DigitalCommons
@WayneState

Journal of Modern Applied Statistical Methods

Volume 7 | Issue 2 Article 28

11-1-2008

Vol. 7, No. 2 (Full Issue)

JMASM Editors

Journal of Modern Applied Statistical Methods
November, 2008, Vol. 7, No. 2, 368-631

1538 – 9472/08/$95.00

Journal Of Modern Applied Statistical Methods

Copyright © 2008 JMASM, Inc.
1538 – 9472/08/$95.00

Editorial Board

בס״ד
Journal of Modern Applied Statistical Methods
November, 2008, Vol. 7, No. 2, 368-631

1538 – 9472/08/$95.00

Journal Of Modern Applied Statistical Methods

Invited Article

Regular Articles

Early Scholars

JMASM is an independent print and electronic journal (http://tbf.coe.wayne.edu/jmasm), 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.

Editorial Assistants: **Michael Lance, Laura Rolands, Julie Smith**
Internet Sponsor: **Paula C. Wood**, Dean, College of Education, Wayne State University

Cushing-Malloy, Inc.	(888) 295-7244 toll-free (Phone)	Sales & Information:
Internet: www.cushing-malloy.com	(734) 663-5731 (Fax)	skehoe@cushing-malloy.com

Errata: Matthew E. Elam's affiliation listed on the Editorial Board in
Vol. 6(2) and 7(1) should be Industrial Engineering, Texas A&M – Commerce.

Journal of Modern Applied Statistical Methods
November, 2008, Vol. 7, No. 2, 368-375

1538 – 9472/08/$95.00

INVITED ARTICLES

Estimating Explanatory Power in a Simple Regression Model Via Smoothers

Rand R. Wilcox
University of Southern California

Consider the regression model $Y = \gamma(X) + \varepsilon$, where $\gamma(X)$ is some conditional measure of location associated with Y, given X. Let $\hat{Y}$ be some estimate of Y, given X, and let $\tau^2(Y)$ be some measure of variation. Explanatory power is $\eta^2 = \tau^2(\hat{Y}) / \tau^2(Y)$. When $\gamma(X) = \beta_0 + \beta_1 X$ and $\tau^2(Y)$ is the variance of Y, $\eta^2 = \rho^2$, where ρ is Pearson's correlation. The small-sample properties of some methods for estimating a robust analog of explanatory power via smoothers is investigated. The robust version of a smoother proposed by Cleveland is found to be best in most cases.

Key words: strength of association, smothers, effect size, robust methods and nonparametric regression.

Introduction

Consider the simple, nonparametric regression model

$$Y = \gamma(X) + \varepsilon, \quad (1)$$

where X and ε are independent random variables, and $\gamma(X)$ is some unknown function that represents some conditional measure of location associated with Y given X. A fundamental goal is measuring the strength of the association between Y and X. Certainly the best-known approach is to assume

$$\gamma(X) = \beta_0 + \beta_1 X,$$

estimate β_0 and β_1 via ordinary least squares, and then use ρ^2, where ρ is Pearson's correlation. It is well known that Pearson's correlation is not robust (e.g., Wilcox, 2005) and can yield a highly misleading sense about the strength of the association among the bulk of the points. Yet another concern is the assumption that the regression line is straight. Situations are encountered where this assumption seems to be a reasonable approximation of reality, but experience with nonparametric regression methods (e.g. Efromovich, 1999; Eubank, 1999;

Rand R. Wilcox (rwilcox@usc.edu) is Professor of Psychology. He is the author of seven textbooks on statistics, the most recent of which is *Basic Statistics: Understanding Conventional Methods and Modern Insights* (2009, New York, Oxford University Press)

Fan & Gijbels, 1996; Fox, 2001; Green & Silverman, 1993; Gyofri et al., 2002; Hardle, 1990; Hastie & Tibshirani, 1990), sometimes called smoothers, suggest that it is common to encounter situations where this is not the case.

Let $\hat{Y}$ be some estimate of Y given X, and let $\tau^2(Y)$ be some measure of variation associated with the marginal distribution of Y. Then a general approach to measuring the strength of the association between Y and X, called explanatory power, is

$$\eta^2 = \frac{\tau^2(\hat{Y})}{\tau^2(Y)} \quad (2)$$

(e.g., Doksum, Blyth, Bradlow, Meng, & Zhao, 1994; Wilcox, 2003, p. 506). If it is assumed that the conditional distribution of Y given X has the form

$$Y = \beta_0 + \beta_1 X + \varepsilon,$$

where $E(\varepsilon) = 0$, and if τ^2 is taken to be the usual variance, $\eta^2 = \rho^2$. It is well-known, however, that the usual variance and Pearson's correlation are not robust. Roughly, small changes in any distribution can substantially alter ρ resulting in a potentially misleading sense about the strength of the association among the bulk of the points. In particular, slight departures from normality can be a practical concern when interpreting ρ.

A simple method for robustifying η^2 is to take τ^2 to be some robust measure of variation. Many such measures have been proposed, comparisons of which are reported by Lax (1985). Based on efficiency, Lax concludes that two so-called A-estimators are best, one of which corresponds to the percentage bend midvariance that was studied by Shoemaker and Hettmansperger (1982). It can be designed to have a reasonably high breakdown point, its efficiency compares well to the usual sample variance under normality, and its standard error can be substantially smaller than the standard error of the sample variance when sampling from a heavy-tailed distribution. For these reasons it is used here, but this is not to suggest that all other measures of variation have no practical value for the problem at hand.

In addition to many robust measures of variation, there are many nonparametric regression methods that might be used when trying to deal with curvature. Here, no attempt is made to examine all such methods when estimating explanatory power, but rather to consider a few methods that appear to deserve serious consideration, with the goal of finding one method that performs well over a fairly broad range of situations when the sample size is small. In particular, three estimates of η^2 are considered that are based on three nonparametric regression estimators: the robust version of the method in Cleveland (1979), a particular version of a kernel regression estimator derived by Fan (1993), and the running interval smoother (e.g., Wilcox, 2003, section 11.4.4). Consideration was given to a variation of the running interval smoother based on bootstrap bagging (e.g., Buhlmann & Yu, 2002), but it performed rather poorly in the simulations reported here, so further details are omitted.

To add perspective, some results are included assuming

$$\gamma(X) = \beta_0 + \beta_1 X$$

with β_0 and β_1 estimated using the robust method derived by the Theil (1950) and Sen (1968) as well as the ordinary least squares estimator. Of course, when there is curvature, any method that assumes

$$\gamma(X) = \beta_0 + \beta_1 X$$

has the potential to perform poorly. The issue here is how much is sacrificed when a nonparametric estimate of the regression line is used and the regression line is indeed straight. As is well known, there are many robust alternatives to the Theil-Sen estimator that have excellent theoretical properties. The Theil-Sen estimator is used because, in terms of efficiency, it seems to perform about as well as the ordinary least squares (OLS) estimator when the error term has a normal distribution, and it continues to perform well in situations where OLS performs poorly (e.g., Wilcox, 2005). If the regression line is straight, perhaps there is some practical advantage to using some other robust estimator, but this issue is not addressed here. The primary goal is to consider methods that can be used when curvature might exist. Although not considered here, another well-known approach to nonparametric regression is based on what are called splines, and so for completeness, some comments seem in order. Some informal comparisons with other smoothers suggest that sometimes splines are not quite as satisfactory as other methods (Hardle, 1990; Wilcox, 2005). For this reason, they are not considered, but in fairness, it seems that an extensive formal comparison with the regression methods used here has not been made.

An attempt could be made to fit a parametric model in a manner that takes into account curvature, but simulating this process is difficult. The results reported here suggest that, even when fitting a correct parametric model, little is gained relative to method C, which is described below.

Methodology

The Percentage Bend Midvariance

The objective now is to summarize how the percentage bend midvariance measure of dispersion is computed. For a recent summary of how this measure of dispersion compares to other robust measures of variation, see Wilcox (2005, section 3.12). Let $X_1, \ldots, X_n$ be a random sample. For some β satisfying 0< β < .5, compute (1-β)n+.5, round the result to the nearest integer, and label the result m. The choice β=.1 results in good efficiency under normality, but a relatively low breakdown point. That is, with β=.1, only 10% of the observations have to be changed to destroy this measure of dispersion. Accordingly, β=.2 is used. Let $W_i = |X_i - M|, i = 1, \ldots, n$, and let $W_{(1)} \leq \ldots \leq W_{(n)}$ be the W_i values written in ascending order. Set $\hat{\omega}_\beta = W_{(m)}$,and

$$U_i = \frac{X_i - M}{\hat{\omega}_\beta}.$$

Let $a_i = 1$ if $U_i < 1$; otherwise $a_i = 0$. The estimated percentage bend midvariance is

$$\tau^2 = \frac{n\omega_\beta^2 \sum \psi^2(U_i)}{(\sum a_i)^2}, (3)$$

where $\psi(x) = \max[-1, \min(1, x)]$.

Fan's Kernel Regression Estimator

The first of the nonparametric regression methods considered here stems from Fan (1993). $(X_1, Y_1), \ldots, (X_n, Y_n)$ be a random sample of n points. Let $K(u)$ be the Epanechnikov kernel given by

$$K(u) = \frac{3}{4}(1 - \frac{1}{5}u^2)/\sqrt{5}$$

If $|u| < \sqrt{5}$; otherwise $K(u)$=0. Let $h = \min(s, IQR/1.34)$, where s is the standard deviation of the X values and IQR is the interquartile range. Bjerve and Doksum (1993) take $h = s$, but it is well known that a robust measure of variation, such as the interquartile range, can have practical value when using a kernel density estimator (e.g., Silverman, 1986).

There is the issue of how to estimate IQR. Many quantile estimators have been proposed, comparisons of which were made by Parrish (1990) as well as Dielman, Lowry, and Pfaffenberger (1994). Here the interquartile range is estimated via the so-called ideal fourths (Frigge, Hoaglin, & Iglewicz, 1989). Perhaps some alternative quantile estimator offers a practical advantage for the problem at hand, but this goes beyond the scope of this paper.

To be more precise, the ideal fourths are computed as follows. Let $X_{(1)} \leq \ldots \leq X_{(n)}$ be the observations written in ascending order. Estimates of the lower quartile typically have the form

$$q_1 = (1 - \ell)X_{(j)} + \ell X_{(j+1)}$$

The ideal fourths are computed by taking j to be the integer portion of (n/4)+(5/12) and

$$\ell = \frac{n}{4} + \frac{5}{12} - j$$

The estimate of the upper quartile is taken to be $q_2 = (1-\ell)X_k + \ell X_{(k-1)}$ where k=n-j+1, in which case the interquartile range is estimated with $IQR = q_2 - q_1$. Let $m(x) = E(Y \mid X = x)$. Then $m(x)$ is estimated with $\hat{m}(x) = b_0 + b_1 x$, where b_0 and b_1 are determined via weighted least squares with weights $w_i = K((X_i - x)/h)$. This will be called method F.

Cleveland's Method

To outline Cleveland's method, for any x, let $\delta_i = \mid X_i - x \mid$. Sort the δ_i values and retain the κn pairs of points that have the smallest δ_i values, where κ is some number between 0 and 1 and is called the span. Let $Q_i = \frac{\mid x - X_i \mid}{\delta_m}$, and if $0 \le Q_i < 1$, set $w_i = (1 - Q_i^3)^3$, otherwise $w_i = 0$. Next, use weighted least squares to estimate m(x) using w_i as weights.

Cleveland (1979) also discussed a robustified version of this method, which is used here. In effect, extreme Y values get little or no weight, and so they have little or no impact on the smooth. (An outline of these additional computations can also be found in Hardle, 1990, p. 192.) Both R and S-PLUS provide access to a function, called lowess, which performs the computations, and the R version was used in the simulations reported here using the default value κ=.75. This will be called method C.

The Running-Interval Smoother

Finally, the so-called running interval smoother was considered. For some constant f, declare x to be close to X_i if

$$\mid X_i - x \mid \le f \times MADN,$$

where MADN=MAD/.6745, MAD is the median of the values $\mid X_1 - M \mid, \ldots, \mid X_n - M \mid$, and M is the usual sample median of the X_i values. Let $N(X_i) = \{j : \mid X_j - X_i \mid \le f \times MADN\}$. That is, $N(X_i)$ indexes the set of all X_j values that are close to X_i. Then m(X_i) is taken to be some measure of location based on all Y_j values such that $j \in N(X_i)$. Here, a 20% trimmed mean is used. It has nearly the same efficiency as the mean under normality, but it continues to have high efficiency, relative to the usual sample mean, when sampling from heavy-tailed distributions. It appears that often a good choice for the span, f, is f=1 (e.g., Wilcox, 2005) and this value is used here. However, results in the next section indicate that this choice can be relatively ineffective for the problem at hand; a smaller value for f seems to be desirable, at least with small sample sizes. But even now, all indications are that Cleveland's method gives superior results. This will be called method R.

The Theil-Sen Estimator

This section reviews how the Theil-Sen estimator is computed. Let X_i and X_j be any two X values such that $X_i > X_j$. Denote the slope corresponding to the two points (X_i, Y_i) and (X_j, Y_j) by b_{1ij}. The median of all such slopes is the Theil-Sen estimate of β_1 and is labeled b_{1ts}. The intercept is estimated with $b_{0ts} = M_y - b_{1ts} M_x$,where M_y and M_x are the sample medians corresponding to the Y and X values, respectively. Estimation of explanatory power via the Theil-Sen estimator will be called method TS.

Estimating Explanatory Power

Based on the regression estimators just described, explanatory power is estimated in an obvious way. For each X_i, compute $\hat{Y}_i$, the estimate of Y given that $X = X_i$. Then explanatory power is estimated with

$$\hat{\eta}^2 = \frac{\hat{\tau}^2(\hat{Y})}{\hat{\tau}^2(Y)},$$

where $\hat{\tau}^2(Y)$ indicates the estimated percentage bend midvariance based on $Y_1,\ldots,Y_n$.

Results

Simulations were used to the check the small sample properties of the methods just described Here, two types of regression lines are considered: $Y = X + \varepsilon$ and $Y = X^2 + \varepsilon$. In both cases, bias was found to be an important issue, as will be seen. It is noted that additional simulations were run with $Y = \varepsilon$, in which case $\eta^2 = 0$, again bias is an issue, but for brevity, no additional details are given. For $Y = X^2 + \varepsilon$, no results are reported when using OLS and method TS, since they are based on the assumption that $Y = \beta_0 + \beta_1 X + \varepsilon$ and are clearly unsatisfactory when in fact $Y = X^2 + \varepsilon$. Both X and ε are assumed to have one of four g-and-h distributions (Hoaglin, 1985), which contains the standard normal distribution as a special case. If Z has a standard normal distribution, and if $g > 0$, then

$$W = \frac{\exp(gZ) - 1}{g} \exp(hZ^2/2)$$

has a g-and-h distribution where g and h are parameters that determine the first four moments. If $g = 0$, this last equation is taken to be $W = Z\exp(hZ^2/2)$. The four distributions were the standard normal ($g = h = 0$), a symmetric heavy-tailed distribution (h=.5, g=0), an asymmetric distribution with relatively light tails (h=0, g=.5), and an asymmetric distribution with heavy tails (g=h=.5). Table 1 shows the theoretical skewness (κ_1) and kurtosis (κ_2) for each distribution considered. When h=.5, the fourth moment is not defined and the value for κ_2 is left blank. Additional properties of the g-and-h distribution are summarized by Hoaglin (1985).

Table 1:
Some properties of the g-and-h distribution

g	h	κ_1	κ_2
0.0	0.0	0.0	3.0
0.0	0.5	0.00	---
0.5	0.0	0.61	9.7
0.5	0.5	2.81	---

There remains the problem of determining the population value of η^2 when and ε have some specified distribution. First consider the case $Y = X + \varepsilon$, where both X and ε are assumed to have one of four g-and-h distributions previously described. Then the correct estimate of Y is $\hat{Y} = X$, in which case $\tau^2(\hat{Y}) = \tau^2(X)$, which was determined by randomly sampling n=100,000 observations from the distribution under consideration. As for $\tau^2(Y)$, the following process was used. First generate 5000 values for both ε and X, which yields 5000 values for Y. Computing τ based on these 5000 values yields an estimate of τ. Here, this process was repeated 5000 times, and the average of the resulting τ values is taken to be the population value of $\tau^2(Y)$. And of course, having determined both $\tau^2(\hat{Y})$ and $\tau^2(Y)$, η^2 is taken be $\tau^2(\hat{Y})/\tau^2(Y)$. As for the case $Y = X^2 + \varepsilon$, the same process was used. For $Y = X + \varepsilon$, the values of η^2 were found to be .499, .409, .338, .314 corresponding to (g,h)=(0,0), (.5,0), (0,.5) and (.5,.5), respectively. As for $Y = X^2 + \varepsilon$, the values were found to be .323, .242, .365 and .330.

Each replication in the simulations consisted of generating n values for X, another n values for ε, computing $Y = X + \varepsilon$ or $Y = X^2 + \varepsilon$, and then applying the estimators described in the previous section. Two sample sizes were considered: n=30 and 100. Here, X and ε have the same g-and-h distribution.

Table 2: Estimated bias

g	h	TS	C	F	R	OLS
$Y = X + \varepsilon$						
0.0	0.0	.017	.007	-.005	-.109	.019
0.5	0.0	.028	.040	.021	-.052	.094
0.0	0.5	.042	.047	.396	-.015	.158
0.5	0.5	.045	.050	.313	.013	.207
$Y = X^2 + \varepsilon$						
0.0	0.0	---	.022	.009	-.112	---
0.5	0.0	---	.086	.021	-.019	---
0.0	0.5	---	.084	-.013	-.003	---
0.5	0.5	---	.121	.047	.077	---

Table 3: Estimated squared standard error

g	h	TS	C	F	R	OLS
$Y = X + \varepsilon$						
0.0	0.0	.031	.034	.037	.035	.038
0.5	0.0	.029	.035	.051	.051	.062
0.0	0.5	.035	.039	83.875	.047	.178
0.5	0.5	.034	.040	6.490	.063	2.452
$Y = X^2 + \varepsilon$						
0.0	0.0	---	.035	.033	.038	---
0.5	0.0	---	.074	.052	.076	---
0.0	0.5	---	.142	.559	.135	---
0.5	0.5	---	.159	1.018	.343	---

This process was repeated 1000 times yielding 1000 estimates of η^2, say $\eta_1^2, \ldots, \eta_{1000}^2$. Bias was estimate with

$$\frac{1}{1000}\sum(\hat{\eta}_i^2 - \eta^2)$$

and the squared standard error of η^2 was estimated with

$$\frac{1}{999}\sum(\eta_i^2 - \overline{\eta}^2)^2 ,$$

where $\overline{\eta}^2 = \sum \eta_i^2 / 1000$. The results are summarized in Tables 2 and 3 for the case n=30.

First consider bias. Method F performs well when the regression line is straight and when both X and ε have symmetric distributions. But when the distributions are skewed, bias can be severe, suggesting that method F be eliminated from consideration. Method R performs reasonably well, except under normality where it performs poorly. Increasing n to 100, it still performs poorly, in terms of bias, for this special case. Only method C has relatively low bias, and it competes well with OLS and method TS, even when the regression line is straight. However, when there is curvature, now the bias of method C is rather high compared to method F. Again, method R is found to be unsatisfactory under normality.

As for the squared standard error of the estimators, Table 3 indicates that method F can be relatively disastrous when the regression line is straight and sampling is from skewed distributions. And for heavy-tailed distributions, OLS does not perform well compared to methods C and R. Method R competes reasonably well with method C, but there are obvious exceptions. Generally, method C performed best among the situations considered.

To provide some sense of how method C improves when $Y = X^2 + \varepsilon$, as n gets large, some additional simulations were run with n=100 for the cases (g, h)=(0.0, 0.5) and (0.5, 0.5). Now the bias of method C was estimated to be .088 and .080, respectively. So for the skewed, heavy-tailed distribution considered here, the reduction in bias is substantial, but for the skewed, light-tailed distribution the amount of bias remains about the same. Method F has small bias for these situations, but its squared standard error is relatively high. Method R has about the same amount of bias as method C and a smaller standard error, but because it performs poorly in other situations, it would seem that it should be used with caution.

Conclusion

One limitation of the results reported here is that, when using a smoother, the span was chosen to be a fixed constant that is often used as the default value. Checks made when using method R indicate that a smaller span can improve its performance considerably. However, it remains unknown how best to adjust the span when estimating explanatory power, and even for the adjustments considered here (f=.7 and .5), it was found that method C remains a bit more satisfactory in most situations.

Although method C offers protection against the deleterious effects of outliers among the Y values, it is known that a sufficient number of outliers can negatively affect its performance relative to method R (Wilcox, 2005).

This was one of the main reasons for considering method R and it might explain why method C can be unsatisfactory when there is curvature and when dealing with extremely heavy-tailed distributions. Perhaps in most practical situations this is not an issue, but the extent to which this is true is difficult to determine.

When the usual variance is used, rather than the percentage bend midvariance, results in Doksum and Samarov (1995) suggest estimating explanatory power with r^2, the square of Pearson's correlation, rather than with the ratio of the variances of $\hat{Y}$ and Y. An analog of this approach is to use the percentage bend correlation (Wilcox, 2005, p. 391). Consideration was given to this approach, but it proved to be unsatisfactory in the simulations described here.

Perhaps the most surprising result is that there is little or no advantage to fitting a straight line to the data, versus using something like method C, when in fact the regression line is straight and when using the percentage bend variance. Consequently, method C is recommended for general use.

References

Bjerve, S. & Doksum, K. (1993). Correlation curves: Measures of association as functions of covariate values. *Annals of Statistics*, *21*, 890-902.

Buhlmann, P. & Yu, B. (2002). Analyzing bagging. *Annals of Statistics*, *30*, 927-961.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, *74*, 829-836.

Davison, A. C., Hinkley, D. V. & Young, G. A. (2003). Recent developments in bootstrap methodology. *Statistical Science*, *18*, 141-157.

Dielman, T., Lowry, C., & Pfaffenberger, R. (1994). A comparison of quantile estimators. *Communications in Statistics--Simulation and Computation*, *23*, 355-371.

Doksum, K., Blyth, S., Bradlow, E., Meng, X.-L., Zhao, H. (1994). Correlation curves as local measures of variance explained by regression. *Journal of the American Statistical Association*, *89*, 571-582.

Doksum, K. A. & Samarov, A. (1995). Nonparametric estimation of global functionals and a measure of the explanatory power of covariates in regression. *Annals of Statistics*, *23*, 1443--1473.

Efromovich, S. (1999). *Nonparametric curve estimation: Methods, theory and applications*. New York: Springer-Verlag.

Eubank, R. L. (1999). *Nonparametric regression and spline smoothing*. New York: Marcel Dekker.

Fan, J. (1993). Local linear regression smoothers and their minimax efficiencies. *Annals of Statistics*, *21*, 196-216.

Fan, J. & Gijbels, I. (1996). *Local polynomial modeling and its applications*. Boca Raton, FL: CRC Press.

Fox, J. (2001). *Multiple and generalized nonparametric regression*. Thousands Oaks, CA: Sage.

Frigge, M., Hoaglin, D. C., & Iglewicz, B. (1989). Some implementations of the boxplot. *American Statistician*, *43*, 50-54.

Green, P. J. & Silverman, B. W. (1993). *Nonparametric regression and generalized linear models: A roughness penalty approach.* Boca Raton, FL: CRC Press.

Gyorfi, L., Kohler, M., Krzyzk, A. Walk, H., & Gyorfi, L. (2002). *A distribution-free theory of nonparametric regression*. New York: Springer Verlag.

Hardle, W. (1990). Applied nonparametric regression. *Econometric Society Monographs No. 19*, Cambridge, UK: Cambridge University Press.

Hastie, T. J. & Tibshirani, R. J. (1990). *Generalized additive models*. New York: Chapman and Hall.

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

Lax, D. A. (1985). Robust estimators of scale: Finite-sample performance in long-tailed symmetric distributions. *Journal of the American Statistical Association*, *80*, 736-741.

Parrish, R. S. (1990). Comparison of quantile estimators in normal sampling. *Biometrics*, *46*, 247-257.

Sen, P. K. (1968). Estimate of the regression coefficient based on Kendall's tau. *Journal of the American Statistical Association*, *63*, 1379-1389.

Shoemaker, L. H. & Hettmansperger, T. P. (1982) Robust estimates and tests for the one- and two-sample scale models. *Biometrika*, *69*, 47-54.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. New York: Chapman and Hall.

Theil, H. (1950). A rank-invariant method of linear and polynomial regression analysis. *Indagationes Mathematicae*, *12*, 85-91.

Wilcox, R. R. (2003). *Applying contemporary statistical techniques*. New York: Academic Press.

Wilcox, R. R. (2005). *Introduction to robust estimation and hypothesis testing*. New York: Academic Press.

Journal of Modern Applied Statistical Methods
November, 2008, Vol. 7, No. 2, 376-384

1538 – 9472/08/$95.00

REGULAR ARTICLES
Comparing Factor Loadings in Exploratory Factor Analysis: A New Randomization Test

W. Holmes Finch
Ball State University

Brian F. French
Purdue University

Factorial invariance testing requires a referent loading to be constrained equal across groups. This study introduces a randomization test for comparing group exploratory factor analysis loadings so as to identify an invariant referent. Results show that it maintains the Type I error rate while providing adequate power under most conditions.

Key words: Exploratory factor analysis, randomization test, multigroup confirmatory factor analysis, invariance testing.

Introduction

Score validity evidence can be considered the primary focus in instrument development and evaluation (AERA, APA, & NCME, 1999). For instance, Standard 1.1 of the *Standards for educational and psychological testing* states “A rationale should be presented for each recommended interpretation and use of test scores, together with a comprehensive summary of the evidence and theory bearing on the intended use or interpretation” (p. 17, AERA et al., 1999). Measurement invariance (MI) or equivalence is one form of validity evidence that is important when scores are used for group comparisons. MI refers to the case where an assessment measures one or more latent constructs identically across groups. The presence of this property helps ensure that the measurement of the specified construct is the same across groups, thus allowing for accurate comparisons in score parameters. Otherwise group comparisons may be meaningless, as observed differences could be the result of ability differences or measurement differences.

Factor invariance is one form of measurement invariance (MI) and is typically established using multi-group confirmatory factor analysis (MCFA). Through MCFA, an *a priori* theoretically specified latent structure of an instrument is evaluated for MI across groups (Alwin & Jackson, 1981; Golembiewski, Billingsley, & Yeager, 1976). The presence of MI is tested using differences in the chi-square goodness-of-fit statistics for more (loadings held equal across groups) and less restrictive (loadings allowed to vary by group) models. If the fit of the models differs significantly, as measured by the chi-square difference test, the researcher concludes a lack of invariance. This method is well documented (e.g., Bollen, 1989; Byrne, Shavelson, & Muthén, 1989; Jöreskog & Sörbom, 1996; Maller & French, 2004; Raju, Laffitte, & Byrne, 2002; Reise, Widaman, & Pugh, 1993).

