302
			0.0483	0.0463	0.0462	0.0409	0.0469	0.0387	0.0502	0.0383
1	7	0.0483	0.0316	0.0312	0.0283	0.0252	0.0292	0.0245	0.0316	0.0235
1	7	0.0521	0.0396	0.0385	0.036	0.0318	0.0374	0.0312	0.0403	0.0307
			0.0491	0.0473	0.0453	0.0399	0.0475	0.0392	0.0515	0.0392
1	3	0.0463	0.0286	0.0281	0.0309	0.0277	0.0321	0.0268	0.0354	0.0276
1	3	0.0514	0.0374	0.0362	0.0398	0.0351	0.0411	0.034	0.0448	0.0345
			0.0473	0.0454	0.0501	0.0444	0.0527	0.0434	0.0555	0.0434
-1	3	0.0461	0.0286	0.028	0.0313	0.0278	0.0326	0.0271	0.0349	0.0264
-1	3	0.0517	0.0371	0.0358	0.0399	0.0355	0.0411	0.0346	0.0438	0.0338
			0.047	0.0449	0.0505	0.0444	0.0517	0.0433	0.055	0.0426
-3	25	0.0441	0.0278	0.0272	0.0228	0.0198	0.0189	0.0153	0.002	0.0013
-1	3	0.0519	0.0357	0.0346	0.0304	0.0264	0.025	0.0203	0.0029	0.0019
			0.045	0.0431	0.0392	0.0343	0.0328	0.0266	0.0043	0.0028
3	25	0.0446	0.0284	0.0279	0.0226	0.0201	0.0199	0.0163	0.0019	0.0012
1	3	0.0512	0.0359	0.0349	0.0293	0.0259	0.026	0.0212	0.0027	0.0018
			0.0455	0.0436	0.0382	0.0329	0.0341	0.0274	0.0043	0.0026
3	25	0.0464	0.0307	0.0302	0.0221	0.0194	0.0217	0.0175	0.0112	0.0079
1	7	0.0521	0.0383	0.0371	0.029	0.0253	0.0288	0.023	0.0151	0.0109
			0.0472	0.0454	0.0375	0.0325	0.0377	0.0306	0.0207	0.0146
-3	25	0.0474	0.0317	0.0312	0.0213	0.0187	0.0219	0.0178	0.0115	0.0081
-1	7	0.0527	0.0389	0.0379	0.028	0.0244	0.0288	0.0235	0.0158	0.011
			0.0482	0.0464	0.0366	0.0318	0.0375	0.0303	0.0212	0.0151
-1	3	0.0482	0.0311	0.0304	0.0293	0.0263	0.0296	0.0244	0.0254	0.019
-1	7	0.0521	0.0389	0.0379	0.0374	0.0331	0.0373	0.0313	0.0333	0.0247
			0.0489	0.0473	0.0471	0.0415	0.0469	0.0392	0.0422	0.0321
1	3	0.0473	0.0301	0.0294	0.0299	0.0267	0.0298	0.025	0.0258	0.0194
1	7	0.0522	0.038	0.0366	0.038	0.0338	0.0379	0.0315	0.0331	0.0251
			0.0481	0.0461	0.0482	0.0424	0.0475	0.0396	0.0419	0.0321

Table 2: Type I Error Rates for Left-Tailed Test, 0.01 Level of Significance

		RHO = 0			RHO = 0.5		RHO = 0.7		RHO = 0.9	
Skewness	Kurtosis	Pearson Spearman	r_L	r_F	r_L	r_F	r_L	r_F	r_L	r_F
3	25	0.0113	0.0039	0.0049	0.0008	0.0011	0.0009	0.001	0.0021	0.0019
3	25	0.0118	0.0069	0.0079	0.0021	0.0023	0.0028	0.0027	0.0063	0.0049
			0.0115	0.0123	0.0049	0.0047	0.0077	0.0067	0.0142	0.011
-3	25	0.0114	0.0033	0.0043	0.0006	0.0008	0.001	0.001	0.0023	0.002
-3	25	0.0119	0.0063	0.0073	0.0019	0.002	0.0028	0.0026	0.0061	0.0047
			0.0117	0.0124	0.0051	0.0049	0.008	0.0068	0.0149	0.0113
-1	7	0.0105	0.0023	0.0033	0.0016	0.002	0.0016	0.0017	0.0017	0.0016
-1	7	0.0117	0.0052	0.0063	0.0037	0.004	0.0036	0.0035	0.004	0.0032
			0.0108	0.0118	0.009	0.0088	0.0086	0.0074	0.0097	0.0073
1	7	0.0105	0.0022	0.0032	0.0015	0.0019	0.0015	0.0017	0.0017	0.0015
1	7	0.0112	0.0051	0.0059	0.004	0.0042	0.0039	0.0038	0.0041	0.0033
			0.0108	0.0118	0.0086	0.0084	0.0088	0.0077	0.0091	0.0069
1	3	0.009	0.0017	0.0024	0.0019	0.0023	0.0018	0.002	0.002	0.0018
1	3	0.0116	0.0041	0.0048	0.0044	0.0047	0.0044	0.0043	0.0046	0.0038
			0.0094	0.0102	0.0099	0.0096	0.0102	0.0089	0.011	0.0083
-1	3	0.0084	0.0016	0.0023	0.0017	0.0021	0.0019	0.0021	0.0023	0.002
-1	3	0.0115	0.0039	0.0047	0.0042	0.0046	0.0046	0.0044	0.0055	0.0045
			0.0087	0.0097	0.0096	0.0094	0.0101	0.0089	0.0123	0.0097
-3	25	0.0087	0.0015	0.0024	0.0012	0.0014	0.0006	0.0007	5E-05	5E-05
-1	3	0.0123	0.0039	0.0049	0.0026	0.0029	0.0019	0.0018	0.0001	0.0001
			0.009	0.0097	0.0065	0.0063	0.0048	0.0041	0.0004	0.0003
3	25	0.0085	0.0019	0.0026	0.0009	0.0011	0.0007	0.0008	6E-05	5E-05
1	3	0.0121	0.0041	0.0049	0.0025	0.0025	0.0018	0.0017	0.0002	0.0002
			0.0088	0.0095	0.0056	0.0054	0.0045	0.004	0.0003	0.0002
3	25	0.0111	0.0028	0.0038	0.0009	0.0012	0.0008	0.0009	0.0003	0.0002
1	7	0.0114	0.0058	0.0068	0.0023	0.0025	0.0022	0.0021	0.0009	0.0006
			0.0116	0.0125	0.0059	0.0057	0.0057	0.005	0.0024	0.0018
-3	25	0.0109	0.0029	0.0036	0.0011	0.0014	0.0008	0.0009	0.0003	0.0003
-1	7	0.0114	0.0058	0.0068	0.0029	0.0031	0.0022	0.0021	0.0009	0.0007
			0.0112	0.012	0.0063	0.0061	0.0055	0.0048	0.0024	0.0017
-1	3	0.0096	0.002	0.0029	0.0017	0.002	0.0016	0.0018	0.0015	0.0013
-1	7	0.0119	0.0047	0.0055	0.0039	0.0042	0.0042	0.004	0.0036	0.0029
			0.0099	0.0107	0.0086	0.0084	0.0091	0.008	0.0077	0.006
1	3	0.0094	0.0019	0.0027	0.0017	0.0021	0.0016	0.0017	0.0011	0.001
1	7	0.0117	0.0043	0.0052	0.0038	0.0041	0.0039	0.0037	0.0026	0.0021
			0.0097	0.0105	0.009	0.0088	0.0089	0.0078	0.007	0.0054

Table 3: Type I Error Rates for Right-Tailed Test, 0.05 Level of Significance

		RHO = 0			RHO = 0.5		RHO = 0.7		RHO = 0.9	
Skewness	Kurtosis	Pearson Spearman	r_L	r_F	r_L	r_F	r_L	r_F	r_L	r_F
3	25	0.0635	0.0479	0.0474	0.1168	0.1171	0.1419	0.142	0.1671	0.1666
3	25	0.0511	0.0555	0.0544	0.1308	0.1303	0.1579	0.1571	0.1835	0.1822
			0.0642	0.0626	0.1458	0.1447	0.1746	0.1733	0.2006	0.1985
-3	25	0.0654	0.05	0.0494	0.1179	0.1181	0.1431	0.1432	0.1664	0.1661
-3	25	0.0524	0.0573	0.0564	0.132	0.1316	0.1578	0.1571	0.1826	0.1816
			0.0662	0.0645	0.1465	0.1454	0.1742	0.1729	0.1998	0.198
-1	7	0.0528	0.0362	0.0357	0.0508	0.051	0.0587	0.0588	0.0674	0.0672
-1	7	0.0532	0.0441	0.0429	0.0616	0.0612	0.0699	0.0694	0.0799	0.0787
			0.0538	0.0517	0.0737	0.0728	0.0828	0.0817	0.0936	0.092
1	7	0.0533	0.0356	0.0348	0.0511	0.0514	0.0598	0.0599	0.0683	0.0681
1	7	0.0512	0.0442	0.043	0.0614	0.0611	0.0707	0.0702	0.0804	0.0795
			0.0542	0.0523	0.0737	0.0728	0.0827	0.0816	0.0937	0.0922
1	3	0.0539	0.0353	0.0347	0.0431	0.0433	0.0461	0.0462	0.0497	0.0495
1	3	0.0532	0.0442	0.0428	0.0525	0.0522	0.0566	0.0563	0.0601	0.0591
			0.055	0.0528	0.064	0.0633	0.0685	0.0676	0.0725	0.0708
-1	3	0.0535	0.0357	0.035	0.0424	0.0424	0.0469	0.047	0.0495	0.0492
-1	3	0.0529	0.0443	0.0431	0.0525	0.0523	0.0569	0.0565	0.0604	0.0598
			0.0544	0.0524	0.0635	0.0625	0.0694	0.0684	0.0728	0.0714
-3	25	0.0569	0.0394	0.0389	0.0666	0.0668	0.0774	0.0775	0.0943	0.0937
-1	3	0.0526	0.0475	0.0465	0.0783	0.078	0.0915	0.0909	0.1151	0.1136
			0.0578	0.0559	0.0921	0.091	0.1078	0.1065	0.139	0.136
3	25	0.0582	0.0401	0.0396	0.0666	0.0668	0.0796	0.0797	0.0968	0.0964
1	3	0.0524	0.0494	0.048	0.0794	0.079	0.0931	0.0926	0.1168	0.115
			0.0591	0.0573	0.0937	0.0927	0.1096	0.1082	0.1412	0.1383
3	25	0.0576	0.0404	0.0399	0.0781	0.0784	0.0925	0.0926	0.1068	0.1063
1	7	0.0533	0.0486	0.0474	0.0906	0.0902	0.1061	0.1055	0.124	0.1226
			0.0585	0.0567	0.1044	0.1032	0.122	0.1207	0.1431	0.1412
-3	25	0.0585	0.0409	0.0403	0.0773	0.0776	0.0925	0.0926	0.1081	0.1078
-1	7	0.0532	0.0491	0.048	0.0897	0.0893	0.1065	0.1059	0.1256	0.1243
			0.0591	0.0575	0.1042	0.103	0.1234	0.122	0.1444	0.1423
-1	3	0.0523	0.0344	0.0337	0.0464	0.0467	0.051	0.0511	0.0561	0.0558
-1	7	0.0523	0.043	0.0418	0.0565	0.0562	0.0618	0.0613	0.0677	0.0667
			0.0533	0.051	0.0681	0.0674	0.0743	0.0731	0.0812	0.0797
1	3	0.0521	0.0349	0.0345	0.0466	0.0467	0.0507	0.0507	0.0562	0.0558
1	7	0.0516	0.0431	0.0419	0.0571	0.0568	0.0607	0.0602	0.0669	0.0661
			0.0529	0.051	0.0686	0.0677	0.0732	0.0721	0.0798	0.0781

Table 4: Type I Error Rates for Two-Tailed Test, 0.05 Level of Significance

		RHO = 0			RHO = 0.5		RHO = 0.7		RHO = 0.9	
Skewness	Kurtosis	Pearson Spearman	r_L	r_F	r_L	r_F	r_L	r_F	r_L	r_F
3	25	0.0648	0.0376	0.0395	0.0757	0.0789	0.0974	0.0997	0.1223	0.1222
3	25	0.0532	0.0499	0.0508	0.0946	0.0961	0.122	0.1218	0.152	0.1484
			0.0659	0.0658	0.1193	0.1184	0.1525	0.1486	0.1879	0.1793
-3	25	0.064	0.0368	0.0389	0.0773	0.0806	0.0998	0.1022	0.1235	0.1232
-3	25	0.0535	0.0494	0.0504	0.0976	0.0989	0.1251	0.1249	0.1523	0.1481
			0.0652	0.065	0.1224	0.1217	0.1561	0.1525	0.1863	0.1782
-1	7	0.0543	0.0253	0.0274	0.0306	0.032	0.0352	0.036	0.0425	0.0416
-1	7	0.0539	0.0378	0.0389	0.045	0.0452	0.051	0.0506	0.0592	0.0564
			0.0554	0.0554	0.0646	0.0629	0.0729	0.0701	0.0823	0.0766
1	7	0.0534	0.0254	0.0272	0.0314	0.033	0.0369	0.0377	0.0418	0.0414
1	7	0.0544	0.0376	0.0387	0.045	0.0457	0.0526	0.0519	0.0599	0.0569
			0.0545	0.0544	0.0651	0.0638	0.0741	0.071	0.0822	0.0763
1	3	0.0513	0.0233	0.025	0.0268	0.0281	0.0301	0.03	0.0324	0.031
1	3	0.054	0.0353	0.0363	0.0407	0.0406	0.0443	0.0429	0.0477	0.0443
			0.0526	0.0524	0.0601	0.0582	0.0637	0.0604	0.0691	0.0626
-1	3	0.0524	0.0238	0.0256	0.0273	0.0285	0.029	0.0291	0.032	0.031
-1	3	0.0555	0.0361	0.037	0.0401	0.0401	0.0436	0.0422	0.0472	0.0439
			0.0537	0.0536	0.06	0.058	0.0643	0.0606	0.0687	0.0619
-3	25	0.0547	0.0265	0.0284	0.0379	0.04	0.0432	0.0448	0.037	0.0397
-1	3	0.0556	0.0388	0.0398	0.0535	0.0545	0.061	0.0613	0.0546	0.0565
			0.0557	0.0556	0.0743	0.0734	0.0836	0.0819	0.0788	0.0788
3	25	0.0541	0.0259	0.0279	0.0374	0.0395	0.0429	0.0446	0.0388	0.0415
1	3	0.0552	0.0384	0.0393	0.0533	0.0539	0.0608	0.0606	0.0571	0.0584
			0.0553	0.0551	0.074	0.0734	0.0837	0.0821	0.0805	0.0805
3	25	0.0571	0.0293	0.0311	0.0451	0.0473	0.055	0.057	0.0559	0.0579
1	7	0.0545	0.0415	0.0426	0.0612	0.0622	0.0739	0.0742	0.0752	0.0754
			0.0581	0.058	0.0827	0.0818	0.098	0.0958	0.1	0.0978
-3	25	0.0566	0.0302	0.0321	0.0448	0.0474	0.055	0.057	0.0551	0.0572
-1	7	0.0543	0.0425	0.0435	0.0618	0.0627	0.0742	0.0743	0.0747	0.0747
			0.0578	0.0577	0.0833	0.0826	0.0978	0.0956	0.0989	0.0967
-1	3	0.0516	0.0235	0.0251	0.0275	0.0289	0.0298	0.0301	0.0317	0.0311
-1	7	0.0536	0.0353	0.0365	0.0413	0.0413	0.0438	0.043	0.0463	0.0444
			0.0528	0.0527	0.0606	0.0589	0.0642	0.0612	0.0676	0.0626
1	3	0.0518	0.0234	0.0254	0.0282	0.0293	0.031	0.0317	0.0316	0.0311
1	7	0.0556	0.0358	0.037	0.042	0.0422	0.0454	0.0444	0.0465	0.0443
			0.0529	0.0529	0.0612	0.0597	0.0663	0.0629	0.0674	0.0621

Table 5: Power Results for Left-Tail Test when ρ = 0.7, 0.05 Level of Significance

		RHO = 0.7		RHO = 0.5		RHO = 0.3		RHO = 0.1	
Skewness	Kurtosis	r_L	r_F	r_L	r_F	r_L	r_F	r_L	r_F
3	25	0.0323	0.0263	0.1658	0.1442	0.3906	0.3583	0.648	0.6175
3	25	0.0421	0.0344	0.1964	0.1734	0.4354	0.4012	0.6858	0.658
		0.0537	0.0444	0.2300	0.2033	0.4781	0.4445	0.7195	0.6934
-3	25	0.0325	0.0261	0.1633	0.1429	0.3891	0.3565	0.6489	0.6194
-3	25	0.0429	0.0348	0.1948	0.1705	0.4338	0.4000	0.6875	0.6583
		0.0549	0.0451	0.2283	0.2021	0.4756	0.4420	0.7212	0.6951
-1	7	0.0289	0.0241	0.1612	0.1424	0.3909	0.3577	0.639	0.6059
-1	7	0.0369	0.0307	0.1919	0.1685	0.4374	0.4023	0.6809	0.6495
		0.0469	0.0387	0.2257	0.1986	0.4824	0.4466	0.7195	0.6891
1	7	0.0292	0.0245	0.161	0.1409	0.391	0.3587	0.6357	0.6046
1	7	0.0374	0.0312	0.1917	0.1682	0.4366	0.4021	0.6784	0.6460
		0.0475	0.0392	0.2245	0.1983	0.4821	0.4460	0.7179	0.6865
1	3	0.0321	0.0268	0.1696	0.1494	0.3985	0.3669	0.6369	0.6059
1	3	0.0411	0.034	0.1998	0.1767	0.443	0.4086	0.6776	0.6469
		0.0527	0.0434	0.2335	0.2069	0.488	0.452	0.7166	0.6859
-1	3	0.0326	0.0271	0.1706	0.1501	0.3986	0.3667	0.6393	0.6078
-1	3	0.0411	0.0346	0.2013	0.1777	0.4428	0.4089	0.6803	0.6492
		0.0517	0.0433	0.2344	0.2081	0.4881	0.4522	0.7184	0.6879
-3	25	0.0189	0.0153	0.1476	0.1285	0.3818	0.3503	0.6333	0.6034
-1	3	0.0250	0.0203	0.1759	0.1542	0.4253	0.3925	0.6744	0.6434
		0.0328	0.0266	0.2066	0.1819	0.4689	0.4346	0.7112	0.6828
3	25	0.0199	0.0163	0.1461	0.1270	0.3800	0.349	0.6366	0.6065
1	3	0.0260	0.0212	0.1746	0.1522	0.4239	0.3905	0.6778	0.6468
		0.0341	0.0274	0.206	0.181	0.4671	0.4327	0.7147	0.6859
3	25	0.0217	0.0175	0.1471	0.1275	0.3773	0.3457	0.6382	0.6062
1	7	0.0288	0.023	0.1768	0.154	0.4222	0.388	0.6793	0.6481
		0.0377	0.0306	0.2087	0.183	0.4672	0.4316	0.7169	0.6879
-3	25	0.0219	0.0178	0.1479	0.1286	0.3798	0.3478	0.6397	0.6078
-1	7	0.0288	0.0235	0.1778	0.155	0.4243	0.3905	0.6795	0.6497
		0.0375	0.0303	0.2094	0.1843	0.4694	0.4341	0.7171	0.6876
-1	3	0.0296	0.0244	0.1642	0.1444	0.3955	0.363	0.6361	0.6045
-1	7	0.0373	0.0313	0.1942	0.1711	0.4399	0.4062	0.6771	0.6463
		0.0469	0.0392	0.2268	0.2008	0.4851	0.4491	0.7167	0.6849
1	3	0.0298	0.025	0.1659	0.1458	0.3942	0.3621	0.6356	0.6032
1	7	0.0379	0.0315	0.1964	0.1731	0.4385	0.4053	0.6768	0.6458
		0.0475	0.0396	0.2289	0.2031	0.4835	0.4482	0.7159	0.6856

Journal of Modern Applied Statistical Methods
November 2011, Vol. 10, No. 2, 710-717

1538 – 9472/11/$95.00

Identifying Outliers in Fuzzy Time Series

S. Suresh K. Senthamarai Kannan
ManonmaniamSundaranar University,
Tirunelveli, India

Time series analysis is often associated with the discovery of patterns and prediction of features. Forecasting accuracy can be improved by removing identified outliers in the data set using the Cook's distance and Studentized residual test. In this paper a modified fuzzy time series method is proposed based on transition probability vector membership function. It is experimentally shown that the proposed method minimizes the average forecasting error compared with other known existing methods.

Key words: Membership functions, fuzzy sets, fuzzy logical relations, outliers, Cook's distance, average forecasting error.

Introduction

Time series analysis plays a vital role in most actuarial related problems. Fuzzy time series is a scientific method that can be applied to time series data and in forecasting future events. Commonly actuarial issues are mainly related to the concept of uncertainty, each observation of a fuzzy time series is assumed to be a fuzzy variable along with an associated membership function. The accuracy of fuzzy time series plays a significant role in forecasting. Conventional methods that deal with forecasting problems show their inefficiency when solving problems related to linguistic values.

Several approaches in the literature have been developed to identify outliers in time series analysis. Fox (1972) introduced the concept of outliers in time series analysis and discussed different types of time series outliers. Tsaur (1986) used an iterative fashion to detect multiple outliers.

S. Suresh is a Ph.D. student in the department of statistics and a recipient of the INSPIRE fellowship provided by the Department of Science and Technology, New Delhi. Email him at: sureshstat22@gmail.com. K. Senthamarai Kannan is a professor in the Department of Statistics. His research interests include stochastic modeling and data mining. Email him at: senkannan2002@gmail.com.

A complete survey and discussion regarding outlier detection can be found in Barnett and Lewis, (1984). The Studentized residual analysis method and Cook's distance can be used to detect outliers in time series. According to Barnett and Lewis (1984), the identified outliers can be either accommodated or removed. Chang and Tiao (1988) discussed estimation of time series parameters in the presence of outliers.

Song and Chissom (1993) introduced definitions of fuzzy time series and its modeling by using fuzzy relational equations and approximate reasoning by Zadeh (1965). Song and Chissom (1993) outlined modeling procedures and implemented time-invariant and time-variant models to forecast enrollments at the University of Alabama. Sullivan and Woodall (1994) reviewed the first-order time-variant fuzzy time series model and the first-order time-invariant fuzzy time series model presented. Chen (1996) developed a basic or simplified method for time series forecasting using arithmetic operations rather than complicated max-min composition operations. Sullivan and Woodall (1999) have discussed three methods for estimating Markov transition matrices when observed state probabilities are not all either zeros or ones and a simulation-based comparison of the performance of the estimators. Huarng (2001) worked on finding the effective lengths of intervals to improve

forecasting accuracy. Chen (2002) developed a fuzzy time series using arithmetic operations.

Song (2003) has proposed the sample autocorrelation functions of fuzzy time series and used in model selection. The main idea is to select a number of different data sets from each fuzzy set and calculate the sample auto correlation function for each data set. Chung and Hsu (2004) proposed a higher order fuzzy time series applied for Taiwan future exchange. Lee et al., (2004) have presented an improved method to forecast university enrollments based on the fuzzy time series. The method proposed not only defines the supports of the fuzzy numbers that represent the linguistic values of the linguistic variable more appropriately, but also makes the RMSE smaller Sah et al., (2005) presented the method for forecasting given high accuracy and comparing existing methods. Tsaur et al., (2005) have proposed fuzzy relation matrix affecting the forecasting performance and proposed an arithmetic procedure for deriving fuzzy relation matrix method using Fuzzy relation analysis in fuzzy time series. Fuzzy relation is a crucial connector in presenting fuzzy time series model. Also the concept of entropy is applied to measure the degrees of fuzziness when a time invariant matrix is derived. Singh (2007) proposed a method for fuzzy time series forecasting using a simple time variant method. Hao-Tien Liu (2007) has proposed improved time-variant fuzzy time series method. The proposed method takes into consideration of Window base, length of interval, degrees of membership values, and existence of outliers. The improved method provides decision makers with more precise forecasted values.

Fuzzy Time Series

Song and Chissom (1993) proposed a procedure for solving fuzzy time series models described as follows: Let *U*be the universe of discourse,

$$U = [V_{\min} - V_1, V_{\max} + V_2],$$

where $U = \{u_1, u_2 ... u_n\}$ is the given historical data, the minimum data is V_{min}, the maximum data is V_{max} and V_1, V_2 are two real numbers. A fuzzy set A_i of U is defined by

$$A_i = f_{Ai}(u_1)/u_1 + f_{Ai}(u_2)/u_2 + ...+ f_{Ai}(u_n)/u_n$$

where f_A is the membership function of fuzzy set A_i. Let $Y(t), (t = 0, 1, 2...)$ be a subset of R. If $Y(t)$ is the universe of interest defined by the fuzzy set $\mu_i(t), i = 1, 2, ...$ then F(t) is called a fuzzy time series of $Y(t)$. If there exists a fuzzy relationship $R(t, t-1)$, such that $F(t) - F(t-1) * R(t, t-1)$, where the symbol * is an operator, then $F(t)$ is said to be induced by $F(t–1)$ the relationship can be denoted by $F(t-1) \rightarrow F(t)$. Suppose $F(t-1)$ by A_i and $F(t)$ by A_j fuzzy logical relationship can be defined by $A_i \rightarrow A_i$ where A_i and A_j are called, respectively, the left hand side and right hand side of the fuzzy logical relationship.

Detection of Outliers

Outlier defines an observation that is numerically distant from the rest of the data, or is any observation in a set of data that is inconsistent with the remainder of the observations in the data set. The outlier is inconsistent in the sense that it is not indicative of possible future behavior of the data sets. Cook's Distance (D_i) defines how much an observation affects a change in a parameter estimate of least square regression analysis:

$$D_i = \frac{e_i^2}{p * MSE}\left(\frac{h_{ii}}{(1-h_{ii})^2}\right).$$

To interpret D_i, compare it to the F-distribution with (p, n–p) degrees of freedom to determine the corresponding percentile; if the percentile value is greater than 50%, then the observation has a major influence on the fitted values and should be examined. Thus, if D_i>F(0.5, p, n–p) then consider influence.

The Studentized residual analysis methods can assist in determining whether outliers exist in historical data. The Studentized test can be employed to examine the outliers as follows: If there are n historical data x_1, x2,…,x_n a square matrixR can be defined as,

$$R = X\left(X^T X\right)^{-1} X^T$$

$$X = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \cdot & \cdot \\ \cdot & \\ \cdot & \cdot \\ 1 & X_n \end{bmatrix}$$

The Studentized residual can be defined by the Studentized Residual Test:

$$\frac{e_i}{S_j},$$

where

$$S_j = \hat{\sigma}(i)\sqrt{1-r_{ii}}.$$

Here, S_j is the estimated variance of the residual, e_i specifies the residual of the i^{th} datum, $\hat{\sigma}(i)$ is the estimated value of the standard deviation σ without the i^{th} observation, and r_i is the i^{th} diagonal element in matrix R. The data is considered to be an outlier where the absolute residual values having Studentized residuals are greater than 2.0.

Discrete Time Markov Chain

A Markov chain is a discrete random process with the property that the next state depends only on the current state; the past states have no influence on the future.

A Markov chain X is said to be time-homogenous if the conditional probability $P[X_{n+1} = j \mid X_n = i] = P_{ij},\ i, j \in S$ is independent of *n*, and *S* is the countable state space. The probabilities of P_{ij} are called the transition probabilities for the Markov chain *X*. It is customary to arrange the P_{ij} or $P(i,j) = P_{ij}$ into a square array and to call the resulting matrix $P=(P_{ij})$ the transition probability matrix of the Markov chain *X*; for any $i, j \in S$, $P_{ij} \geq 0$, and $\sum_{j \in E} P_{ij} = 1$ for any $m \in N$, $P[X_{n+m} = j \mid X_n = i] = P_{ij}^{m},\ i, j \in S$. Here $P_{ij}^{(m)}$ denotes the probability that the process goes from state *i* to state *j* in m transitions. The transition probabilities P_{ij} can be exhibited as a square matrix

$$P = P_{ij} = \begin{bmatrix} P_{00} & P_{01} & P_{02} & P_{03} & \cdots \\ P_{10} & P_{11} & P_{12} & P_{13} & \cdots \\ P_{20} & P_{21} & P_{22} & P_{23} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ P_{i0} & P_{i1} & P_{i2} & P_{i3} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \end{bmatrix}$$

which is called the transition probability matrix of the chain. If the number of states is finite, for example *n*, then there will be n rows and n columns in the matrix P; otherwise the matrix will be infinite.As it is known, $P_{ij} \geq 0$, and $\sum_{j=0}^{\infty} P_{ij} = 1.$ for every $i, j = 0,1,2,...$

Modified Method of Forecasting

This article aims to provide better forecasting accuracy using fuzzy time series with forecasts using only historical data. The step by step forecasting procedure is as follows:

1. First identify outliers from the historical data using Cook's distance and the Studentized residual test.

2. After identifying the outlier, compute the appropriate length of interval *l* using the distribution based method by Chen (2002).

3. Compute the number of intervals *m* as follows:

$$m = \frac{(V_{max} + V_2) - (V_{min} - V_1)}{l}$$

where V_{max} is the maximum value of the historical data, V_2 is the positive integer, V_{min} is the minimum value of the historical data, V_1 is the positive integer and *l* is the appropriate length of interval.

4. Let *U* be the universe of discourse, $U = [V_{min} - V_1,\ V_{max} + V_2]$ and partition

into *m* equal length intervals $\{u_1, u_2, u_3 \ldots u_m\}$.

5. Fuzzify the variations of the historical data and determine the fuzzy logical relationships.

6. If A_i is the fuzzified value of current year n and A_j is the fuzzified value of next year *n+1*, then fuzzy logical relation is denoted by $A_i \rightarrow A_j$.

7. Define Fuzzy sets A_i on universe of discourse *U*, then determined how many linguistic variables to be fuzzy sets.

8. Define the linguistic terms of A_i represented by the fuzzy sets are as follows:

$$A_1 = \{u_1/0.667, u_2/0.337, u_3/0, \ldots, u_m/0\}$$

$$A_2 = \{u_1/0.25, u_2/0.5, u_3/0.25, \ldots, u_m/0\}$$

$$A_3 = \{u_1/0, u_2/0.25, u_3/0.50, \ldots, 0, u_m/0\}$$

$$\ldots$$

$$A_m = \{u_1/0, \ldots, u_{m-1}/0.333, u_m/.0.667\}$$

9. Fuzzify the historical data are as follows: If the value belongs to u_1, then fuzzified membership into $0.667/A_1 + 0.333/A_2 + 0/A_3$ denoted by A_1.If the value belongs to u_i,i=2,3,…,n-1, then the fuzzified membership values into $0.25/A_{i-1} + 0.5/A_i + 0/A_{i+1}$ denoted by A_i. If the value belongs to u_n then the fuzzified membership values $0/A_{n-2} + 0.333/A_{n-1} + 0.667/A_n$ denoted by A_n.

10. Identify the fuzzy logical relationship of first order fuzzy time series is as follows: $A_{j-1} \rightarrow A_j$.

11. Determine the fuzzy logical relationship $R_i = A_{i-1}^T \times A_i$, i = 1,2,…,n and obtain the transition probability matrix is $P_m = \bigcup_{i=1}^{n} R_i$.

12. Calculate the forecast outputs using transitions state probability membership function as $P'_{t+1} = P'_t \times P_m$, where, P'_{t+1} is the current year historical data is obtained from previous year vector probability membership P'_t and probability matrix P_m.

13. Obtain the average forecasting error using actual and forecasted values:

$$Forcast\ error = \frac{|forcasted\ value - actual\ value|}{(actual\ value)} \times 100\,\%$$

Numerical Example

The proposed approach isdescribed with actual data corresponding to the number of accidentsoccurring in India. The original data set is shown in Table 1.

Table 1: Identifying Outliers Using Cook's Distance and Studentized Residual Test

Year	Number of Accidents	CooksDistance	Student Residual
1985	20700	0.001	-0.168
1986	21550	0.008	-0.415
1987	23400	0.000	0.026
1988	24670	0.000	0.067
1989	27000	0.020	0.837
1990	28260	0.019	0.870
1991	29340	0.014	0.777
1992	26030	**0.115**	**-2.597**
1993	28010	0.067	-1.890
1994	32040	0.000	0.142
1995	34890	0.037	1.292
1996	37120	0.097	2.126
1997	37370	0.048	1.345
1998	38500	0.051	1.290
1999	38640	0.010	0.524
2000	39140	0.000	0.037
2001	40560	0.002	0.182
2002	40750	0.016	-0.528
2003	40670	0.040	-1.504
2004	42990	0.035	-0.667
2005	43920	0.070	-0.882
2006	46090	0.006	-0.085
2007	47920	0.135	0.371

Table.1 shows an unusual residual value (–2.597) in 1992, which has a Studentized absolute value residual greater than 2.0. Studentized residuals measure how many standard deviations each observed value deviates from a model fitted using all of the data except that observation.In this case, there is one Studentized residual greater than 2.0, but none greater than 3.0. The step by step procedure is as follows:

1. First, the appropriate length of interval l is computed using distribution based length procedure to obtain an interval length of $l = 2000$.

2. The calculated number of intervals, $m = \frac{48000 - 20000}{2000} = 14.$

3. Define the universe of discourse or universal set, U = [20000, 48000], and partition U into 14 equal length of intervals, u_i, i=1, 2,…, 14, u_1 =[20000, 22000), u_2 =[22000, 24000), u_3 =[24000, 26000),..., u_{13} =[44000, 46000), and u_{14}= [46000, 48000].

4. It is assumed that the linguistic variable of the historical data can take fuzzy values are as follows: A_1 (very big decrease), A_2 (big decrease), A_{13} (big increase) and A_{14} (very big increase). Then, for the given intervals u_i, i = 1, 2….14, each u_i belongs to a particular A_j, j=1,2,…,14 and is expressed by the real value within the range [0,1]. The complete sets of relationship are shown in Table 2.

$A_1 \rightarrow A_1$	$A_1 \rightarrow A_2$	$A_2 \rightarrow A_3$
$A_3 \rightarrow A_4$	$A_4 \rightarrow A_5$	$A_5 \rightarrow A_5$
$A_5 \rightarrow A_4$	$A_4 \rightarrow A_5$	$A_5 \rightarrow A_7$
$A_7 \rightarrow A_8$	$A_8 \rightarrow A_9$	$A_9 \rightarrow A_9$
$A_9 \rightarrow A_{10}$	$A_{10} \rightarrow A_{10}$	$A_{10} \rightarrow A_{10}$
$A_{10} \rightarrow A_{11}$	$A_{11} \rightarrow A_{11}$	$A_{11} \rightarrow A_{11}$
$A_{11} \rightarrow A_{12}$	$A_{12} \rightarrow A_{12}$	$A_{12} \rightarrow A_{14}$
$A_{14} \rightarrow A_{14}$	$A_{13} \rightarrow A_{14}$	$A_{14} \rightarrow A_{14}$
$A_{13} \rightarrow A_{14}$		

5. The fuzzy relationships are combined into fuzzy logical relations starting from identical left-hand sides. Then R_i,i=1, 2…22 is calculated as a sum of logical relationships in each group. Here, the relation matrix R_i is converted into a transition probability matrix P_m is shown in Figure 1.

6. Table 3 illustrates the defuzzified forecast outputs using transition state probability membership function. The outputs are multiplied with corresponding mid values of the fuzzy interval over the period of years and its overall summation leads the predicted values. For example, year 2004 is forecasted using fuzzified values of 2003 .The midpoints of the intervals u_1,u_2 ,..., u_{14} are multiplied into corresponding defuzzified probability values and its overall summation. The actual and predicted value of number of accidents in India is shown in Figure 2.

7. Finally, the average forecasting error is obtained using actual and forecasted values, when compared with the other existing methods. The result is shown in Table 3.

Conclusion

This article is mainly focused on improving the forecasting accuracy by removing the identified outlier in the data set. This proposed method first predicts the fuzzy time series using transition probability vector membership functions, then, the average forecasting error is calculated based on after removing the outliers in the data. The experimental results show that the average forecasting error is 2.86% for the historical data. After removing the outlier, the method produces 2.60% of average forecasting error. Thereby, the proposed method improves average forecasting accuracy by approximately 9%.The results indicate that the proposed method is more appropriate compared to other existing methods. It is supported by numerical and graphical representations.