The requirement of an equality constraint of a referent indicator across groups in MCFA calls for methodological attention (Millsap, 2005). Comparison of a latent factor model can only occur if the same coordinate system is used for all groups in question (Wilson, 1981). Model identification procedures

W. Holmes Finch is Professor of Psychology in the Department of Educational Psychology, and Educational Psychology Director of Research in the Office of Charter School. Email: whfinch@bsu.edu. Brian F. French is Associate Professor and Co-Director Learning and Performance Research Center Washington State University. Email: frenchb@wsu.edu

ensure this required comparability by assigning the same units of measurement to the latent variables for groups in question (Jöreskog & Sörbom, 1996). Model identification is often accomplished by assigning the latent factors to a scale based on a common indicator across groups, typically either a factor variance or a factor loading for a single variable. The most common practice is to set one of these parameter values to 1.0 across groups, with the factor loading method being the most common (Brown, 2006; Vandenberg & Lance, 2000). This factor loading referent approach requires the assumption that the referent loading is equal for all groups in the population (i.e. the loading is assumed to be invariant).

When the referent parameter is not invariant, estimates of other model parameters, such as factor loadings, may be distorted and hypothesis tests for the group invariance of these other parameters may be inaccurate (Bollen, 1989; Cheung & Rensvold, 1999; Millsap, 2001). Therefore, a circular situation exists where (a) the referent loading must be invariant, (b) invariance of the referent (or any other) loading cannot be established without estimating a model, and (c) model estimation requires an invariant referent loading. Thus, we return to the original invariant referent assumption, which is commonly not assessed in practice, most likely due to the fact that there is not a relatively straight forward way of doing so. A procedure to locate an invariant referent variable would be useful to ensure the remainder of invariance assessment is accurate.

Heretofore, this assumption of referent invariance could not be directly tested (Bielby, 1986; Cheung & Rensvold, 1999; Wilson, 1981). A search procedure, the factor-ratio test and stepwise partitioning procedure, has been suggested (Rensvold & Cheung, 2001). The procedure uses each variable as the referent in a set of models with each other variable constrained to be invariant. The iterative procedure tests all pairs of variables (i.e., p (p – 1) / 2 pairs) and becomes quite complex as the number of indicator variables increases, making it not "user-friendly" for practitioners (Vandenberg, 2002). For example, a moderate length instrument (i.e., 30 indicators) requires 435 individual invariance tests to fully identify which loadings could be used as a referent in the final MCFA analysis. Evaluation of this procedure demonstrated adequate (e.g., acceptable false and true positives) but not perfect performance (French & Finch, 2006a).

Exploratory factor analysis (EFA) has been suggested as an alternative approach for identifying an invariant referent loading. In its relative simplicity, EFA overcomes the limitations associated with the factor-ratio test and search procedure. The EFA based approach involves conducting a single EFA for each group separately and descriptively comparing their respective loading estimates to ascertain which appear to be invariant in the sample. Such an analysis may be considered a weak test of factorial invariance (Zumbo, 2003) and is in accord with suggestions that EFA be used to examine loadings with an "interocular eyeball test" (Vandenberg, 2002, p. 152) to judge the similarity of loadings to identify referent variables. Evaluation of this procedure has been favorable (Finch & French, *in press*), though it does not offer a formal hypothesis test of invariance, instead allowing for the comparison of parameter estimates across groups in order to provide a sense of factor loading differences without the need to conduct a large number of analyses. Specifically, pattern coefficients appearing most similar would be eligible for serving as a referent variable in the MCFA. The obvious limitation to the current EFA procedure is the lack of a statistical test to give a formal determination about the differences between factor loadings.

The purpose of this study was to develop a randomization test based on EFA and to assess its utility in identifying invariant factor loadings between two groups. This procedure would be used prior to conducting the actual MCFA, as a purification process for identifying a loading that is likely to be group invariant and thus eligible for use as the referent parameter. The procedure entails conducting one EFA per group and then comparing the factor loadings (i.e., pattern coefficients) from the separate analyses via the test statistic to determine differences of individual loadings. Loadings that are significantly different would not be used as a referent.

Factor loading invariance randomization test (FLIRT)

Statisticians have developed exact tests for a number of applications involving group comparisons (see Good, 1994, for a thorough description of exact tests). Regardless of the context, every exact test for group comparison involves finding all possible permutations of the data, with respect to group membership. For each of these permutations the test statistic of interest is calculated and the collection of these statistics across all permutations forms a sampling distribution. The test statistic for the observed sample is also calculated and, if it is more extreme than a predetermined (e.g., 95th) percentile of the permutation distribution, the null hypothesis of no group difference can be rejected.

One common problem in the actual application of permutation tests is that, even for modestly sized samples, the number of permutations that must be determined can be large. For example, for a simple two group comparison with a total sample of 30 individuals (15 per group), the number of permutations would be 155,117,520. The computer time necessary to conduct analyses for each of these permutations would be prohibitive for any real application. An alternative approach to using all possible permutations is known as randomization, or Monte Carlo, testing (Edgington, 1980). With this methodology, a random sample of the permutations is selected and the test statistic of interest is calculated for each to create the sampling distribution as described above. As with the full permutation testing approach, the test statistic value obtained from the observed data is compared with this distribution and, if it is more extreme than some predetermined (e.g. 95th) percentile, the null hypothesis of no group difference is rejected. The description of the specific randomization test statistic for comparing two groups' factor loadings appears below.

The factor loading invariance randomization test (FLIRT) for comparing two groups' factor loadings is based upon the supposition that there exists configural invariance for the two groups; i.e., the basic factor structure is the same, though the actual factor loading values may not be. To test the null hypothesis of equal (invariant) group loadings for a single indicator variable, EFA is run separately for the two groups and the difference in the loadings for the target indicator is calculated. Next, 100 random samples are taken from the population of all possible permutations and for each of these EFA is conducted by group. The difference in the target loadings is calculated for each permutation to develop a distribution against which the group loading difference for the observed data is compared. If this observed difference is larger than the 95th percentile from the randomization distribution, the null hypothesis of no group differences on the target loading is rejected. The current study evaluated FLIRT through the use of a Monte Carlo simulation, as well as the analysis of a real dataset. The performance of the test was judged in terms of power and Type I error under a variety of conditions (e.g., sample size, factor model) in the simulation study, and by comparing hypothesis test results for the observed data with those presented in Thompson (2004).

Methodology

Simulated data were used to control variables that could influence the magnitude of factor loading estimates, with 1,000 replications for each combination of conditions described below. Simulations and analyses were completed in *SAS, V9.1* (The SAS Institute, 2003). Conditions were held as consistent as possible with previous studies (e.g., Finch & French, 2008 *in press*) for comparability of results. Second, a real data set, the LibQUAL+ study (Thompson, 2004), was employed to provide an applied example.

Number of Factors and Indicators

Data were simulated from both 1- and 2-factor models, with interfactor correlations set at .50 to represent moderately related factors, and simple structure for continuous and normally distributed subtest level data. The number of indicators per factor was 6.

Sample Size

The necessary sample size to obtain reasonable estimates in factor analysis varies depending on the data conditions. Four sample size conditions were simulated: 100, 250, 500, and 1,000 per group in order to reflect small, medium and large samples. These values are consistent with other factor analysis simulation studies (Cheung & Rensvold, 2002; Lubke & Muthén, 2004; Meade & Lautenschlager, 2004), ranging from poor (*n* = 100) to excellent (*n* = 1,000) (Comery & Lee, 1992), and may not be of much concern here as communalities were high (MacCallum, Widaman, Zhang, & Hong, 1999).

Magnitude of Difference with the Non-Invariant Indicators

Six levels of factor loading values for the non-invariant indicator were simulated. A baseline condition was established where no group differences in loadings were present, with all variables having a loading value of 0.75, including the target. The remaining 5 conditions were characterized by declines in the target loading from 0.10 to 0.50 in increments of 0.10 (i.e., 0.65, 0.55, 0.45, 0.35, and 0.25). This wide range of levels was selected since there is no effect size, at least to our knowledge, for what represents a meaningful difference (Millsap, 2005) and the range covers previously used values in MCFA simulation work (e.g., French & Finch, 2006b; Meade & Lautenschlager, 2004).

Contamination

The location of invariant parameters may be influenced by the number of indicators that lack invariance (Millsap, 2005; Yoon & Millsap, 2007). Thus, the presence of a factor loading, other than for the target indicator, exhibiting a group difference was varied as either present or absent. In other words, for half of the simulated conditions only the target indicator loading was contaminated, while for the other half of the simulations a second target indicator loading also was contaminated at the same difference as the target indicator. This allowed assessment of the influence of additional contaminated variables.

Analysis

All analyses were conducted by group using maximum likelihood factoring with PROMAX rotation in the 2-factor condition. These settings follow recommendations for using EFA for a referent indicator search and are more consistent with educational and psychological data (e.g., presence of measurement error, correlated factors; (Vandenberg, 2002).

Evaluation Criteria

The outcomes of interest for this study were the power and Type I error rates of the FLIRT. Specifically, the Type I error rate was calculated as the proportion of simulation replications for which the test statistic rejected the null hypothesis when the groups' loadings on a target indicator did not differ. In similar fashion, power was calculated as the proportion of the simulation replications for which the test statistic rejected the null hypothesis when the groups' loadings on the target indicator did in fact differ. To determine which conditions influenced the outcomes of interest, ANOVA and variance components analysis were used with each of the manipulated factors serving as an independent variable. For the applied data set results are presented in terms of locating differences in factor loadings as would be for an application.

Results

Simulation study

Type I error

None of the manipulated factors, or their interactions, was identified by the ANOVA as being significantly related to the Type I error rate of the FLIRT. Table 1 contains these Type I error rates by each of the manipulated variables. Overall, there is a very slight elevation of the error rate above the nominal 0.05, with the most notable difference between the 1 and 2 factor conditions. However, none of the sample differences evident in this table were statistically significant, suggesting that they may not be present in the population as a whole.

Power

Based on the results of the ANOVA and variance components analysis, the interaction of sample size by the difference in the groups' target loadings, as well as the main effects of

Table 1: Type I Error Rates by Sample Size, Number of Factors, and Level of Contamination

Sample size	Type I error rate
100	0.067
250	0.064
500	0.059
1000	0.060
Factors	
1	0.069
2	0.057
Contamination	
No	0.061
Yes	0.064

Table 2: Power by Sample Size and Group Difference in Target Loading

Sample size per group	Difference	Power
100	0.1	0.23
	0.2	0.61
	0.3	0.87
	0.4	0.96
	0.5	0.97
250	0.1	0.49
	0.2	0.92
	0.3	0.96
	0.4	1.00
	0.5	1.00
500	0.1	0.80
	0.2	1.00
	0.3	1.00
	0.4	1.00
	0.5	1.00
1000	0.1	0.97
	0.2	1.00
	0.3	1.00
	0.4	1.00
	0.5	1.00

sample size and difference in target loadings were statistically significant and contributed more than 10% of the variance to the power of the test statistic. Specifically, the interaction accounted for 38.4% of the variance as did the main effect of difference in loading values, while the main effect of sample size contributed an additional 20.2% to the variation of power. contains power rates by the interaction of sample size and group loading differences.

For the largest sample size condition, power was well above 0.95 regardless of the difference between the groups' loadings. Thus, even when the target loadings only differed by 0.1 the test statistic would virtually always identify this divergence. On the other hand, for samples of 100 per group, the test had power rates below 0.8 for differences of 0.1 and 0.2. In general, across the lower sample size conditions (100 and 250 most particularly), power was somewhat low for a difference of 0.1 but rose to above 0.8 for discrepancies in target loadings of 0.3 or more.

Table 3 shows power rates by the number of factors and level of contamination. Neither of these terms contributed more than 3% to the variance in power. A perusal of the results in this table shows that there were essentially no differences in power for 1 and 2 factors or when another loading beyond the target loading differed between the groups.

Table 3: Power by Number of Factors and Contamination

Number of factors	Power
1	0.90
2	0.88
Contamination	
No	0.89
Yes	0.89

Analysis of real data

To demonstrate the FLIRT in real world conditions, data taken from the LibQUAL+ study were analyzed. For a more complete discussion of this dataset and the study from which it was drawn, the interested reader is encouraged to consult Thompson (2004). The 12 items included on this survey could be divided into three factors, including service provided by library staff, the environment of the library and the quality of the library's holdings. Each factor was represented by 4 items, which were on a rating scale with response options ranging from 1 to 9. The dataset used, which is available in Thompson (2004), included a random sample of 200 survey respondents, 100 of whom were graduate students and 100 who were faculty members.

Thompson described differences in factor loading values between graduate students and faculty members for item 6, "A meditative place". To demonstrate the utility of the FLIRT with real data, the faculty and student loadings for item 6 were compared using this new statistic. The factor loading values by group were 0.7587 for graduate students and 0.9079 for faculty, leading to an observed loading difference of 0.1492. The distribution of differences across the 100 randomized datasets appears in Figure 1, a visual examination of which shows that the observed difference falls in the 99th percentile of the randomized values. Thus, if $\alpha = 0.05$, we would conclude that there is a statistically significant difference between the loading values for the two groups, which is in line with the conclusion reached by Thompson. The two groups loadings for item 5, "A haven for quiet and solitude", were also compared. This was not identified by Thompson as differing between the groups. The loading for the students was 0.9114, and 0.9342 for the faculty, leading to an observed difference of 0.0228. This value fell at the 46th percentile of the randomized loading differences, which would lead to a conclusion of no significant difference between group loadings at the aforementioned level of 0.05.

The purpose of this analysis with previously analyzed real data using MCFA was to demonstrate the potential utility of FLIRT. If FLIRT had been used as a step prior to the MCFA in this example, item 6 would not have been selected as a referent variable whereas item 5 could have been. The results presented are in accord with those of Thompson (2004), thus providing further evidence, beyond the simulation study, that this new statistic does appear to be reasonably accurate in correctly identifying group loading differences, even for samples as small as 100 per group.

Conclusion

The results suggest that in many instances, the FLIRT may be a useful tool for identifying potential indicator variables with invariant factor loadings across groups for use in a subsequent MCFA. This outcome was especially evident when the differences between loadings and/or the sample sizes were large. However, even for differences in loadings as small as 0.2 and samples of 100 per group, FLIRT was able to find differences more than 60% of the time. In all but one case, when sample size was 250 or more per group, the rates for correctly detecting loading differences were at least 0.8, and often near 1.0. Furthermore, the Type I error rates (identifying loadings as differing when they do not) were very close to the nominal rate of 0.05

for all studied conditions. The combination of these results supports the use of the new FLIRT statistic in conjunction with EFA for accurately detecting a non-invariant loading that could then be used as the referent in a subsequent MCFA.

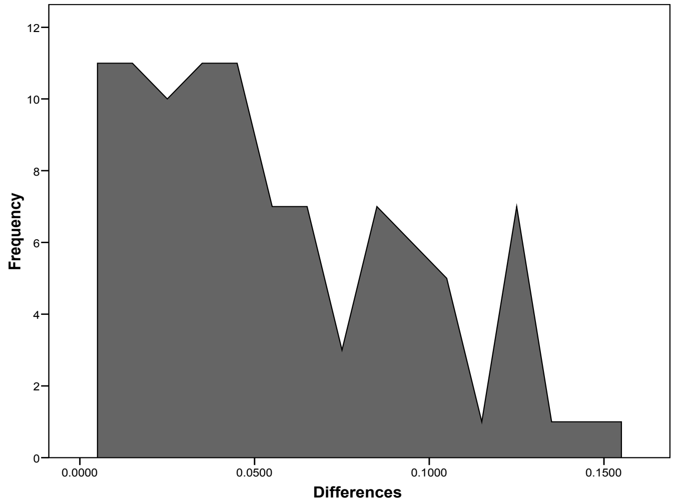
Correct specification of an invariant referent loading is a crucial step in MCFA. Failure to do so could lead to biased parameter estimates and, in turn, compromise other analyses, such as latent mean comparisons. The primary method suggested for identifying invariant indicators is the factor-ratio test and SP procedure (Rensvold & Cheung, 2001), which can be a very complex and time consuming multi-step technique. While this procedure does work reasonably well in identifying invariant referent loadings, it can become intractably time consuming with increasing model complexity (French & Finch, 2006a). To overcome such limitations, EFA is one approach that has been advocated for use in practice and involves comparison of factor loading estimates between two groups (Vandenberg, 2001; Zumbo, 2003). While this method does not have the advantage of significance testing that is offered by the factor-ratio test, it is much simpler to conduct. We have attempted to overcome the inference limitation of EFA, while maintaining its advantage of simplicity, by developing the FLIRT.

The results seem to indicate that in need to locate an invariant referent for use in MCFA they may find that this simple approach performs well in a fairly wide variety of study FLIRT generally provides an accurate conditions such as those simulated; EFA with assessment of identifying the variables that may lack invariance. Therefore, when practitioners conditions. The FLIRT is more accurate (i.e., greater power) with larger sample sizes and a greater magnitude of difference between loadings and appears to have Type I error rates that are always close to the nominal level.

Limitations and directions for future research

The generalizability of the results is limited to the conditions simulated in this study. First, the factor models examined were fairly simple (1 or 2 factors with 6 indicators each). Thus, in future research the FLIRT should be evaluated with more complex models and data (e.g., greater number of factors, different variables, various levels of communalities). Second, a related area that deserves attention is the combination of loadings for the observed variables. In this study, all of the loadings were set at 0.75 (unless contaminated). Given that this is the first investigation of the randomization test to accurately identify invariant referent variables, clarity of result interpretation was considered paramount, and thus non-target loadings were not varied. However, further investigation should be carried out for a more complex combination of loading values and factor models, as well as data conditions (e.g., ordinal variables) before the test is applied unequivocally.

Figure 1: Distribution of randomized loading differences for item 6



References

Finch, W. H., & French, B. F. (2008). Using exploratory factor analysis for locating invariant referents in factor invariance studies. *Journal of Modern and Applied Statistical Methods, 7(1), 223-233*.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Alwin, D. F. & Jackson, D. J. (1981). Applications of simultaneous factor analysis to issues of factorial invariance. In D. Jackson & E. Borgatta (Eds.), *Factor analysis and measurement in sociological research: A multi-dimensional perspective* (pp. 249-279). Beverly Hills, CA.: Sage.

Bielby, W. T. (1986). Arbitrary metrics in multiple-indicator models of latent variables. *Sociological Methods & Research*, *15*, 3-23.

Bollen, K. A. (1989). *Structural equations with latent variables.* New York, NY: Wiley.

Brown, T.A. (2006). *Confirmatory Factor Analysis for Applied Research*. The New York, NY: Guilford Press.

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures. *Psychological Bulletin*, *105*, 456-466.

Cheung, G. W., & Rensvold, R. B. (1999). Testing factorial invariance across groups: A reconceptualization and proposed new method. *Journal of Management*, *25*, 1-27.

Comrey, A. L., & Lee, H. B. (1992). *A first course in factor analysis.* Hillsdale, N.J.: L.Erlbuam Associates.

Erlbaum. Edgington, E.S. (1980). *Randomization Tests.* New York, NY: Marcel and Dekker, Inc.

French, B. F., & Finch, W. H. (2006a, June). *Locating the Invariant Referent in Multi-Group Confirmatory Factor Analysis.* Paper presented at the International Psychometric Society meeting in Montreal, Canada.

French, B. F., & Finch, W. (2006b). Confirmatory factor analytic procedures for the determination of measurement invariance. *Structural Equation Modeling*, *13*, 378-402.

Golembiewski, R.T., Billingsley, K. & Yeager, S. (1976). Measuring change and persistence in human affairs: Types of change generated by OD designs. *Journal of Applied Behavioral Science, 12,* 133-157.

Good, P. (1994). *Permutation Tests: A practical guide to resampling methods for testing hypotheses.* New York, NY: Springer-Verlag.

Jöreskog, K.G., & Sörbom, D. (1996). *LISREL 8: User's reference guide.* Chicago: Scientific Software.

Lubke, G. H., & Muthén, B. O. (2004). Applying multigroup confirmatory factor models for continuous outcomes to likert scale data complicates meaningful group comparisons. *Structural Equation Modeling*, *11*, 514-534.

MacCallum, R.C., Widaman, K.F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods, 4,* 84-99.

Maller, S. J., & French, B. F. (2004). Factor invariance of the UNIT across deaf and standardization samples. *Educational and Psychological Measurement*, *64*, 647-660.

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.

Millsap, R. E. (2001). When trivial constraints are not trivial: The choice of uniqueness constraints in confirmatory factor analysis. *Structural Equation Modeling*, *8*, 1-17.

Millsap, R.E. (2005). Four unresolved problems in studies of factorial invariance. In A. Maydeu-Olivares & J.J. McArdle (Eds.) *Contemporary Psychometrics* (pp. 153-172). Mahwah, NJ: Lawrence Erlbaum Associates.

Raju, N. S., Laffitte, L. J., & Byrne, B. M. (2002). Measurement equivalence: A comparison of methods based on confirmatory factor analysis and item response theory, *Journal of Applied Psychology*, *87*, 517-529.

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches to exploring measurement invariance. *Psychological Bulletin*, *114*, 552-566.

Rensvold, R. B., & Cheung, G. W. (2001). Testing for metric invariance using structural equation models: Solving the standardization problem. In C. A. Schriesheim & L. L. Neider (Eds.), *Research in management: Equivalence in measurement* (pp. 25-50). Greenwich, CT: Information Age Publishing.

SAS Institute (2004) *SAS version 9.1.3*. Cary, NC: SAS Institute, Inc.

Thompson, B. (2004). *Exploratory and confirmatory factor analysis: Understanding concepts and applications.* Washington, D.C.: American Psychological Association.

Vandenberg, R. J. (2002). Toward a further understanding of and improvement in measurement invariance methods and procedures. *Organizational Research Methods*, *5*, 139-158.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, *3*, 4-69.

Wilson, K. L. (1981). On population comparisons using factor indexes or latent variables. *Social ScienceResearch*, *10*, 301-313.

Yoon, M., & Millsap, R. E. (2007). Detecting violations of factorial invariance using data-based specification searches: A monte carlo study. *Structural Equation Modeling*, *14*, 435-463.

Zumbo, B. D. (2003). Does Item-Level DIF Manifest Itself in Scale-Level Analyses?: Implications for Translating Language Tests. *Language Testing*, *20*, 136-147.

Journal of Modern Applied Statistical Methods
November, 2008, Vol. 7, No. 2, 385-397

Type I Error Rates of the Kenward-Roger *F*-test for a Split-Plot Design with Missing Values and Non-Normal Data

Miguel A. Padilla
Old Dominion University

YoungKyoung Min
Korea Foundation for the
Advancement of Science and Creativity

Guili Zhang
East Carolina University

The Type I error of the Kenward-Roger (KR) *F*-test was assessed through a simulation study for a between- by within-subjects split-plot design with non-normal ignorable missing data. The KR-test for the between- and within-subjects main effect was robust under all simulation variables investigated and when the data were missing completely at random (MCAR). This continued to hold for the between-subjects main effect when data were missing at random (MAR). For the interaction, the KR *F*-test performed fairly well at controlling Type I under MCAR and the simulation variables investigated. However, under MAR, the KR *F*-test for the interaction only provided acceptable Type I error when the within-subjects factor was set at 3 and 5% missing data.

Keywords: missing values, Kenward-Roger *F*-test, robustness, mixed models, split-plot design, non-normal data, and covariance heterogeneity.

Introduction

Linear mixed-effects, or mixed models, have become increasingly popular in analyzing data from split-plot designs such as longitudinal research designs. The increased popularity can be attributed to at least three factors. Linear mixed-effects models (LMEM) offer modeling flexibility in that the fixed effects, random effects, and the covariance structure can all be modeled. Also, parameters of LMEMs are estimated via maximum likelihood and hence have the asymptotic properties of being unbiased and efficient. In addition, because LMEM parameters are estimated through ML, the parameters can still be consistently estimated with missing data as long as the data are missing completely at random (MCAR) or missing at random (MAR) (Rubin, 1976). It is this last property which may ultimately account for the increased popularity of LMEMs. Even so, it is unclear exactly under which conditions LMEMs will have consistent parameter estimates when there are missing data.

When applying LMEM to split-plot designs, it is usually inferences about the fixed effects that are of main interest. Within this endeavor, a typical strategy is to try to fit a model for the means and select an appropriate covariance structure. The model is then tested for fit and appropriate modifications are made if required in order to test for inferences of interest (Wolfinger, 1993). A likelihood ratio, score, or Wald test can be used to test hypothesis about the fixed effect, but the Wald test is more commonly used (Schaalje, McBride, & Fellingham, 2002b; Brown & Prescott, 2006). The Wald test has good large sample properties, but they begin to dwindle with smaller sample sizes. However, using Satterthwaite-type degrees of freedom (Fai & Cornelius, 1996) can improve Wald test small sample properties. In addition to adjusting the degrees of freedom, the Wald test's small sample properties can further be enhanced by adjusting the covariance matrix (Kenward & Roger, 1997). Several simulation studies have shown that tests based on the Satterthwaite (SW) and Kenward-Roger (KR)

Miguel A. Padilla is Assistant Professor of Quantitative Psychology. Email: mapadill@odu.edu. YoungKyoung Min is Senior Research Scientist. Email: ykymin@yahoo.com. Guili Zhang is Assistant Professor of Research and Evaluation Methodology. Email: zhangg@ecu.edu.

adjustments tend to behave well (Keselman et al., 1998; Schaalje, McBride, & Fellingham, 2002a; Padilla & Algina, 2007). In particular, the KR-test tends to behave well even with missing data (Padilla et al., 2007).

The small sample situation can further be complicated by missing data. It is a common occurrence in research and can have dramatic affects on the properties of standard statistical models, such as ordinary least squares regression. The way in which missing data will affect statistical models largely depends on the type of missing data mechanism and the way in which the missing data is handled. As an example, by far the most common method for handling missing data is to perform listwise deletion, also known as complete case analysis. This is most likely because it is the default in most popular statistical packages (e.g., SAS, SPSS, etc.). Nevertheless, if the data are MAR, parameter estimates can be biased and hence inference can be inaccurate. Additionally, there will be some loss of power in that participants with at least one missing value will be completely discarded from the analysis. If the small sample condition is added to this situation then the problems only worsen, adding another layer of uncertainty about inferences being drawn.

There are two major alternatives to handling missing data: multiple imputation (MI) and maximum likelihood (ML). Although both methods are a vast improvement over listwise deletion – and virtually any other method for handling missing data – the focus here will be on ML within the framework of split-plot designs and LMEMs. The reader interested in MI is referred to Schafer (1997) and Little & Rubin (2002).