Figure 1: Transition Probability Matrix from Relation Matrix

$$P_m = \begin{bmatrix}
0.39 & 0.39 & 0.18 & 0.04 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\
0.22 & 0.28 & 0.28 & 0.18 & 0.04 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\
0.00 & 0.06 & 0.25 & 0.38 & 0.25 & 0.06 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\
0.00 & 0.00 & 0.04 & 0.25 & 0.38 & 0.21 & 0.08 & 0.04 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\
0.00 & 0.00 & 0.00 & 0.18 & 0.36 & 0.25 & 0.14 & 0.07 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\
0.00 & 0.00 & 0.00 & 0.13 & 0.25 & 0.19 & 0.19 & 0.19 & 0.06 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.17 & 0.42 & 0.33 & 0.08 & 0.00 & 0.00 & 0.00 & 0.00 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.05 & 0.25 & 0.40 & 0.25 & 0.05 & 0.00 & 0.00 & 0.00 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.09 & 0.31 & 0.38 & 0.19 & 0.03 & 0.00 & 0.00 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.02 & 0.16 & 0.34 & 0.32 & 0.14 & 0.02 & 0.00 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.05 & 0.20 & 0.34 & 0.25 & 0.10 & 0.06 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.07 & 0.25 & 0.29 & 0.20 & 0.19 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.08 & 0.15 & 0.31 & 0.47 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.33 & 0.67
\end{bmatrix}$$

Table 3: Forecasting Number of Accidents from 1985-2007

Year	Actual	Fuzzy Output Vectors														Predicted
		A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13	A14	
1985	20700															
1986	21550	0.33	0.35	0.22	0.09	0.01	0	0	0	0	0	0	0	0	0	23202
1987	23400	0.33	0.35	0.22	0.09	0.01	0	0	0	0	0	0	0	0	0	23202
1988	24670	0.20	0.26	0.25	0.19	0.08	0.02	0	0	0	0	0	0	0	0	24488
1989	27000	0.05	0.10	0.21	0.29	0.23	0.08	0.02	0.01	0	0	0	0	0	0	26858
1990	28260	0	0.02	0.08	0.26	0.34	0.18	0.08	0.04	0	0	0	0	0	0	28952
1991	29340	0	0	0.01	0.18	0.33	0.22	0.14	0.09	0.02	0	0	0	0	0	30280
1993	28010	0	0	0.01	0.18	0.33	0.22	0.14	0.09	0.02	0	0	0	0	0	30280
1994	32040	0	0	0.01	0.18	0.33	0.22	0.14	0.09	0.02	0	0	0	0	0	30280
1995	34890	0	0	0	0.03	0.06	0.05	0.14	0.32	0.28	0.10	0.01	0	0	0	34958
1996	37120	0	0	0	0	0	0	0.07	0.25	0.36	0.24	0.07	0.01	0	0	37042
1997	37370	0	0	0	0	0	0	0.01	0.12	0.30	0.34	0.19	0.05	0.01	0	38477
1998	38500	0	0	0	0	0	0	0.01	0.12	0.30	0.34	0.19	0.05	0.01	0	38477
1999	38640	0	0	0	0	0	0	0	0.03	0.17	0.32	0.29	0.14	0.04	0.02	39996
2000	39140	0	0	0	0	0	0	0	0.03	0.17	0.32	0.29	0.14	0.04	0.02	39996
2001	40560	0	0	0	0	0	0	0	0.03	0.17	0.32	0.29	0.14	0.04	0.02	39996
2002	40750	0	0	0	0	0	0	0	0.01	0.06	0.21	0.31	0.23	0.11	0.08	41656
2003	40670	0	0	0	0	0	0	0	0.01	0.06	0.21	0.31	0.23	0.11	0.08	41656
2004	42990	0	0	0	0	0	0	0	0.01	0.06	0.21	0.31	0.23	0.11	0.08	41656
2005	43920	0	0	0	0	0	0	0	0	0.01	0.09	0.23	0.24	0.20	0.23	43441
2006	46090	0	0	0	0	0	0	0	0	0.01	0.09	0.23	0.24	0.20	0.23	43441
2007	47920	0	0	0	0	0	0	0	0	0	0	0.02	0.05	0.32	0.60	46001

Figure 2: Actual and Predicted Values of Accidents in India

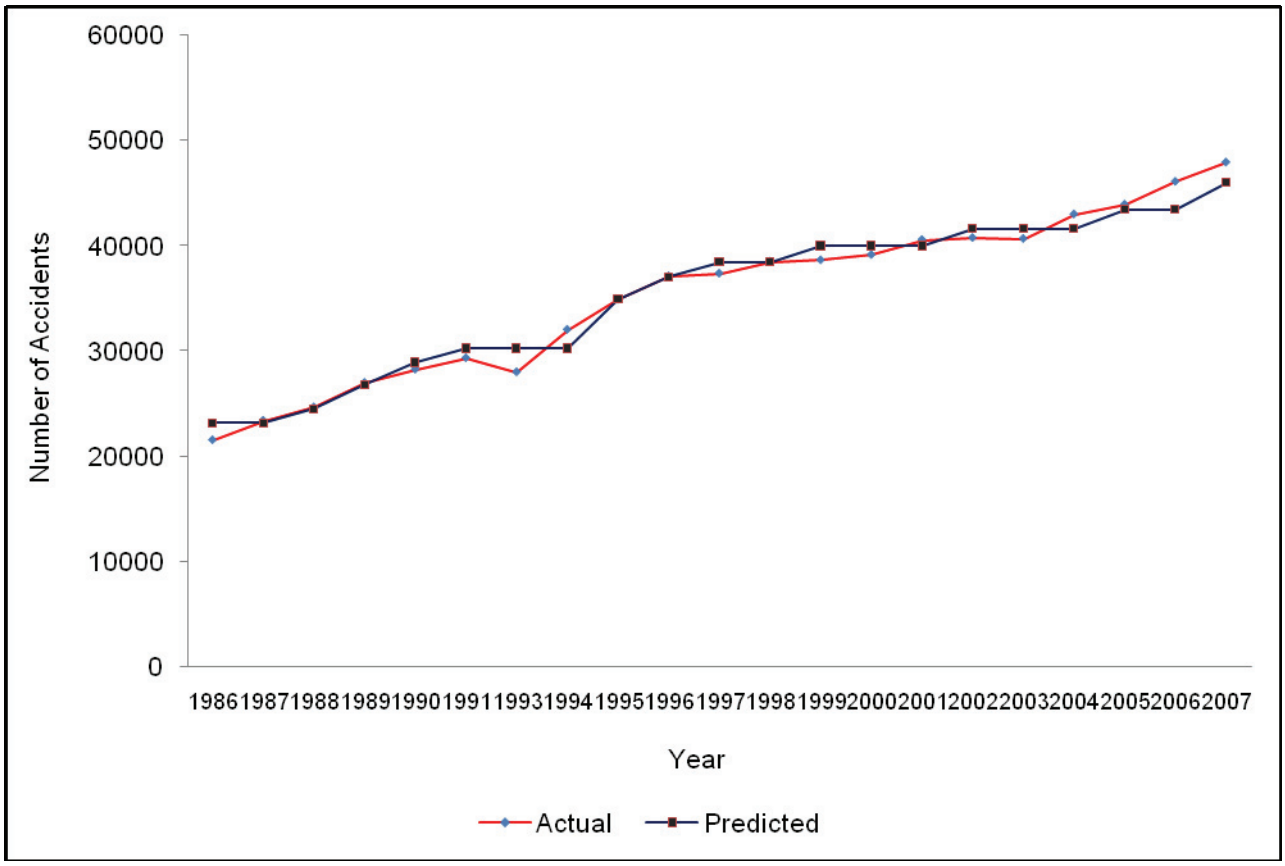


Table 4: Comparison of Average Forecasting Error with Existing Methods

Methods	Chen&Hwang (2000)	Lee, et.al., (2004)	Singh(2007)	Proposed
AFE	3.90%	3.43%	2.89%	2.60 %

References

Abraham, B., & Chuang, A. (1989).Outlier detection and time series modeling. *Technometrics*, *31*(*2*), 241-248.

Barnett, V., & Lewis, T. (1984).*Outliers in statistical data* (*3rd Ed*.). NY: John Wiley & Sons.

Chang, I., and Tiao,G. C. (1988).Estimation of time series parameters in the presence of outliers.*Technometrics*,*30*(*2*), 193-204.

Chen, S. M. (1996). Forecasting enrollments based on fuzzy time series. *Fuzzy Sets and Systems*, *81*, 311-319.

Chen, S. M. (2002). Forecasting enrollments based on high-order fuzzy time series. *International Journal of Cybernetics and Systems*, *33*, 1-16.

Chen, S.M., & Hsu, C-C.(2004). A new method to forecast enrollments using fuzzy time series.*International Journal of Applied Science and Engineering*, *2*(*3*), 234-244.

Own, C-M., & Yu, P-T. (2005). Forecasting fuzzy time series on a heuristic high-order model. *International Journal of Cybernetics and Systems*, *336*, 705-717.

Fox, A. J. (1972). Outliers in time series. *Journal of Royal Statistical Society*, *B34*, 350-363.

Hao-Tien Liu (2007). An Improved Fuzzy Time Series Forecasting Method using Trapezoidal Fuzzy Numbers**,** *Fuzzy Optimum Decision Making,* 6, 63–80.

Huarng, K. (2001). Effective lengths of intervals to improve forecasting in fuzzy time series.*Fuzzy Sets and Systems*, *123*, 387-394.

Lee. A, & Ming-Tao Chou (2004). Fuzzy Forecasting Based on Fuzzy Time Series, *International Journal of Computer Mathematics 81(7),* 781–789.

Sah, M., & Degtiarev, K.Y.(2005). Forecasting enrollment model based on first-order fuzzy time series.*Proceedings of the World Academy of Science, Engineering and Technology.*

Tsaur, R-Y., &Yang, O. (2005).Fuzzy relation analysis in fuzzy time series model.*Computers and Mathematics with Application*, *49*, 539-548.

Tsaur, R-Y. (1986). Time series model specification in the presence of outliers.*Journal of the American Statistical Association*, *81*(*393*), 132- 141.

Singh S.R.(2007). A Simple time variant method for fuzzy time series forecasting.*International Journal Cybernetics and Systems*, *38*, 305-321.

Song, Q., &Chissom, B. (1993).Fuzzy time series and its models.*Fuzzy Sets and Systems*,*54*, 269-277.

Song .Q, (2003). A Note on Fuzzy Time Series Model Selection with Sample Autocorrelation Functions, *Cybernetics and Systems*: *An International Journal*, 34, 93-107.

Sullivan. J., & Woodall, W.H. (1994).A comparison of fuzzy forecasting and Markov modeling.*Fuzzy Sets and Systems*, *64*, 279-293.

Sullivan J & Woodall, W.H. (1999). Estimating Markov Transition Matrices using Uncertain Observed States, *Stochastic Analysis and Applications*, *17(2),* 253-274.

Suresh, S., & Senthamarai Kannan, K. (2009).Forecasting model on fuzzy Markov chain.*International Journal of Statistics and Systems*, *4*(*2*), 137-147.

Zadeh, L.A. (1965). Fuzzy Sets, *Information and Control*, *8(3),* 338-353.

Journal of Modern Applied Statistical Methods
November 2011, Vol. 10, No. 2, 718-729

1538 – 9472/11/$95.00

Identification of Optimal Autoregressive Integrated Moving Average Model on Temperature Data

Olusola Samuel Makinde Olusoga Akin Fasoranbaku
Federal University of Technology,
Akure, Nigeria

Autoregressive Integrated Moving Average (ARIMA) processes of various orders are presented to identify an optimal model from a class of models. Parameters of the models are estimated using an Ordinary Least Square (OLS) approach. ARIMA (*p, d, q*) is formulated for maximum daily temperature data in Ondo and Zaira from January 1995 to November 2005. The choice of ARIMA models of orders *p* and *q* is intended to retain persistence in a natural process. To determine the performance of models, Normalized Bayesian Information Criterion is adopted. The ARIMA (*1, 1, 1*) is adequate for modeling maximum daily temperature in Ondo and Zaira; model parameters are estimated and redundant variables are removed. Causality and the invertibility behavior of some optimal models are also presented.

Key words: Autoregressive Integrated Moving Average, optimal, causality, invertibility, redundancy.

Introduction

A time series of T successive observations is regarded as a sample from an infinite population of a time-series that could have been generated by the stochastic process under study. A powerful way to extract useful information on the underlying process - solely on the basis of the past behavior of the time series itself - is the univariate Box-Jenkins approach. Although originally developed for forecasting purposes (Box & Jenkins, 1976; Nelson, 1976), Box-Jenkins models are useful tools for describing the time dependent structure of stationary and non-stationary time series. Box-Jenkins models for stationary time series, or ARIMA models, have been applied in many areas of research, for example in tree-ring chronologies (Meko, 1981), in the evolution of the unemployment rate (Dobre & Alexandru, 2008), and in the analysis of UK Pounds/US Dollar exchange rate (Shittu and Yaya, 2009).

O. S. Makinde is Graduate/Research Assistant in the Department of Mathematical Sciences. Email him at: osmakinde@futa.edu.ng. O. A. Fasoranbaku is Senior Lecturer in the Department of Mathematical Sciences. Email him at: olusogaf@yahoo.com.

Models are initialized using observed data. As proposed by Lorenz (1963), long range forecasts - those made at a range of two weeks or more - are impossible to definitively predict the state of the atmosphere owing to the chaotic nature of the mechanism involved. Forecast models are used to determine future conditions. However, in real life research and practice, patterns of data are unclear and individual observations involve considerable error; thus, it is necessary to not only uncover the hidden patterns in the data but also to forecast. The ARIMA methodology (Box & Jenkins, 1976) provides a method to accomplish these tasks.

Considering estimates of times series model parameters, Pham-Dinh (1978) computed the exact log likelihood of a time series model and also proposed and justified an asymptotic approximation of the model. Bobba, et al (2006) formulated a stochastic model simulating trends in hydrological and meteorological variables: Their choice of ARIMA model of orders p and q was intended to retain any persistence in the natural processes and they claimed that an ARIMA (1, 0, 1) model was adequate for modeling three variables of temperature, precipitation and stream flow on a seasonal basis in the North East Pond River Watershed. Ojo (2009) compared subsets of autoregressive integrated moving average models to full

autoregressive integrated moving average models. The parameters of these models were estimated and the statistical properties of the derived estimates were investigated. In his study, he showed that subset autoregressive integrated moving average models performed better than full autoregressive integrated moving average models. Makinde (2011) investigated the behavioural pattern of invertibility parameter π_i of the ARIMA (p, d, q) model for various *p* and *d*. He showed that behaviour of π_i depends on the order of autoregressive part (*p*), the order of integrated part (*d*), positive and negative values of moving average parameter (θ). Similarly, Fasoranbaku & Makinde (2011) investigated causality parameter of ARMA model. From their findings, It is deduced that the behaviour of causality parameter ψ_i depends on positive and negative values of autoregressive parameter ϕ and moving average parameter ϑ.

In this study, we shall evaluate parameters of ARIMA(p,d,q) for various values of p and d using an ordinary least squares (OLS) method and Crammer's rule; identify optimal model in a class of ARIMA models for temperature profile of two cities in Nigeria and check for redundant variables in the models using a t-test.

Stationarity and Test of Stationarity

A process is said to be strictly stationary if, for any value of $j_1, j_2, ..., j_n$, the joint distribution of $(y_t, y_{t+1}, y_{t+2}, ..., y_{t+j})$ depends only on the interval separating the dates $(j_1, j_2, ..., j_n)$, and not on the date (*t*) itself. If a process is strictly stationary with finite second moments, then it must be covariance stationary (Hamilton, 1994).

In short, if a time series is stationary, its mean, variance, and autocovariance (at various lags) remain the same regardless of the point at which they are measured; that is, they are time invariant. There are several tests of stationarity; which include: (1) graphical analysis, (2) a correlogram, and (3) unit root test, e.t.c. For a stationary time series, a correlogram tapers quickly; whereas for non-stationary time series it dies off gradually. If autocorrelations start high and decline slowly, then the series is nonstationary and should be differenced. Similarly, an ARIMA process is said to be stationary if spikes decay to zero after a few lags. In this study, correlogram use was adopted to test for stationarity of temperature data.

Test for Model Adequacy

To test the adequacy of the model, the Ljung-Box (1978) statistic will be used; this is a statistical test for determining whether any of a group of autocorrelations of a time series is different from zero. As opposed to testing randomness at each distinct lag, it tests the overall randomness based on a number of lags, and is therefore a portmanteau test. The Ljung-Box Statistic is:

$$Q(\hat{r}) = n(n+2)\sum_{k=1}^{h}\frac{\hat{r}_k}{n-k}.$$

Specification of ARIMA in Terms of A Lag Operator

When the models are specified in terms of the lag operator L, the AR (p) model is given by

$$\varepsilon_t = \left(1 - \sum_{i=1}^{p}\phi_i L^i\right) y_t = \phi(L)\, y_t,$$

where

$$\phi(L) = 1 - \sum_{i=1}^{p}\phi_i L^i,$$

and the MA(q) model is given by

$$y_t = \left(1 + \sum_{i=1}^{p}\theta_i L^i\right)\varepsilon_t = \theta(L)\,\varepsilon_t,$$

where

$$\theta(L) = 1 + \sum_{i=1}^{p}\theta_i L^i.$$

ARIMA (*p*,0, *q*) is

$$\left(1 - \sum_{i=1}^{p}\phi_i L^i\right) y_t = \left(1 + \sum_{i=1}^{p}\theta_i L^i\right)\varepsilon_t \qquad (1)$$

or more concisely:

$$\phi(L)y_t = \theta(L)\varepsilon_t,$$

which implies $y_t = \psi(L)\varepsilon_t$, where

$$\psi(L) = \frac{\theta(L)}{\phi(L)} = \frac{1+\theta_1 L+\theta_2 L^2+\ldots+\theta_p L^p}{1-\phi_1 L-\phi_2 L^2-\ldots-\phi_p L^p}. \quad (2)$$

The ARIMA process is stationary if This occurs if the series $\psi(Z)$ converges for every Z with $|Z|\le 1$. Because $\psi(Z)$ is a rational function, the series converges for every Z with $|Z|\le 1$ if the complex zeros of $\phi(Z)$ lie outside the unit circle. If a process is stationary, then because $y_t = \psi(L)\varepsilon_t$, and the expected values of ε_t are all 0, the expected value of y_t is also 0.

Causality of Some ARIMA Processes

Some ARIMA processes of various orders are shown in causal form to provide a useful way of generating a random sequence. That is, a linear process y_t, as a linear combination of white noise variates ε_t. For an ARIMA (1, 0, 1) process, $y_t = c+\phi y_{t-1}+\varepsilon_t+\theta\varepsilon_{t-1}$ implies that $y_t = (1-\phi L)^{-1}c+(1-\phi L)^{-1}(\varepsilon_t+\theta\varepsilon_{t-1})$ which gives

$$y_t = \frac{c}{(1-\phi)}+\varepsilon_t+\sum_{i=1}^{\infty}(\phi^i+\theta\phi^{i-1})\varepsilon_{t-i},$$

where $i = 1,2,\ldots$ with $E(y_t) = \mu = \frac{c}{(1-\phi)}$; this holds only if $\phi \neq 1$.

For an ARIMA (1, 0, 2) process, $y_t = c+\phi y_{t-1}+\varepsilon_t+\theta_1\varepsilon_{t-1}+\theta_2\varepsilon_{t-2}$ which gives

$$y_t = \frac{c}{(1-\phi)}+\varepsilon_t+\sum_{i=1}^{\infty}(\phi^i+\theta_1\phi^{i-1}+\theta_2\phi^{i-2})\varepsilon_{t-i},$$

$$y_t = \frac{c}{(1-\phi)}+\sum_{i=0}^{\infty}\psi_i\varepsilon_{t-i}, \quad (3)$$

where $\psi_0 = 1$, $\psi_1 = \phi+\theta$ and $\psi_j = \phi^j+\theta_1\phi^{j-1}+\theta_2\phi^{j-2}$, $j = 2,3,4,\ldots$; this is valid if $\phi \neq 1$.

Fasoranbaku & Makinde (2011) has shown that the causality parameter ψ_i is skewed to the right and sinusoidal for positive and negative values of ϕ respectively. Absolute value of causality parameter ψ_i of ARIMA (1, 0, q) increases as the value of q increases for positive values of ϕ. The behavioural pattern of the causality parameters for $d = 0$ and $|\phi| < 1$ is well studied in Fasoranbaku & Makinde (2011).

Representation ARIMA Models in Inverted Form

An ARIMA (p, d, q) process is said to be invertible if the series converges in mean to ε_t as $v \to \infty$. This happens when $\theta(Z) = 0$ lie outside the unit circle. An ARIMA (p, d, q) process is invertible if the absolute value of the parameters of ARIMA (p, d, q) model satisfy $|\theta_i| < 1$ for $i = 1, \ldots, q$.

ARIMA (1, 0, 1)

$$y_t = c+\phi y_{t-1}+\varepsilon_t+\theta\varepsilon_{t-1}, \quad (4)$$

$$\varepsilon_t = -\frac{c}{(1+\theta)}+y_t+\sum_{i=1}^{\infty}(-1)^i(\theta^i+\phi\theta^{i-1})y_{t-i},$$

$$y_t = \frac{c}{(1+\theta)}+\sum_{i=1}^{\infty}(-1)^{i+1}[\phi\theta^{i-1}+\theta^i]y_{t-i}+\varepsilon_t,$$

$$y_t = \frac{c}{(1+\theta)}+\sum_{i=1}^{\infty}\pi_i y_{t-i}+\varepsilon_t, \quad (5)$$

where $\pi_i = (-1)^{i+1}\left[\phi\theta^{i-1}+\theta^i\right]$, $i = 1,2,3,\ldots$; this holds if $\theta \neq -1$.