The split-plot design is commonly used in behavioral research, such as educational and psychological research (Keselman et al., 1998b). It is, in essence, a hybrid of a between- and within-subjects designs incorporating elements of both. A longitudinal study is a typical split-plot design in that it has a between-subjects factor represented by subjects that are randomly assigned to treatment groups and a within-subjects factor represented by the measured multiple time points for each subject. Split-plot designs have various ways in which to analyze the data they generate and each of those methods have their strengths and limitations in terms of analyzing the data and how they handle missing values or data. However, the one promising technique for analyzing data from a split-plot with missing values is the linear mixed or mixed-effects model estimated through ML. Before delving on, the three missing data mechanisms are described.

Missing Data Mechanisms

The three general definitions of missing data, ordered from most restrictive to least restrictive, are missing completely at random (MCAR), missing at random (MAR), and not missing at random (NMAR) (Rubin, 1976; Little & Rubin, 2002, p. 12). As described by Verbeke & Molenberghs (2000), let $f\left(\boldsymbol{r}_i \mid \boldsymbol{y}_i, \boldsymbol{X}_i, \boldsymbol{\psi}\right)$ denote the distribution of the missing data indicator or missing data mechanism for the i^{th} participant, where $\boldsymbol{r}_i$ is a $K \times 1$ vector containing zero for missing and one for observed scores in the corresponding $K \times 1$ $\boldsymbol{y}_i$ vector of repeated measurements or variables, $\boldsymbol{X}_i$ is the design matrix for the factors, and $\boldsymbol{\psi}$ contains the parameters of the relationship of $\boldsymbol{r}_i$ to $\boldsymbol{y}_i$ and $\boldsymbol{X}_i$. Furthermore, $\boldsymbol{y}_i$ can be partitioned as $\boldsymbol{y}_i = \left(\boldsymbol{y}'_{i(obs)} \quad \boldsymbol{y}'_{i(miss)}\right)'$ where $\boldsymbol{y}'_{i(obs)}$ has observed scores and $\boldsymbol{y}'_{i(miss)}$ has missing scores for the i^{th} participant. The full data density can then be factorized as:

$$f\left(\boldsymbol{y}_i, \boldsymbol{r}_i \mid, \boldsymbol{X}_i, \boldsymbol{\theta}, \boldsymbol{\psi}\right) = f\left(\boldsymbol{y}_i \mid \boldsymbol{X}_i, \boldsymbol{\theta}\right) f\left(\boldsymbol{r}_i \mid \boldsymbol{y}_i, \boldsymbol{X}_i, \boldsymbol{\psi}\right) \tag{1}$$

where $\boldsymbol{\theta} = \left(\boldsymbol{\beta}', \boldsymbol{\sigma}'\right)'$, $\boldsymbol{\beta}$ contains the fixed effects parameters, and $\boldsymbol{\sigma}$ contains the nonredundant parameters of the covariance matrix. This factorization is the foundation of selection modeling because the factor to the far right corresponds to the selection of individuals into observed or missing groups. The missing data are MCAR if $f\left(\boldsymbol{r}_i \mid \boldsymbol{y}_i, \boldsymbol{X}_i, \boldsymbol{\psi}\right) = f\left(\boldsymbol{r}_i \mid \boldsymbol{X}_i, \boldsymbol{\psi}\right)$, that is, the distribution of the missing data indicators does not depend on the repeated measures or variables. The missing data are

MAR if $f\left(\boldsymbol{r}_i \mid \boldsymbol{y}_i, \boldsymbol{X}_i, \boldsymbol{\psi}\right) = f\left(\boldsymbol{r}_i \mid \boldsymbol{y}_{i(obs)}, \boldsymbol{X}_i, \boldsymbol{\psi}\right)$, that is, the distribution of the missing data indicator does not depend on the variables in which the i^{th} participant has missing scores. In general, missing data are NMAR if they are not MCAR or MAR. However, it is generally defined as $f\left(\boldsymbol{r}_i \mid \boldsymbol{y}_i, \boldsymbol{X}_i, \boldsymbol{\psi}\right) = f\left(\boldsymbol{r}_i \mid \boldsymbol{y}_{i(miss)}, \boldsymbol{X}_i, \boldsymbol{\psi}\right)$, that is, the distribution of the missing data indicator depends on the missing values in the data.

A general method for consistent ML estimation of $\boldsymbol{\theta}$ is obtained by including both the missing data indicators ($\boldsymbol{r}_\text{i}$) and the parameters of their relationship to $\boldsymbol{y}_i$ and $\boldsymbol{X}_i$ ($\boldsymbol{\psi}$) in the likelihood. The likelihood of the full data density can then be written as:

$$L\left(\boldsymbol{\theta}, \boldsymbol{\psi} \mid \boldsymbol{X}_i, \boldsymbol{y}_i, \boldsymbol{r}_i\right) \propto f\left(\boldsymbol{y}_i, \boldsymbol{r}_i \mid \boldsymbol{X}_i, \boldsymbol{\theta}, \boldsymbol{\psi}\right) \qquad (2)$$

If the missing data mechanism is MCAR or MAR and if $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ are disjoint, ML estimators of $\boldsymbol{\theta}$ will be consistent if $\boldsymbol{r}_\text{i}$ and $\boldsymbol{\psi}$ are excluded from the analysis (Rubin, 1976). Dropping $\boldsymbol{r}_\text{i}$ and $\boldsymbol{\psi}$ is referred to as ignoring the missing data mechanism. Hence, MCAR or MAR missing data mechanisms are ignorable when model parameters ($\boldsymbol{\theta}$) are estimated via ML. If data are MCAR, listwise deletion and ML ignoring the missing data mechanism will produce consistent estimators, but ML estimators will be more precise because they use all available data.

In addition, Rubin (1976) showed that MCAR missing data mechanisms are ignorable for inferences based on sampling distributions. Thus, listwise deletion or ML ignoring the missing data mechanism can be used for inferences if the data are MCAR, but ML will result in more powerful inferences and narrower confidence intervals because it does not delete individuals with only partially observed scores on $\boldsymbol{y}_i$.

On the other hand, the validity of ML based inferences for a MAR missing data mechanism will depend on how the sampling covariance matrix is estimated. When the missing data mechanism is MAR, it will be ignorable if inferences are based on the sampling covariance obtained from the observed information matrix (Kenward & Molenberghs, 1998). This is in line with arguments from Efron & Hinkley (1978) in that the observed information matrix provides much better precision than the expected information matrix; that is, better variance component estimates. If ML inferences are based on the sampling covariance obtained from the expected information matrix, the MAR missing data mechanism may not be ignorable. The expected information matrix must take into account the actual sampling process implied by the MAR mechanisms in order for inferences to be valid (Kenward et al., 1998).

When the missing data mechanism is NMAR, then it is non-ignorable for purposes of ML estimation. In order to obtain consistent ML estimates in this particular case, the pattern of the missing values must be taken into account. A selection model that incorporates the missing values indicators ($\boldsymbol{r}_\text{i}$) or using a pattern mixture model that stratifies the data on the basis of the pattern of missing values can be used to obtain consistent ML estimates under an NMAR framework (Albert & Follmann, 2000; Diggle & Kenward, 1994; Fitzmaurice, Laird, & Shneyer, 2001; Kenward et al., 1998; Kenward, 1998; Troxel, Harrington, & Lipsitz, 1998; Algina & Keselman, 2004a; Algina & Keselman, 2004b; Little, 1995).

Linear Mixed-Effects Model

The linear mixed-effects model (LMEM) can be written as

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Zu} + \boldsymbol{\varepsilon} \qquad (3)$$

where $\boldsymbol{X}$ and $\boldsymbol{\beta}$ are the design matrix and its corresponding fixed effects vector, $\boldsymbol{Z}$ and $\boldsymbol{u}$ are the design matrix and its corresponding random effects vector, and $\boldsymbol{\varepsilon}$ is the vector of random errors. It is generally assumed that $\boldsymbol{u}$ and $\boldsymbol{\varepsilon}$ are independent, hence

$$\begin{bmatrix} \boldsymbol{u} \\ \boldsymbol{\varepsilon} \end{bmatrix} \sim N\left(\begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{0} \end{bmatrix}, \begin{bmatrix} \boldsymbol{G} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{R} \end{bmatrix} \right) \qquad (4)$$

Based on this assumption, $E(\boldsymbol{y}) = \boldsymbol{X\beta}$ and $Var(\boldsymbol{y}) = \boldsymbol{V} = \boldsymbol{ZGZ}' + \boldsymbol{R}$. A common estimator for $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}'\hat{\boldsymbol{V}}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\hat{\boldsymbol{V}}^{-1}\boldsymbol{y} \tag{5}$$

Also, $Var\left(\hat{\boldsymbol{\beta}}\right) = \left(\boldsymbol{X}'\hat{\boldsymbol{V}}^{-1}\boldsymbol{X}\right)^{-1}$ is the estimated generalized least-squares covariance of $\hat{\boldsymbol{\beta}}$.

Let $\boldsymbol{L}$ be a contrast matrix of full row rank *r*. Then the main effect and interaction hypothesis about the between- and within-subjects factors can be expressed as $H_0 : \boldsymbol{L\beta} = \boldsymbol{0}$. The common test statistic for this hypothesis is the Wald

$$F_{r,\,ddf} = \frac{\left(\boldsymbol{L}\hat{\boldsymbol{\beta}}\right)'\left[\boldsymbol{L}\left(\boldsymbol{X}'\hat{\boldsymbol{V}}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{L}'\right]^{-1}\left(\boldsymbol{L}\hat{\boldsymbol{\beta}}\right)}{r} \tag{6}$$

where *ddf* is the denominator degrees of freedom. It should be noted that, under the null hypothesis, the Wald $F_{r,\,ddf}$ approximately follows an *F* distribution. However, there are times when it follows an *F* distribution exactly. Even so, when there is no missing data, $\left(\boldsymbol{X}'\hat{\boldsymbol{V}}^{-1}\boldsymbol{X}\right)^{-1}$ tends to underestimate $\left(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X}\right)^{-1}$ and hence is a biased estimate because it fails to take into account the uncertainty introduced by using $\hat{\boldsymbol{V}}$ (Booth & Hobert, 1998; Kackar & Harville, 1984; Prasad & Rao, 1990).

Kenward-Roger *F*-Test

Better estimates were developed as a response to the poor statistical properties of $Var\left(\hat{\boldsymbol{\beta}}\right)$. The first estimate, denoted as $Var\left(\hat{\boldsymbol{\beta}}^{@}\right) = \hat{\boldsymbol{m}}^{@}$, was proposed by Harville & Jeske (1992). Subsequently, Kenward & Roger (1997) developed an alternative estimator, denoted as $Var\left(\hat{\boldsymbol{\beta}}_A\right) = \hat{\boldsymbol{\Phi}}_A$. Additionally, Kenward & Roger derived the test statistic

$$F^{*}_{r,\,d} \simeq \lambda\frac{\left(\boldsymbol{L}\hat{\boldsymbol{\beta}}\right)'\left(\boldsymbol{L}\hat{\boldsymbol{\Phi}}_A\boldsymbol{L}'\right)^{-1}\left(\boldsymbol{L}\hat{\boldsymbol{\beta}}\right)}{r} \tag{7}$$

where λ is a scaling factor and *d* is the approximate denominator degrees of freedom. As in the case of $F_{r,\,ddf}$, $F^{*}_{r,\,d}$ is assumed to follow an *F* distribution under the null hypothesis. Both λ and *d* are calculated from the data. First, $\hat{\boldsymbol{\Phi}}_A$ is estimated to account for small sample bias in $\left(\boldsymbol{X}'\hat{\boldsymbol{V}}^{-1}\boldsymbol{X}\right)^{-1}$ and variability introduced by using $\hat{\boldsymbol{V}}$ (Kackar et al., 1984). Then *d* is approximated by using the spectral decomposition of $\left(\boldsymbol{L}\hat{\boldsymbol{\Phi}}_A\boldsymbol{L}'\right)^{-1}$ concurrently with repeated applications of the single degree of freedom *t*-test (Fai et al., 1996; Giesbrecht & Burns, 1985). The Kenward-Roger (KR) *F*-test is implemented in SAS PROC MIXED, but uses $\hat{\boldsymbol{m}}^{@}$ instead of $\hat{\boldsymbol{\Phi}}_A$. (See Padilla & Algina, 2007) for how to specify model parameters using the mean vector and an indicator matrix for the missing values.)

Some research has been conducted investigating the Type I error rate of the KR method (Fai et al., 1996; Kenward & Roger, 1997; Kowalchuk, Keselman, Algina, & Wolfinger, 2004; Gomez, Schaalje, & Fellingham, 2005). However, very little research is available on the Type I error rate of the KR method when there are missing values. To date, Padilla & Algina (2007) is the only work investigating the Type I error rate of the KR *F*-test when the missing values are MAR.

Fai & Cornelius (1996) derived four test statistics (F_1, F_2, F_3, F_4) for hypothesis testing on the means in multivariate data. The F_1 and F_2 statistics use $\left(\boldsymbol{X}'\hat{\boldsymbol{V}}^{-1}\boldsymbol{X}\right)^{-1}$ whereas F_3 and F_4 use $\hat{\boldsymbol{m}}^{@}$ to estimate $Var\left(\hat{\boldsymbol{\beta}}\right)$. Additionally, F_2 and F_4 have scaling factors λ_2 and λ_4, respectively. The F_1 statistic is available in SAS PROC MIXED when the Satterthwaite option is used for DDFM. The F_4 statistic is similar to the PROC MIXED KR *F*-test, but uses a different formula for the scaling factor and denominator degrees of freedom. (See Fai & Cornelius for further details.)

Fai & Cornelius (1996) applied their tests to simulated data from four unbalanced 3 (between) × 4 (within) split-plot designs with a compound symmetric covariance structure. Imbalance was created by varying the number of subjects of the between-subjects factor without generating some combinations of subjects and the within-subjects factor. Missing data were never actually generated; hence the missing data mechanism is MCAR. The four unbalanced designs had total sample sizes of $N = 25, 34, 40, 48$. Estimated Type I error rate and power were reported for the between-subjects main effects. All tests controlled the Type I error rate reasonably well. The results of F_1 and F_3 were similar, and power and Type I error were always larger for F_4 than for F_3.

In their initial work, Kenward & Roger (1997d) investigated the Type I error rate of the KR *F*-test in simulated data from four research designs: (a) a four-treatment, two-period cross-over, (b) a row-column-α design, (c) a random coefficients regression model for repeated measures data, and (d) a split-plot design. Design (c) had MCAR missing values and (d) had missing values with an unspecified missing data mechanism. Estimated Type I error rates were reported for the between-subjects main effect. In all situations, the KR *F*-test Type I error rate was well controlled.

Kowalchuk, Keselman, Algina, & Wolfinger (2004) compared the Type I error rates of the KR and Welch-James (WJ) *F*-tests under several simulation conditions for a 3 (between) × 4 (within) split-plot design. Investigated conditions were (a) type of covariance structure, (b) group size inequality, (c) positive and negative parings of covariance matrices with group sample sizes, (d) shape of data distribution, and (e) type of covariance structure fit to data. A heterogeneous covariance structure with a 1:3:5 ratio was used for all simulation conditions, and missing values were not investigated. Estimated Type I error rates were reported for the main effects and interaction. Under all conditions with small sample sizes (total *N* = 30, 40), the Type I error rate of the KR *F*-test were closer to the target value $(\alpha = .05)$ than the WJ *F*-test. Additionally, the Type I error rates of the KR *F*-test were always comparable when using an unstructured covariance matrix to modeling the true covariance matrix.

Gomez, Schaalje, & Fillingham (2005) investigated the Type I error rate of the KR *F*-test when using AIC (Akaike, 1974) and BIC (Schwarz, 1978) to select the covariance structure. Investigated conditions were (a) type of covariance structures with within- and between-subjects heterogeneity (1:3:5 ratio for between-subjects), (b) equal $(total\ N = 9, 15)$ and unequal group sample sizes $(n = 3, 5, 7)$, (c) positive and negative paring for unequal group sample sizes, (d) and levels of the within-subjects factor $(K = 3, 5)$. The between-subjects factor was fixed at 3 and no missing values were investigated. Estimated Type I error rates were reported for the main effects only. In general the Type I error rate was close to the target value when the correct covariance structure was used. However, the Type I error rate becomes inflated with complex covariance structures and small sample sizes. Additionally, the Type I error rate increased with heterogeneity within- and between-subjects, and even more so with negative pairings. In general, the success rate of choosing the correct covariance structure was low for both the AIC and BIC. At most, the success rate was 73.91%. Even so, the success rate was higher for the larger sample sizes and simpler covariance structures. Lastly, the AIC had better success with complicated covariance structures and the BIC with simpler ones.

Padilla & Algina (2007) studied the Type I error rate of the KR *F*-test with missing values and heterogeneity of covariance matrices $(1:3:5\ ratio)$. Investigated conditions were (a) level of between-subject factor (*J*), (b) level of within-subject factor (*K*), (c) $n_{\min}/(K-1)$, (d) sample size inequality, (e) degree of sphericity, (f) covariance and group sample size pairing, (g) missing data mechanism (MCAR or MAR), and (h) percent of missing data. Estimated Type I error rates were reported for the main effects and interaction. In general, the Type I error rates of the KR *F*-test were close to the target value of $\alpha = .05$ for the between- and within-subjects main effects and the between- by within-subjects interaction. The best Type I error control was attained by the between-subjects main effect with the between- by within-subjects interaction attaining the worst. However, the distribution of the data was normal.

The previous studies demonstrate that the Type I error rate of the KR *F*-test remains close to the target value $(\alpha = .05)$ under a variety of repeated measures designs and simulation conditions, which included MCAR unbalanced data. However, Padilla & Algina (2007) is the only study to investigate the Type I error rate of the KR *F*-test under the MAR condition in normal data. This study builds on Padilla & Algina and investigates the Type I error rate of the KR *F*-test under several simulation conditions. Of particular interest is the KR *F*-test Type I error rate when data are non-normal with missing values as it is implemented in SAS PROC MIXED.

Methodology

Design

The simplest of the split-plot design with one between- and one within-subjects factor (*i.e.*, $J \times K$) with heterogeneity between the j^{th} covariance matrix and non-normal data was investigated. In this type of design subjects are randomly assigned to the levels of the between-subjects factor $\left(j = 1, 2, \ldots, n; \sum_j n_j\right)$ and measured under all levels of the within-subjects factor $(k = 1, 2, \ldots, K)$. The heterogeneity between the j^{th} covariance matrices was set at 1:3:5; that is $\boldsymbol{\Sigma}_1 = 1/3\,\boldsymbol{\Sigma}_2$ and $\boldsymbol{\Sigma}_3 = 5/3\,\boldsymbol{\Sigma}_2$ (Algina & Keselman, 1997; Keselman, Algina, Kowalchuk, & Wolfinger, 1999; Padilla et al., 2007; Keselman, Carriere, & Lix, 1993). The non-normal data were generated from a multivariate lognormal distribution under the null using the methods outlined in Algina & Oshima (1994) with skewness set at 1.75 and kurtosis at 5.90 (Keselman, Algina, Wilcox, & Kowalchuk, 2000; Kowalchuk, Keselman, Algina, & Wolfinger, 2004).

All simulations and analyses were done on SAS 9.1. The PROC MIXED code for estimating the Kenward-Roger *F*-test can be found in Padilla and Algina (2007).

Simulation Variables

Eight variables were investigated. The variables of interest are (a) number of levels of the between-subjects factor (*J*), (b) number of levels of the within-subjects factor (*K*), (c) sample size, (d) sample size inequality across the j^{th} groups, (e) degree of sphericity, (f) pairing of the j^{th} group sizes with covariance matrices, (g) type of missing data, and (h) percent of missing data. Because this study builds on Padilla & Algina (2007), the simulation variables here are similar to theirs.

Between- and Within-Subjects Factors

The between- and within-subjects factors each had two levels with $J, K = 3, 6$.

Sample Size

Sample sizes were based on the $n_{\min}/(K-1)$ ratio (Keselman, Carriere, & Lix, 1993b). The ratios were set as in Padilla & Algina (2007) and for the same reasons. The actual sample sizes used, in combination with sample size inequality, are displayed in Tables 1 and 2.

Table 1:
Groups Sizes for Each Level of *J* at *K* = 3

	Sample Size Inequality			
J	C ≈ .16	C ≈ .33	C ≈ .16	C ≈ .33
	$n_{\min}/(K-1) = 4.0$		$n_{\min}/(K-1) = 6.0$	
3	8	8	12	12
	10	14	15	20
	12	20	18	28
	$n_{\min}/(K-1) = 5.0$		$n_{\min}/(K-1) = 7.7$	
6	10	10	15	15
	13	17	19	25
	16	24	23	35
	10	10	15	15
	13	17	19	25
	16	24	23	35

Table 2:
Groups Sizes for Each Level of J at $K = 6$

	Sample Size Inequality			
J	C ≈ .16	C ≈ .33	C ≈ .16	C ≈ .33
	$n_{\min}/(K-1) = 4.0$		$n_{\min}/(K-1) = 6.0$	
3	20	20	30	30
	25	34	37	50
	30	48	44	70
	$n_{\min}/(K-1) = 5.0$		$n_{\min}/(K-1) = 7.7$	
6	25	25	38	38
	31	42	47	64
	37	59	56	90
	25	25	38	38
	31	42	47	64
	37	59	56	90

Sample Size Inequality

Unequal sample sizes are common in split-plot designs and hence were investigated here (Keselman et al., 1998). The unequal group sample size were investigated through the coefficient of variation as defined by Keselman et al. (1993):

$$C = \left(\bar{n}\sqrt{J}\right)^{-1}\sqrt{\sum_{j=1}^{J}\left(n_j - \bar{n}\right)^2} \qquad (8)$$

where $C \simeq .16, .33$ describes moderate and severe group sample size inequality, respectively.

Covariance Sphericity

Sphericity as quantified by Box's epsilon (1954) was investigated with $\varepsilon = .60, .75, .90$. Here, $\varepsilon = .60$ represents a relatively severe departure from sphericity whereas $\varepsilon = .75$ a moderate one. Epsilon values were chosen based on the argument that $\varepsilon = .75$ represent the lower limit of ε found in educational and psychological data (Huynh & Feldt, 1976). (See Padilla & Algina (2007) for the actual covariance matrices.)

Group Pairing with Covariance

Pairing of the unequal group samples sizes and heterogeneous covariance matrices were investigated. The two conditions investigated were positive and negative pairings because positive pairing tend to produce conservative Type I error rates whereas negative pairings tend to produce liberal ones (Keselman & Keselman, 1990). A positive pairing occurs when the largest n_j is paired with the covariance matrix with the largest elements and a negative pairing occurs when the largest n_j is paired with the covariance matrix with the smallest elements. For positive pairings, the ratios of group sample size to heterogeneity of covariance matrices was set at $5:3:1$ for $J = 3$ and $5:3:1:5:3:1$ for $J = 6$. For negative pairings, it was set at $1:3:5$ for $J = 3$ and $1:3:5:1:3:5$ for $J = 6$.

Missing Data Mechanism

Both MCAR and MAR missing data mechanisms were investigated. The missing data mechanisms were simulated as described by Padilla & Algina (2007). NMAR was not investigated because it negatively impacts the Type I error rate of the KR F-test in a repeated measures designs with no between-subjects factor and normal data (Padilla & Algina, 2004).

Percent of Missing Data

Five percent (5%) and 15% probability of missing data at each level of the within-subjects factor were investigated. The exception here is that there was no missing data in the first level. Higher missing data probabilities were not investigated because the sample sizes are considerably small (see Table 1) and this will impede the convergence of the Newton-Raphson algorithm.

Analysis

The p-values of KR F-test were available from 5,000 replications for each combination of the simulation variables. The Type I error for each of the p-values was defined as

$$\textit{Type I Error} = \begin{cases} 0 \; \textit{if} \;\; p - \text{value} < .05 \\ 1 \; \textit{otherwise} \end{cases}.$$

Logistic regression models were used to analyze the between-subjects main effect, within-subjects main effect, and the between- by within-subjects interaction of the KR *F*-test separately. In each logistic model the Type I error variable was used as the dependent variable with the simulation variables as the independent variables. A forward selection approach was used to select appropriate models beginning with the intercept-only model and moving up to main effect only, main effect with two-way interaction, etc. A model adequately fit the data if the χ^2 goodness of fit test was non-significant or if $CFI \geq .95$ (Bentler, 1990). With large sample sizes (i.e., number of replications), the χ^2 goodness of fit statistic is sensitive to small effects, hence a fit index was used to supplement the χ^2. In this context, the CFI is calculated as follows:

$$CFI = 1 - \left(\lambda / \lambda_i\right) \tag{9}$$

where $\lambda = \max\left(\chi^2 - df, 0\right)$ with χ^2 (the test statistic) and *df* (the degrees of freedom) for the fitted model and $\lambda_i = \max\left(\chi_i^2 - df_i, \chi^2 - df, 0\right)$ with χ_i^2 and *df_i* for the intercept-only model.

Bradley's (1978) liberal criterion was used to assess the Type I error rates. The liberal criterion is $.5\alpha \leq \tau \leq 1.5\alpha$ where α is the nominal Type I error and τ is the empirical Type I error. With $\alpha = .05$ the liberal range is $.025 \leq \tau \leq .075$. Hence if the Type I error is within the range, the test is considered to be robust.

Results

Between-Subjects Main Effect

The logistic model with main effects and two-way interactions had $\chi^2(339) = 388.40$, $p = .0331$ and $CFI = .98$. Inspection of all two-way interaction tables indicated that for the between-subjects main effects all Type I error rates were within Bradley's liberal criterion. In fact the range of the Type I error rates across all two-way interaction tables was [.051, .071]. Even though the KR *F*-test for the between-subjects main effect does appear to be slightly liberal, it is not too strongly affected by the simulation variables.

Within-Subjects Main Effect

The logistic model with main effects and three-way interactions had $\chi^2(262) = 261.76$, $p = .4925$ and $CFI = 1.00$. Therefore, the three-way interaction model was selected for further analysis. Wald tests of the logistic model indicated that levels of the *within-subjects factor* (*K*), *group pairing with covariance*, *missing data mechanism*, and *percent of missing data* had significant main effects and also entered into the most significant three-way interactions.

Mean Type I error rates are displayed in Table 3. The range of mean Type I error rates under MCAR was [.054, .067]. Although slightly liberal, the mean Type I error rates are well within Bradley's liberal criterion. Under MAR, the situation changes dramatically. In fact, the mean Type I error rates were all liberal ranging from $[.079, .158]$ and above Bradley's liberal criterion. Furthermore, the mean Type I error rate increases as both the levels of the *within-subjects factor* (*K*) and *percent of missing data* increases. On the other hand, under MAR, the mean Type I error rate decreases as the *group pairing with covariance* changes from positive to negative (consistent with Keselman et al., 1990).

Table 3: Within-Subjects Main Effect

Missing Data Mechanism	% Missing	*K*	Group Pairing	
			Positive	Negative
MCAR	5	3	.0625	.0670
		6	.0543	.0572
	15	3	.0634	.0670
		6	.0607	.0631
MAR	5	3	**.0794**	**.0794**
		6	**.0938**	**.0880**
	15	3	**.1078**	**.0986**
		6	**.1580**	**.1389**

Note: Type I error rate above Bradley's liberal criterion are in bold type.

Between- by Within-Subjects Interaction

The logistic model with main effects and three-way interactions had $\chi^2(262) = 308.64$, $p = .0252$ and $CFI = 1.00$. Hence, the three-way interaction model was selected for additional analysis. Wald tests of the logistic model indicated that *K*, *J*, *sample size*, *group pairing with covariance*, *covariance sphericity*, and *percent of missing data* had significant main effects. However, *K*, *J*, *sample size inequality*, *group pairing with covariance*, *missing data mechanism*, and *percent of missing data* entered into the most significant three-way interactions. Thus, these latter simulation variables were selected for further analysis.

Mean Type I error rates under MCAR are displayed in Table 4. With the exception of 15% missing data, a negative pairing, and a severe group sample size inequality, the majority of mean Type I error rates are within Bradley's liberal criterion. However, the mean Type I error rates increase as the *percent of missing data*, *K*, and *J* increases and as *group pairing* changes from positive to negative. As noted above, the situation becomes more aggravated under the most severe conditions of the simulation variables.

Mean Type I error rates under MAR are presented in Table 5. Here, most of the mean Type I error rates are outside of the range of the Bradley's liberal criterion. The only time the mean Type I error rate is controlled is under the simplest of conditions for *group pairing with covariance*, *K*, and *J*. Nevertheless, as was the case for the MCAR condition, the mean Type I changes from positive to negative. The one difference is that mean Type I error rate error rates tend to increase as *percent of missing data*, *K*, and *J* increases and as *group pairing* increases as the *sample size inequality* becomes more severe. Not surprising the mean Type I error rates become more liberal under the more severe conditions of the simulation variables.

Conclusion

The results indicate that sampling distribution based inferences on the means for the between-subjects factor of a split-plot design using ML estimates can control the Type I error rate under an MCAR and MAR missing data mechanism and non-normal data. Furthermore, the Type I error control can be achieved with relatively small to moderate sample sizes when using the KR *F*-test. The same cannot be said of inferences about the within-subjects factor or the within- by between-subjects interaction.

The Type I error rates of the KR *F*-test for the latter two cases are impacted by several conditions of the simulation variable with the most dramatic being the MAR condition. This is most clearly seen in inferences about the within-subjects factor, in which case none of the Type I error rates were acceptable. Under MCAR, increasing the percent of missing data and switching from a positive to negative pairing of groups with covariance matrices tended to increase the Type I error rate, but the Type I error rate was still within Bradley's (1978) liberal criterion. Although the same pattern of increase in Type I error rate is observed under MAR, the increase in Type I error rate was

Table 4: MCAR for Interaction

% Missing	Group Pairing	Sample Size Inequality	$K = 3$		$K = 6$	
			$J = 3$	$J = 6$	$J = 3$	$J = 6$
5	Positive	Moderate	.0495	.0575	.0446	.0549
		Severe	.0468	.0527	.0456	.0513
	Negative	Moderate	.0642	**.0769**	.0569	.0607
		Severe	**.0787**	**.0864**	.0614	.0673
15	Positive	Moderate	.0503	.0582	.0562	.0584
		Severe	.0517	.0564	.0542	.0629
	Negative	Moderate	.0679	**.0774**	.0637	.0709
		Severe	**.0781**	**.0910**	.0715	**.0769**

Note: Type I error rate above Bradley's liberal criterion are in bold type.

Table 5: MAR for Interaction

% Missing	Group Pairing	Sample Size Inequality	$K = 3$		$K = 6$	
			$J = 3$	$J = 6$	$J = 3$	$J = 6$
5	Positive	Moderate	.0504	.0637	.0665	**.0859**
		Severe	.0523	.0661	.0728	**.0994**
	Negative	Moderate	.0672	**.0843**	.0696	**.0977**
		Severe	**.0823**	**.1041**	**.0820**	**.1181**
15	Positive	Moderate	.0561	**.0789**	**.0987**	**.1582**
		Severe	.0668	**.0963**	**.1247**	**.2009**
	Negative	Moderate	.0699	**.0991**	**.0954**	**.1665**
		Severe	**.0916**	**.1338**	**.1200**	**.2208**

Note: Type I error rate above Bradley's liberal criterion are in bold type.

sharper and obvious when switching from MCAR to MAR in which case none of the Type I error rates were within Bradley's liberal criterion.

With regard to the within- by between-subjects interaction, the KR *F*-test is once again severely impacted by several of the simulation conditions, but more dramatically by the MAR condition. Under the MCAR condition the majority of the Type I error rates are within Bradley's liberal criterion. When the within-subjects factor is 3, the same pattern is observed for 5% and 15% missing data: a negative pairing of groups with covariance matrices coupled with severe sample size inequality increased the Type I error rate above the liberal criterion. When the within-subject factor is 6, the Type I error rate was above the liberal criterion only under the more severe simulation conditions. Under MAR, most of the Type I error rates were above the liberal criterion. The only time the Type I error rates were consistently within the liberal criterion was when the within-subjects factor was 6, 5% of the data were missing, and there was positive pairing of groups with covariance matrices. The remaining acceptable Type I error rates tended to occur when the between-subjects factor was 3 and under the least severe of the simulation conditions. Even so, the Type I error rate tended to increase as the simulation conditions switched into the more severe conditions investigated.

By far the MAR condition had the largest impact on the Type I error rate of the KR *F*-test for the within-subjects factor and the within- by between-subjects interaction. It is clear that missing values coupled with non-normal data impact the accuracy of the *F*-distribution as an approximation to the sampling distribution of the KR *F*-test. The KR *F*-test uses an adjusted estimator of the covariance which is then used to estimate Satterthwaite type degrees of freedom. The procedure provides a better approximation to the *F*-distribution with small sample sizes (Kenward & Roger, 1997). This seemed to be the case for the between-subjects factor under all the simulation variables of this study. However, for the within-subjects factor and the within- by between-subjects interaction, it appears that the MAR condition coupled with non-normal data severely limited the KR *F*-test's ability to control the Type I error.

Two potential reasons exist for this result. First, SAS PROC MIXED does not compute the covariance matrix by inverting the Hessian (information matrix) for the fixed effects and the covariance parameters. According to Verbeke & Molenberghs (2000), the observed Hessian should be used and not the expected Hessian. Again, the observed Hessian provides more precision than the expected Hessian (Efron & Hinkley, 1978). Second, sample sizes were too small; particularly when the within- and between-subjects factors were both set at six. Although the samples sizes were based on the recommendations set by Keselman et al. (1993a) and Algina & Keselman (1997), those studies did not have missing values, which is not the case here. Here it appears that missing values coupled with data non-normality put a heavy burden on the analysis. A simple solution is to increase the sample sizes. However, doing so will increase the computation time of PROC MIXED's KR procedure, but it should provide more information for the procedure to use. However, increasing the sample sizes is not easy in practice.

The KR *F*-test for the between-subjects factor appears to be robust, in terms of controlling the Type I error, to non-normal data under the simulation variables investigated. Also, the KR *F*-test for the within-subjects factor is robust to non-normal data under the simulation variables investigated as long as the missing data mechanism is MCAR. The KR *F*-test for the within- by between-subjects interaction performed fairly well under MCAR, but care should be taken when using it when the within-subjects factor is three and the more extreme conditions of the simulation variables. Unfortunately, the KR *F*-test for the within- by between-subjects interaction is not robust under MAR and the simulation variables investigated. The only time the KR *F*-test for the interaction provided acceptable Type I error rates was when the within-subjects factor was set at 3 and only 5% of the data were missing. More work is required in order to fully assess the KR *F*-test's Type I error rate under missing values and non-normal data.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transaction on Automatic Control*, *19*, 716-723.

Albert, P. S., & Follmann, D. A. (2000). Modeling repeated count data subject to informative dropout. *Biometrics*, *56*, 667-677.

Algina, J., & Keselman, H. J. (1997). Testing repeated measures hypotheses when covariance matrices are heterogeneous: Revisiting the robustness of the Welch-James test. *Multivariate Behavioral Research*, *32*, 255-274.

Algina, J., & Keselman, H. J. (2004a). A comparison of methods for longitudinal analysis with missing data. *Journal of Modern Applied Statistical Methods*, *3*, 13-26.

Algina, J., & Keselman, H. J. (2004b). Assessing treatment effects in randomized longitudinal two-group designs with missing observations. *Journal of Modern Applied Statistical Methods*, *3*, 271-287.

Algina, J., & Oshima, T. C. (1994). Type-I error rates for Huynh general approximation and improved general approximation tests. *British Journal of Mathematical & Statistical Psychology*, *47*, 151-165.

Booth, J. G., & Hobert, J. P. (1998). Standard errors of prediction in generalized linear mixed models. *Journal of the American Statistical Association*, *93*, 262-272.

Box, G. E. P. (1954). Some theorems on quadratic forms applied in the study of analysis of variance problems, II. Effects of inequality of variance and of correlation between errors in the two-way classification. *The Annals of Mathematical Statistics*, *25*, 484-498.

Bradley, J. V. (1978). Robustness? *The British Journal of Mathematical & Statistical Psychology*, *31*, 144-152.

Brown, H., & Prescott, R. (2006). *Applied mixed models in medicine*. (2nd ed.) Wiley: New York.

Diggle, P. D., & Kenward, M. G. (1994). Informative dropout in longitudinal data analysis. *Applied Statistics*, *43*, 49-93.

Efron, B., & Hinkley, D. V. (1978). Assessing the accuracy of the maximum likelihood estimator: Observed versus expected Fisher information. *Biometrika*, *65*, 457-481.

Fai, A. H. T., & Cornelius, P. L. (1996). Approximate F-tests of multiple degree of freedom hypotheses in generalized least squares analyses of unbalanced split-plot experiments. *Journal of Statistical Computation and Simulation*, *54*, 363-378.

Fitzmaurice, G. M., Laird, N. M., & Shneyer, L. (2001). An alternative parameterization of the general linear mixture model for longitudinal data with non-ignorable drop-outs. *Statistics in Medicine*, *20*, 1009-1021.

Giesbrecht, F. G., & Burns, J. C. (1985). Two-stage analysis based on a mixed model: Large-sample asymptotic theory and small-sample simulation results. *Biometrics*, *41*, 477-486.

Gomez, E. V., Schaalje, G. B., & Fellingham, G. W. (2005). Performance of the Kenward-Roger method when the covariance structure is selected using AIC and BIC. *Communications in Statistics-Simulation and Computation*, *34*, 377-392.

Harville, D. A., & Jeske, D. R. (1992). Mean squared error of estimation or prediction under a general linear-model. *Journal of the American Statistical Association*, *87*, 724-731.

Huynh, H., & Feldt, L. S. (1976). Estimation of the Box correction for degrees of freedom from sample data in randomized block and split-plot designs. *Journal of Educational Statistics*, *1*, 69-82.

Kackar, R. N., & Harville, D. A. (1984). Approximations for standard errors of estimators of fixed and random effects in mixed linear models. *Journal of the American Statistical Association*, *79*, 853-862.

Kenward, M. G. (1998). Selection models for repeated measurements with non-random dropout: An illustration of sensitivity. *Statistics in Medicine*, *17*, 2723-2732.

Kenward, M. G., & Molenberghs, G. (1998). Likelihood based frequentist inference when data are missing at random. *Statistical Science*, *13*, 236-247.

Kenward, M. G., & Roger, J. H. (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics*, *53*, 983-997.

Keselman, H. J., Algina, J., Kowalchuk, B. K., & Wolfinger, R. D. (1999). A comparison of recent approaches to the analysis of repeated measurements. *British Journal of Mathematical & Statistical Psychology*, *52*, 63-78.

Keselman, H. J., Algina, J., Wilcox, R. R., & Kowalchuk, R. K. (2000). Testing repeated measures hypotheses when covariance matrices are heterogeneous: Revisiting the robustness of the Welch-James test again. *Educational and Psychological Measurement*, *60*, 925-938.

Keselman, H. J., Carriere, K. C., & Lix, L. M. (1993). Testing repeated-measures hypotheses when covariance matrices are heterogeneous. *Journal of Educational Statistics*, *18*, 305-319.

Keselman, H. J., Huberty, C. J., Lix, L. M., Olejnik, S., Cribbie, R. A., Donahue, B., et al. (1998). Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA, and ANCOVA analyses. *Review of Educational Research*, *68*, 350-386.

Keselman, H. J., & Keselman, J. C. (1990). Analyzing unbalanced repeated measures designs. *British Journal of Mathematical and Statistical Psychology*, *43*, 265-282.

Kowalchuk, R. K., Keselman, H. J., Algina, J., & Wolfinger, R. D. (2004). The analysis of repeated measurements with mixed-model adjusted F tests. *Educational and Psychological Measurement*, *64*, 224-242.

Little, R. J. A. (1995). Modeling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association*, *90*, 1112-1121.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data*. (2nd ed.) New York: Wiley.

Padilla, M. A., & Algina, J. (2004). Type I error rates for a one factor within-subjects design with missing values. *Journal of Modern Applied Statistical Methods*, *3*, 406-416.

Padilla, M. A., & Algina, J. (2007). Type I error rates of the Kenward-Roger adjusted degree of freedom F-test for a split-plot design with missing values. *Journal of Modern Applied Statistical Methods*, *6*.

Prasad, N. G. N., & Rao, J. N. K. (1990). The estimation of the mean squared error of small-area estimators. *Journal of the American Statistical Association*, *85*, 163-171.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*, 581-592.

Schaalje, G. B., McBride, J. B., & Fellingham, G. W. (2002a). Adequacy of approximations to distributions of test statistics in complex mixed linear models. *Journal of Agricultural Biological and Environmental Statistics*, *7*, 512-524.

Schaalje, G. B., McBride, J. B., & Fellingham, G. W. (2002b). Adequacy of approximations to distributions of test statistics in complex mixed linear models. *Journal of Agricultural Biological and Environmental Statistics*, *7*, 512-524.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. London: Chapman & Hall/CRC.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*, 461-464.

Troxel, A. B., Harrington, D. P., & Lipsitz, S. R. (1998). Analysis of longitudinal data with non-ignorable non-monotone missing values. *Journal of the Royal Statistical Society Series C-Applied Statistics*, *47*, 425-438.

Verbeke, G., & Molenberghs, G. (2000). *Linear mixed models for longitudinal data*. New York: Springer-Verlag.

Wolfinger, R. (1993). Covariance structure selection in general mixed models. *Communications in Statistics-Simulation and Computation*, *22*, 1079-1106.

Journal of Modern Applied Statistical Methods
November, 2008, Vol. 7, No. 2, 398-407

A Randomization Method to Control the Type I Error Rates in Best Subset Regression

Yasser A. Shehata
Productivity and Quality Institute

Paul White
University of the West of England

A randomization method for the assessment of statistical significance for best subsets regression is given. The procedure takes into account the number of potential predictors and the inter-dependence between predictors. The approach corrects a non-trivial problem with Type I errors and can be used to assess individual variable significance.

Key words: best subset regression, randomization, Type I error, bias.

Introduction

Subset selection in multiple linear regression is long established: computational algorithms for forward selection techniques date back at least to the 1950's, (see Kramer, 1957), and Canning (1959) gave an example of backward elimination. The use of subset selection techniques is widespread and continuing. George (2000) wrote "The problem of variable selection is one of the most pervasive model selection problems in statistical application. The use of variable selection procedures will only increase as the information revolution brings us larger data sets with more and more variables. The demand for variable selection will be strong and it will continue to be a basic strategy for data analysis."

The use of automated computer techniques for model building is rife. Some researchers use automated search algorithms as a data mining exercise (Lovell, 1983), examining a research question by collecting data on virtually every variable that could possibly be related to the phenomenon under investigation and attempting to obtain a parsimonious model based on patterns in sample data. In recognition of this type of problem Larzelere and Mulaik (1977) suggested basing inferences on the total number of potential predictors rather than the number of predictors in a given subset.

It is commonly argued that a purpose of automated selection techniques is to obtain a simple, high-quality representation of the phenomenon under investigation. This is accomplished by not including potential predictors deemed to be uninformative in a final model. Models based on smaller numbers of predictor variables are comparatively easier to understand and it is hoped that a parsimonious model will give greater insight into the underlying processes that generated the data. In some instances smaller subsets may lead to greater economy (Derksen & Keselman, 1992).

Problems relating to variable selection from using backward elimination, forward selection, best subset regression and other automated model building techniques are well documented in the context of multiple linear regression. Investigations have generally been through simulation work in which the theoretical underpinning model assumptions are satisfied and any deviation between simulation results and anticipated theoretical results is therefore attributable to the variable selection technique. For instance, the simulation work of Derksen &

Yasser A. Shehata is Lecturer at the Arab Academy for Science, Technology and Maritime Transport, Alexandria, Egypt. Email: Yasser.Shehata@live.uwe.ac.uk. Paul White is Senior Lecturer and a member of the Applied Statistics and Quantitative Methods Consultancy in the Department of Mathematics & Statistics, Bristol, UK. Email: Paul.White@uwe.ac.uk.

Keselman (1992) gave broad the conclusions that automated selection techniques overly capitalize on false associations between potential predictors and the criterion variable with too many purely random (noise) variables being wrongly classified as authentic (true) predictors. The inclusion of noise variables in a final model necessarily implies a model misspecification or misidentification and incorrect inferences are drawn. Derksen & Keselman (1992) additionally found that the incidence with which noise and authentic variables find, or do not find, their way into a final model depends upon the degree of correlation between predictor variables. As such, it would seem that controlling the error rate may require a solution which explicitly utilizes within sample correlation information.

Hurvich & Tsai (1990) pointed out that, if a model is not fully pre-specified and, if a model selection technique is used, then the number of regression parameters is a random variable. Moreover, once a model has been decided upon by some technique, the model estimation and the associated hypothesis tests usually proceed on the assumption that the data driven and technique selected model is the true model. In other words, the data is analyzed "as though they were a fresh data set generated by the selected model" (Hurvich & Tsai, 1990, p. 214). Under these conditions, as pointed out by Miller (1984), the regression estimators may be biased and standard hypothesis tests may not be valid.

Automated model building techniques, such as stepwise regression, proceed on the basis of performing many statistical tests and do so in instances whereby the hypothesis test procedure may not be valid. Multiplicity of testing contributes to model selection problems. In the context of stepwise regression Derksen & Keselman (1992) wrote "when many tests of significance are computed in a given experiment, the probability of making at least one Type I error in the set of tests, that is, the maximum familywise Type I error rate (MFWER), is far in excess of the probability associated with any one of the tests" (p. 269). In subset selection there are a potentially large number of statistical tests to be performed to drive the algorithms. The number of such tests is not known in advance and simple Bonferroni corrections may be too liberal in correcting this problem, especially when potential predictors are not orthogonal. Paradoxically, others have suggested that a more liberal approach is appropriate. Bendel & Afifi (1977) advocated the use of nominal significance levels between $\alpha = 0.15$ and $\alpha = 0.25$ in forward selection so as to include all authentic variables at the expense of an increased risk of including additional noise variables in a final model.

The all subsets approach searches through all possible subsets for each subset size of $1, 2, ..., J$ and best subsets chooses the one that has the best summary statistics for a given subset size. A possible best summary statistic is the R^2 statistic (the coefficient of determination). An advantage of the best subsets and all subsets approach over sequential procedures is that this approach, by definition, will not miss finding the best fitting subset of any given size. Indeed, Mantel (1970) pointed out, and gave instances and explanations that a multivariate combination of variables might produce the best fit, but these multivariate combinations might not be identified by sequential procedures. Further, Kuk (1984) pointed out a relative weakness of sequential procedures in that "they lead to a single subset of variables and do not suggest alternative good subsets" unlike all subsets and further points out that sequential procedures have "the possibility of premature termination" (Kuk, 1984, p. 587). Identification of best subsets need not necessarily be computationally burdensome as the identification of the best subset does not require the calculation of all possible subsets (Furnival & Wilson, 1974).

The above provides a strong rationale for considering best subsets regression. The standard inferential approach for best subsets regression has problems arising from using standard hypothesis tests based on a global null hypothesis of no effect for a model determined by sample data. Motivated by the stance of Larzelere & Mulaik (1977) the use of randomization to control Type I error rates in best subsets regression is considered, and the approach takes into account the total number of predictors under consideration. Derksen & Keselman (1992) concluded that the extent of the problem with automated techniques depends upon the degree of correlation between predictor variables. The use of randomization permits the correlation structure between potential predictor variables to be accounted for. The approach adopted is to compute *p*-values for overall model significance and for each variable under a global null model (as per standard approaches) but which will correct the bias associated with the procedural aspects of best subsets regression. Randomization additionally permits a like-for-like comparison for individual variables that comprise a best subset solution; topics which are expanded in this article.

A brief overview of the traditional least squares approach to determine overall model significance of a best subset regression solution in addition to the individual significance of the variables that comprise the model is first given. Next, a randomization approach that empirically estimates overall and individual significance of best subset regression is described. Descriptions of two models are given, namely a global null-model and a non-null model. These two models, under certain conditions, are used to compare the performance of the randomization algorithm with the traditional approach. Results of the simulation, effects of number of predictors and effects of sample size are provided. The discussion addresses issues concerning the paradoxical problems associated with judging inference in best subsets regression.

Methodology

Best Subsets Regression

Consider the classic linear regression model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_J X_J + \varepsilon \ (1)$$

where Y is the dependent variable with J predictors $X_1, X_2, ..., X_J$ and where ε denotes a normally distributed random variable. Let $y_i, x_{1i}, x_{2i}, ..., x_{Ji}, \ (i = 1,2,......, I)$ denote I independent cases generated from the above model.

In best subsets regression, the best subset of size j is the subset of j predictor variables that maximizes the within sample prediction of the dependent variable, *y*, in a linear least squares regression. This is the percentage of variation in *y* that is accounted for by a regression equation is the usual R^2 statistic. In the following, R_j^2 will be used to denote the R^2 statistic for the best subset of size *j*. Overall significance of the best subset of size *j* is judged using the standard F statistic, $F = S_R^2 \big/ S_E^2$ where S_R^2 is the mean square due to regression, S_E^2 is the mean square error and overall model significance is judged by making reference to the F distribution with $(\upsilon_1, \upsilon_2) = (j, I - j - 1)$ degrees of freedom.

The relative magnitude of the observed value of the *F* statistic is quantified by the *p*-value and contemporary practice is to declare a statistically significant subset of predictors whenever $p < 0.05$. In addition, let S_p^2 denote the change in the error sum of squares for deleting a variable X_p from a regression model. An assessment of the statistical significance of X_p in the model is made by referring $F = S_p^2 / S_E^2$ to the *F* distribution with degrees of freedom $(\upsilon_1, \upsilon_2) = (1, I - j - 1)$. For a detailed explanation of best subsets of regression see Draper & Smith (1981, p. 303).

If the potential predictor variables $X_j, (j = 1,2,......, J)$, are noise variables, i.e. unrelated to Y in as much as $\beta_j = 0, (j = 1,2,......, J)$, then the *p*-values for judging overall model significance for any subset of size *j*, should be uniformly distributed *U*(0, 1). Thus, if a researcher works at the α significance level and, if none of the potential predictor variables are related to *Y*, then a Type I error in assessing significance of the overall best subset model should only be made $\alpha\%$ of the time for any value $\alpha \in (0,1)$. Arguably, the same requirement should also apply to individual predictor variables. An alternative procedure for assessing the overall significance of any best subset of size *j* and for assessing the statistical significance of each variable included in the best subset model is proposed. This alternative procedure, a randomization method, does not make explicit use of the properties of the *F* distribution. Ordering the variables that comprise a best subset solution in terms of their individual *F* values is also considered along with deriving an estimate of their *p*-value by considering similarly ordered *F* values under randomization.

Randomization

Consider sample data $y_i, x_{1i}, x_{2i}, ..., x_{Ji}$, $(i = 1, 2, ..., I)$, and let R_j^2 denote the coefficient of determination for the best subset of size $j, (j = 1, 2, ..., J)$. Next consider where the order of cases for the predictor variables in the data is randomly permuted but with the response variable held fixed at $y_i, x_{1i}, x_{2i}, ..., x_{Ji} \rightarrow y_i, x_{1k}, x_{2k}, ..., x_{Jk}$. This random permutation of predictor records ensures that the sample correlation structure between the predictors in the original data set is precisely preserved in the newly created randomized data set. The random permutation also ensures that the predictor variables in the randomized data set are stochastically independent of the response, *Y*, but may be correlated with *Y* in any sample through a chance arrangement.

Best subsets regression can be performed on the newly created randomized data set. Let S_j^2 denote the coefficient of determination for the best subset of size $j, (j = 1, 2, ..., J)$ for the randomized data set. If for subset $j, S_j^2 > R_j^2$, then the randomized chance solution may be viewed as having better within sample predictability than the observed data.

For any given data set many permutations of the original data set may be generated by taking another random permutation. In what follows the proportion of instances that $S_j^2 > R_j^2$ is estimated through simulation. This estimate is taken to be an estimate of the *p*-value for determining the statistical significance of R_j^2 for any subset of size *j*. For a given data set, an increase in the number of random permutations will serve to increase the accuracy of the estimated value. The above procedure may be summarized as follows:

For given data set and for a subset of size *j*:

```
1. Determine the best subset of predictors of size $j$ and record the coefficient of
   determination $R_j^2$
2. Set KOUNT = 0
3. DO n = 1 TO N
   a. Randomly permute $x_{1i}, x_{2i}, ......, x_{Ji}$ independently of $y_i$ i.e.
      $y_i, x_{1i}, x_{2i}, ......, x_{Ji} \rightarrow y_i, x_{1k}, x_{2k}, ......, x_{Jk}$
   b. For the newly created fake data set determine the best subset of size $j$ and
      record the coefficient of determination $S_j^2$
   c. If $S_j^2 > R_j^2$ then KOUNT = KOUNT+1
4. ENDDO
5. Estimated p-value = KOUNT/N
```

The counting process effectively estimates rank position of the original solution in relation to randomization solutions. Under the randomization process all permutations are equally likely. Likewise if the original predictors are generated under a system whereby none of them are related to the outcome then the observed value of R_j^2 is just a likely to be as large as any value of S_j^2 obtained from random permutation.