ARIMA (1, 1, 1)

$$\Delta y_t = c + \phi \Delta y_{t-1} + \varepsilon_t + \theta \varepsilon_{t-1} \quad (6)$$

$$\varepsilon_t = -\frac{c}{(1+\theta)} + \sum_{i=0}^{\infty} \pi_i y_{t-i}$$

$$\pi_i = \begin{cases} 1, & i = 0 \\ -(\theta + \phi + 1), & i = 1 \\ (-1)^i[(\theta+\phi)(\theta^{i-1} + \theta^{i-2})], & i = 2, \ldots \end{cases} \quad (7)$$

ARIMA (2, 0, 1)

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \varepsilon_t + \theta \varepsilon_{t-1}, \quad (8)$$

$$y_t = \begin{bmatrix} \frac{c}{(1+\theta)} + (\theta + \phi_1) y_{t-1} \\ + \sum_{i=2}^{\infty} (-1)^{i+1} [\theta^i + \phi_1 \theta^{i-1} \\ + (-1)^{i+1} \phi_2 \theta^{i-2}] y_{t-i} + \varepsilon_t \end{bmatrix},$$

$$y_t = \frac{c}{(1+\theta)} + \sum_{i=1}^{\infty} \pi_i y_{t-i} + \varepsilon_t, \quad (9)$$

where $\pi_1 = \phi_1 + \theta$ and

$$\pi_i = (-1)^{j+1}[\theta^j + \phi_1 \theta^{j-1} + (-1)^{j+1} \phi_2 \theta^{j-2}],$$

$j = 2, 3, 4, \ldots$; this holds if $\theta \neq -1$.

Makinde (2011) has shown that invertibility parameter π_i of ARIMA (p, d, 1) for various integer values of d are sinusoidal, the absolute value of the invertibility parameter, $|\pi_i|$ increases as d increases for positive values of θ and the lower the integer value of d, the faster $|\pi_i|$ converges to zero. The behavioural pattern of the invertibility parameter is well discussed in Makinde (2011).

Evaluation of ARIMA (p, d, q)

ARIMA(1, 0, 0)

$$y_t = c + \phi y_{t-1} + \varepsilon_t \quad (10)$$

If $c = 0$, then $\sum_{t=1}^{T} y_t y_{t-1} - \phi \sum_{t=1}^{T} y_{t-1}^2 = 0$, which implies that

$$\hat{\phi} = \frac{\sum_{t=1}^{T} y_t y_{t-1}}{\sum_{t=1}^{T} y_{t-1}^2}. \quad (11)$$

If c ≠ 0, equation (8) gives

$$\begin{pmatrix} n & \sum_{t=1}^{T} y_{t-1} \\ \sum_{t=1}^{T} y_{t-1} & \sum_{t=1}^{T} y_{t-1}^2 \end{pmatrix} \begin{pmatrix} c \\ \phi \end{pmatrix} = \begin{pmatrix} \sum_{t=1}^{T} y_t \\ \sum_{t=1}^{T} y_t y_{t-1} \end{pmatrix}$$

Using Crammer's rule to solve for c and ϕ results in

$$\hat{c} = \frac{\nabla_1}{\nabla} \text{ and } \hat{\phi} = \frac{\nabla_2}{\nabla} \quad (12)$$

where

$$\nabla = n \sum_{t=1}^{T} y_{t-1}^2 - \left[\sum_{t=1}^{T} y_{t-1} \right]^2,$$

$$\nabla_1 = \left[\sum_{t=1}^{T} y_t \right] \left[\sum_{t=1}^{T} y_{t-1}^2 \right] - \left[\sum_{t=1}^{T} y_{t-1} \right] \left[\sum_{t=1}^{T} y_t y_{t-1} \right]$$

and

$$\nabla_2 = n \sum_{t=1}^{T} y_t y_{t-1} - \left[\sum_{t=1}^{T} y_t \right] \left[\sum_{t=1}^{T} y_{t-1} \right].$$

ARIMA (1, 1, 0)

$$\Delta y_t = c + \phi \Delta y_{t-1} + \varepsilon_t \quad (13)$$

when $c = 0$, $\Delta y_t = \phi \Delta y_{t-1} + \varepsilon_t$ which gives

$$y_t - y_{t-1} = \phi(y_{t-1} - y_{t-2}) + \varepsilon_t$$

$$\hat{\phi} = \frac{\sum_{t=1}^{T} (y_t - y_{t-1})(y_{t-1} - y_{t-2})}{\sum_{t=1}^{T} (y_{t-1} - y_{t-2})^2} \quad (14)$$

When $c \neq 0,$ equation (11) gives

$$\begin{pmatrix} n & \sum_{t=1}^{T}\Delta y_{t-1} \\ \sum_{t=1}^{T}\Delta y_{t-1} & \sum_{t=1}^{T}\Delta y_{t-1}{}^{2} \end{pmatrix}\begin{pmatrix} c \\ \phi \end{pmatrix}=\begin{pmatrix} \sum_{t=1}^{T}\Delta y_{t} \\ \sum_{t=1}^{T}\Delta y_{t}\Delta y_{t-1} \end{pmatrix}$$

Using Crammer's rule, results in

$$\hat{c}=\frac{\nabla_1}{\nabla} \text{ and } \hat{\phi}=\frac{\nabla_2}{\nabla} \qquad (15)$$

where

$$\nabla = n\sum_{t=1}^{T}\Delta y_{t-1}{}^{2}-\left[\sum_{t=1}^{T}\Delta y_{t-1}\right]^{2},$$

$$\nabla_1=\left[\sum_{t=1}^{T}\Delta y_{t}\right]\left[\sum_{t=1}^{T}\Delta y_{t-1}{}^{2}\right]-\left[\sum_{t=1}^{T}\Delta y_{t-1}\right]\left[\sum_{t=1}^{T}\Delta y_{t}\Delta y_{t-1}\right]$$

$$\nabla_2 = n\sum_{t=1}^{T}\Delta y_{t}\Delta y_{t-1}-\left[\sum_{t=1}^{T}\Delta y_{t}\right]\left[\sum_{t=1}^{T}\Delta y_{t-1}\right]$$

ARIMA (2, 0, 0)

$$y_t = c+\phi_1 y_{t-1}+\phi_2 y_{t-2}+\varepsilon_t \qquad (16)$$

If $c=0,$ then

$$\hat{\phi}_1=\frac{\sum_{t=1}^{T}y_t y_{t-1}\sum_{t=1}^{T}y_{t-2}{}^{2}-\sum_{t=1}^{T}y_t y_{t-2}\sum_{t=1}^{T}y_{t-1}y_{t-2}}{\sum_{t=1}^{T}y_{t-1}{}^{2}\sum_{t=1}^{T}y_{t-2}{}^{2}-(\sum_{t=1}^{T}y_{t-1}y_{t-2})^{2}}$$

$$\hat{\phi}_2=\frac{\sum_{t=1}^{T}y_{t-1}{}^{2}\sum_{t=1}^{T}y_t y_{t-2}-\sum_{t=1}^{T}y_t y_{t-1}\sum_{t=1}^{T}y_{t-1}y_{t-2}}{\sum_{t=1}^{T}y_{t-1}{}^{2}\sum_{t=1}^{T}y_{t-2}{}^{2}-(\sum_{t=1}^{T}y_{t-1}y_{t-2})^{2}} \qquad (17)$$

If $c \neq 0,$ then

$$\begin{pmatrix} n & \sum_{t=1}^{T}y_{t-1} & \sum_{t=1}^{T}y_{t-2} \\ \sum_{t=1}^{T}y_{t-1} & \sum_{t=1}^{T}y_{t-1}{}^{2} & \sum_{t=1}^{T}y_{t-1}y_{t-2} \\ \sum_{t=1}^{T}y_{t-2} & \sum_{t=1}^{T}y_{t-1}y_{t-2} & \sum_{t=1}^{T}y_{t-2}{}^{2} \end{pmatrix}\begin{pmatrix} c \\ \phi_1 \\ \phi_2 \end{pmatrix}$$

$$=\begin{pmatrix} \sum_{t=1}^{T}y_t \\ \sum_{t=1}^{T}y_t y_{t-1} \\ \sum_{t=1}^{T}y_t y_{t-2} \end{pmatrix}$$

Using Crammer's rule, $c=\frac{\nabla_1}{\nabla}$, $\phi_1=\frac{\nabla_2}{\nabla}$ and $\phi_2=\frac{\nabla_3}{\nabla}$ where

$$\nabla=\begin{vmatrix} n & \sum_{t=1}^{T}y_{t-1} & \sum_{t=1}^{T}y_{t-2} \\ \sum_{t=1}^{T}y_{t-1} & \sum_{t=1}^{T}y_{t-1}{}^{2} & \sum_{t=1}^{T}y_{t-1}y_{t-2} \\ \sum_{t=1}^{T}y_{t-2} & \sum_{t=1}^{T}y_{t-1}y_{t-2} & \sum_{t=1}^{T}y_{t-2}{}^{2} \end{vmatrix},$$

$$\nabla_1=\begin{vmatrix} \sum_{t=1}^{T}y_t & \sum_{t=1}^{T}y_{t-1} & \sum_{t=1}^{T}y_{t-2} \\ \sum_{t=1}^{T}y_t y_{t-1} & \sum_{t=1}^{T}y_{t-1}{}^{2} & \sum_{t=1}^{T}y_{t-1}y_{t-2} \\ \sum_{t=1}^{T}y_t y_{t-2} & \sum_{t=1}^{T}y_{t-1}y_{t-2} & \sum_{t=1}^{T}y_{t-2}{}^{2} \end{vmatrix},$$

$$\nabla_2 = \begin{vmatrix} n & \sum_{t=1}^{T} y_t & \sum_{t=1}^{T} y_{t-2} \\ \sum_{t=1}^{T} y_{t-1} & \sum_{t=1}^{T} y_t y_{t-1} & \sum_{t=1}^{T} y_{t-1} y_{t-2} \\ \sum_{t=1}^{T} y_{t-2} & \sum_{t=1}^{T} y_t y_{t-2} & \sum_{t=1}^{T} y_{t-2}^{\,2} \end{vmatrix},$$

and

$$\nabla_3 = \begin{vmatrix} n & \sum_{t=1}^{T} y_{t-1} & \sum_{t=1}^{T} y_t \\ \sum_{t=1}^{T} y_{t-1} & \sum_{t=1}^{T} y_{t-1}^{\,2} & \sum_{t=1}^{T} y_t y_{t-1} \\ \sum_{t=1}^{T} y_{t-2} & \sum_{t=1}^{T} y_{t-1} y_{t-2} & \sum_{t=1}^{T} y_t y_{t-2} \end{vmatrix}.$$

ARIMA (2, 1, 0)

$$\Delta y_t = c + \phi_1 \Delta y_{t-1} + \phi_2 \Delta y_{t-2} + \varepsilon_t \qquad (18)$$

when, $c = 0$, $\Delta y_t = \phi_1 \Delta y_{t-1} + \phi_2 \Delta y_{t-2} + \varepsilon_t$ which is

$$y_t - y_{t-1} = \phi_1 (y_{t-1} - y_{t-2}) + \phi_2 (y_{t-2} - y_{t-3}) + \varepsilon_t .$$

The result is $\hat{\phi}_1 = \dfrac{\nabla_1}{\nabla}$ and $\hat{\phi}_2 = \dfrac{\nabla_2}{\nabla}$, where

$$\nabla = \sum_{t=1}^{T} (y_{t-1} - y_{t-2})^2 \sum_{t=1}^{T} (y_{t-2} - y_{t-3})^2 - \left[\sum_{t=1}^{T} (y_{t-1} - y_{t-2})(y_{t-2} - y_{t-3}) \right]^2$$

$$\nabla_1 = \sum_{t=1}^{T} (y_t - y_{t-1})(y_{t-1} - y_{t-2}) \sum_{t=1}^{T} (y_{t-2} - y_{t-3})^2 - \sum_{t=1}^{T} (y_t - y_{t-1})(y_{t-2} - y_{t-3}) \sum_{t=1}^{T} (y_{t-1} - y_{t-2})(y_{t-2} - y_{t-3})$$

and

$$\nabla_2 = \sum_{t=1}^{T} (y_{t-1} - y_{t-2})^2 \sum_{t=1}^{T} (y_t - y_{t-1})(y_{t-2} - y_{t-3}) - \left[\begin{array}{l} \sum_{t=1}^{T} (y_t - y_{t-1})(y_{t-1} - y_{t-2}) \\ * \sum_{t=1}^{T} (y_{t-1} - y_{t-2})(y_{t-2} - y_{t-3}) \end{array} \right]$$

ARIMA (P, 0, 0)

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \phi_3 y_{t-3} + \ldots + \phi_P y_{t-P} + \varepsilon_t \qquad (19)$$

given $c = 0$, $A\Psi = B$. That is

$$\begin{pmatrix} \sum_{t=1}^{T} y_{t-1}^{\,2} & \sum_{t=1}^{T} y_{t-2} y_{t-1} & \sum_{t=1}^{T} y_{t-3} y_{t-1} & \cdots & \sum_{t=1}^{T} y_{t-p} y_{t-1} \\ \vdots & & & \vdots & \vdots \\ \sum_{t=1}^{T} y_{t-1} y_{t-p} & \sum_{t=1}^{T} y_{t-2} y_{t-p} & \sum_{t=1}^{T} y_{t-3} y_{t-p} & \cdots & \sum_{t=1}^{T} y_{t-p}^{\,2} \end{pmatrix} \begin{pmatrix} \phi_1 \\ \vdots \\ \phi_p \end{pmatrix} = \begin{pmatrix} \sum_{t=1}^{T} y_t y_{t-1} \\ \vdots \\ \sum_{t=1}^{T} y_t y_{t-p} \end{pmatrix}$$

where A is $p \times p$ matrix and Ψ is a column matrix, that is $\Psi = \left(\phi_1 \;\; \phi_2 \;\; \phi_3 \ldots \phi_p \right)'$. B is a column matrix:

$$\begin{pmatrix} \sum_{t=1}^{T} y_t y_{t-1} \\ \vdots \\ \sum_{t=1}^{T} y_t y_{t-p} \end{pmatrix}.$$

The expression for each parameter $\phi_i, i = 1, 2, \ldots, p$ can thus be determined using Crammer's rule or the Gauss-Schidel method.

Given $c \neq 0,\ A\Psi = B,$ where $\Psi = \left(c, \phi_1\ \phi_2\ \ \phi_3 \ldots \phi_p\right)'$. That is,

$$\begin{pmatrix} n & \sum_{t=1}^{T} y_{t-1} & \sum_{t=1}^{T} y_{t-2} & \cdots & \sum_{t=1}^{T} y_{t-p} \\ & \vdots & & \vdots & \vdots \\ \sum_{t=1}^{T} y_{t-1}y_{t-p} & \sum_{t=1}^{T} y_{t-2}y_{t-p} & \sum_{t=1}^{T} y_{t-3}y_{t-p} & \cdots & \sum_{t=1}^{T} y_{t-p}^{2} \end{pmatrix} \begin{pmatrix} c \\ \phi_1 \\ \vdots \\ \phi_p \end{pmatrix}$$

$$= \begin{pmatrix} \sum_{t=1}^{T} y_t \\ \sum_{t=1}^{T} y_t y_{t-1} \\ \vdots \\ \sum_{t=1}^{T} y_t y_{t-p} \end{pmatrix}.$$

ARIMA (P, 1, 0)

$$\Delta y_t = c + \phi_1 \Delta y_{t-1} + \phi_2 \Delta y_{t-2} + \phi_3 \Delta y_{t-3} + \ldots + \phi_P \Delta y_{t-P} + \varepsilon_t \quad (20)$$

When, $c = 0$ (see Formula 1).

For the estimate of parameters in ARIMA (p, 0, 0) and (p, 1, 0), it is deduced that every term y_{t-j} in ARIMA (p, 0, 0) is replaced by $y_{t-j} - y_{t-j-1}$ in ARIMA (p, 1, 0). Also, Ψ is a p column matrix for $c = 0$, and Ψ is a $(p+1)$ column matrix for $c \neq 0$.

Results

Daily temperature data for the maximum daily temperature of Ondo, Nigeria and Zaira, Nigeria from January 1995 to November 2005 are used in this study. Stationarity of a series is determined by the use of a correlogram for describing both autocorrelation and partial autocorrelation functions for the series. The series is non-stationary, it is therefore differenced once (i.e., $d = 1$) to ensure stationarity. Figures 1 and 2 show the correlograms for the series after differencing each once (stationary at $d = 1$). Also, the residual terms (white noise process or innovation series (Bobba, et al., 2006)) are independently and identically distributed because the autocorrelation function at various lags hover around zero (see Figure 1) (Gujarati, 2004). Similarly, Figures 3a and 3b show that residuals are normally distributed, thus, $\varepsilon_t \sim iid\ N(0, \sigma^2)$.