In a similar way for best subset of size *j*, consider the *F*-values for each predictor variable arranged in order, $F_{(1)} > F_{(2)} > .... > F_{(j)}$. The *F*-values from a random permutation may be ordered in a similar way, i.e. $F_{(1)}^* > F_{(2)}^* > .... > F_{(j)}^*$. The proportion of times $F_{(p)}^* > F_{(p)}$ provides an estimate of the *p*-value of the *p*-th ordered variable in the observed best subset solution.

Simulation Design

For a specific application consider the model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon \, . \quad (2)$$

To illustrate the properties of the proposed technique, four specific parameter settings (referred to in the following as Model A and Model B) with two different correlation structures have been considered.

Model A is a genuine null model with $\beta_0 = 1$ and $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$ so that all proposed predictors are noise variables and are unrelated to the outcome *Y*. For Model B consider $\beta_0 = 1$, $\beta_1 = 0.5$, $\beta_2 = \beta_3 = \beta_4 = 0$ (i.e., one authentic variable and three noise variables).

In the following simulations each model is considered with potential predictor variables being (i) Case 1, stochastically independent in which their correlation matrix is the identity matrix, and (ii) Case 2, strongly correlated with elements of the correlation matrix being $\rho(X_1, X_2)$ = 0.708, $\rho(X_1, X_3)$ = 0.802, $\rho(X_1, X_4)$ = –0.655, $\rho(X_2, X_3)$ = 0.757, $\rho(X_2, X_4)$ = –0.582, $\rho(X_3, X_4)$ = –0.593, where $\rho(X_l, X_m)$ denotes Pearson's correlation coefficient between X_l and X_m.

In all instances the error terms are independent, identically distributed realizations from the standard normal distribution $(\mu = 0,\ \sigma^2 = 1)$, so that the underpinning assumptions for the OLS linear regression models are satisfied. Simulations herein are reported based on *I* = 30 cases per simulation instance and increasing sample size and increasing the number of potential predictors are considered.

Results

Figure 1 is a percentile-percentile plot of the *p*-values obtained from implementing the aforementioned algorithm for step *j* = 1, 2, 3 in best subsets regression for Model A with potential predictor variables being stochastically independent. The vertical axis denotes the theoretical percentiles of the uniform distribution *U*(0, 1) and the horizontal axis represents the empirically derived percentiles based on 500 simulations with each simulation based on 1,000 randomization instances. Note that the *p*-values based on the traditional method are systematically smaller than required, indicating that the true Type I error rate for overall model significance is greater than any pre-chosen nominal significance level α. By contrast the estimated *p*-values based on the randomization algorithm have an empirical distribution that is entirely consistent with the uniform distribution *U*(0, 1) for any subset of size 1, 2, or 3 out of 4 predictors.

Under Model A, qualitatively similar results are obtained for *j* = 1, 2, 3 for potential predictors being correlated, Case 2. For *j* = 4 there is no subset selection under the simulations and in these cases both the traditional method and the randomization method have *p*-values uniformly distributed *U*(0, 1).

Simulations under Model B for step *j* = 1, 2 in best subsets regression with independent predictors, Case 1, or with correlated predictors, Case 2, correctly show that the proposed method retains power at any level of α; the power is marginally lower than the power under the traditional method (see Figure 2), but this is expected due to the liberal nature of the traditional method.

Once overall model significance has been assessed, a normal practice is to assess the individual significance of each variable alone. Figure 3 is a percentile-percentile plot of the *p*-values for the variables that comprise the best subset of size *j* = 3 of 4 under Model A, Case 1. In this instance the three variables included in the model have been ordered according to their *F*-values. The traditionally computed *p*-value for the variable with the largest *F*-value is typically too small when judged against the uniform distribution, *U*(0, 1). Contrary, for the variable with the smallest *F*-value the *p*-values calculated using the standard method are typically too large when judged against the uniform distribution, *U*(0, 1). By contrast, the *p*-value under the randomization method, for all ordered effects, is entirely consistent with the uniform distribution *U*(0, 1).

Qualitatively similar results are obtained for Model A but for potential predictors being correlated, Case 2.

Simulations under a true null model (i.e. with all potential predictors being noise variables), for *J* = 4, 8, 16, 32, 64 keeping the number of cases fixed, *I* = 30, have been performed. In all of these cases the simulations show that the *p*-value for overall subset

Figure 1: Percentile – Percentile plot for *p*-values for overall significance for best subset of size *j* = 1, 2, 3 from 4 independent predictors, Model A.

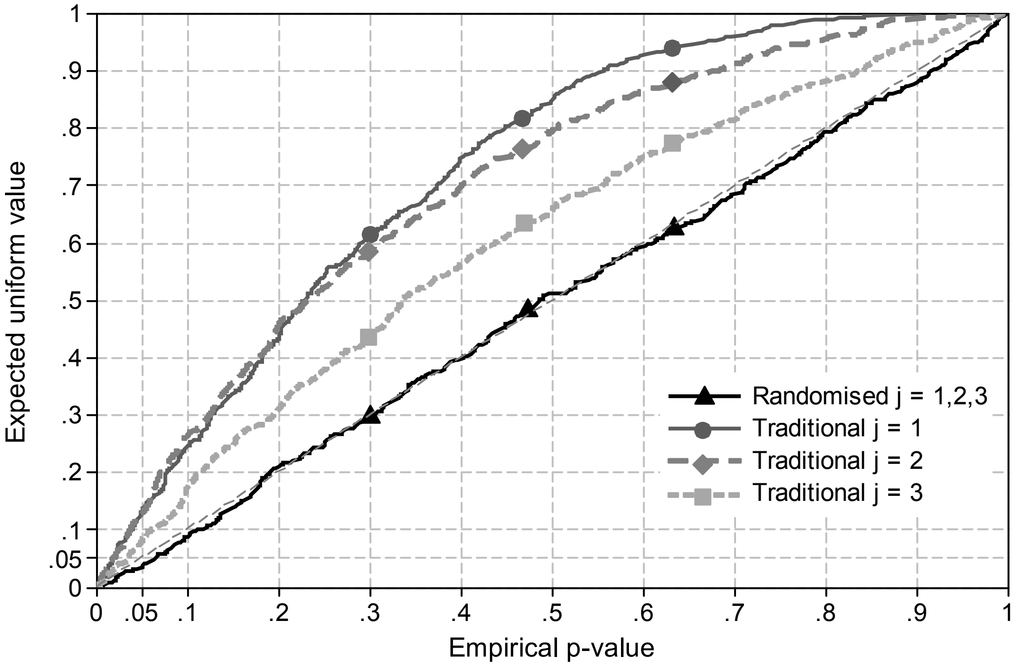


Figure 2: Percentile – Percentile plot for *p*-values for best subset of size *j* = 1, 2 from 4 independent predictors, Model B.

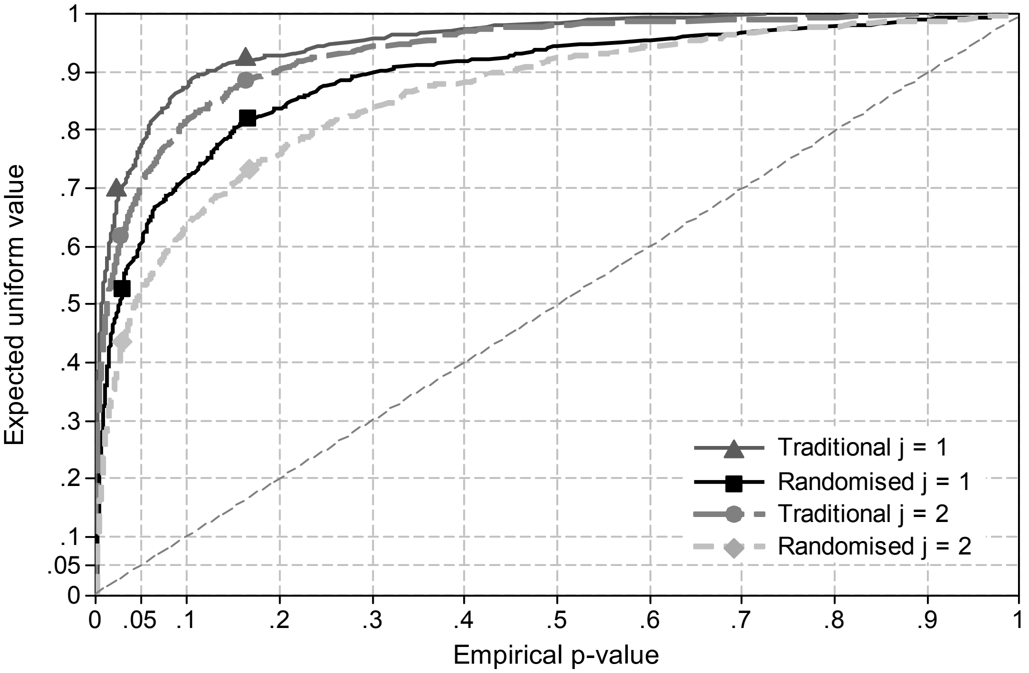
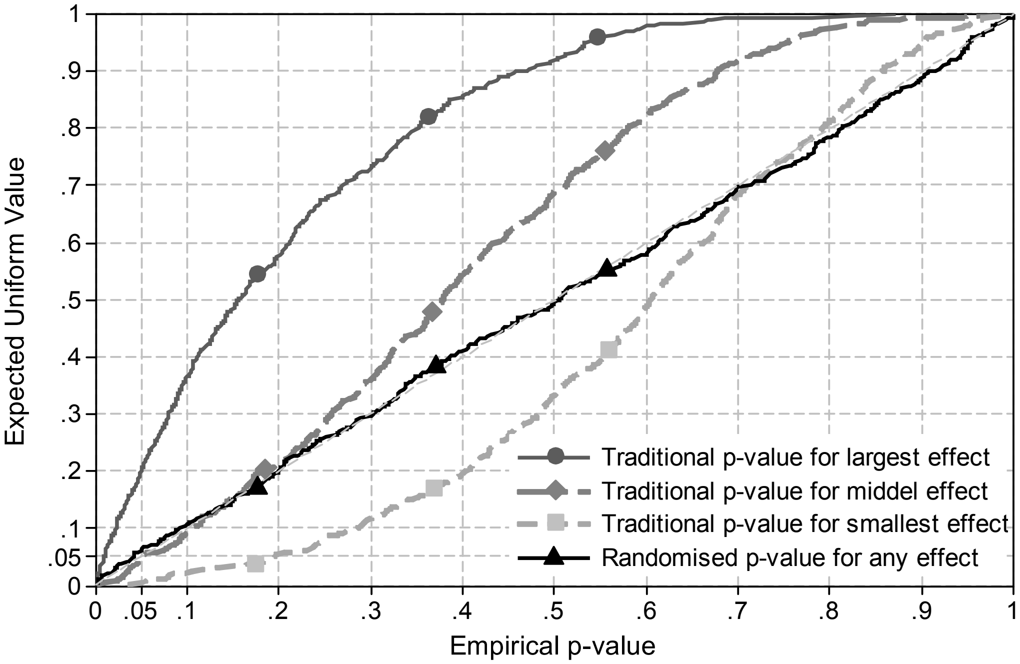


Figure 3: Percentile – Percentile plot for *p*-values for each variable in a best subset of size j = 3 from 4 independent predictors when the effect size is order by magnitude, Model A.



significance using the proposed randomization method is uniformly distributed $U(0, 1)$.

In every simulation instance the estimated *p*-value in the randomization method for overall model significance was not less than the *p*-value under the traditional method. The distribution of the differences for j = 1 and J = 4, 8, 16, 32, 64 is summarized in Figure 4. Note that the discrepancy tends to increase with increasing values of J and that this discrepancy is a substantive non-trivial difference.

Simulations under a true null model (i.e., with all potential predictors being noise variables), for J = 4, 8, 16, 32, 64, but with different sample sizes, I = 30, 60, 90, 120 have been performed. In all of these cases the simulations show that the distribution of *p*-value for overall subset significance using the proposed randomization method is uniform $U(0, 1)$. In every simulation instance the estimated *p*-value using the randomization method is not less than the *p*-value under the traditional method. Figure 5 summarizes the extent of the differences.

Figure 4: Discrepancy between randomized and traditional *p*-values for best subset of size *j* = 1 with *I* = 30 and different number of predictors.

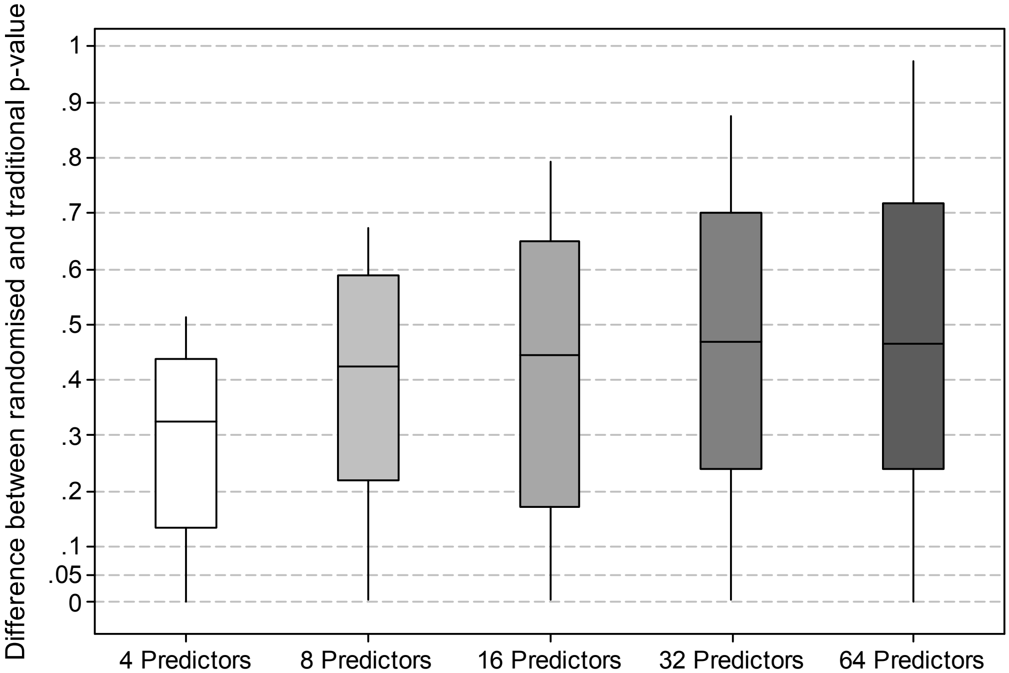
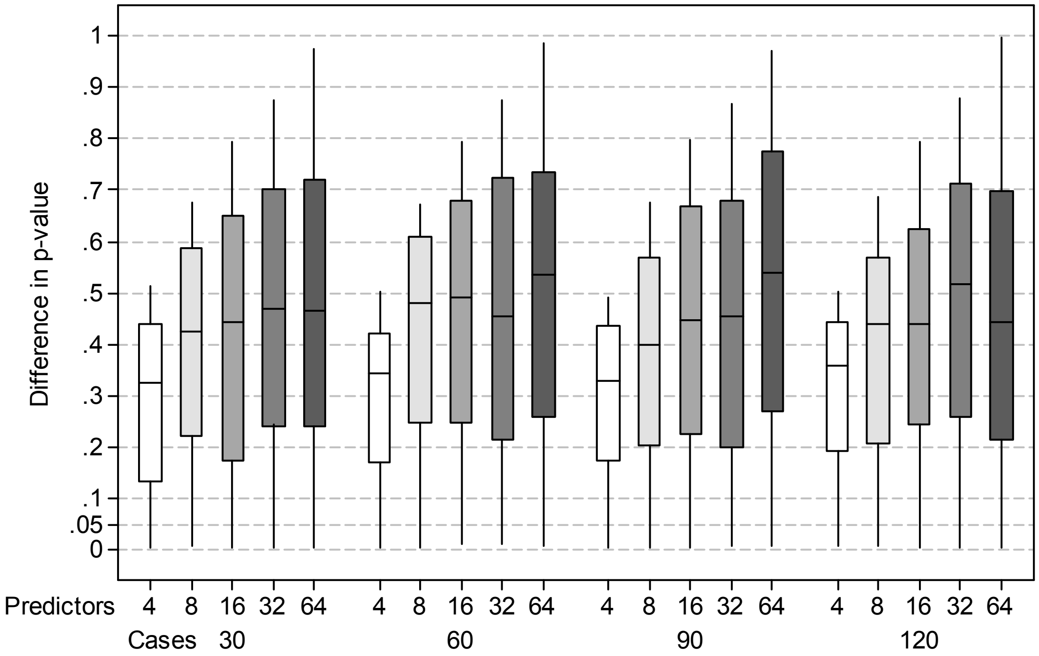


Figure 5: Distribution of the difference in *p*-values for overall model significance under both the randomization and the traditional methods for Model A for subset of size *j* =1.



Conclusion

A computer based heuristic that uses randomization has been described. The algorithm allows control of Type I error rate for the overall statistical significance of a best subsets regression model and control for the variables that comprise the model based on their relative order. This randomization algorithm permits the Type I error rate to be controlled at any pre-determined nominal significance level, α. The data sets created under the randomization procedure, each precisely retained the correlation structure observed in the original data and, as such, the approach takes into account the data set dependent problems that arise due to the correlation structure between potential predictor variables (see Derksen & Keselman, 1992). For the *j*-th best subset the procedure produces *p*-values indirectly based on the number of potential predictor variables (*J*) rather than the number of predictor variables in a given subset (*j*) and, as such, retains some similarity with the stance of Larzelere & Mulaik (1977). Their approach, however, does not take into account the correlation structure between potential predictor variables. By contrast, the algorithm outlined in this article establishes the *p*-value for overall model significance based on the effective number of predictors. For example consider J potential predictors, and consider an extreme case whereby $J-1$ of the predictors are mutually orthogonal but the other predictor is perfectly correlated with one of the other predictors in the orthogonal set. In this extreme case the number of predictors is *J* but the number of effective predictors is $J-1$.

The simulation work demonstrates that the randomization algorithm corrects a non-trivial problem. This correction also applies in those particularly problematic cases whereby the number of predictors exceeds the number of cases (subject to subset size *j* being less than sample size *I*).

Significance tests in classical least squares regression are based on the assumption that the underpinning error terms are independent, identically distributed normal random variables. When these assumptions are satisfied the *p*-value for overall model significance for a best subsets regression of size j still displays a bias. By contrast, the corresponding *p*-value estimation using the randomization algorithm does not suffer from this bias. In practice the underpinning normality assumptions are likely to be violated to some extent, and these violations may lead to additional biases in the estimated *p*-values for overall model significance in a best subsets regression using the standard approach. The randomization approach is based on the sample data and the estimation of the *p*-value does not explicitly rely upon distributional assumptions. Indeed, the algorithm is not peculiar to ordinary least squares regression and could be applied to other classes of model, including those models that rely upon inferential tests of significance based upon large sample asymptotic theory (e.g. binary logistic regression).

The approach for assessing individual significance of variables that comprise a final best subset is to consider a rank ordering on the variables in the model according to the value of their corresponding *F* statistic. This imposition of an ordering allows for a fair comparison with similarly ordered variables in the randomized solutions. It is recognized that this may produce a seemingly paradoxical outcome in some situations. For instance, and for simplicity of exposition, consider a two variable subset $j=2$ with a variable, X_1 with *F*-value $F_{(1)}$ and a variable, X_2, with *F*-value $F_{(2)}$. Without loss of generality assume $F_{(1)}>F_{(2)}$. In evaluating the statistical significance of X_1, the value $F_{(1)}$ will be compared against similarly ordered values $F^*_{(1)}$ and the value $F_{(2)}$ will be compared with similarly ordered values $F^*_{(2)}$. No condition is imposed to ensure that the proportion of times $F^*_{(1)} > F_{(1)}$ is less than the proportion of times $F^*_{(2)} > F_{(2)}$. However it should be borne in mind that X_1 and X_2 were not specified in advance; rather the significance tests alluded to are tests of significance for the variable with the largest value $F_{(1)}$ and for the variable with the second largest value $F_{(2)}$. In practice, interest would focus on those final solutions where all variables in the model met some pre-defined nominal level of significance (e.g. $\alpha = 0.05$).

A motivation behind this research was to help develop a sound methodological process to assist researchers in constructing valid and good initial models in exploratory research. However, the use of automated techniques is not in itself a substitute for quality of thought in determining what may be a good predictor of an outcome variable. An understanding of the procedural aspects involved in assessing statistical significance through the use of randomization may have an added benefit of focusing a researcher to determine seemingly good predictors at the outset rather than a researcher collecting data on all conceivable predictors and using these without penalty as per procedures currently offered by standard statistical software.

References

Bendel, R. B., & Afifi, A. A. (1977). Comparison of stopping rules in forward 'stepwise' regression. *Journal of the American Statistical Association*, *72*, 46-53.

Canning, F. L. (1959). Estimating load requirements in a job shop. *Journal of Industrial Engineering*, *10*, 447.

Derksen, S., & Keselman, H. J. (1992). Backward, forward and stepwise automated subset selection algorithms: frequency of obtaining authentic and noise variables. *British Journal of Mathematical and Statistical Psychology*, *45*, 265-282.

Draper, N. & Smith, H. (1981). *Applied regression analysis*. John Wiley & Sons: New York, NY.

Furnival, G. M., & Wilson, R. W. (1974). Regression by leaps and bounds. *Technometrics*, *16*, 499-511.

George, E. (2000). The variable selection problem. *Journal of the American Statistical Association*, *95*, 1304-1307.

Hurvich, C. M., & Tsai, C. L. (1990). The impact of model selection on inference in linear regression. *The American Statistician*, *44, 3*, 214-217.

Kramer, C. Y. (1957). Simplified computations for multiple regression. *Industrial Quality Control*, *13, 8*, p8.

Kuk, A. (1984). All subset regression in a proportional hazards model. *Biometrika*, *71, 3,* 587-592.

Larzelere, R. E., & Mulaik, S. A. (1977). Single-sample tests for many correlations. *Psychological Bulletin*, *84*, 557 – 569.

Lovell, M. C. (1983). Data mining. *The Review of Economics and Statistics*, *65*, 1-12.

Mantel, N. (1970). Why stepdown procedures in variable selection. *Technometrics*, *12, 3*, 621-625.

Miller, A. J. (1984). Selection of subsets of regression variables (with discussion). *Journal of the Royal Statistical Society*, *A*, *147*, 389-425.

Correlation between the Sample Mean and Sample Variance

Ramalingam Shanmugam
Texas State University-San Marcos

This article obtains a general formula to find the correlation coefficient between the sample mean and variance. Several particular results for major non-normal distributions are extracted to help students in classroom, clients during statistical consulting service.

Key words: Skewness, kurtosis, non-normal data, count and continuous distributions.

Introduction

Interest about the relationship between the sample descriptive measures is growing among the statisticians. For example, Zhang (2007) using a lengthy combinatorial argument obtained an expression for computing the covariance of sample mean and sample variance without the assumption of normality to help teachers explain to students. Such a tedious combinatorial new derivation is obsolete as it is a direct consequence of the results in Stuart and Ord (1994). Their result is helpful to find additionally the covariance between the sample mean and any even moment about the mean. However, no formula appears for computing the correlation between the sample mean and variance of a non-normal sample. Yet, almost all students in statistics courses and the clients during statistical consulting service curiously seek to know an estimate of such correlation in their data. So, there is a need to have a list of expressions for non- normal data which the statisticians can readily use to answer the clients' query.

Ramalingam Shanmugam is a Professor in the School of Health Administration. His research interests include multivariate data modeling, informatics, optimal decision support systems, and health/medical application of statistics. Email: rs25@txstate.edu.

Let $\bar{X} = n^{-1}\sum_{i=1}^{n} X_i = m'_1$ and $(\frac{n}{n-1})m_2 =$

$S^2 = (n-1)^{-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ be the sample mean and variance of a random sample drawn from a given population, where m_1 and m_2 are the notations in Stuart and Ord (1994, volume 1, page 350). A consequence of their results is that

$$Cov[\bar{X}, S^2] = Cov[m'_1, m_2) = \frac{\mu_3}{n}. \qquad (1)$$

Methodology

In the next two sections, the general formula for finding the correlation and results for particular specified non-normal samples are obtained. The results for Poisson, geometric, and Bernoulli samples are illustrated with data from the literature for better understanding.

Derivation of formula for $Corr(\bar{X}, S^2)$

Because $Var[\bar{X}] = Var[m'_1] = \frac{\sigma^2}{n}$ (2)

and

$$Var[S^2] = Var[(\frac{n}{n-1})m_2]$$
$$= \frac{\mu_4 - (\sigma^2)^2}{n} + \frac{2(\sigma^2)^2}{n(n-1)}$$

due to (10.9) in Stuart and Ord (1994) where σ^2 is the population variance. The kurtosis is measure of tail flatness of the frequency trend of the data. Because the kurtosis is defined to be $K_u = \frac{\mu_4}{(\sigma^2)^2}$, rewrite

$$Var[S^2] = \frac{(\sigma^2)^2}{n}[K_u - 1 + \frac{2}{(n-1)}] \quad (3).$$

Similarly, the skewness is a measure of the lack of symmetry in the frequency trend of the data. The skewness is defined as $S_k = \frac{\mu_3^2}{(\sigma^2)^3}$. Using (1) through (3), the correlation coefficient between the sample mean and sample variance is obtained and after algebraic simplifications, it is

$$Corr[\bar{X}, S^2] = \sqrt{\left|\frac{S_k}{K_u - 1 + \frac{2}{n-1}}\right|}, n \geq 2 \quad (4).$$

With only one observation (that is, n = 1), the correlation between the sample mean and variance cannot be determined if the skewness is non zero, according to (4) as it requires multiple observations. Also, from expression (4), notice that the correlation is zero when the skewness is zero and it occurs in a random sample from a symmetric population. The t, Laplace, error distribution, the discrete and continuous uniform probability distributions in addition to normal distribution are symmetric population with zero skewness. Hence, the zero correlation between the sample mean and sample variance does not necessarily mean only the normal population due to (4). Furthermore, notice in (4) that the correlation weakens as the sample size increases. The skewness and kurtosis moderate the correlation coefficient in a way. The details are discussed, listed and illustrated below in several cases. In the next section, the results for particular non-normal cases which are come across in graduate courses and statistical consulting service.

Special non-normal cases

The power of mathematical statistics enables to group several probability distributions under one “umbrella” as they possess a common property.

(Modified) power series family sample

One such property is power series nature of the probability mass function (pmf). The pmf of power series distribution is defined (it seems earliest by Kosambi, 1949) to be

$$\Pr[x] = \frac{a_x \theta^x}{\eta(\theta)} \quad (5)$$

with a non-negative and differentiable function $\eta(\theta)$ of a natural parameter θ. The variance in this family is

$$\sigma^2 = \theta^2 \partial_{\theta\theta}^2 \ln \eta(\theta) + \theta\partial_\theta \ln \eta(\theta) \quad (6)$$

where $\partial_{\theta\theta}^k$ means the k-th derivative with respect to the natural parameter. The skewness is

$$S_k = (\theta\partial_\theta \sigma^2)^2 / \sigma^6 \quad . \quad (7)$$

The kurtosis is

$$K_u = (\theta\partial_\theta \mu_3 + 3\sigma^4) / \sigma^4 \quad . \quad (8)$$

Substituting (6), (7), and (8) in (4), the correlation for the power series family could be readily computed.

This family is modified in several ways. One modification is by Gupta (1974), who introduced a modified power series distribution (MPSD) with pmf

$$\Pr[x] = \frac{a_x [u(\theta)]^x}{\eta(\theta)}$$

The variance, skewness, and kurtosis in (6), (7), and (8) change to

$$\sigma^2 = \frac{\partial_\theta [\frac{\partial_\theta \ln \eta(\theta)}{\partial_\theta \ln u(\theta)}]}{\partial_\theta \ln u(\theta)},$$

$$S_k = (\frac{\partial_\theta \sigma^2}{\partial_\theta \ln u(\theta)})^2 / \sigma^6,$$

and

$$K_u = (\partial_\theta [\frac{\partial_\theta \ln \eta(\theta)}{\partial_\theta \ln u(\theta)}] + 3\sigma^4) / \sigma^4 .$$

for MPSD. By substituting in (4), the correlation for the modified power series family can be computed.

Binomial sample (with replacement)

For binomial sample, one need to consider $\eta(\theta) = (1+\theta)^r$ with $r \geq 1$ denotes the number of trials and the natural parameter $\theta = p/(1-p)$. By substituting the skewness

$$\hat{S}_k = (1 - \frac{2\overline{x}}{r})^2 (\overline{x}[1 - \frac{\overline{x}}{r}])^{-1}$$

and the kurtosis

$$\hat{K}_u = 3 + (1 - \frac{6\overline{x}}{r}[1 - \frac{\overline{x}}{r}])(\overline{x}[1 - \frac{\overline{x}}{r}])^{-1},$$

in (4), the correlation of the binomial sample mean and variance is noticed. When the number of trials is large (that is, $r \longrightarrow \infty$), the correlation diminishes but not to zero.

Bernoulli trials

With r=1 in the above binomial results, note that the correlation for Bernoulli sample mean and variance is

$$\begin{aligned} &C\hat{o}rr[\overline{X}, S^2] \\ &= (1 - 2\overline{x})[1 + (\frac{2n}{n-1} - 6)\overline{x}(1 - \overline{x})]^{-1} \end{aligned} \quad . \qquad (9)$$

This is useful in discussions of the logistic regression data. Consider the following partial data (Dalal, et al., 1989) of n = 5 observations with respect to failure (X = 1) and non failure (X = 0) of O-rings in space rockets. The shuttle challenger exploded after its launch on 28 January 1986 with a loss of seven lives. A commission was charged with determining the causes of that tragedy. Their report concluded that the failure of O-rings in nozzle joints due to thermal stress was the reason. The gas went through the cracks in the stressed O-rings caused the explosion.

Table 1. Date and O-ring failure (X = 1) or non-failure (X =0) of n = 5 cases

Date	21 April 81	12 Nov 1981	8 Nov 1984	30 Aug 1984	21 Jan 1986
X =	0	1	0	1	1

Note: the sample mean $\overline{x} = 0.6$ and sample variance $s^2 = 0.24$ with n = 5. Substituting in (9), the correlation coefficient between the Bernoulli sample mean and sample variance is computed and it is $C\hat{o}rr[\overline{X}, S^2] = 0.489$.

Poisson sample

The Poisson distribution is a limiting case of binomial distribution when the Bernoulli chance p is small but the number of trials is large enough to make a finite mean θ. If the number of O-rings to be investigated is large and the chance of any failure is very slim, then the expected number of O-ring failures is $\theta > 0$ and it is the mean of Poisson frequency trend. For such a Poisson sample, note that $\eta(\theta) = e^{\theta}$ with the natural parameter θ denoting the incidence rate in the power series family. Substituting the Poisson skewness and kurtosis

$$S_k = \theta^{-1} = K_u - 3$$

in (4), the correlation of the Poisson sample mean and variance could be obtained. It is

$$\begin{aligned} &C\hat{o}rr[\overline{X}, S^2] \\ &= [2(\hat{\theta} + \frac{1}{n-1}) - 1]^{-1/2} \, . \\ &= [2(\overline{x} + \frac{1}{n-1}) - 1]^{-1/2} \end{aligned} \qquad (10)$$

With the larger incidence rate, the Poisson correlation diminishes.

Incidence rate restricted Poisson sample

In spite of rarity in the Poisson data, sometimes the data might not be well governed by the above described Poisson distribution. A modification in the Poisson probability distribution is necessary. One such modification is due to Shanmugam's (1991). When the regular Poisson distribution does not fit a given data, one could consider the incidence rate restricted Poisson distribution (IRRPD) because it is versatile enough to fit the data. The pmf of IRRPD is

$$\Pr[x] = \frac{(1+\gamma x)^{x-1}(\theta e^{-\gamma\theta})^x}{x! e^{\theta}}$$

where the incidence rate $\theta \leq \frac{1}{\gamma}$ and γ is the restriction parameter. The restriction is removed when γ approaches zero and in which case, it reduces to the Poisson distribution in section (3.4). The skewness and kurtosis are

$$S_k = (1+2\gamma\theta)^2[\theta(1-\gamma\theta)]^{-1}$$

and

$$K_u = 3 + (1+8\gamma\theta+6\gamma^2\theta^2)[\theta(1-\gamma\theta)]^{-1}$$

The estimates of the IRRPD parameters are $\hat{\gamma}\hat{\theta} = 1-\sqrt{\frac{\bar{x}}{s^2}}$ and $\hat{\theta} = \bar{x}\sqrt{\frac{\bar{x}}{s^2}}$. Substituting these estimates in (4), the correlation of the Poisson sample mean and variance can be obtained. When γ approaches zero, the above results reduce to those in Section 3.4 for the Poisson distribution.

To illustrate, consider the following data in the literature about the number of tram accidents, X in Belgrade during 1965 and 1970 from Shanmugam and Singh (2001) as re-displayed in the Table 2 below. The estimate of the IRRPD parameters with data on n =134 drivers are $\hat{\theta} = 3.724$ and $\hat{\gamma} = 0.101$.

Table 2. # Tram Accidents in Belgrade

X=	0	1	2	3	4	5	6	7	8	9	≥ 10
f=	1	8	14	17	16	19	16	9	6	6	21

With these estimates, the skewness, kurtosis, and hence the correlation between the sample mean and variance are obtained and they are $\hat{S}_k = 1.321, \hat{K}_u = 5.089$ and $C\hat{o}rr[\bar{X}, S^2] = 0.567$, respectively.

For another example, consider the number of injury accidents that occurred in the Interstate-95 during January 1, 1969 through October 31, 1970 as reported in Shanmugam and Singh (1981).

Table 3. # Injury Accidents in Virginia State during January 1, 1969 & October 31, 1970

X=	0	1	2	3	4	5 +
f=	286	216	92	30	14	1

With n = 639, the estimates of IRRPD are $\hat{\theta} = 0.06$ and $\hat{\gamma} = 13.5$. Hence, the skewness, kurtosis, and the correlation between the IRRPD sample mean and variance are

$$\hat{S}_k = 287.3, \hat{K}_u = 1003.403$$

and

$$C\hat{o}rr[\bar{X}, S^2] = 0.535,$$

respectively.

Inverse binomial sample

With $\eta(\theta) = (1-\theta)^{-r}$ in the power series family, $r \geq 1$ denoting the number of cases to be of a particular, and the natural parameter $\theta = p$ be the probability of outcome of a type, the number of cases to be investigated is an inverse binomial random variable. Substituting its skewness

$$\hat{S}_k = (2-\hat{p})^2(r[1-\hat{p}])^{-1}$$

and the kurtosis

$$\hat{K}_u = 3 + \frac{\hat{p}^2 + 6[1-\hat{p}]}{r[1-\hat{p}]}$$

with $\hat{p} = \frac{\bar{x}-1}{r+\bar{x}-1}$. Substituting in (4), the correlation of the inverse binomial sample mean and variance is obtainable.

Geometric sample

With $r = 1$ in the above results for the inverse binomial sample, the correlation between the geometric mean and sample variance of a geometric sample is obtained and it is

$$Corr[\bar{X}, S^2] = \frac{\bar{x}+1}{\sqrt{\bar{x}(7\bar{x}+\bar{x}^{-1}+2[n-1]^{-1})}}. \quad (11)$$

For illustration, consider the geometric sample data on the number of heart failures experienced by a random sample of n = 15 cardiology patients in Alabama state as used in Bartolucci et al (1999). The sample mean is 1.2 in that geometric data. Substituting in (11) the sample mean, the correlation between the sample mean and variance of the geometric data is obtained and it is $Corr[\bar{X}, S^2] = 0.657$.

Log series sample

For sample from a logarithmic series distribution, note that $\eta(\theta) = -\ln(1-\theta)$ with the natural parameter θ in power series family. Substituting its skewness

$$S_k = \frac{[(1+\theta) + \frac{3\theta}{\ln(1-\theta)} + \frac{2\theta^2}{(\ln(1-\theta))^2}]^2}{[\frac{\theta}{\ln(1-\theta)}]^2[1+\frac{\theta}{\ln(1-\theta)}]^3}$$

and the kurtosis

$$K_u = \frac{[1 + 4\theta + \theta^2 + \frac{4\theta(1+\theta)}{\ln(1-\theta)} + \frac{6\theta^2}{(\ln(1-\theta))^2} + 3\frac{\theta^3}{(\ln(1-\theta))^3}]}{[\frac{-\theta}{\ln(1-\theta)}][1+\frac{\theta}{\ln(1-\theta)}]^2}$$

in (4), the correlation between the binomial sample mean and variance could be obtained.

Hypergeometric sample (without replacement)

In many health applications, random sampling is done without replacement. For an example, once the virus infected individuals are identified in the population and are kept removed from the community. Suppose that Np individuals are suspected to be infected where $0 < p < 1$ indicates the prevalence level of the virus in the community. The number of infected individuals in a random sample of n persons without replacement is a hypergeometric outcome. Its skewness and kurtosis in data of this type are respectively

$$S_k = \frac{(1-\frac{1}{N})}{np(1-p)(1-\frac{n}{N})}[\frac{(1-2p)(1-\frac{2n}{N})}{(1-\frac{2}{N})}]^2$$

and

$$K_u = 3 + \frac{\begin{array}{l}[(N-1)[N(N+1-6n)\\ +3p(1-p)(n-2)+6(n/N)^2\\ +3p(1-p)(n/N)(6-n)\\ -18p(1-p)(n/N)^2]\end{array}}{np(N-n)(N-2)(N-3)}$$

Substituting in (4), the correlation between the hypergeometric sample mean and variance could be obtained.

Katz's family sample

The pmf of the Katz's family is denoted by $\Pr[x+1] = (\frac{\alpha+\beta x}{1+x})\Pr[x]$. Its skewness and kurtosis are

$$S_k = (\frac{2}{\beta} - 1)/\sigma^2$$

and

$$K_u = 3 + (\frac{6}{(1-\beta)^2} - \frac{6}{(1-\beta)} + 1)/\sigma^2$$

respectively. Substituting in (4), the correlation between the sample mean and variance of Katz's family can be obtained.

Log-normal sample

Consider a random sample is drawn from a log-normal population with the threshold, location, and scale parameters θ, ξ, σ^2 respectively. The skewness and kurtosis are

$$S_k = (\varpi + 2)^2(\varpi - 1)$$

and

$$K_u = \varpi^4 + 2\varpi^3 + 3\varpi^2 - 3$$

with $\varpi = e^{\sigma^2}$. Substituting in (4), the correlation between the log-normal sample mean and variance could be obtained.

Gamma sample

The skewness and kurtosis of the gamma population with pdf

$$f(x) = (\frac{\mu}{\sigma^2})(\frac{x\mu}{\sigma^2})^{(\frac{\mu}{\sigma})^2-1} e^{-(\frac{x\mu}{\sigma^2})} / \Gamma((\frac{\mu}{\sigma})^2)$$

are $S_k = 4(\frac{\mu}{\sigma})^{-2}$ and $K_u = 3 + 6(\frac{\mu}{\sigma})^{-2}$ where μ and σ^2 denote the mean and variance. Substituting in (4), the correlation between the gamma sample mean and variance is obtained. In the gamma case, it is

$$Corr[\bar{X}, S^2] = \sqrt{\frac{2\sigma^2}{\mu^2 + 3\sigma^2 + [n-1]^{-1}}}\ .$$

Exponential sample

When $\mu = \sigma$ in the above results, they reduce to those for exponential population. The exponential population is an interesting special case. Then, the correlation between the exponential sample mean and variance is

$$C\hat{o}rr[\overline{X}, S^2] = \sqrt{\frac{2\overline{x}^2}{4\overline{x}^2 + [n-1]^{-1}}} \qquad (12)$$

For an illustration, consider Zelen's exponential data below about the number of weeks a random sample of n = 11 tumor patients survived in a health clinic. The data are well fit by an exponential distribution as it is verified in Shanmugam (1991).

Table 4. Zelen's data of survival weeks of n = 11 tumor patients

3	4	5	8	8	10	12	16	17	30	33

The sample mean is equal to 13.6 weeks. According to (12), the correlation between the exponential sample mean and variance of this exponential data is $Corr[\overline{X}, S^2] = 0.71$.

Inverse gaussian sample

The inverse Gaussian distribution is considered as an alternate model for positive but skewed data. Its skewness and kurtosis are

$$\hat{S}_k = \frac{9\overline{x}[\sum_{i=1}^{n} x_i^{-1} - (\overline{x})^{-1}]}{n-1}$$

and

$$\hat{K}_u = 3 + \frac{15\overline{x}[\sum_{i=1}^{n} x_i^{-1} - (\overline{x})^{-1}]}{n-1}$$

respectively. Substituting in (4), the correlation between the sample mean and variance of inverse Gaussian data can be obtained.

Pareto sample

The Pareto distribution is considered another alternate model for positive but skewed data. Its skewness and kurtosis are

$$\hat{S}_k = 4[\frac{(\sqrt{1+(\frac{\overline{x}}{s})^2} + 2)^2(\sqrt{1+(\frac{\overline{x}}{s})^2} - 1)}{(\sqrt{1+(\frac{\overline{x}}{s})^2} - 2)^2(\sqrt{1+(\frac{\overline{x}}{s})^2} + 1)}]$$

and

$$\hat{K}_u = \frac{3(\sqrt{1+(\frac{\overline{x}}{s})^2} - 1)[3(\sqrt{1+(\frac{\overline{x}}{s})^2} + 1)^2 + \sqrt{1+(\frac{\overline{x}}{s})^2} - 2]}{(\sqrt{1+(\frac{\overline{x}}{s})^2} + 1)(\sqrt{1+(\frac{\overline{x}}{s})^2} - 2)(\sqrt{1+(\frac{\overline{x}}{s})^2} - 3)}$$

respectively. Substituting in (4), the correlation between the sample mean and variance of Pareto data can be obtained.

Beta sample

The beta distribution is considered suitable for percentage data. Its skewness and kurtosis are

$$\hat{S}_k = \frac{4(\hat{\varpi} - \hat{\upsilon})^2(\hat{\varpi} + \hat{\upsilon} + 1)}{(\hat{\varpi} + \hat{\upsilon} + 1)^2 \hat{\varpi}\hat{\upsilon}}]$$

and

$$\hat{K}_u = \frac{3(\hat{\varpi} + \hat{\upsilon})(\hat{\varpi} + \hat{\upsilon} + 1)(\hat{\upsilon} + 1)(2\hat{\varpi} - \hat{\upsilon})}{\hat{\varpi}\hat{\upsilon}(\hat{\varpi} + \hat{\upsilon} + 2)(\hat{\varpi} + \hat{\upsilon} + 3)} + \frac{\hat{\upsilon}(\hat{\upsilon} - \hat{\varpi})}{(\hat{\varpi} + \hat{\upsilon})}$$

respectively. Substituting in (4), the correlation between the sample mean and variance of beta data can be obtained where $\hat{\upsilon} = \overline{x}\{\frac{\overline{x}(1-\overline{x})}{s^2} - 1\}$

and

$$\hat{\upsilon} = (1-\overline{x})\{\frac{\overline{x}(1-\overline{x})}{s^2} - 1\}.$$

(Non) central chi-squared sample

The non-central chi-squared sample is considered and analyzed in the discussion of statistical power calculation of hypothesis testing or analysis of variance. Its skewness and kurtosis are

$$\hat{S}_k = \frac{64(s^2 - \overline{x})^2}{s^6}$$

and

$$\hat{K}_u = 3 + \frac{24(3s^2 - 4\bar{x})}{s^4}$$

respectively. Substituting these in (4), the correlation between the sample mean and variance of non-central chi-squared observations can be obtained. When $s^2 = 2\bar{x}$, all the above results reduce to those of central chi-squared sample.

(Non) central F sample

The non-central F sample is considered in the discussion of statistical power calculation of hypothesis testing or analysis of variance. The results are too messy to mention. However, the skewness and kurtosis of the central F sample are

$$\hat{S}_k = \frac{8(2\hat{\upsilon} + \hat{\varpi} - 2)(\hat{\varpi} - 4)}{\hat{\upsilon}(\hat{\varpi} - 6)(\hat{\upsilon} + \hat{\varpi} - 2)}$$

and

$$\hat{K}_u = 3 + \frac{12[(\hat{\varpi} - 2)^2(\hat{\varpi} - 4) + \hat{\upsilon}(\hat{\upsilon} + \hat{\varpi} - 2)(5\hat{\varpi} - 22)]}{\hat{\upsilon}(\hat{\varpi} - 6)(\hat{\varpi} - 8)(\hat{\upsilon} + \hat{\varpi} - 2)}$$

respectively, where

$$\hat{\varpi} = \frac{2\bar{x}}{\bar{x} - 1}$$

and

$$\hat{\upsilon} = \frac{(\bar{x} - 1)^2 s^2 (2 - \bar{x})}{\bar{x}^2} - 2.$$

Substituting these in (4), the correlation between the sample mean and variance of central F observations can be obtained.

(Non) central t sample

The non-central t sample is considered in the discussion of testing one sided hypothesis. Its skewness and kurtosis are

$$\hat{S}_k = \hat{\delta}(3 + 1.25\frac{\hat{\delta}^2}{\hat{\upsilon}})^2 / \hat{\upsilon}$$

and

$$\hat{K}_u = \frac{1.406(\hat{\upsilon} - 3.2)}{(\hat{\upsilon} - 4)}\beta_1 + \frac{3(\hat{\upsilon} - 2)}{(\hat{\upsilon} - 4)}$$

Where

$$\hat{\delta} = (s^2 + \bar{x}^2)(\frac{\hat{\upsilon}^2 - 2}{\hat{\upsilon}} - 1)^{1/2}$$

and

$$\frac{\bar{x}}{\hat{\delta}} = \frac{\sqrt{\frac{\hat{\upsilon}^2}{2}}\Gamma(\frac{\hat{\upsilon}^2 - 1}{2})}{\Gamma(\frac{\hat{\upsilon}^2}{2})}$$

Substituting these in (4), the correlation between the sample mean and variance of noncentral t observations can be obtained.

Power function sample

In financial studies, random sample of observations is well fit by a power function distribution. The skewness and kurtosis in power function distribution are estimated using

$$\hat{S}_k = \frac{4(1 - \hat{c})^2(2 + \hat{c})}{(3 + \hat{c})^2 \hat{c}}$$

and

$$\hat{K}_u = \frac{3(\hat{c} + 2)[2(\hat{c} + 1)^2 + \hat{c}(\hat{c} + 5)]}{\hat{c}(\hat{c} + 3)(\hat{c} + 4)}$$

where $\hat{c} = \sqrt{1 + (\frac{\bar{x}}{s})^2} - 1$.

Substituting the skewness and kurtosis in formula (4), the correlation between the sample mean and variance of the power function sample can be obtained.

References

Bartolucci, A., Shanmugam, R., & Singh, K. (1999). Geometric Model Generalized for Cardiovascular Studies. *Epidemiology, Health, and Medical Research*, *2*. L. Oxley, F. Scrimgeour, & M. McAleer, (Eds.). Modeling and Simulation Society of Australia Inc., p. 501-506.

Dalal, S. R., Fowkles, E. B., & Hoadley, B. (1989). Risk analysis of the space shuttle: Pre-Challenger prediction of failures. *Journal of the American Statistical Association*, *84*, 945-957.

Gupta, R. C. Modified power series distributions and some of its applications, *Sankhya*, *35*, 288-298.

Kosambi, D. D. (1949). Characteristic properties of series distributions. *Proceedings of the National Institute of Science, India*, *15*, 109-113.

Shanmugam, R., & Singh, K. (2001). Testing of Poisson incidence rate restriction. *International Journal of Reliability and Applications*, *2*, 263-268.

Shanmugam, R. (1991). Testing guaranteed exponentiality. *Naval Research Logistics*, *38,* 877-892.

Shanmugam, R., & Singh, J. (1981) Some bivariate probability models applicable to traffic accidents and fatalities data, *Statistical Distributions in Scientific Work. C. Tallie, et al.* (Eds.), *4*, 95-103.

Stuart, A., & Ord, J. K. (1994) *Kendall's advanced theory of statistics, Volume 1.* London: Arnold Publisher.

Zhang, L. (2007) Sample mean and sample variance: Their covariance and their (in)dependence. *American Statistician*, *61*, 159-160.

Journal of Modern Applied Statistical Methods
November 2008, Vol. 7, No. 2, 416-434

1538 – 9472/08/$95.00

Constructing Confidence Intervals for Spearman's Rank Correlation with Ordinal Data: A Simulation Study Comparing Analytic and Bootstrap Methods

John Ruscio
The College of New Jersey

Research shows good probability coverage using analytic confidence intervals (CIs) for Spearman's rho with continuous data, but poorer coverage with ordinal data. A simulation study examining the latter case replicated prior results and revealed that coverage of bootstrap CIs was usually as good or better than coverage of analytic CIs.

Key words: Spearman's rank correlation, confidence intervals, bootstrap.

Introduction

Spearman's (1904) rank correlation[1] (r_S) is a nonparametric statistic that allows an investigator to describe the strength of an association between two variables X and Y without making the more restrictive assumptions of the Pearson product-moment correlation (r). To calculate r_S, one converts each variable to ranks, assigning equal ranks to any tied scores (but see Gonzales & Nelson, 1996, for alternative approaches to handling ties), and then uses the usual formula for r or this computational shortcut

$$r_S = 1 - \frac{6\sum d_i^2}{N(N^2 - 1)}, \qquad (1)$$

where the d_i are the differences in the ranked scores on X and Y for each pair of cases and N is the sample size. Because this statistic is sensitive only to the order of differences between adjacent scores, and not their magnitudes, it belongs to the family of ordinal statistics (Cliff, 1996).

Cliff (1996) argues that ordinal statistics such as r_S are better able to answer ordinal research questions than more conventional parametric statistics. For example, asking whether higher self-esteem is associated with higher academic achievement poses an ordinal question. Using r to address it requires assumptions that may be unrelated to the research question and can be difficult to satisfy. Whereas r measures the strength of a linear relationship between X and Y, r_S assesses how well an arbitrary monotonic function describes the relationship. Testing for the strictly linear relationship between self-esteem and academic achievement will underestimate the strength of a relationship if it is nonlinear. Also, the insensitivity of r_S to monotonic transformations of the data can be a significant strength when it is safer to presume a monotonic relationship between one's measure of a variable and the underlying construct than to presume a linear relationship (Cliff, 1996). Whereas r assumes bivariate normality, r_S makes no assumptions about the distribution of either variable. Wilcox (2003) discusses the sensitivity of parametric statistics to extreme scores and, in many instances, even small departures from their assumptions. Caruso and Cliff (1997) suggest that r_S should be less sensitive to extreme scores and a more inferentially robust measure than r.

In addition to the fact that r_S does not require assumptions of linearity or bivariate normality, r_S can be used with ordinal data. According to Stevens (1946), a variable is classified as ordinal if scores can be scaled as rank-ordered categories but the absolute distances between them are unknown. Cliff (1996) observed that many variables of interest to psychologists are ordinal in nature. When one

John Ruscio is an Associate Professor in the Department of Psychology. Email him at ruscio @tcnj.edu.

or both of a pair of variables is ordinal, using r_S enables researchers to study relationships using variables that do not meet the interval scaling requirement of r.

Methods for evaluating the statistical significance of r_S are based on its sampling distribution under the null hypothesis (H_0) of $\rho_S = 0$. A randomization test (Edgington, 1987) may be the best way to test H_0, and many textbooks present tables of critical values for relatively small sample sizes (e.g., critical values in Zar, 1972, have been reprinted). With sufficiently large samples, one can use an approximation to the t distribution with $df = N–2$:

$$t = \frac{r_S}{\sqrt{(1-r_S^2)/(N-2)}}. \qquad (2)$$

This is the same approximation that is ordinarily used to test the statistical significance of r.

Although null hypothesis significance testing remains popular in the social and behavioral sciences, guidelines provided by the APA’s Task Force on Statistical Inference (Wilkinson et al., 1999) and its *Publication Manual* (American Psychological Association, 2009) recommended constructing a confidence interval (CI) instead. This is usually more informative because a CI allows an assessment of the null hypothesis (i.e., if the CI includes 0, one would retain H_0, otherwise one would reject H_0) and provides additional information, such as the precision with which a population parameter has been estimated. The more narrow the CI, the greater the precision of the estimate.

Testing the statistical significance of r_S is possible because the sampling distribution under H_0 is asymptotically normal and the variance of r_S can be estimated as $1/(N – 1)$ (Higgins, 2004). To construct a CI, however, one cannot assume that $\rho_S = 0$, and when $\rho_S \neq 0$ the variance of r_S is more complex. Techniques have been developed to estimate the variance of Fisher-transformed r_S such that, when transformed back into r_S units, the coverage of CIs constructed in this manner will approximate the nominal level. Several approaches have been developed and studied, and each is an adjustment to the technique used with r. After Fisher-transforming r, where $z_r = \tanh^{-1}(r)$, the usual estimate of the variance of z_r is $1/(N – 3)$. With this estimate of the sampling error of z_r and the assumption that these errors are normally distributed, one can construct a CI as follows:

$$CI(\rho) = \tanh\left[z_r \pm \sqrt{\frac{1}{N-3}}(z_{(1+CL)/2})\right], \qquad (3)$$

where CL is the desired confidence level (e.g., .95) and $z_{(1+CL)/2}$ is the percentile point of a standard normal distribution below which the subscripted proportion of scores lies. For example, constructing a 95% CI for $r = .50$ and $N = 50$ would proceed as follows: $z_r = \tanh^{-1}(.50) = .5493$, $z_{(1+CL)/2} = z_{.025} = 1.96$, and CI(ρ) = tanh(.5493 ± .1459 × 1.96) = .26 to .68. Note that for $r \neq 0$, this technique yields a CI asymmetric about r.