Formula 1: ARIMA (P, 1, 0) when $c = 0$

$$\begin{pmatrix} \sum_{t=1}^{T}(y_{t-1} - y_{t-2})^2 & \sum_{t=1}^{T}(y_{t-1} - y_{t-2})(y_{t-2} - y_{t-3}) & \cdots & \sum_{t=1}^{T}(y_{t-1} - y_{t-2})(y_{t-p} - y_{t-p-1}) \\ & \vdots & \vdots & \vdots \\ \sum_{t=1}^{T}(y_{t-1} - y_{t-2})(y_{t-p} - y_{t-p-1}) & \sum_{t=1}^{T}(y_{t-2} - y_{t-3})(y_{t-p} - y_{t-p-1}) & \cdots & \sum_{t=1}^{T}(y_{t-p} - y_{t-p-1})^2 \end{pmatrix} \begin{pmatrix} \phi_1 \\ \vdots \\ \phi_p \end{pmatrix}$$

$$= \begin{pmatrix} \sum_{t=1}^{T}(y_t - y_{t-1})(y_{t-1} - y_{t-2}) \\ \vdots \\ \sum_{t=1}^{T}(y_t - y_{t-1})(y_{t-p} - y_{t-p-1}) \end{pmatrix}$$

Figure 1: Correlogram after Difference for Ondo, Nigeria

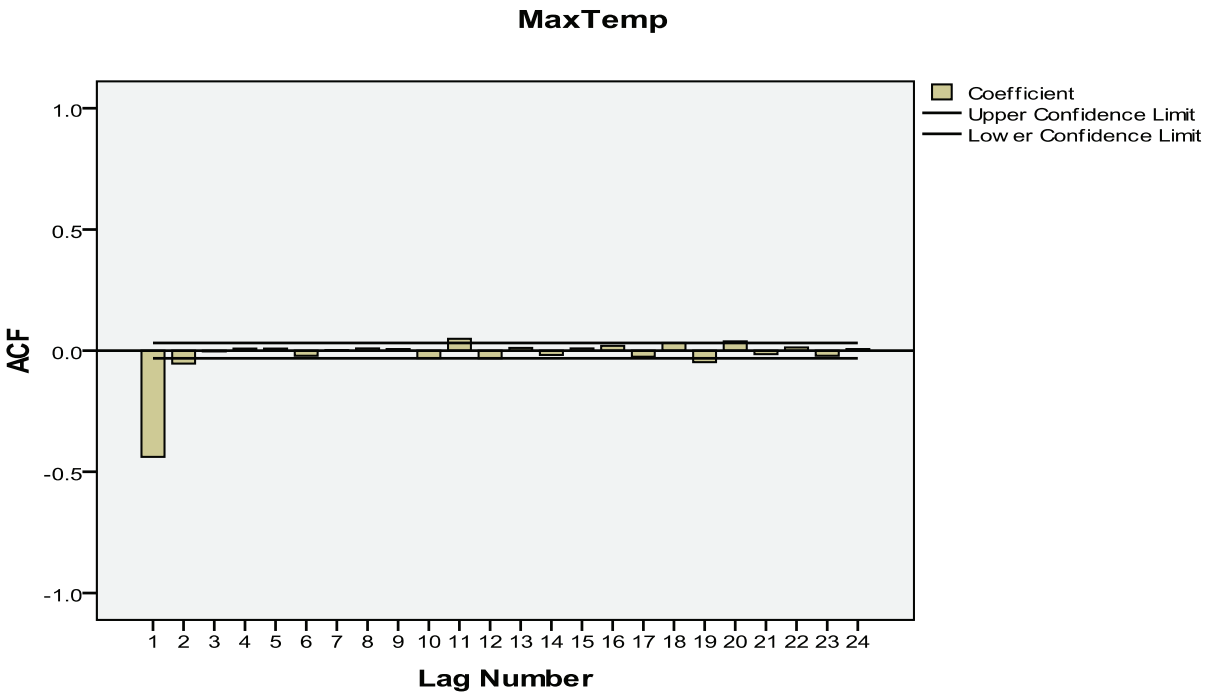


Figure 2: Correlogram after Difference for Zaira, Nigeria

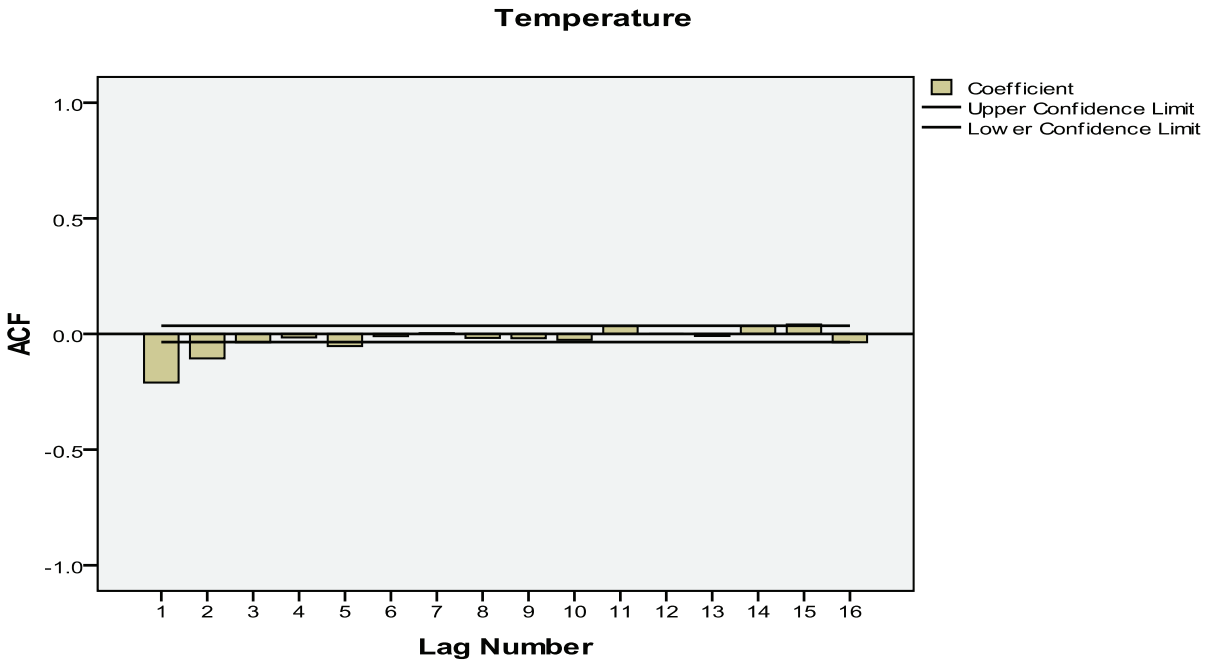


Figure 3(a): Histogram of Residuals for Ondo, Nigeria

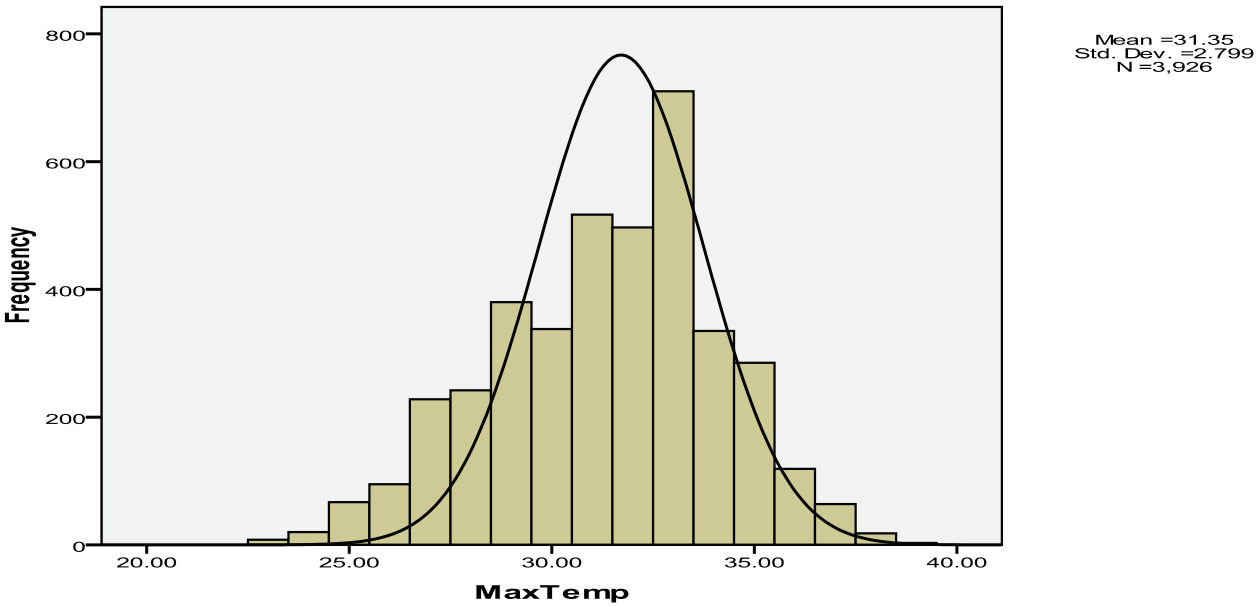


Figure 3(b): Histogram of Residuals for Zaira, Nigeria

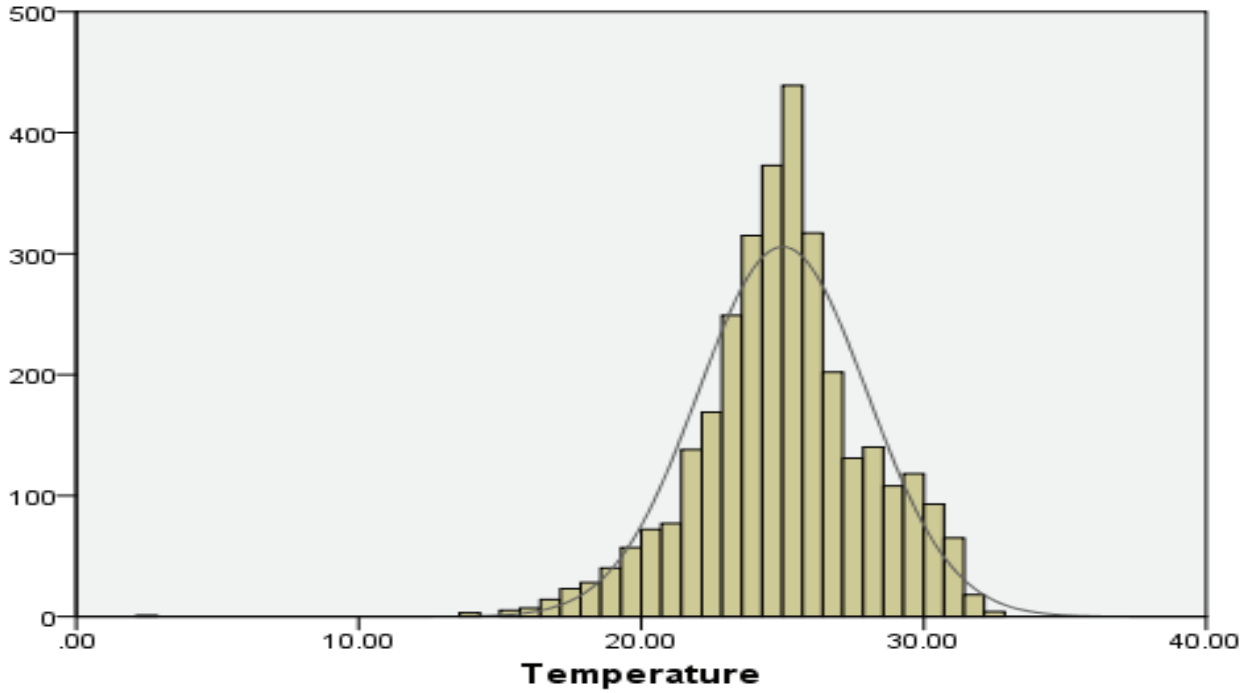


Table 1: Various ARIMA Model Fits with Normalized BIC Values

Model Selection Criteria	Normalized BIC Values for Model Fit							
	ARIMA (0,1,0)	ARIMA (0,1,1)	ARIMA (1,1,0)	ARIMA (1,1,1)	ARIMA (1,1,2)	ARIMA (2,1,0)	ARIMA (2,1,1)	ARIMA (2,1,2)
Ondo	1.354	0.917	1.144	0.915	0.917	1.049	0.917	0.918
Zaira	0.62	0.558	0.578	0.515	0.518	0.556	0.518	0.52

Table 2: Adequacy Test Results of the Model

Model Statistics						
City	Number of Predictors	Model Fit Statistics: Stationary R-Squared	Ljung-Box Q(18)			Number of Outliers
			Statistics	DF	Sig.	
Ondo	0	0.105	28.999	16	.024	0
Zaira	0	0.105	28.999	16	.024	0

Table 3: Parameter Estimates

		Ondo, Nigeria				Zaira, Nigeria			
		Estimate	SE	t	Sig.	Estimate	SE	t	Sig.
Maximum Temperature	Constant	0	0.005	0.052	0.959	0.00001	0.008	0.055	0.956
	AR Φ	0.087	0.02	4.391	0	0.55	0.028	19.724	0
	Difference	1				1			
	MA θ	0.833	0.011	75.779	0	0.836	0.018	45.597	0

Using a normalized BIC as the model selection criterion (that is, to test for goodness of fit) for various values of *p* and *q* (see Table 1), ARIMA (*1, 1, 1*) has the least normalized BIC value, which equals 0.915 for Ondo and 0.515 for Zaira. Hence, the ARIMA (*1, 1, 1*) is considered the best model for the maximum daily temperature data. for both Ondo and Zaira. To test for the adequacy of the model, the Ljung-Box Statistic is used to test the randomness of residuals. The *p*-values of the Ljung-Box Statistic at various lags (in ACF and PACF) are less than 0.05; this shows that the data are random. The Ljung-Box Statistic for the model is 28.999 with a *p*-value = 0.024, this establishes that the model is adequate (see Table 2).

Table 3 presents the estimates of parameters of the ARIMA(*1, 1, 1*) model for Ondo and Zaira. The model for Ondo is:

$$y_t = 0.00001 + 0.087 y_{t-1} + \varepsilon_t + 0.836\varepsilon_{t-1}.$$

that is, $c = 0.00001$, $\phi = 0.550$, and $\theta = 0.836$ (see Table 5). Also, in testing for significance of the parameter estimates, Table 3 shows the t-statistics for the parameter estimates of the model. It is shown that a *c* with $t = 0.052$ and a *p*-value = 0.959 is not significantly different from zero; thus, *c* is redundant.

To improve the model result, *c* was removed because it is redundant. This removal had no effect on the estimates of other parameters or on the Ljung-Box value of the model; rather it results in a smaller normalized BIC value (=0.912). Hence, the optimal model for maximum temperature of Ondo is:

$$y_t = 0.087 y_{t-1} + \varepsilon_t + 0.833\varepsilon_{t-1}.$$

The invertibility behavior of the optimal model for Ondo, Nigeria is $\varepsilon_t = \sum_{i=0}^{\infty} \pi_i y_{t-i}$, because *c* is redundant, where

$$\pi_i = \begin{cases} 1, & i = 0 \\ -(1.920), & i = 1 \\ (-1)^i \left[0.920(0.833^{i-1} + 0.833^{i-2}) \right], & i = 2,3,4,\ldots \end{cases}$$

The model for Zaira is

$$y_t = 0.00001 + 0.550 y_{t-1} + \varepsilon_t + 0.836\varepsilon_{t-1},$$

that is, $c = 0.00001$, $\phi = 0.550$, and $\theta = 0.836$ (see Table 3). Also, in testing for significance of the estimates of parameters, results show that a *c* with $t = 0.055$ and a *p*-value = 0.956 is not significantly different from zero (see Table 3). Hence, *c* is redundant.

To improve the model result, *c* was removed; this had no effect on the estimates of other parameters or on the Ljung-Box value of the model, instead, it results in a smaller normalized BIC value (=0.512). Hence, the optimal model for the maximum temperature of Ondo is:

$$y_t = 0.550 y_{t-1} + \varepsilon_t + 0.836\varepsilon_{t-1}.$$

The invertibility behavior of the optimal model for Zaira is $\varepsilon_t = \sum_{i=0}^{\infty} \pi_i y_{t-i}$, because *c* is redundant, where

$$\pi_i = \begin{cases} 1, & i = 0 \\ -(2.386), & i = 1 \\ (-1)^i \left[1.386(0.836^{i-1} + 0.836^{i-2}) \right], & i = 2,3,4,\ldots \end{cases}$$

Conclusion

Autoregressive Integrated Moving Average (ARIMA) processes of various orders are presented with the goal of identifying an optimal model from a class of models. ARIMA (*p, d, q*) model is formulated for daily maximum temperature data of Ondo, Nigeria and Zaira, Nigeria from 1995 to 2005. A normalized Bayesian Information Criteria (BIC) is used to

measure performance of the models. ARIMA (*1, 1, 1*) is optimal and adequate for modeling the daily maximum temperatures because it has the least normalized BIC, parameters of the model are estimated and the redundant variable is removed. The behavioral pattern of the optimal model for each of the cities is reported.

References

Bobba, A. G., Rudra, R. P., & Diiwu, J. Y. (2006). A stochastic model for identification of trends in observed hydrological and meteorological data due to climate change in watersheds. *Journal of Environmental Hydrology*, *14*(*10*), 1-11.

Dobre, I., & Alexandru, A. A. (2008). Modeling unemployment rate using Box-Jenkins procedure. *Journal of Applied Quantitative Methods*, *3*(*2*), 156-166.

Fasoranbaku, O. A. & Makinde, O. S. (2011). Behavioural pattern of causality parameter of autoregressive moving average model. *Journal of Nigerian Association of Mathematical Physics*, *19*, 583-590.

Gujarati, D. N. (2004). *Basic econometrics*, *4th Ed*. McGraw-Hill: The McGraw-Hill Companies.

Hamilton, J. D. (1994). *Time series analysis*. Princeton, NJ: Princeton University Press.

Ljung, G. M., & Box, G. E. P. (1978). On a measure of a lack of fit in time series models. *Biometrika*, *65*(*2*), 297-303

Lorenz, E. N. (1963). Deterministic nonperiodic flow. *Journal of Atmospheric Science*, *20*, 130-141.

Makinde, O. S. (2011). Behavioural Pattern of Invertibility Parameter of ARIMA Model. *Journal of Nigerian Association of Mathematical Physics*, *19*, 591-606.

Meko D. M. (1981) Application of Box Jenkins Methods of Time Series Analysis to reconstruction of drought from tree rings. *Ph.D. Dissertation, University of Arizona*.

Nelson, C. R. (1973) *Applied time series analysis*. San Francisco, CA: Holden-Day.

Ojo, J. F. (2008). Identification of optimal models in higher order of integrated autoregressive models and autoregressive integrated moving average models in the presence of 2^k-1 subsets. *Journal of Modern Mathematics and Statistics*, *2*(*1*), 7-11.

Ojo, J. F. (2009). On the estimation and performance of subset autoregressive integrated moving average models. *European Journal of Scientific Research*, *28*(*2*), 287-293.

Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*(*2*), 461-464.

Shittu, O. I., & Yaya, O. S. (2009). Measuring forecast performance of ARIMA and ARFIMA: An application to US Dollar/UK Pound foreign exchange rate. *European Journal of Scientific Research*, *32*(*2*), 167-176.

Tuan, P. (1978). Establishment of parameters in the ARIMA model when the characteristic polynomial of MA operator has a unit zero. *Annals of Statistics*, *6*(*6*), 1369-1389.

Journal of Modern Applied Statistical Methods
November 2011, Vol. 10, No. 2, 730-740

1538 – 9472/11/$95.00

LQ-Moments for Regional Flood Frequency Analysis: A Case Study for the North-Bank Region of the Brahmaputra River, India

Abhijit Bhuyan Munindra Borah
Tezpur University,
Napaam, India

The LQ-moment proposed by Mudholkar, et al. (1998) is used for regional flood frequency analysis of the North-Bank region of the river Brahmaputra, India. Five probability distributions are used for the LQ-moment: generalized extreme value (GEV), generalized logistic (GLO) and generalized Pareto (GPA), lognormal (LN3) and Pearson Type III (PE3). The same regional frequency analysis procedure proposed by Hosking (1990) for the L-moment is used for the LQ-moment. Based on the LQ-moment ratio diagram and $\left|Z_i^{dist}\right|$ -statistic criteria, the PE3 distribution is identified as the robust distribution for the study area. For estimation of floods of various return periods for both gauged and ungauged catchments of the study area, regional flood frequency relationships are developed using the LQ-moment based PE3 distribution.

Key words: Regional flood frequency analysis, PE3 distribution, LQ-moment ratio diagram.

Introduction

Hosking (1990) introduced the concept of L-moment parameter estimation methods for regional frequency analysis. The performance of a particular model depends on the accuracy of the estimation of the parameters. Many parameter estimation methods are described in statistical literature. The unbiased estimation of parameters depends mainly on the parameter estimation method used and the data availability. Regional frequency analysis overcomes the difficulties arising from at-site frequency analysis. In many countries, the L-moments procedure for regional flood frequency analysis has been used and various researches are ongoing. In India, L-moments based regional flood frequency analysis was conducted by Paradia, et al. (1998) and Kumar et al. (1999, 2003 and 2005) to develop a flood frequency relationship for both gauged and ungauged catchments for different regions. Additionally, some recent application of regional flood frequency analysis include: Atiem and Harmancioglu (2006), Modarres (2007), Saf (2008) and Hussain, et al. (2008).

Kumar, et.al. (2005) used L-moments to develop a regional flood frequency relationship for both gauged and ungauged catchments of the North Brahmaputra region of India. Mudholkar, et al. (1998) introduced the concept of LQ-moment analogs of L-moments of Hosking (1990). LQ-moments are linear functions of the medians, trimeans, or Gastwirth's location estimators of the distributions of certain order statistics and reduce to weighted averages for certain population quantiles. LQ-moments are often easier to evaluate and estimate than L-moments and, in general, behave similarly to the L-moments when the latter exist. (Modhulkar, et al., 1998). Modhulkar, et al. (1998) used an LQ-moment in the context of generalized extreme value distribution for flood frequency analysis of the river Blackstote and Feather. Zin Wan, et al. (2008) used LQ-moments to determine the best fitting probability distribution for annual maximum rainfall in Peninsular Malaysia.