To construct a CI for ρ_S in a parallel fashion, one begins with the Fisher transformation $z_{r_S} = \tanh^{-1}(r_S)$ and then uses its estimated variance in much the same way shown in Eq. 3. Whereas the z distribution is used to form CIs for ρ, Woods (2007) recommended using the t distribution (with $df = N – 2$) to form CIs for ρ_S. Because Woods found that the observed coverage of CIs for ρ_S often was below the nominal level, and sometimes substantially so, the t distribution will be used in the present study. (Using the z distribution would produce narrower CIs than using the t distribution, hence coverage even further below the nominal level.) Thus, the CI for ρ_S is constructed as follows:

$$CI(\rho_S) = \tanh[z_{r_S} \pm \sigma(z_{r_S}) \times t_{(1+CL)/2}], \qquad (4)$$

with formulas to estimate $\sigma^2(z_{r_S})$, the variance of the Fisher-transformed r_S, developed by three sets of investigators: Fieller, Hartley, and Pearson (1957), Caruso and Cliff (1997), and Bonnett and Wright (2000). Each represents an ad hoc adjustment to the formula used to estimate the variance of z_r (recall that this is 1 / [N – 3]) that performed well under the conditions studied by its creators:

$$\sigma_F^2(z_{r_S}) = \frac{1.06}{N-3}, \quad (5)$$

$$\sigma_{CC}^2(z_{r_S}) = \frac{1}{N-2} + \frac{|z_{r_S}|}{6N+4\sqrt{N}}, \quad (6)$$

$$\sigma_{BW}^2(z_{r_S}) = \frac{1+r_S^2/2}{N-3}. \quad (7)$$

Caruso and Cliff (1997) studied CIs with ρ_S ranging from .00 to .89 using bivariate normal data with $N = 10$ to 200. Their technique (based on Eq. 6) achieved the nominal coverage levels. Bonnett and Wright (2000) studied CIs constructed using each of the three formulas shown above (Eqs. 5-7) with ρ_S ranging from .10 to .95 using bivariate normal data with $N = 25$ to 200. Their technique (Eq. 7) achieved good coverage even at large ρ_S (.80 to .95), where the other methods became liberal (i.e., coverage dropped below the nominal level). These results suggest that 95% CIs for ρ_S provide fairly accurate coverage for bivariate normal variables, with tendencies toward liberal coverage at large ρ_S and small N, and that the Bonnett and Wright formula for $\sigma^2(z_{r_S})$ may be the most useful of the three evaluated in these studies.

To date, only Woods (2007) investigated the coverage of CIs for ρ_S using ordinal data. Woods examined CIs constructed using each of the three formulas for $\sigma^2(z_{r_S})$ shown above using populations based on empirical data in which variables with either 4 or 5 categories correlated with one another from near-zero to large values of ρ_S; sample sizes in the simulation study ranged from $N = 25$ to 100. In the corrected results[2], Woods found that the Bonnett and Wright (2000) formula provided CIs with slightly better coverage than its rivals, but there remained room for improvement. For example, the coverage of nominally 95% CIs was below 90% for many conditions. Coverage dropped further below the nominal level for larger values of ρ_S, which is consistent with the findings of research using ratio scale data.

At least two factors that may constrain the performance of the analytic method of constructing a CI by using a formula for $\sigma^2(z_{r_S})$, at least under conditions that diverge from bivariate normality. First, each of the three formulas was developed as an ad hoc adjustment to the formula for estimating the variance of z_r. Because data may diverge substantially from bivariate normality (e.g., ordinal data will not be distributed in this way), it may not be possible to adjust the formula for the variance of z_r in a way that works well for a broad variety of data conditions. Second, constructing CIs for ρ_S using any of these formulas involves an assumption about the shape of the sampling distribution that may not be satisfied. Specifically, the *t* distribution is used to construct the CI. Whenever the sampling distribution does not follow the *t* distribution, the coverage of these CIs may deviate from the nominal level.

Bootstrap methods for constructing CIs avoid both of these potential problems (Efron & Tibshirani, 1993). Rather than using a formula to estimate the variance of a statistic and making an assumption about the shape of its sampling distribution, one treats the available data as the best estimate of the population, draws random samples from it a large number of times (this is known as resampling, which provides what are called bootstrap samples), and calculates the statistic in each of these bootstrap samples. The distribution of the statistic across the bootstrap samples constitutes an empirical sampling distribution.[3] The empirical sampling distribution is generated without recourse to assumptions such as bivariate normality, no formula is needed to estimate the variance of the statistic in the relevant population, and no assumptions are made about the shape of the sampling distribution. The strengths - and weaknesses - of bootstrap methods involve their heavy reliance on the empirical data rather than standard parametric assumptions (Kline, 2005).

Once one has generated an empirical sampling distribution, CIs can be obtained in several ways. The simplest, although not always the best, method for constructing a bootstrap CI is to record the values of the statistic in the sampling distribution that span the desired proportion of results, with the remainder lying beyond the CI in equal proportions in both tails. For example, suppose a sample of $N = 50$ cases of ordinal data yielded $r_S = .72$. Treating these data as the population of pairwise scores, one can draw cases at random (with replacement) to obtain a new sample of N = 50, calculate r_S in this bootstrap sample, and repeat this procedure B times, where B is the number of bootstrap samples. When this was done B = 2,000 times and the results were rank-ordered, values of r_S = .53 and .86 spanned the middle 95% of the empirical sampling distribution. These constitute the lower and upper limits of a 95% CI for ρ_S using what is called the percentile bootstrap method (Efron & Tibshirani, 1993).

The percentile bootstrap operates by sorting the B values in the empirical sampling distribution and identifying the CI limits as the values indexed at the positions $B \times \alpha_L$ (for the lower limit) and $B \times \alpha_U$ (for the upper limit), where α_L and α_U are calculated as follows:

$$\alpha_L = (1 - CL)/2, \tag{8}$$

$$\alpha_U = (1 + CL)/2. \tag{9}$$

If either position is not a whole number, the next whole number toward the end of the range is used (e.g., if $B \times \alpha_L$ = 47.6 and $B \times \alpha_U$ = 1943.1, the values at positions 47 and 1944 would be used). For many statistics, percentile bootstrap CIs provide good coverage. When empirical sampling distributions are asymmetric, however, the bias-corrected and accelerated (BCA) bootstrap method often provides better coverage (Chan & Chan, 2004; Efron & Tibshirani, 1993). The BCA bootstrap method calculates α_L and α_U as follows:

$$\alpha_L = \Phi\left(z_0 + \frac{z_0 + z_{(1-CL)/2}}{1 - a(z_0 + z_{(1-CL)/2})} \right) \tag{10}$$

$$\alpha_U = \Phi\left(z_0 + \frac{z_0 + z_{(1+CL)/2}}{1 - a(z_0 + z_{(1+CL)/2})} \right), \tag{11}$$

where Φ is the standard normal cumulative distribution function and z_0 and a index median bias and skewness, respectively. Formulas for the latter two values appear below.

$$z_0 = \Phi^{-1}\left(\frac{\#(r_S^* < r_S)}{B} \right), \tag{12}$$

where r_S is the correlation in the replication sample, r_S^* is a correlation in a bootstrap sample, # is the count function (applied across all bootstrap samples), and Φ^{-1} is the inverse standard normal cumulative distribution function. The closer r_S is to the median of the empirical sampling distribution, the closer the proportion in parentheses will be to .5 and the closer z_0 will be to 0.

$$a = \frac{\sum (r_{S(\cdot)} - r_{S(i)})^3}{6\left(\sum (r_{S(\cdot)} - r_{S(i)})^2\right)^{3/2}}, \tag{13}$$

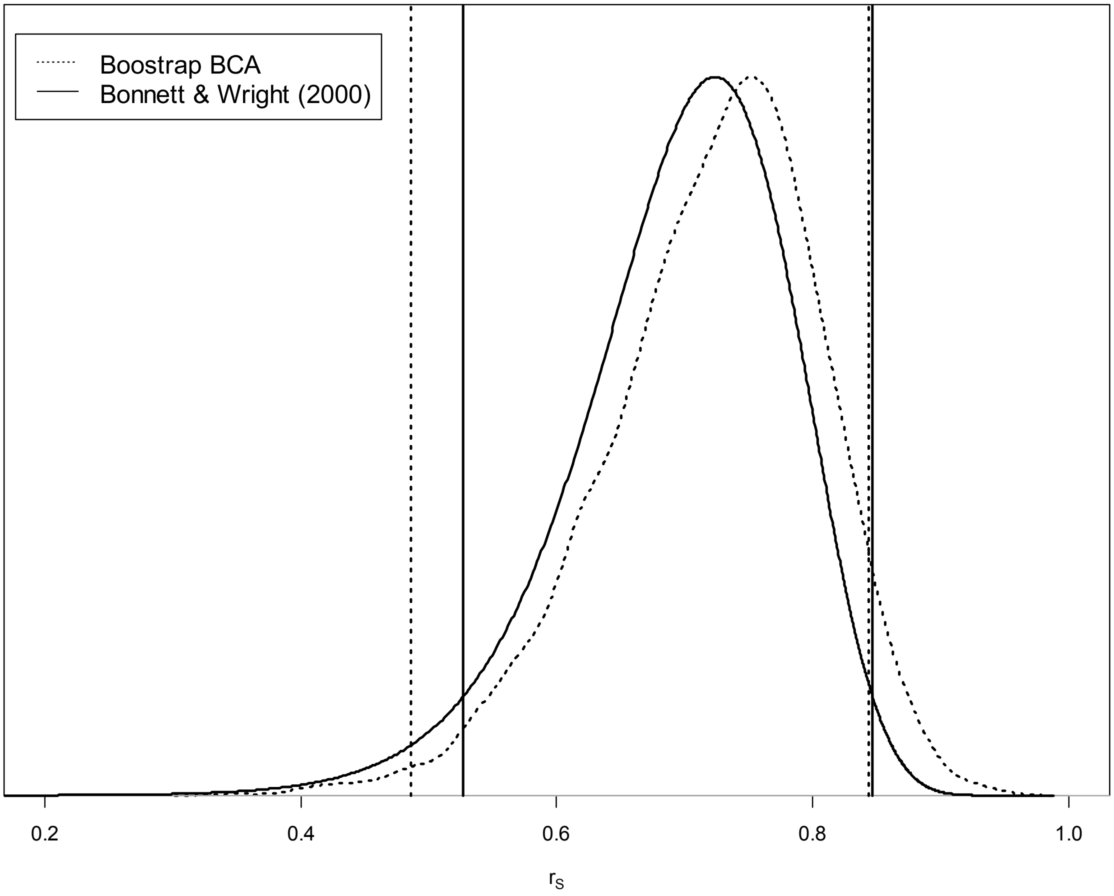
where $r_{S(i)}$ is a jackknife value of r_S calculated using all but the ith case and $r_{S(\cdot)}$ is the mean of all jackknife values. As is evident in the form of Eq. 13, a is related to skewness and indexes what is referred to in the bootstrap literature as acceleration, or the rate of change in the standard error of a statistic relative to its true parameter value. When $a = z_0 = 0$, Eqs. 10 and 11 simplify to Eqs. 8 and 9, in which case the BCA bootstrap method yields the same CI as the percentile bootstrap method. When $a \neq 0$ or $z_0 \neq 0$, Eqs. 10 and 11 involve adjustments to the values of α_L and α_U.

By indexing median bias and skewness to adjust α_L and α_U, BCA bootstrap CIs often provide better coverage than percentile bootstrap CIs. For example, in a study of CIs for ρ under conditions of range restriction, Chan and Chan (2004) found that the BCA bootstrap method yielded CIs with better coverage than did other bootstrap methods. Because the sampling distribution of Spearman's rank correlation is expected to be asymmetric when $\rho_S \neq 0$, the BCA bootstrap was included in the present study and the percentile bootstrap was not.

To illustrate the difference between conventional and bootstrap approaches, Figure 1 displays sampling distributions generated analytically, using the Bonnett and Wright (2000) estimate of $\sigma^2(z_{r_S})$, and empirically, using the BCA bootstrap method. Whereas the shape of the former is assumed (prior to transformation from Fisher-transformed r_S back to ordinary r_S units, it followed the *t* distribution with 48 *df*), the latter is based on the observed results for B = 2,000 bootstrap samples drawn from the original data. The Bonnett and Wright 95% CI ranged from .53 to .85, which is nearly the same as the percentile bootstrap CI of .53 to .86. The BCA bootstrap method adjusted these limits downward, and this CI ranged from .49 to .84. Only the BCA bootstrap CI included the correct value of ρ_S = .50, so it appears that the adjustments for median bias (z_0 = -.085) and skewness (a = -.038) were helpful in this instance.

Because the construction of bootstrap CIs does not require a formula to estimate the standard error of r_S (or Fisher-transformed r_S) and does not assume the shape of the sampling distribution, it may provide better coverage than the analytic method for constructing CIs. On the other hand, bootstrap methodology for constructing CIs treats the sample data as the best estimate of the population and resamples from this bivariate distribution. Any irregularities in the sample can be magnified in bootstrap applications, and this can be especially problematic with small samples (Kline, 2005). The present study was designed to compare the coverage of analytic and bootstrap CIs for ρ_S across a wide range of ordinal data conditions, including small sample sizes.

Figure 1: Sampling distributions for r_S in analyses of a sample of N = 50 cases drawn from a population in which both variables were distributed asymmetrically across 5 categories and ρ_S = .50. Vertical lines represent the limits of 95% confidence intervals constructed from each sampling distribution.



Methodology

Design

Four factors were studied. First, the marginal frequencies of variables in the populations were either derived from empirical data or simulated using asymmetric, symmetric, or uniform distributions. Second, the size of the contingency table for a bivariate relationship was either 5 × 5 or 4 × 5, which limited each variable to a small number of ordered categories and allowed for equal or unequal numbers of categories. Third, ρ_S varied from zero to a very large value (.90). Fourth, sample size varied from small (N = 25) to modestly large values (N = 200).

Population Data

Four types of bivariate population distributions were included in the study. First, the distributions in Woods (2007) were used so that results for BCA bootstrap CIs could be compared to those for the methods in prior research. Because Woods focused primarily on measures of ordinal association in the gamma family, populations were selected such that Γ ranged from near zero (-.01 to .01) through small (.35 to .39), medium (.55 to .59), and large (.85 to .89) levels. Populations were not selected for values of ρ_S, and consequently these do not vary as widely or discretely as the four levels of Γ. At each level of Γ, the number of categories was selected such that variables had equal or unequal numbers of categories.

Specifically, both 5 × 5 and 4 × 5 contingency tables were used. Woods studied four sample sizes (N = 25, 50, 75, and 100), and each sample size had a corresponding population distribution from which cases were sampled (with replacement). The variables' marginal distributions generally were asymmetric. Figures 2 and 3 show the population distributions for all 32 conditions (4 sample sizes × 4 levels of Γ × 2 table sizes) in Woods' study.[4] In addition to Γ for each condition, ρ_S is shown. All samples drawn from the Woods populations had the same sizes as in the original study (N = 25, 50, 75, and 100).

Because Woods (2007) selected populations for study from an empirical data set, there is a degree of realism to the data conditions. However, the finite number of variable pairs available in these data may have precluded an orthogonal manipulation of the design factors. For example, marginal distributions are not independent of sample size or ρ_S. To supplement the distributions analyzed by Woods, three additional types of population distributions were created in which design factors were manipulated orthogonally. First, marginal distributions were similar to those used by Woods in that they were asymmetric.

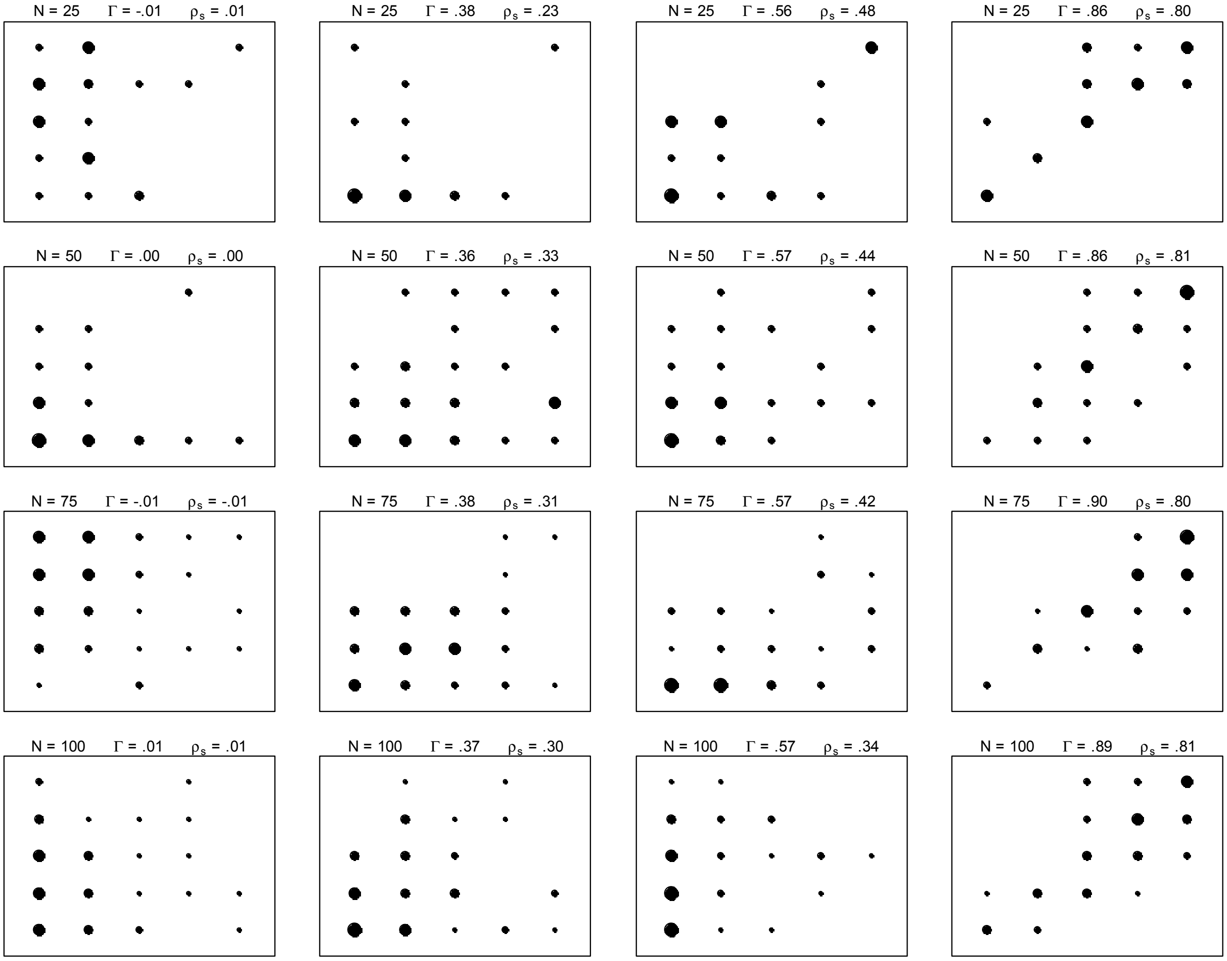
Values for variables with 5 categories were sampled with probabilities of .55, .20, .12, .08, and .05; values for variables with 4 categories were sampled with probabilities of .60, .22, .11, and .07. These distributions approximated the asymmetry observed in many of Woods' populations. Second, marginal distributions were symmetric (and unimodal), with probabilities calculated using thresholds of -1.5, -.5, .5, and 1.5 in a standard normal distribution to create 5 categories and thresholds of -1, 0, and 1 to create 4 categories; these correspond to probability distributions of .07, .24, .38, .24, and .07 for 5 categories and .16, .34, .34, and .16 for 4 categories. Third, marginal distributions were uniform.

For each type of distribution, both 5 × 5 and 4 × 5 tables were created at each of six levels of ρ_S (.00, .10, .30, .50, .70, and .90). To generate each of these 36 bivariate population distributions (3 types of marginal distribution × 2 table sizes × 6 levels of ρ_S), the iterative technique developed by Ruscio, Ruscio, and Meron (2007), and subsequently generalized with improved efficiency by Ruscio and Kaczetow (2008), was used. This technique generates multivariate data sets with user-specified marginal distributions and correlation matrix. Both of the papers cited above demonstrate that this technique reproduces the desired distributions and correlations with good precision, especially at large sample sizes. In the present study, data were generated such that each of the 36 populations possessed 100,000 cases, which enabled a very close match between ρ_S as specified in the study design and as calculated in the finite population from which replication samples were drawn: With one exception, these values were within .005 of each other.[5] From each population, samples were drawn with $N = 25$, 50, 100, and 200, yielding a total of 144 cells in this portion of the study design (36 populations × 4 sample sizes).

Replication Sample

Within each cell of the design, including the 32 conditions created by Woods (2007) and the 144 new conditions involving asymmetric, symmetric, and uniform populations, 1,000 replication samples were drawn for analysis. Whereas previous studies of CIs for ρ_S have used larger number of replication samples, this was not feasible in the present study due to the inclusion of a bootstrap method that required extensive resampling and analysis for each replication sample. For each replication sample, B = 2,000 bootstrap samples were drawn and analyzed. Using 1,000 replication samples per condition – the same number used in Chan and Chan's (2004) study of bootstrap CIs for ρ in situations of range restriction – was both feasible, given the inclusion of a computationally intensive bootstrap method, and adequate for informative comparisons among the four types of CI studied. Each replication sample was checked to ensure that the variance for each variable was greater than zero so that a correlation could be calculated. In a small number of instances, primarily when drawing small samples from asymmetric populations, all values for a variable were identical and that sample was not included in the study.

Figure 2: Population distributions for data conditions with 5 × 5 tables in Woods (2007). The area of each plotting symbol is proportional to the frequency in that cell of the contingency table.



Data Analysis

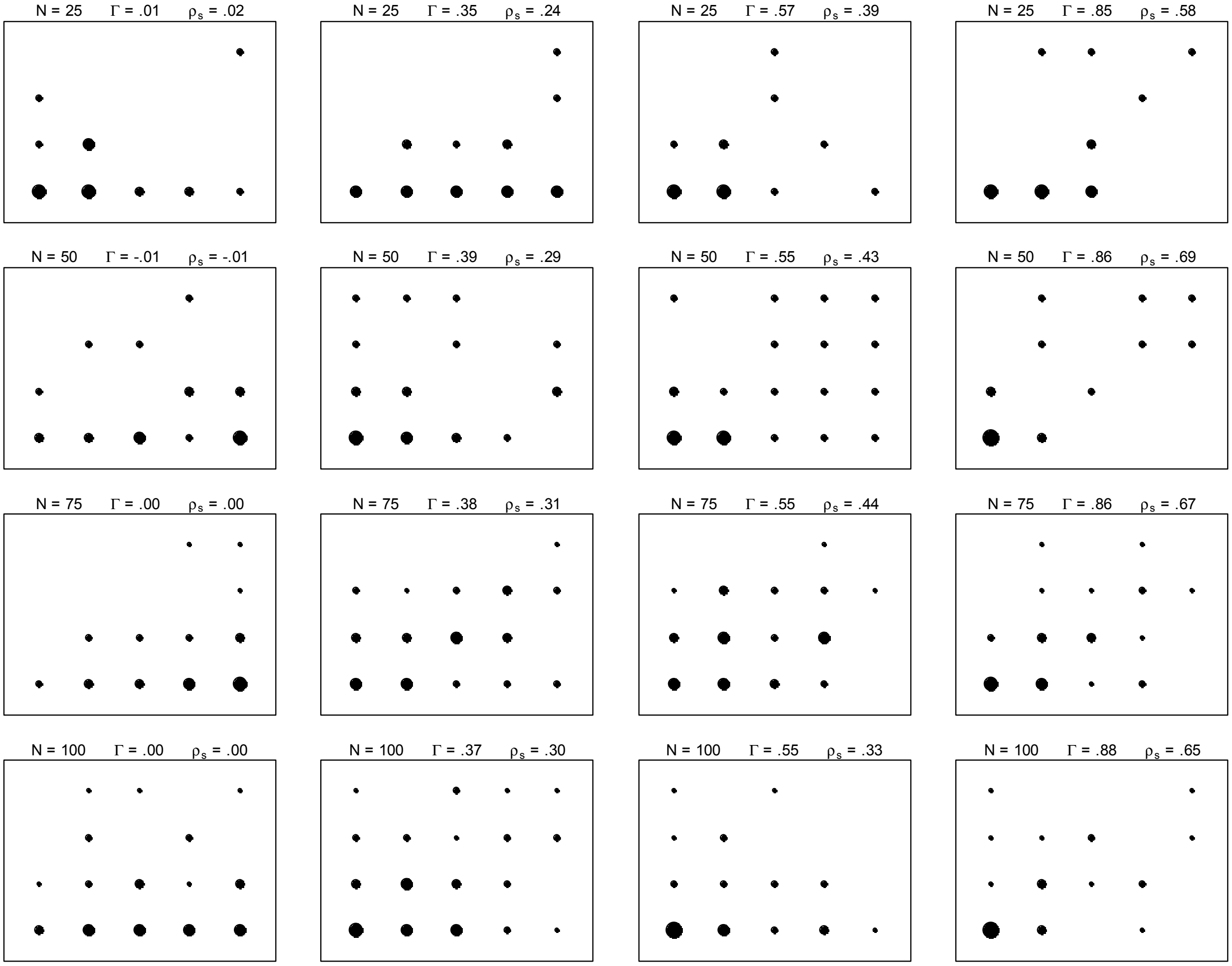
For each replication sample, r_S was calculated and Eqs. 5-7 were used to estimate the variance of $\sigma^2(z_{r_S})$ and construct CIs according to the methods of Fieller et al. (1957), Caruso and Cliff (1997), and Bonnett and Wright (2000). Then, B = 2,000 bootstrap samples - a quantity recommended by DiCiccio and Efron (1996) and also used by Chan and Chan (2004) - were drawn from each replication sample and r_S was calculated for each to construct a bootstrap BCA CI. The nominal level of all CIs was .95 (95%). Each bootstrap sample was checked to ensure that a correlation could be calculated (i.e., that both variables' variances were greater than zero); in a small number of instances, a new sample was drawn to replace one that was discarded because a correlation could not be calculated.