Various studies have found that LQ-moments are widely used to study at-site flood

Abhijit Bhuyan is a Ph.D. student in the Department of Mathematical Sciences. Email: abhijit@tezu.ernet.in. Munindra Borah is a Professor in the Department of Mathematical Sciences. Email: mborah@tezu.ernet.in.

frequency analysis and at-site rainfall frequency analysis in different countries of the world. But in the case of flood frequency analysis, data availability is difficult for estimating floods for desired return periods. Therefore, this study uses regional frequency analysis as an alternative to at-site frequency analysis based on LQ-moments. The linear quantile estimator as a sample quantile estimator and trimean functional as quick estimator are also used in this study of regional flood frequency analysis.

Five probability distributions that are generally used for regional flood frequency analysis by using L-moments are used in this study: generalized extreme value (GEV), generalized Pareto (GPA), generalized normal (GNO), generalized logistic (GLO) and Pearson Type III (PE3). This study employs the LQ-moment as a parameter estimation method for regional flood frequency analysis of nine sites in the North-Bank region of the Brahmaputra River in India. The same procedure for regional frequency analysis for L-moments proposed by Hosking (1990) is used for LQ-moment. The relationship between LQ-skewness and LQ-kurtosis has been developed for each of the probability distributions used for this study.

LQ-Moments

Let $X_1, X_2, ..., X_n$ be a sample from a continuous distribution function $F_X(\cdot)$ with quantile function $Q_X(u) = F_X^{-1}(u)$. If $X_{1:n} \le X_{2:n} \le .... \le X_{n:n}$ denote the order statistics, then the r[th] LQ-moments ζ_r of X proposed by Mudholkar, et al. (1998) are given by

$$\zeta_r = r^{-1}\sum_{k=0}^{r-1}(-1)^k \binom{r-1}{k} \tau_{p,\alpha}(X_{r-k:r}), \quad r = 1,\ 2,\ ... \tag{1}$$

where $0 \le \alpha \le 1/2,\ 0 \le p \le 1/2$, and

$$\tau_{p,\alpha}(X_{r-k:r}) = pQ_{X_{r-k:r}}(\alpha) + (1-2p)Q_{X_{r-k:r}}(1/2) + pQ_{X_{r-k:r}}(1-\alpha). \tag{2}$$

The linear combination $\tau_{p,\alpha}$ is a quick measure of the location of the sampling distribution of order statistic $X_{r-k:r}$. With appropriate combinations of α and p, estimators for $\tau_{p,\alpha}(\cdot)$ can be found which are functions of commonly used estimators such as median, trimean and Gastwirth. This study considers the trimean-based estimator, defined as:

$$Q_{X_{r-k:r}}\left(\frac{1}{4}\right)/4 + Q_{X_{r-k:r}}\left(\frac{1}{2}\right)/2 + Q_{X_{r-k:r}}\left(\frac{3}{4}\right)/4.$$

The first four LQ-moments of the random variable X are given by:

$$\zeta_1 = \tau_{p,\alpha}(X),$$

$$\zeta_2 = \frac{1}{2}[\tau_{p,\alpha}(X_{2:2}) - \tau_{p,\alpha}(X_{1:2})],$$

$$\zeta_3 = \frac{1}{3}[\tau_{p,\alpha}(X_{3:3}) - 2\tau_{p,\alpha}(X_{2:3}) + \tau_{p,\alpha}(X_{1:3})],$$

$$\zeta_4 = \frac{1}{4}\begin{bmatrix}\tau_{p,\alpha}(X_{4:4}) - 3\tau_{p,\alpha}(X_{3:4}) \\ +3\tau_{p,\alpha}(X_{2:4}) - \tau_{p,\alpha}(X_{1:4})\end{bmatrix}.$$

The LQ-CV, LQ-skewness and LQ-kurtosis are defined by

$$\eta = \varsigma_2/\varsigma_1,\ \eta_3 = \zeta_3/\zeta_2$$

and

$$\eta_4 = \zeta_4/\zeta_2.$$

If $Q_X(.) = F_X^{-1}(.)$ is the quantile function of the random variable X then the quick location measure (2) defined by Modhulkar, et al. (1998) is

$$\tau_{p,\alpha}(X_{r-k:r}) = pQ_X[B^{-1}_{r-k:r}(\alpha)] + (1-2p)Q_X[B^{-1}_{r-k:r}(1/2)] + pQ_X[B^{-1}_{r-k:r}(1-\alpha)]$$

where $B^{-1}_{r-k:r}(\alpha)$ denotes the corresponding α^{th} quantile of a beta random variable with parameters $r-k$ and $k+1$.

Sample Estimates of LQ-Moments

Modhular, et al. (1998) defines sample estimates of LQ-moments as follows. Let $X_{1:n} \le X_{2:n} \le \ldots \le X_{n:n}$ denote the sample order statistics then the quantile estimator of $Q_X(u)$ is given by

$$\hat{Q}_X(u) = (1-\varepsilon)X_{[n'u]:n} + \varepsilon X_{[n'u]+1:n},$$

where $\varepsilon = n'u - [n'u]$ and $n' = n+1$. Thus for samples of size n, the r[th] sample LQ-moment is given by

$$\hat{\zeta}_r = r^{-1}\sum_{k=0}^{r-1}(-1)^k \binom{r-1}{k} \hat{\tau}_{p,\alpha}(X_{r-k:r}),$$

where $\hat{\tau}_{p,\alpha}(X_{r-k:r})$, the quick estimator of the location for the distribution of $X_{r-k:r}$ in a random sample of size r.

The first four sample LQ-moments are given by

$$\hat{\zeta}_1 = \hat{\tau}_{p,\alpha}(X)$$

$$\hat{\zeta}_2 = \frac{1}{2}[\hat{\tau}_{p,\alpha}(X_{2:2}) - \hat{\tau}_{p,\alpha}(X_{1:2})]$$

$$\hat{\zeta}_3 = \frac{1}{3}\left[\begin{matrix}\hat{\tau}_{p,\alpha}(X_{3:3}) - 2\hat{\tau}_{p,\alpha}(X_{2:3}) \\ +\hat{\tau}_{p,\alpha}(X_{1:3})\end{matrix}\right]$$

$$\hat{\zeta}_4 = \frac{1}{4}\left[\begin{matrix}\hat{\tau}_{p,\alpha}(X_{4:4}) - 3\hat{\tau}_{p,\alpha}(X_{3:4}) \\ +3\hat{\tau}_{p,\alpha}(X_{2:4}) - \hat{\tau}_{p,\alpha}(X_{1:4})\end{matrix}\right],$$

where, the quick estimator $\hat{\tau}_{p,\alpha}(X_{r-k:r})$ of the location of the order statistic $X_{r-k:r}$ is given by

$$\begin{aligned}\hat{\tau}_{p,\alpha}(X_{r-k:r}) &= p\hat{Q}_{X_{r-k:r}}(\alpha) + (1-2p)\hat{Q}_{X_{r-k:r}}(1/2) + p\hat{Q}_{X_{r-k:r}}(1-\alpha) \\ &= p\hat{Q}_X[B^{-1}_{r-k:r}(\alpha)] + (1-2p)\hat{Q}_X[B^{-1}_{r-k:r}(1/2)] + p\hat{Q}_X[B^{-1}_{r-k:r}(1-\alpha)]\end{aligned}$$

$0 \le \alpha \le 1/2,\ 0 \le p \le 1/2, B^{-1}_{r-k:r}(\alpha)$, is the α^{th} quantile of beta random variable with parameters $r-k$ and $k+1$, and $\hat{Q}_X(.)$ denotes the linear interpolation estimator shown above.

Probability Distributions and Parameters Based on Trimean Function: Generalized Extreme Value Distribution (Modhulkar, et al., 1998)

The probability distribution function (PDF) for the generalized extreme value distribution is defined as:

$$f(x) = \frac{1}{\alpha}\left[1-k\left(\frac{x-\xi}{\alpha}\right)\right]^{\frac{1}{k}-1} \exp\left[-\left\{1-k\left(\frac{x-\xi}{\alpha}\right)\right\}^{\frac{1}{k}}\right].$$

Its quantile function is given by

$$Q(u) = \xi + \alpha Q_0(u)$$

where

$$\begin{aligned}Q_0(u) &= [1-(-\log u)^k]/k, \quad k \ne 0 \\ &= -\log(-\log u), \qquad k = 0.\end{aligned}$$

The shape parameter k can be estimated with good accuracy by using the approximation equation

$$k = 0.2985 - 2.0234\eta_3 + 0.3732\eta_3^2 - 0.1429\eta_3^3 + 0.0449\eta_3^4.$$

The estimates of the parameters ξ and α are then given by:

$$\xi = \varsigma_1 - \alpha[Q_0(1/4)/4 + Q_0(1/2)/2 + Q_0(3/4)/4]$$

and

$$\alpha = 8\varsigma_2 / \begin{bmatrix} 2Q_0(0.707) - 2Q_0(0.293) \\ +Q_0(0.866) - Q_0(0.134) \end{bmatrix}.$$

Probability Distributions and Parameters Based on Trimean Function: Generalized Pareto Distribution

The PDF of the generalized Pareto distribution is:

$$f(x) = \frac{1}{\alpha}\left[1 - k\left(\frac{x-\xi}{\alpha}\right)\right]^{\frac{1}{k}-1}.$$

Its quantile function is given by

$$Q(u) = \xi + \alpha Q_0(u)$$

where

$$Q_0(u) = [1-(1-u)^k]/k, \quad k \neq 0$$
$$= -\log(1-u), \qquad k = 0.$$

The shape parameter k can be estimated with good accuracy by using the approximation equation

$$k = 0.9998 - 3.4965\eta_3 + 1.4681\eta_3^2 - 0.6243\eta_3^3 + 0.1535\eta_3^4$$

The estimates of the parameters ξ and α are then given by

$$\xi = \varsigma_1 - \alpha[Q_0(1/4)/4 + Q_0(1/2)/2 + Q_0(3/4)/4]$$

$$\alpha = 8\varsigma_2 / \begin{bmatrix} 2Q_0(0.707) - 2Q_0(0.293) \\ +Q_0(0.866) - Q_0(0.134) \end{bmatrix}.$$

Probability Distributions and Parameters Based on Trimean Function: Generalized Logistic Distributions

The PDF of the generalized logistic distribution is given by Rao and Hamed (2000) as

$$f(x) = \frac{1}{\alpha}\left[1 - k\left(\frac{x-\xi}{\alpha}\right)\right]^{\frac{1}{k}-1}\left[1 + \left\{1 - k\left(\frac{x-\xi}{\alpha}\right)\right\}^{\frac{1}{k}}\right]^{-2}.$$

Its quantile function is given by

$$Q(u) = \xi + \alpha Q_0(u)$$

where

$$Q_0(u) = [1-\{(1-u)/u\}^k]/k, \quad k \neq 0$$
$$= -\log\{(1-u)/u\}, \qquad k = 0.$$

The shape parameter k can be estimated with good accuracy by using the approximation equation

$$k = -1.3328\eta_3 - 0.0286\eta_3^3 + 0.0166\eta_3^5.$$

The estimates of the parameters ξ and α are then given by

$$\xi = \varsigma_1 - \alpha[Q_0(1/4)/4 + Q_0(1/2)/2 + Q_0(3/4)/4]$$

and

$$\alpha = 8\varsigma_2 / \begin{bmatrix} 2Q_0(0.707) - 2Q_0(0.293) \\ +Q_0(0.866) - Q_0(0.134) \end{bmatrix}.$$

Probability Distributions and Parameters Based on Trimean Function: Generalized Lognormal Distribution

The PDF of the generalized lognormal distribution is

$$f(x) = \frac{\exp^{-\log\{1-k(x-\xi)/\alpha\}-\frac{1}{2}[-\frac{1}{k}\log\{1-k(x-\xi)/\alpha)\}]^2}}{\alpha\sqrt{2\pi}}.$$

Its cumulative distribution function is

$$F(x) = \Phi\left[\frac{\{\log(x-\varsigma)-\mu\}}{\sigma}\right]$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution, and its quantile function is given by

$$Q(u) = \varsigma + \exp(\mu)Q_0(u)$$

where

$$Q_0(u) = \exp[\sigma\Phi^{-1}(u)]$$

and $\Phi^{-1}(.)$ has a standard normal distribution with mean zero and unit variance. The σ can be approximated by the

$$\sigma = 2.3284\eta_3 - 0.0002\eta_3^2 + 0.1220\eta_3^3 + 0.0009\eta_3^4 - 0.0332\eta_3^5.$$

Estimates of the parameters ς and $\exp(\mu)$ are then given by

$$\varsigma = \varsigma_1 - \exp(\mu)[Q_0\left(\frac{1}{4}\right)/4 + Q_0\left(\frac{1}{2}\right)/2 + Q_0\left(\frac{3}{4}\right)/4]$$

$$\mu = 8\varsigma_2 / \begin{bmatrix} 2Q_0(0.707) - 2Q_0(0.293) \\ +Q_0(0.866) - Q_0(0.134) \end{bmatrix}$$

The parameters k, α and ξ can be obtained from the relation given below after determining the parameter values ς, μ and σ for the standard cumulative lognormal distribution. The relations between the parameter are $k = -\sigma$, $\alpha = \sigma e^{\mu}$ and $\xi = \varsigma + e^{\mu}$.

Probability Distributions and Parameters Based on Trimean Function: Pearson Type III Distribution (PE3)

The PDF of the Pearson Type III distribution is given by

$$f(x) = \frac{(x-\xi)^{\beta-1}e^{-(x-\xi)/\beta}}{\alpha^{\beta}\Gamma(\beta)}.$$

The cumulative distribution function is given as

$$F(x) = \frac{1}{\alpha^{\beta}\Gamma(\beta)}\int_{\xi}^{x}(x-\xi)^{\beta-1}e^{-(x-\xi)/\beta}dx,$$

and the quantile function can be given as

$$Q(u) = \xi + \alpha Q_0(u)$$

where

$$Q_0(u) = \beta\left[1 - \frac{1}{9\beta} + \Phi^{-1}(u)\sqrt{\frac{1}{9\beta}}\right]^3.$$

The location (μ), scale (σ) and shape (k) can be represented in terms of α, β and ξ as: $\beta = \frac{4}{k^2}$, $\alpha = \frac{1}{2}\sigma|k|$ and $\xi = \mu - 2\sigma/k$.

The regression equation developed for estimating the shape parameter k in terms of LQ-skewness (η_3) is now given as

$$k = 6.9839\eta_3 + 0.0001\eta_3^2 - 6.6634\eta_3^3 - 0.0035\eta_3^4,$$

and the estimates of the parameters ξ and α are then given by

$$\xi = \varsigma_1 - \alpha[Q_0(1/4)/4 + Q_0(1/2)/2 + Q_0(3/4)/4]$$

and

$$\alpha = 8\varsigma_2 / \begin{bmatrix} 2Q_0(0.707) - 2Q_0(0.293) \\ +Q_0(0.866) - Q_0(0.134) \end{bmatrix}.$$

Relationship between LQ-Skewness and LQ-Kurtosis based on Trimean Functionals

The relationship between η_3 and η_4 are developed for the probability distribution used in this study are given as follows:

$$\eta_4^{GEV} = 0.1080 + 0.1130\eta_3 + 0.8178\eta_3^2 - 0.0314\eta_3^3 - 0.0103\eta_3^4 - 0.0015\eta_3^5 + 0.0069\eta_3^6 - 0.0037\eta_3^7$$

$$\eta_4^{GLO} = 0.1585 + 0.8190\eta_3^2 - 0.0117\eta_3^4 - 0.0045\eta_3^6$$

$$\eta_4^{GPA} = -0.0019 + 0.2228\eta_3 + 0.8606\eta_3^2 - 0.0618\eta_3^3 - 0.0590\eta_3^4 + 0.0501\eta_3^5 + 0.0059\eta_3^6 - 0.0160\eta_3^7$$

$$\eta_4^{LN3} = 0.1201 + 0.7934\eta_3^2 - 0.0001\eta_3^3 - 0.0064\eta_3^4 + 0.0005\eta_3^5 - 0.0059\eta_3^6$$

$$\eta_4^{PE3} = 0.1227 - 0.0007\eta_3 + 0.4179\eta_3^2 + 0.0019\eta_3^3 - 0.5133\eta_3^4$$

Methodology

Study Area and Data Availability

Regional flood frequency analysis is carried out for North Bank region of the Brahmaputra River of India. The Brahmaputra River basin extends over an area of 580,000 km^2 and lies in Tibet, Bhutan, India and Bangladesh. The drainage area of the basin lying in India is 194,413 km^2, which forms nearly 5.9% of the total geographical area of the country. The mean annual rainfall over the basin (excluding Tibet and Bhutan) is approximately 2,300 mm. Annual maximum peak flood data for nine stream flow gauging sites lying in the North Bank region of the Brahmaputra River and varying between 11-36 years in record length were used in this study.

Steps in Regional Flood Frequency Analysis

The steps involved in the regional flood frequency analysis by L-moments proposed by Hosking and Wallis (1997) are:

1. Screening of the data;
2. Formation of homogeneous region;
3. Selection of appropriate distribution;
4. Estimation of parameters of the probability distribution; and
5. Development of regional flood frequency relationship for gauged and ungauged catchments of the region.

This procedure has been applied for LQ-moment for the study area described.

Data Screening

Hosking and Wallis (1997) proposed a discordancy measure (D_i) based on L-moments, to recognize sites that are grossly discordant with the group as a whole. The discordancy measure (D_i) for the LQ-moment is defined as if there are N sites in the group. Let $u_i = [\eta^{(i)}\ \eta_3^{(i)}\ \eta_4^{(i)}]^T$ be a vector containing the sample LQ-moment ratios η, η_3 and η_4 for site *i*, and T denote transposition of a vector or matrix. Let

$$\bar{u} = N^{-1}\sum_{i=1}^{N} u_i ,$$

be the (unweighted) group average. The matrix of sums of squares and cross product is then defined as:

$$S = \sum_{i=1}^{N} (u_i - \bar{u})(u_i - \bar{u})^T ,$$

and the discordancy measure for site i is defined as:

$$D_i = \frac{1}{3} N (u_i - \bar{u})^T S^{-1} (u_i - \bar{u}).$$

Site *i* is declared to be discordant if D_i is greater than the critical value of the discordancy statistic D_i given in a tabular form for the L-moment by Hosking and Wallis (1997). Based on such

discordancy measures for LQ-moment, no discordance site was found for this study region. The discordancy measure, site names, sample sizes and LQ-moments are shown in Table 1.

Regional Homogeneity

The procedure proposed by Hosking and Wallis (1997) for L-moments, with required modification for an LQ-moment was used to test for regional homogeneity. The regional average LQ-CV, LQ-skewness and LQ-kurtosis, weighted proportional to the sites' record length were calculated and, similar to the regional average mean considered as 1 in the L-moment method, the regional average first LQ-moment ratio is also considered as 1. For this the LQ-moments and the parameters of the Kappa distribution based on the Trimean function have been developed and fit the developed Kappa distribution to the regional average LQ-moment ratios for 500 simulations. The values of heterogeneity measure computed by carrying out 500 simulations using the Kappa distribution based on the data for the 9 sites are provided in Table 2. Based on the heterogeneity measure the 9 site study area was found to be homogeneous.

Table 1: North Brahmaputra Region Site Information, Sample Statistics and Discordancy Measures

Site No.	Site Name	Sample Size	Catchment Area (km^2)	$\hat{\zeta}_1$	LQ-CV	LQ-Skewness	LQ-Kurtosis	D_i
1	Monas	17	30,100	5965.56	0.1739	0.1437	0.1008	0.25
2	Nonai	11	148	91.32	0.2159	0.1580	0.2021	0.14
3	Borolia	15	310	194.22	0.2540	-0.0345	0.1656	0.88
4	Dhansin	21	530	1275.50	0.1715	0.2039	0.3430	1.74
5	Pachnoi	22	198	196.82	0.2930	0.2915	0.1767	2.06
6	Jiabharali	36	11,000	4015.77	0.2607	0.0856	0.0412	0.54
7	Subansiri	27	25,886	8498.75	0.1777	0.2060	0.1198	0.36
8	Beki	13	1,331	748.60	0.2957	-0.0512	0.1490	1.26
9	Sankush	12	9,799	1865.99	0.1418	0.0703	-0.0942	1.76

Table 2: Heterogeneity Measure Based on LQ-Moment

Site No.	Heterogeneity Measures	Values
1	Heterogeneity Measure H(1) (a) Observed standard deviation of group LQ-CV (b) Simulated mean of standard deviation of group LQ-CV (c) Simulated standard deviation of standard deviation of group LQ-CV (d) Standardized test value H(1)	 0.0522 0.0457 0.0108 0.6000
2	Heterogeneity Measure H(2) (a) Observed average of LQ-CV/LQ-Skewness distance (b) Simulated mean of average LQ-CV/LQ-Skewness distance (c) Simulated standard deviation of average LQ-CV/LQ-Skewness distance (d) Standardized test value H(2)	 0.1015 0.1363 0.0327 -1.0600
3	Heterogeneity Measure H(3) (a) Observed average of LQ-Skewness/LQ-Kurtosis distance (b) Simulated mean of average LQ-Skewness/LQ-Kurtosis distance (c) Simulated standard deviation of average LQ-Skewness/LQ-Kurtosis distance (d) Standardized test value H(3)	 0.1331 0.1953 0.0403 -1.5400

Goodness-of-Fit Measure: $\left|Z_i^{dist}\right|$ Statistic Criteria

The same $\left|Z_i^{dist}\right|$-statistic criteria for the L-moment proposed by Hosking and Wallis (1997) was used as the goodness-of-fit measure for the LQ-moment to select the best fit distribution for the study region. The Z_i^{dist} statistic for the various three parameter distributions is shown in Table 3.

Table 3: Z_i^{dist} Statistic for Various Distributions for the Study Area

Distribution	Z_i^{dist} -statistic
GEV	0.77
LN3	0.71
GLO	1.38
GPA	-0.87
PE3	0.64

It may be observed from Table 3 that $\left|Z_i^{dist}\right|$-statistic values of all the five distributions are less than the critical value 1.64. Further, the $\left|Z_i^{dist}\right|$-statistic is found to be the lowest for PE3 distribution than all other distribution used for this study. Thus, the $\left|Z_i^{dist}\right|$-statistic criteria for the LQ-moment identifies the PE3 distribution as the best fitting distribution for the study region.

Goodness-of-Fit Measure: LQ-Moment Ratio Diagram

The LQ-moment ratio diagram is another goodness-of-fit measure for identifying the best fitting distribution for the study region. The relationships, given above between η_3 and η_4 for the five distributions are used to draw the theoretical curves in the LQ-moment ratio diagram. It can be observed from the LQ-moment ratio diagram (see Figure 1) that the regional values of LQ-skewness ($\eta_3 = 0.1332$) and LQ-kurtosis ($\eta_4 = 0.1324$) lie closest to PE3 distribution, thus the $\left|Z_i^{dist}\right|$-statistic criteria as well as the LQ-moment ratio diagram show that the PE3 distribution is the best fitting distribution for the study region.

Parameters and Quantile Estimates for the Region

The regional parameters and quantiles for the various distributions are given in Tables 4 and 5 respectively.

Regional Flood Frequency Relationship Based on LQ-Moments: Gauged Catchments

The regional flood frequency relationship for gauged catchments was developed, by using the identified best fitting distribution for the study area. The PE3 distribution was identified as the best fitting distribution for the study region in LQ-moment; thus, the relationship was developed using the PE3 distribution. The cumulative density function of the three parameter PE3 distribution as parameterized by Hosking and Wallis (1997) is: If $\alpha = 4/\gamma^2$, $\beta = \frac{1}{2}\sigma|\gamma|$, and $\xi = \mu - 2\sigma/\gamma$, where μ, σ and γ are its location, scale and shape parameters, respectively, then

$$F(x) = G(\alpha, \frac{x-\xi}{\beta}) / \Gamma(\alpha), \text{ if } \gamma > 0$$

and

$$f(x) = 1 - G\left(\alpha, \frac{\xi - x}{\beta}\right) / \Gamma(\alpha), \text{ if } \gamma < 0.$$

When, $\gamma = 0$, it becomes a normal distribution with μ and σ. In each case this distribution has no explicit analytical inverse form. Floods of various return periods *T* may be computed by multiplying $\hat{\zeta}_1$ (the first LQ-moment) of a catchment by the corresponding values of growth factors of the PE3 distribution.

Figure 1: LQ-Moments Ratio Diagram for the North-Bank Region of the Brahmaputra River

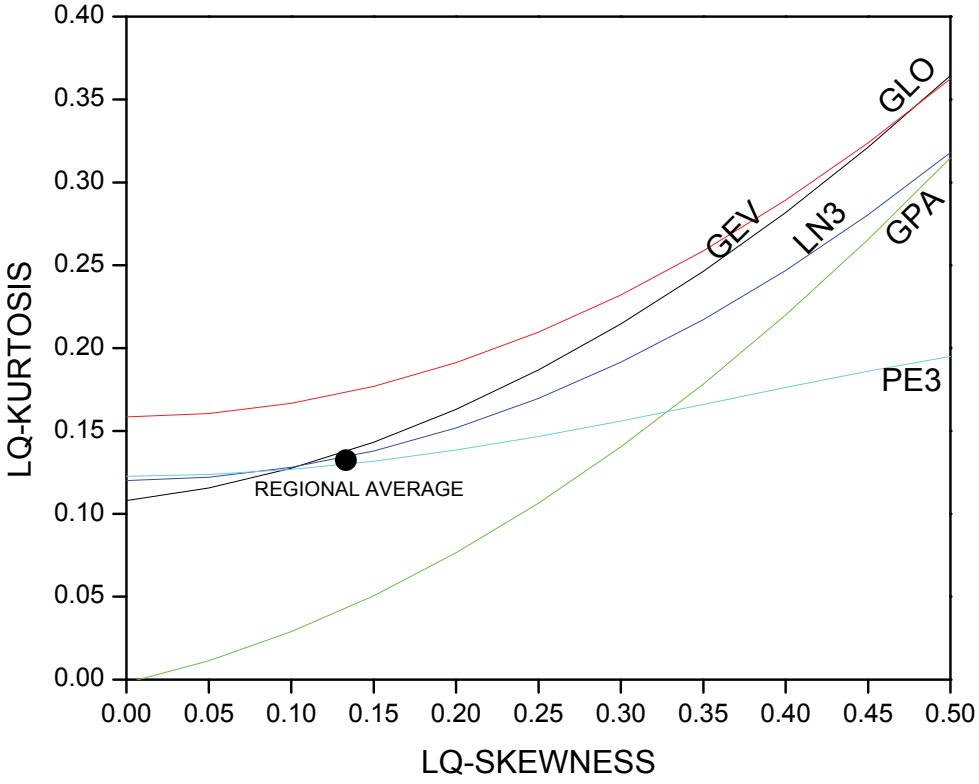


Table 4: Regional Parameters for Various Distributions Based on LQ-Moments

Distribution	Distribution Parameters		
GEV	ξ=0.857	α=0.353	k=0.035
GLO	ξ=0.987	α=0.244	k=-0.178
GPA	ξ=0.519	α=0.808	k=0.559
LN3	ξ=0.738	α=0.259	k=0.310
PE3	ξ=1.049	α=0.212	k=0.915

Table 5: Regional Quantile Estimation Based on LQ-Moments

Distribution	Return Periods (years)								
	2	5	10	25	50	100	200	500	1000
GEV	0.986	1.373	1.621	1.925	2.145	2.357	2.563	2.828	3.023
GLO	0.987	1.371	1.643	2.030	2.357	2.722	3.133	3.758	4.303
GPA	0.983	1.377	1.565	1.725	1.802	1.854	1.890	1.920	1.934
LN3	0.738	0.930	1.012	1.088	1.132	1.167	1.198	1.231	1.253
PE3	1.017	1.212	1.333	1.478	1.580	1.678	1.773	1.894	1.984

Regional Flood Frequency Relationship Based on LQ-Moments: Ungauged Catchments

In this case a relationship between the $\hat{\zeta}_1$ (the first LQ-moments) of gauged catchments in the region and their physiographic catchment characteristics is developed and is used to estimate first LQ-moments for an ungauged site. The relationship developed for the region in log domain using least squares approach based on the data of the study area is given as:

$$\hat{\zeta}_1 = 4.317 * (A)^{0.719} \quad (3)$$

where, A is the catchment area, in square kilometers (km^2) and is the $\hat{\zeta}_1$ first LQ-moments in meters per second (m^3/s). For equation (1), the correlation coefficient is $r = 0.947$. By coupling the regional flood frequency relationship for gauged catchment and the relationship between first LQ-moments and catchment area given by equation (1), the regional flood frequency relationship for ungauged catchments is obtained as:

$$Q_T = C_T A^{0.719} \quad (4)$$

where, Q_T is the flood estimate in m^3/s for return period T, A is the catchment area in km^2 and C_T is a regional coefficient. In Table 7 values of C_T are given for different return periods T for the study area.

Conclusion

The following conclusions can be drawn from the regional flood frequency analysis of the study area using LQ-moments:

1. In the initial screening step of the data the discordancy measure is used, the discordancy measure (Table 1) shows that data for the nine gauging sites of the study area are suitable for using regional flood frequency analysis.

2. For testing homogeneity of the region, the LQ-moment based heterogeneity measure was used, the LQ-moment based heterogeneity measure shows that the region is homogeneous.

3. The regional flood frequency analysis was performed using various frequency distributions: GLO, GEV, LN3, PE3 and GPA and KAP. The LQ-moment ratio diagram and $\left|Z_i^{dist}\right|$-statistic criteria (see Table 2) were used to identify best fitting distribution PE3 for the region.

4. The regional flood frequency relationship for gauged and ungauged catchments was developed for the region. The regional quantile estimates with different return periods *T* for the PE3, LN3, GPA, GLO and GEV distributions were calculated. To estimate floods of various return periods *T* for gauged catchments of the study area, the first LQ-moment of the catchment may be multiplied by corresponding values of the growth factors, computed using the PE3 distribution; however, more accurate results for ungauged sites can be obtained if more physiographic characteristics other than catchments area are available.

Table 7: Values of Regional Coefficient C_T

Return Periods (years)								
2	5	10	25	50	100	200	500	1000
PE3 Growth Factors								
4.390	5.232	5.755	6.381	6.821	7.244	7.654	8.176	8.565

References

Atiem, I. A., & Harmancioglu, N. B. (2006). Assessment of regional floods using L-moments approach: the case of the River Nile. *Water Resources Management*, *20*, 723-747.

Hosking, J. R. M. (1990). L-moments: Analysis and estimation of distributions using linear combinations of order statistics. *Journal of the Royal Statistical Society*, *Series B*, *52*(*2*), 105-124.

Hosking, J. R. M., & Wallis, J. R. (1997). *Regional frequency analysis: An approach based on L-moments*. New York, NY: Cambridge University Press.

Hussain, Z, & Pasha, G. R. (2008). Regional flood frequency analysis of the seven sites of Punjab, Pakistan, using L-moments. *Water Resources* Management, doi: 10.1007/s11269-008-9360-7.

Kumar, R., Singh, R. D.,& Seth, S. M. (1999). Regional flood formulas for seven subzones of zone 3 of India. *Journal of Hydrologic Engineering*, *4*(*3*), 240-244.

Kumar, R., Chatterjee, C., Kumar, S., & Lohani, A. K. (2003). Development of regional flood frequency relationship using L-moments for Middle Ganga Plains Subzone 1(f) of India. *Water Resources Management*, *17*, 243-257.

Kumar, R., & Chatterjee, C. (2005). Regional flood frequency analysis using L-moments for North Brahmaputra Region of India. *Journal of Hydrologic Engineering*, *4*(*3*), 240-244.

Mudholkar, G. S., & Hutson, A. D. (1998). LQ-moments: Analogs of L-moments. *Journal of Statistical Planning and Inference*, *71*, 191-208.

Paradia, B. P., Kachroo, R. K., & Shrestha, D. B. (1998). Regional flood frequency analysis of Mahi-Sabarmati Basin (Subzone 3-a) using index flood procedure with L-moments, *Water Resources Management*, *12*, 1-12.

Rao, A. R., & Hamed, H. K. (2000). *Flood frequency analysis*. Boca Raton, FL: CRC Press.

Saf, B. (2008). Regional Flood frequency analysis using L-moments for the West Mediterranean region of Turkey. *Water Resources Management*, doi 10.1007/s11269-008-9287-z.

Zin Wan, Z. W., Jemain, A. A., & Ibrahim, K. (2008). The best fitting distribution of annual maximum rainfall in Peninsular Malaysia based on methods of L-moment and LQ-moment. *Theories of Applied Climatology*, doi 10.1007/s00704-008-0044-2.

Journal of Modern Applied Statistical Methods
November 2011, Vol. 10, No. 2, 741-750

1538 – 9472/11/$95.00

JMASM Algorithms and Code
JMASM31: MANOVA Procedure for Power Calculations (SPSS)

Alan Taylor
Macquarie University
Sydney, Australia

D'Amico, Neilands & Zambarano (2001) showed how the SPSS MANOVA procedure can be used to conduct power calculations for research designs. This article demonstrates a simple way of entering data required for power calculations into SPSS and provides examples that supplement those given by D'Amico, Neilands & Zambarano.

Key words: Power, sample size, MANOVA, SPSS.

Introduction

Most researchers acknowledge the importance of conducting power calculations prior to embarking on research projects to ensure that there is a good chance that effects regarded as theoretically or practically important will be determined statistically significant. Despite this acknowledgement, many studies are under-powered. Over the years, a number of writers have called on researchers to consider the power of their research designs and have attempted to facilitate power calculations. Cohen (1992), for example, followed up his book on power (1988) by providing a primer aimed at facilitating power calculation and sample size estimation in the behavioral sciences. Others have provided software for the same purpose; *Gpower*, for example, is a free program which allows the calculation of power and sample size in a variety of designs intended to "address the weaknesses of existing power analysis tools" (Erdfelder, Faul & Buchner, 1992, p. 1). In the same spirit, D'Amico, Neilands & Zambarano (2001), noted that the MANOVA procedure in SPSS (now called IBM SPSS Statistics, but referred to as SPSS in the remainder of this article) can be used to calculate power in repeated measures and multivariate designs, which may not be readily accommodated by dedicated power programs. Their suggestion was particularly valuable because SPSS is a package that is accessible to many researchers.

D'Amico, Neilands & Zambarano's (2001) method takes advantage of the fact that the SPSS MANOVA procedure can both write and read datasets that are in a matrix form. This format allows a user to easily experiment with different numbers of cases, means, standard deviations and correlations between dependent and independent variables. In their examples, the authors read the data into SPSS in a matrix format and then read the matrix into MANOVA for the power calculation. This article suggests a minor extension to D'Amico, Neilands & Zambarano's procedure based on MANOVA's ability to write, as well as read, data in matrix format, a suggestion which facilitates power calculations with this method. Further examples of the use of the procedure to illustrate its usefulness in a number of design contexts are also provided; these examples supplement those given by D'Amico, Neilands & Zambarano (2001), which included ANCOVA, MANOVA and a repeated measures ANOVA.

Alan Taylor is a Senior Lecturer in the Department of Psychology. He teaches statistics and data analysis and the use of statistical software. Email him at: alan.taylor@mq.edu.au.

A Summary of the Procedure

1. Set up a dummy dataset in SPSS that is similar to the one for which power estimates

are to be obtained. The means and standard deviations can be drawn from previous research or may be guesswork. It may be easiest to use standardized measures so that effects (e.g., differences between group means) can be specified in terms of standard deviations.

2. Run the MANOVA procedure in order to save the data in matrix form.

3. Alter the values (the number of cases, differences between means, correlations) and run the MANOVA procedure to conduct the power calculations. Continue altering the relevant values (usually the number of cases) and re-running the MANOVA analysis to observe what values are necessary to obtain an acceptable level of power.

Detailed Examples

Example 1: Pre-Post Design with Two Groups

Consider a pre/post design with two groups, treatment and control. The dependent variable is anxiety, measured on a 10-point scale. The goal is to determine whether anxiety decreases more for the treatment group than for the control group. In other words, if there is a significant interaction between group and time. From previous research or from guesswork it is hypothesized that the mean score for each group (to which participants are randomly assigned from a waiting list of people who have come to an anxiety clinic) at pre-test will be about 6 and that the control group score will decline somewhat without treatment to around 5.5, but that the effect of treatment will be strong so the post-test mean for the treatment group will be approximately 4. The standard deviation of this anxiety measure is known to be about 1.5. With such changes – 2 versus 0.5 – a researcher would want to have good chance of finding the interaction significant at alpha 0.05. A dummy dataset, such as that shown in Figure 1, is first created in SPSS; the variable names were entered in the Variable View and the numbers were entered in the Data View.

Note that there are only two observations per case for this dummy dataset and that the required means are obtained by having one case in each group one unit lower than the mean, and the other case one unit higher than the mean. The following MANOVA commands are now run:

Figure 1: Dummy Data Entered into an SPSS Dataset

	group	pre	post
1	.00	5.00	4.50
2	.00	7.00	6.50
3	1.00	5.00	3.00
4	1.00	7.00	5.00

```
manova pre post by group(0,1)/
wsfactor=time(2)/
matrix=out(*)/
design.
```

The MANOVA procedure can only be run using syntax. The matrix subcommand asks for the data to be saved in matrix format (see Figure 2).

The data are now in a form that allows the various values to be altered to simulate the data that might be obtained. In the present case, the goal is to determine if having 10 cases per group provides enough power, thus, the *N* in the top row is increased to 20 and the *N*s for each group to 10. It is also necessary to change the standard deviation to 1.5 at each time point and to reduce the correlation between pre and post scores to a more realistic value, such as 0.5. The altered dataset is shown in Figure 3.

The following new set of MANOVA commands are run to obtain the power values:

```
manova pre post by group(0,1)/
wsfactor=time(2)/
matrix=in(*)/
power=f(.05) exact/
design.
```

This syntax reads the matrix dataset and requests a power analysis with an alpha of 0.05. The relevant section of the MANOVA output (see Table 1), indicates that the power for the interaction is too low to be acceptable.

Figure 2: Example 1 Dummy Data in Figure 1 Shown in Matrix Format

	ROWTYPE_	group	VARNAME_	pre	post
1	N	.		4.0000000	4.0000000
2	MEAN	.00		6.0000000	5.5000000
3	N	.00		2.0000000	2.0000000
4	MEAN	1.00		6.0000000	4.0000000
5	N	1.00		2.0000000	2.0000000
6	STDDEV	.		1.4142136	1.4142136
7	CORR	.	pre	1.0000000	1.0000000
8	CORR	.	post	1.0000000	1.0000000

Figure 3: Matrix Form of the Example 1 Dummy Dataset with Altered *N*s and Standard Deviations

	rowtype_	group	varname_	pre	post
1	N	.		20.0000000	20.0000000
2	MEAN	.00		6.0000000	5.5000000
3	N	.00		10.0000000	10.0000000
4	MEAN	1.00		6.0000000	4.0000000
5	N	1.00		10.0000000	10.0000000
6	STDDEV	.		1.5000000	1.5000000
7	CORR	.	PRE	1.0000000	.5000000
8	CORR	.	POST	.5000000	1.0000000

Table 1: SPSS MANOVA Output Showing the Results of the Power Calculation for Example 1 Dummy Dataset

```
Observed Power at the .0500 Level

Source of Variation   Noncentrality   Power

TIME                        13.889        0.941
GROUP BY TIME                5.000        0.562
```

The number of subjects can be increased and the analysis repeated until an acceptable value is obtained. In this case, if the group size is doubled the power for detecting the interaction is 0.869, which is much more acceptable. In fact, 15 per group, which gives a power of 0.753, may be considered sufficient. Note that, in repeated measures analyses, changes in the correlation between measures may have a dramatic effect on power. For example, if the correlation between the pre- and post-test measures in this example was a still-realistic 0.7 as opposed to 0.5, the power with only 10 cases per group is 0.779 rather than 0.562.

If no information is available regarding the values to expect, the standard deviation can be set to 1 (as for a standard score), and differences between the means specified in terms of the standard deviation. Consider a simple example in which two groups are to be compared. With a standard deviation of 1, if one group had a mean of zero and the other a mean of 0.5 in the dummy dataset this would represent a moderate effect size in terms of Cohen's (1992) classification.

Another strategy that can be adopted when there is uncertainty about the magnitude of differences is to perform a series of analyses with various combinations of *N* and effect sizes (and correlations in a repeated measures design). If the power is reasonable over a range of approximate realistic combinations of values, then the research has a good chance of obtaining a significant result. If not, it may be considered that the research is not worth doing with the number of subjects available.

Example 2: Oneway ANOVA with Contrasts and Unknown Means and Standard Deviation

This example has three groups of subjects, a control group (group = 1) and two treatment groups (2 and 3). In this case expected means and standard deviations are unknown, so the standard deviation is set equal to one, the mean of the control group to zero and the means of groups 2 and 3 to 0.5 and 0.8 respectively. The difference between the means of groups 1 and 2 is therefore (0.5 – 0) = 0.5, and Cohen's d = 0.5/1 = 0.5, which Cohen (1992) terms a medium effect size. The difference between the control group and group 3 is 0.8, and d = 0.8, a large effect size.

Again, a dummy dataset with the necessary structure is entered into SPSS using, in this case, arbitrary numbers for the dependent variable (see Figure 4).

Figure 4: Dummy Dataset for Example 2

	group	score
1	1.00	1.00
2	1.00	2.00
3	2.00	3.00
4	2.00	4.00
5	3.00	5.00
6	3.00	6.00

The MANOVA commands to create the matrix version of the dataset (see Figure 5) are:

```
manova score by group(1,3)/
matrix=out(*)/
design.
```

The standard deviation of 0.7071 is replaced with 1 and the means are given the values noted above. It is assumed that 60 subjects can be recruited, 20 in each group. The revised matrix dataset is shown in Figure 6.

It is assumed that there is no interest in the overall ANOVA result, but rather in two *a priori* contrasts: group 2 versus group 1, and group 3 versus group 1. MANOVA can give the power for each of the contrasts with the commands:

```
manova score by group(1,3)/
contrast(group)=simple(1)/
matrix=in(*)/
power=f(.025) exact/
design=group(1) group(2).
```

The simple (1) option asks for the required contrasts and, in the design statement, group (1) represents the group 2 versus group 1 contrast and group (2) represents the group 3 versus

group 1 contrast. To hold the overall Type I error at 0.05, alpha is set at 0.025 for the two contrasts. The relevant output, (see Table 2) indicates that the power for the first contrast is very low, 0.24, but that for the second is 0.59. At this stage more cases could be added to determine how many more cases would be needed to achieve an acceptable level of power.

Figure 5: Matrix Form of the Data for Example 2

	rowtype_	group	varname_	score
1	N	.		6.0000000
2	MEAN	1.00		1.5000000
3	N	1.00		2.0000000
4	MEAN	2.00		3.5000000
5	N	2.00		2.0000000
6	MEAN	3.00		5.5000000
7	N	3.00		2.0000000
8	STDDEV	.		.7071000
9	CORR	.	SCORE	1.0000000
10				

Figure 6: Matrix Format of the Dataset for Example 2 with Amended Values of *N*, Means and Standard Deviations

	rowtype_	group	varname_	score
1	N	.		60.0000000
2	MEAN	1.00		0E-7
3	N	1.00		20.0000000
4	MEAN	2.00		.5000000
5	N	2.00		20.0000000
6	MEAN	3.00		.8000000
7	N	3.00		20.0000000
8	STDDEV	.		1.0000000
9	CORR	.	SCORE	1.0000000
10				

Table 3: MANOVA output showing the results of the power calculation for Example 2

```
Observed Power at the .0250 Level

                          Noncen-
 Source of Variation      trality   Power

 GROUP(1)                   2.500    .244
 GROUP(2)                   6.400    .592
```

Example 3: Correlation

Assume that a researcher seeks to assess the correlation between measures of anger and narcissism, which is expected to be very low. It is desirable to have a good chance (power at least 0.80) of obtaining a significant result if the correlation in the population is 0.30 or higher. The dummy dataset (see Figure 7) is created and then the following commands are used to produce a matrix version of the data (see Figure 8):

```
manova anger with narciss/
matrix=out(*)/
design.
```

The correlation is replaced with 0.30, and 30 subjects are used at first (the values of the means and standard deviations are immaterial in this case). The revised version of the dataset is shown in Figure 9 and relevant output is shown in Table 4. The value of 0.362 is unacceptably low; thus, further experimentation was conducted to show that 85 subjects are needed to achieve a power of 0.80.

Figure 7: Dummy Dataset for Example 3

	anger	narciss
1	1.00	2.00
2	2.00	1.00
3	3.00	4.00
4	4.00	3.00
5	5.00	5.00
6		

Figure 8: Matrix Form of the Data for Example 3

Undo a user action

	rowtype_	varname_	anger	narciss
1	N		5.0000000	5.0000000
2	MEAN		3.0000000	3.0000000
3	STDDEV		1.5811388	1.5811388
4	CORR	ANGER	1.0000000	.8000000
5	CORR	NARCISS	.8000000	.8000000
6				

Figure 9: Amended Matrix Dataset for Example 3

	rowtype_	varname_	anger	narciss
1	N		30.0000000	30.0000000
2	MEAN		3.0000000	3.0000000
3	STDDEV		1.5811388	1.5811388
4	CORR	ANGER	1.0000000	.3000000
5	CORR	NARCISS	.3000000	1.0000000
6				

Table 4: MANOVA output showing the results of the power calculation for Example 3

Observed Power at the .0500 Level

Source of Variation	Noncentrality	Power
Regression	2.769	.362

Example 4: Using an Existing Dataset

It is often not sensible to calculate the power for an existing dataset (if the effects are significant, the power will be viewed as adequate; if the effects are not significant, the power may be considered too low). However, it can be sensible to ask, for an effect which was not significant: How many more cases would be needed to have a good chance of finding a significant effect if the population characteristics are the same as those of my sample?

This example uses a partly synthetic dataset called *glmdemo.sav* (available for download from: http://www.psy.mq.edu.au/psystat/download.htm). Suppose a multivariate analysis with three variables, *test1* to *test3*, as the dependent variables and *group* (with four categories) as the grouping variable has been conducted. The results for the 99 cases in the dataset are shown in Table 5.

The MANOVA commands to produce the matrix version of the data are:

```
manova test1 to test3 by group(1,4)/
matrix=out(*)
```

The matrix version of the dataset is shown in Figure 10.

The following commands can be used to calculate the observed power:

```
manova test1 to test3 by group(1,4)/
matrix=in(*)/
power=f(.05) exact.
```

The power for the Wilks' Lambda statistic is 0.62; now, determine what improvement would result if five cases were added to each group for an addition of 20 more subjects overall. The subject numbers could be added manually, but the following commands will complete the task and will make it easier to add more in the future:

```
do if (rowtype_ eq "N" and sysmis(group)).
compute test1=test1 + 20.
compute test2=test2 + 20.
compute test3=test3 + 20.
else if (rowtype_ eq "N" and ~sysmis(group)).
compute test1=test1 + 5.
compute test2=test2 + 5.
compute test3=test3 + 5.
end if.
execute.
```

When the MANOVA commands are run again, the resulting power is 0.72. If the above syntax is used with the addition of another five subjects to each group – 139 subjects overall – the power is found to be exactly 0.80.

Table 5: MANOVA results of a multivariate analysis of the *glmdemo* dataset used in Example 4

EFFECT .. GROUP
Multivariate Tests of Significance (S = 3, M = -1/2, N = 45 1/2)

Test Name	Value	Approx. F	Hypoth. DF	Error DF	Sig. of F
Pillais	.14202	1.57358	9.00	285.00	.123
Hotellings	.15739	1.60308	9.00	275.00	.114
Wilks	.86122	1.59333	9.00	226.49	.118
Roys	.11361				

- -
EFFECT .. GROUP (Cont.)
Univariate F-tests with (3,95) D. F.

Variable	Hypoth. SS	Error SS	Hypoth. MS	Error MS	F	Sig. of F
TEST1	2.04393	59.70197	.68131	.62844	1.08412	.360
TEST2	5.62424	57.69468	1.87475	.60731	3.08695	.031
TEST3	2.08582	69.51679	.69527	.73176	.95014	.420

Example 5: Multiple Regression

This analysis regresses a dependent variable *y* on *x1* and *x2* (numeric variables), and *x3* (dichotomous variable). A dummy dataset is created as usual using arbitrary numbers, but with one exception: with the dichotomous variable, *x3*, it is a good idea to insert zeroes and ones (e.g., female = 0, male = 1) in the same proportion as they would be expected to occur in the sample. Because the mean and standard deviation of a proportion are linked, this will help avoid the need to change one or both after creating the matrix dataset. In this example, a 50/50 distribution is assumed. The initial dataset is shown in Figure 11.

The MANOVA commands used to produce the matrix version of the data are:

```
manova y with x1 x2 x3/
matrix=out(*).
```

Figure 12 shows the initial matrix version of the dataset.

Assume, based on past research or theory, that *x1* is moderately (0.5) correlated with *y*, but *x2* and *x3* are only weakly correlated with *y* (both 0.3). Furthermore assume that *x1* and *x2* are highly correlated (0.6), but neither is correlated with *x3* (0.1) and that 50 subjects can be obtained for the research. The amended dataset is shown in Figure 13. The MANOVA commands below provide the output shown in Table 6:

```
manova y with x1 x2 x3/
matrix=in(*)/
power=f(.05).
```

There is ample power (0.97) for the overall regression, but the power for each independent variable is also of interest. The results show the power for an alpha of 0.05; a different alpha, e.g., 0.05/3 = 0.0167, may be used for the predictors. The MANOVA command is run again, with power = f(0.0167), to obtain the results shown in Table 7 (only part of the output is shown).

The power for *x1* remains adequate, but a bigger sample would be needed to achieve acceptable values for *x2* and *x3*. A sample of 100 gives a value of 0.70 for *x3*; *x2* is a lost cause due to its correlation with *x1*.

Conclusion

As D'Amico, Neilands, & Zambarano (2001) noted, the SPSS MANOVA procedure provides a way of conducting power and sample size calculations for multivariate and repeated measures designs that may be impossible or difficult with dedicated power and sample size software. This article illustrated a simple method of creating the matrix dataset, which is at the core of procedure described by D'Amico, Neilands, & Zambarano, and provided additional examples of the method to supplement their work.

Comprehensive power and sample size software, such as NCSS PASS (Hintz, 2012), is available and programs such as *Mplus* (Muthén & Muthén, 1998-2010) offer sophisticated simulation facilities for models that are outside the scope of those that can be handled by the methods described herein. The variety of designs for which power can be calculated using the SPSS MANOVA procedure, together with the ubiquity of the package, make it a valuable contribution to facilitating the routine calculation of power during research design.

Acknowledgement

The author thanks his late friend and colleague Dr. David Cairns for informing him of the D'Amico, Neilands & Zambarano (2001) article.

References

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*, *2nd Ed.* Hillsdale, N.J.: L. Erlbaum Associates.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, 112, 155-159.

D'Amico, E., Neilands, T., & Zambarano, R. (2001). Power designs for multivariate and repeated measures designs: A flexible approach using the SPSS MANOVA procedure. *Behavior Research Methods, Instruments, & Computers*, *33*, 479-484.

Erdfelder, E., Faul, F., & Buchner, A. (1996). GPOWER: A general power analysis program. *Behavior Research Methods, Instruments and Computers*, *28*, 1-11.

Hintz, J. (2011). PASS 11 (computer software). NCSS: Kayesville, Utah.

Muthén, L. K., & Muthén, B. O. (1998-2010). *Mplus user's guide*, *6th Ed*. Los Angeles, CA: Muthén & Muthén.

Figure 10: Matrix Version of the *glmdemo* Dataset Used in Example 4

	ROWTYPE_	group	VARNAME_	test1	test2	test3
1	N	.		99.0000000	99.0000000	99.0000000
2	MEAN	1		4.4244974	4.5246307	4.2067255
3	N	1		29.0000000	29.0000000	29.0000000
4	MEAN	2		4.6060099	4.8869344	4.2621851
5	N	2		24.0000000	24.0000000	24.0000000
6	MEAN	3		4.7934063	5.1225354	4.2714072
7	N	3		31.0000000	31.0000000	31.0000000
8	MEAN	4		4.6287784	4.9799732	4.6443188
9	N	4		15.0000000	15.0000000	15.0000000
10	STDDEV	.		.7927432	.7793025	.8554272
11	CORR	.	test1	1.0000000	.0415152	.1233156
12	CORR	.	test2	.0415152	1.0000000	.2428773
13	CORR	.	test3	.1233156	.2428773	1.0000000

Figure 11: Dummy Dataset Used in Example 5

	y	x1	x2	x3
1	1.00	2.00	4.00	.00
2	2.00	1.00	1.00	.00
3	3.00	4.00	2.00	.00
4	4.00	3.00	5.00	1.00
5	5.00	6.00	3.00	1.00
6	6.00	5.00	6.00	1.00

Figure 12: Initial Matrix Form of the Dataset Used in Example 5

	rowtype_	varname_	y	x1	x2	x3
1	N		6.0000000	6.0000000	6.0000000	6.0000000
2	MEAN		3.5000000	3.5000000	3.5000000	.5000000
3	STDDEV		1.8708287	1.8708287	1.8708287	.5477226
4	CORR	Y	1.0000000	.8285714	.5428571	.8783101
5	CORR	X1	.8285714	1.0000000	.3714286	.6831301
6	CORR	X2	.5428571	.3714286	1.0000000	.6831301
7	CORR	X3	.8783101	.6831301	.6831301	1.0000000

Figure 13: Amended Version of the Matrix Dataset used in Example 5

	rowtype_	varname_	y	x1	x2	x3
1	N		50.0000000	50.0000000	50.0000000	50.0000000
2	MEAN		3.5000000	3.5000000	3.5000000	.5000000
3	STDDEV		1.8708287	1.8708287	1.8708287	.5477226
4	CORR	Y	1.0000000	.5000000	.3000000	.3000000
5	CORR	X1	.5000000	1.0000000	.6000000	.1000000
6	CORR	X2	.3000000	.6000000	1.0000000	.1000000
7	CORR	X3	.3000000	.1000000	.1000000	1.0000000

Table 6: MANOVA output for Example 5

```
Observed Power at the .0500 Level
                        Noncen-
Source of Variation     trality   Power

Regression               20.986    .970

- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
Regression analysis for WITHIN CELLS error term
--- Individual Univariate .9500 confidence intervals
--- two-tailed observed power taken at .0500 level
Dependent variable .. Y

COVARIATE            B        Beta   Std. Err.     t-Value   Sig. of t

X1               .48418      .00000        .153       3.166        .003
X2              -.01582     -.11076        .153       -.103        .918
X3               .86472     6.05305        .420       2.059        .045

COVARIATE   Lower -95%  CL- Upper     Noncent.       Power

X1                .176        .792      10.025       .873
X2               -.324        .292        .011       .051
X3                .019       1.710       4.240       .522
```

Table 7: Output Obtained for the Amended Matrix Dataset Used in Example 5

```
two-tailed observed power taken at .0167 level

COVARIATE   Lower -95%  CL- Upper     Noncent.       Power
X1               .176        .792      10.025        .750
X2              -.324        .292        .011        .017
X3               .019       1.710       4.240        .345
```

Journal of Modern Applied Statistical Methods
November 2011, Vol. 10, No. 2, 403-750

1538 – 9472/11/$95.00

Instructions For Authors

Follow these guidelines when submitting a manuscript:

1. *JMASM* uses a modified American Psychological Association style guideline.
2. Submissions are accepted via e-mail only. Send them to the Editorial Assistant at ea@jmasm.com. Provide name, affiliation, address, e-mail address, and 30 word biographical statements for all authors in the body of the email message.
3. There should be no material identifying authorship except on the title page. A statement should be included in the body of the e-mail that, where applicable, indicating proper human subjects protocols were followed, including informed consent. A statement should be included in the body of the e-mail indicating the manuscript is not under consideration at another journal.
4. Provide the manuscript as an external e-mail attachment in MS Word for the PC format only. (Wordperfect and .rtf formats may be acceptable - please inquire.) Please note that Tex (in its various versions), Exp, and Adobe .pdf formats are designed to produce the final presentation of text. They are not amenable to the editing process, and are **NOT** acceptable for manuscript submission.
5. The text maximum is 20 pages double spaced, not including tables, figures, graphs, and references. Use 11 point Times New Roman font.
6. Create tables without boxes or vertical lines. Place tables, figures, and graphs "in-line", not at the end of the manuscript. Figures may be in .jpg, .tif, .png, and other formats readable by Adobe Photoshop.
7. The manuscript should contain an Abstract with a 50 word maximum, following by a list of key words or phrases. Major headings are Introduction, Methodology, Results, Conclusion, and References. Center headings. Subheadings are left justified; capitalize only the first letter of each word. Sub-subheadings are left-justified, indent optional. Do not number headings or subheadings.
8. Number all formulas, tables, figures, and graphs, but do not use italics, bold, or underline.
9. Do not use underlining in the manuscript. Do not use bold, except for (a) matrices, or (b) emphasis within a table, figure, or graph. Do not number references. Do not use footnotes or endnotes.
10. In the References section, do not put quotation marks around titles of articles or books. Capitalize only the first letter of books. Italicize journal or book titles and volume numbers. Use "&" instead of "and" in multiple author listings.
11. *Suggestions for style*: Instead of "I drew a sample of 40" write "A sample of 40 was selected." Use "although" instead of "while", unless the meaning is "at the same time". Use "because" as opposed to "since," unless the meaning is "after." Instead of "Smith (1990) notes" write "Smith (1990) noted". Do not strike spacebar twice after a period.

Print Subscriptions

Print subscriptions including postage for professionals are US $95 per year; for graduate students are US $47.50 per year; and for libraries, universities and corporations are US $195 per year. Subscribers outside of the US and Canada pay a US $10 surcharge for additional postage. Online access is currently free at http://www.jmasm.com/. Mail subscription requests with remittances to JMASM, P. O. Box 48023, Oak Park, MI, 48237. Email journal correspondence, other than manuscript submissions, to ea@jmasm.com.

Notice To Advertisers

Send requests for advertising information to ea@jmasm.com.