Within each cell of the design, observed coverage was recorded as the proportion of the CIs that included ρ_S (the value observed in the finite population from which replication samples were drawn). The absolute deviance between nominal and observed coverage was also recorded for each cell.

Results

Figure 4 displays the mean absolute deviance between nominal and observed coverage ($\overline{D}$) for each of the four types of CI. These graphs

Figure 3: Population distributions for data conditions with 4 × 5 tables in Woods (2007). The area of each plotting symbol is proportional to the frequency in that cell of the contingency table.



aggregate the results within types of population for all conditions, for each table size, for each level of ρ_S, and for each level of N. For the populations studied by Woods (2007), displayed in the upper-left panel, the results for the three types of analytic CIs are comparable to those in the original study; minor discrepancies are attributable to sampling variation between studies. $\overline{D}$ increased across levels of ρ_S for the analytic methods, reaching substantial values when ρ_S was large.

Because values of ρ_S did not vary discretely across the four levels in the design (recall that, strictly speaking, these were levels of Γ, not ρ_S), results were replotted in Figure 5 as observed coverage levels by ρ_S. This graph shows more clearly the tendency for coverage to fall below the nominal level with larger values of ρ_S. Relative to the coverage observed for the analytic methods, coverage for the bootstrap method was as good or better under most conditions, and much better for $\rho_S > .50$. Coverage for the bootstrap CIs remained within the control limits - the expected range of coverage results at $\alpha = .05$ with 1,000 replication samples, which is [.9365, .9635] - at even for the largest values of ρ_S. As expected, the bootstrap method yielded its largest values of $\overline{D}$ with the smallest samples ($N = 25$). Figure 5 shows that coverage for bootstrap CIs was outside of the control limits for only 4 of the 32 data conditions, each of which corresponded to an instance when $N = 25$. Different conditions seem to impair the performance of CIs for ρ_S constructed using analytic methods - in which case coverage falls below the nominal level as ρ_S increases - and the bootstrap method - in which case coverage is more erratic with smaller N.

Results for asymmetric populations (Figure 4) follow the same general pattern observed for the Woods (2007) populations. Here, the orthogonal manipulation of design factors helps to disentangle the effect of increasing ρ_S from the effects of different marginal distributions. As ρ_S increased, coverage remained closer to the nominal level for the bootstrap method than for the analytic methods; the difference was slight to nonexistent at $.00 \leq \rho_S \leq .30$, modest at $\rho_S = .50$, substantial at $\rho_S = .70$, and very large at $\rho_S = .90$. Once again, larger values of $\overline{D}$ were observed when the bootstrap method was used with smaller samples ($N = 25$) than with larger sample sizes ($50 \leq N \leq 200$).

With symmetric and uniform populations (Figure 4), perhaps the most striking result is that coverage for all methods approximated nominal levels fairly well under most conditions. Relative to the results for asymmetric populations, each of the methods achieved comparable or lower values of $\overline{D}$ under all conditions studied; note that that scaling of the y axes was held constant across panels in Figure 4 to facilitate this comparison. Nonetheless, the pattern of results across levels of ρ_S was similar to that observed for other populations: The bootstrap method maintained good coverage levels even at the highest values of ρ_S, whereas the analytic methods did not.

So far, results have focused primarily on absolute differences between observed and nominal coverage levels, and these discrepancies were averaged across cells in the design. To put more flesh on the bones of these results, for each CI method within each cell of the design, coverage was classified into one of seven categories using the control limits for $\alpha = .05$ (specified earlier), control limits of [.9322, .9678] for $\alpha = .01$, and control limits of [.9273, .9727] for $\alpha = .001$.

This classification indicates whether coverage was within all control limits, liberal (observed coverage less than the nominal level) to one of three extents ($\alpha = .05$, $\alpha = .01$, or $\alpha = .001$), or conservative (observed coverage greater than the nominal level) to one of these three extents. Figure 6 displays the results for the Woods (2007) populations, with results for each CI method in each cell of the design symbolized as within control limits (solid circle), liberal (downward-pointing triangles), or conservative (upward-pointing triangles); the size of a triangle corresponds to the most extreme α level at which the results fell beyond the control limits, with larger triangles indicative of greater deviance between observed and nominal coverage levels. Table 1 summarizes these results by tallying the frequency with which results fell into each of the seven

Figure 4: Mean absolute deviation between nominal (.95) and observed coverage.

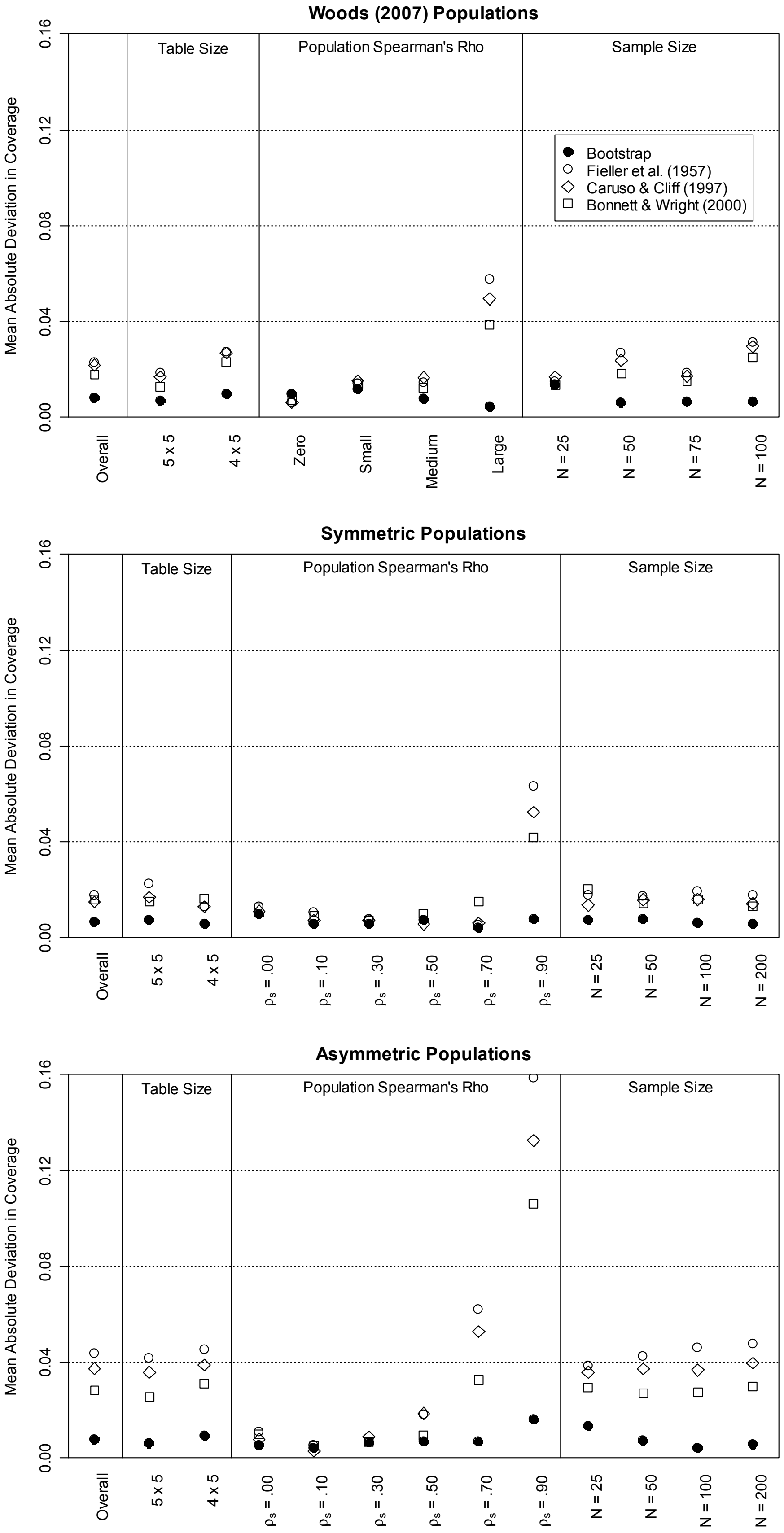
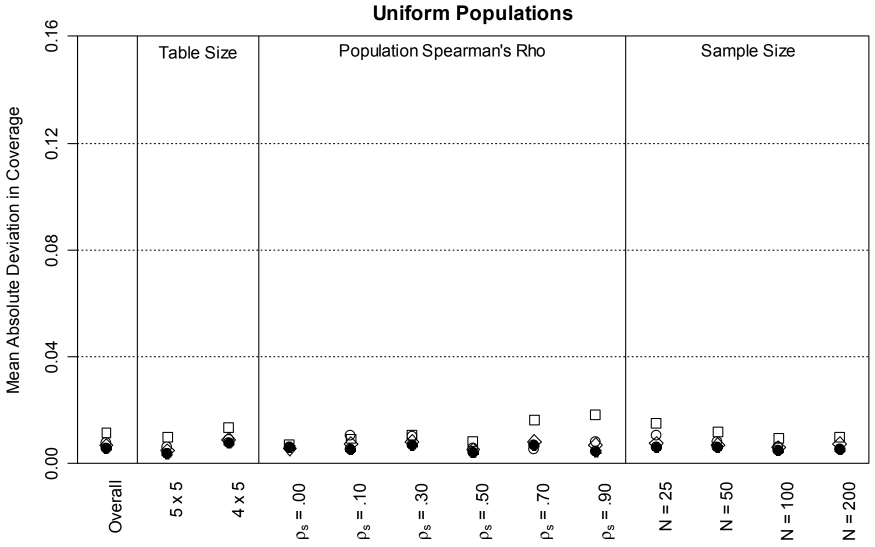


Figure 4 (continued): Mean absolute deviation between nominal (.95) and observed coverage.



categories for each CI method and population type.

Whereas the bootstrap method provided CIs whose coverage was within control limits for α = .05 88% of the time (28 of 32 conditions), the analytic methods provided CIs whose coverage was within these limits only 50% to 53% of the time. As noted earlier, the 4 exceptions for the bootstrap method occurred when N = 25 and exceptions for the analytic methods occurred more often as ρ_S increased.

Figure 7 displays the results for the asymmetric, symmetric, and uniform populations, and Table 1 summarizes these results as tallied frequencies. The bootstrap method provided CIs whose coverage was within control limits for 92%, 85%, and 94% of the conditions in these three types of populations, respectively. The corresponding figures for the analytic methods were lower, often substantially lower, coverage erred on the liberal side two to three times as often as it erred on the conservative side, and most deviances exceeded even the α = .001 level. Across all populations and data conditions (i.e., all 176 cells of the study design), the bootstrap method provided CIs whose coverage was within control limits 90% of the time, whereas the figures for analytic methods were 64% (Fieller, et al., 1957), 67% (Caruso & Cliff, 1997), and 56% (Bonnett & Wright, 2000).

One potential explanation for the generally liberal coverage of the analytic methods is that r_S is a biased statistic, usually underestimating the value of ρ_S (Cliff, 1996). To the extent that r_S is a biased estimator of ρ_S, it should not be surprising that CIs constructed around this statistic do not contain the population value sufficiently often to attain the nominal coverage level. In the present study, however, the magnitude of bias was rather small. The mean level of bias ($r_S - \rho_S$) was calculated across the 1,000 replication samples within each of the 176 cells of the design, and the distribution of these values is shown in Figure 8 (M = -.0024, Mdn = -.0020). It seems unlikely that such a slight bias contributed substantially to the deviance between observed and nominal coverage levels for the analytically derived CIs. Instead, the two factors identified earlier - ad hoc formulas for estimating $\sigma^2(z_{r_S})$ and the use of the t distribution in constructing the CI - remain plausible candidates for the source of this deviance.

Conclusion

This article reveals some important similarities and differences in the coverage of CIs for ρ_S with ordinal data constructed using four methods

Table 1: Frequencies of Observed Coverage Levels Within and Beyond Control Limits.

CI Method	Population Type	− − −	− −	−	CL	+	+ +	+ + +
Bootstrap	Woods (2007)	2	0	1	28	1	0	0
	Asymmetric	2	0	0	44	1	1	0
	Symmetric	0	0	3	41	2	1	1
	Uniform	0	0	0	45	2	1	0
	All Populations	4	0	4	158	6	3	1
Fieller, et al. (1957)	Woods (2007)	9	3	0	16	1	1	2
	Asymmetric	18	2	4	21	2	0	1
	Symmetric	4	0	1	32	3	3	5
	Uniform	0	0	0	44	1	2	1
	All Populations	31	5	5	113	7	6	9
Caruso & Cliff (1997)	Woods (2007)	9	2	0	16	1	2	2
	Asymmetric	19	3	2	23	0	1	0
	Symmetric	4	0	0	35	2	2	5
	Uniform	0	0	0	44	2	2	0
	All Populations	32	5	2	118	5	7	7
Bonnett & Wright (2000)	Woods (2007)	6	0	4	17	1	2	2
	Asymmetric	14	2	3	27	1	0	1
	Symmetric	4	0	0	25	5	9	5
	Uniform	0	0	0	29	8	8	3
	All Populations	24	2	7	98	15	19	11

Notes: There were 32 data conditions for the Woods (2007) populations and 48 data conditions for each of the other three populations (asymmetric, symmetric, and uniform), for a total of 176 data conditions. − − − = coverage < .95 at α = .001; − − = coverage < .95 at α = .01; − = coverage < .95 at α = .05; CL = coverage within control limits for .95 at α = .05; + = coverage > .95 at α = .05; + + = coverage > .95 at α = .01; + + + = coverage > .95 at α = .001.

Figure 5: Scatterplot of observed coverage by ρ_S for the Woods (2007) populations. Dashed lines show the control limits for nominal coverage of .95 at $\alpha = .05$, which are [.9365, .9635].

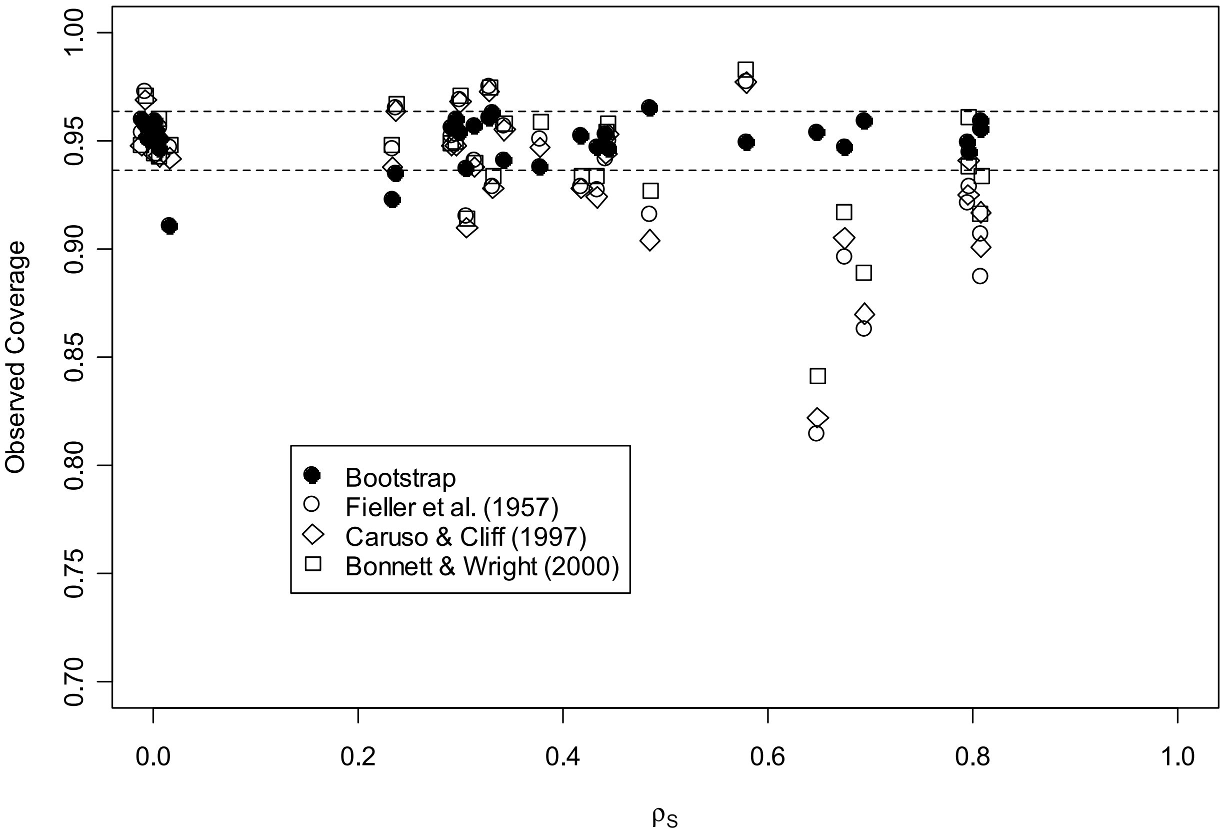
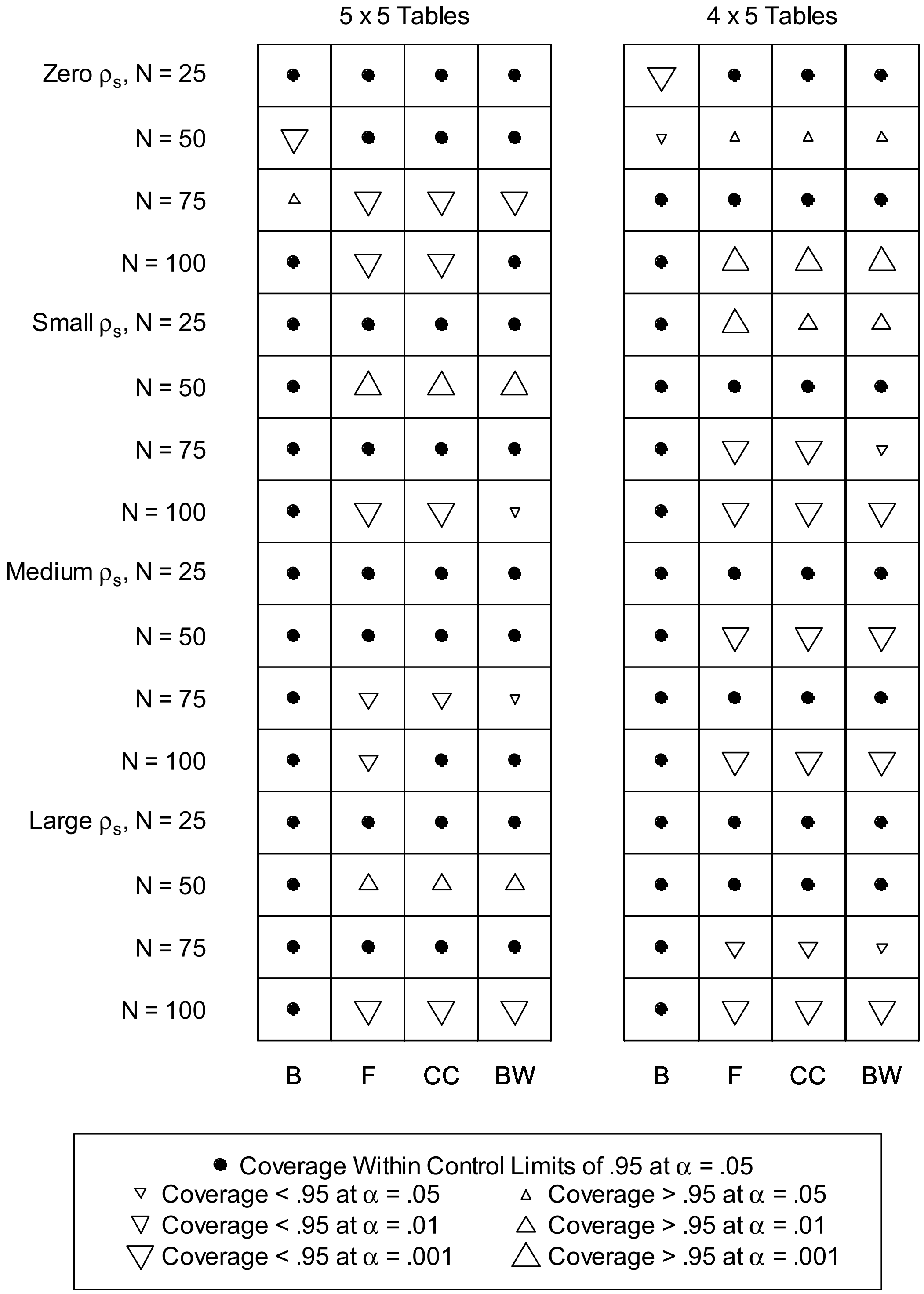


Figure 6: Chart indicates whether coverage was within the control limits of .95. These limits are [.9365, .9635] for α = .05, [.9322, .9678] for α = .01, and [.9273, .9727] for α = .001. B = bootstrap. F = Fieller, et al. (1957). CC = Caruso and Cliff (1997). BW = Bonnett and Wright (2000).



	5 x 5 Tables				4 x 5 Tables			
	B	F	CC	BW	B	F	CC	BW
Zero ρ_s, N = 25	●	●	●	●	**▽**	●	●	●
N = 50	**▽**	●	●	●	▿	▵	▵	▵
N = 75	▵	**▽**	**▽**	**▽**	●	●	●	●
N = 100	●	**▽**	**▽**	●	●	**△**	**△**	**△**
Small ρ_s, N = 25	●	●	●	●	●	**△**	△	△
N = 50	●	**△**	**△**	**△**	●	●	●	●
N = 75	●	●	●	●	●	**▽**	**▽**	▿
N = 100	●	**▽**	**▽**	▿	●	**▽**	**▽**	**▽**
Medium ρ_s, N = 25	●	●	●	●	●	●	●	●
N = 50	●	●	●	●	●	**▽**	**▽**	**▽**
N = 75	●	▽	▽	▿	●	●	●	●
N = 100	●	▽	●	●	●	**▽**	**▽**	**▽**
Large ρ_s, N = 25	●	●	●	●	●	●	●	●
N = 50	●	△	△	△	●	●	●	●
N = 75	●	●	●	●	●	▽	▽	▿
N = 100	●	**▽**	**▽**	**▽**	●	**▽**	**▽**	**▽**

● Coverage Within Control Limits of .95 at α = .05

▿ Coverage < .95 at α = .05	▵ Coverage > .95 at α = .05
▽ Coverage < .95 at α = .01	△ Coverage > .95 at α = .01
▽ Coverage < .95 at α = .001	**△** Coverage > .95 at α = .001

Figure 7: Chart indicates whether coverage was within the control limits of .95. These limits are [.9365, .9635] for α = .05, [.9322, .9678] for α = .01, and [.9273, .9727] for α = .001. B = bootstrap. F = Fieller, et al. (1957). CC = Caruso and Cliff (1997). BW = Bonnett and Wright (2000).

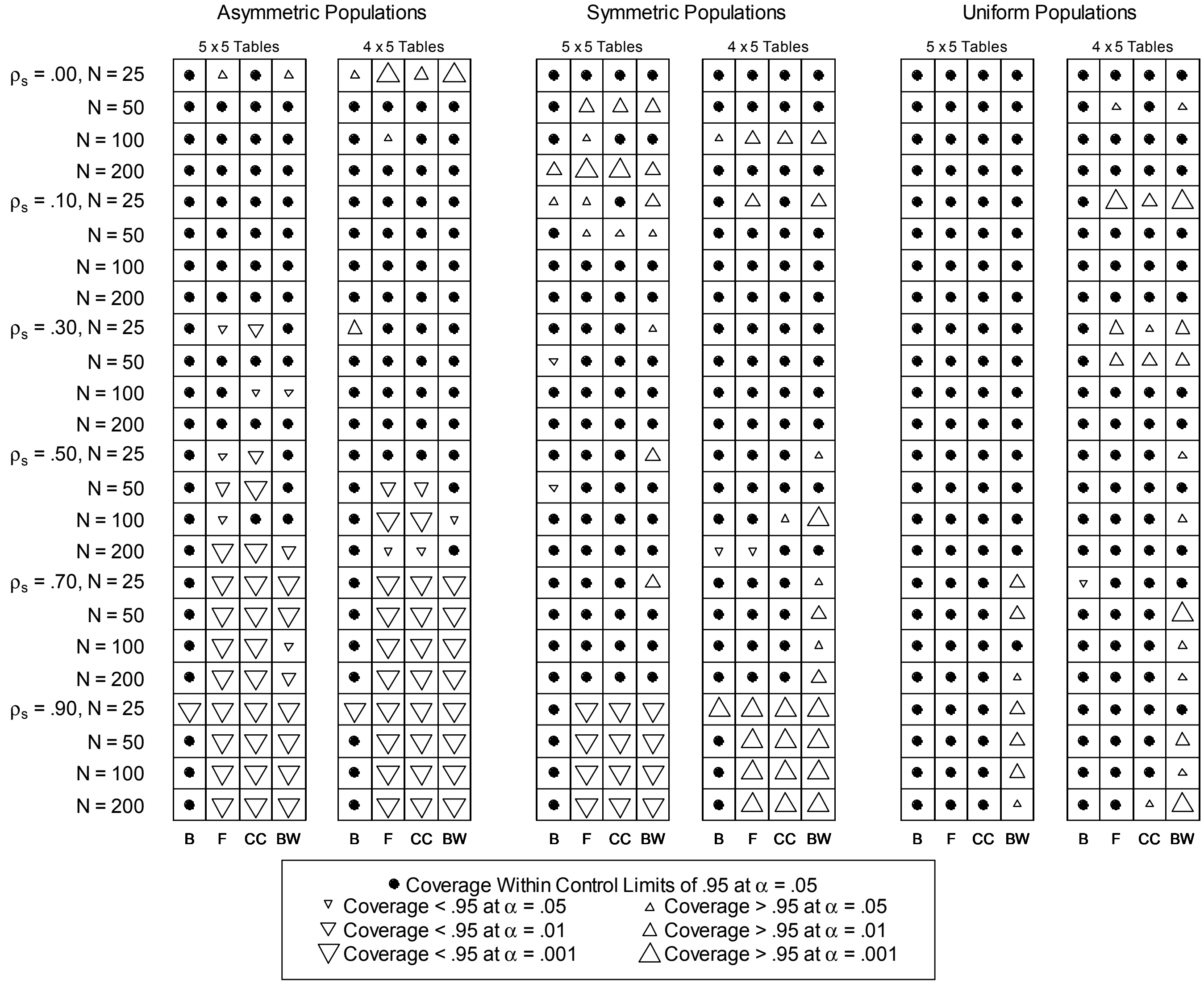
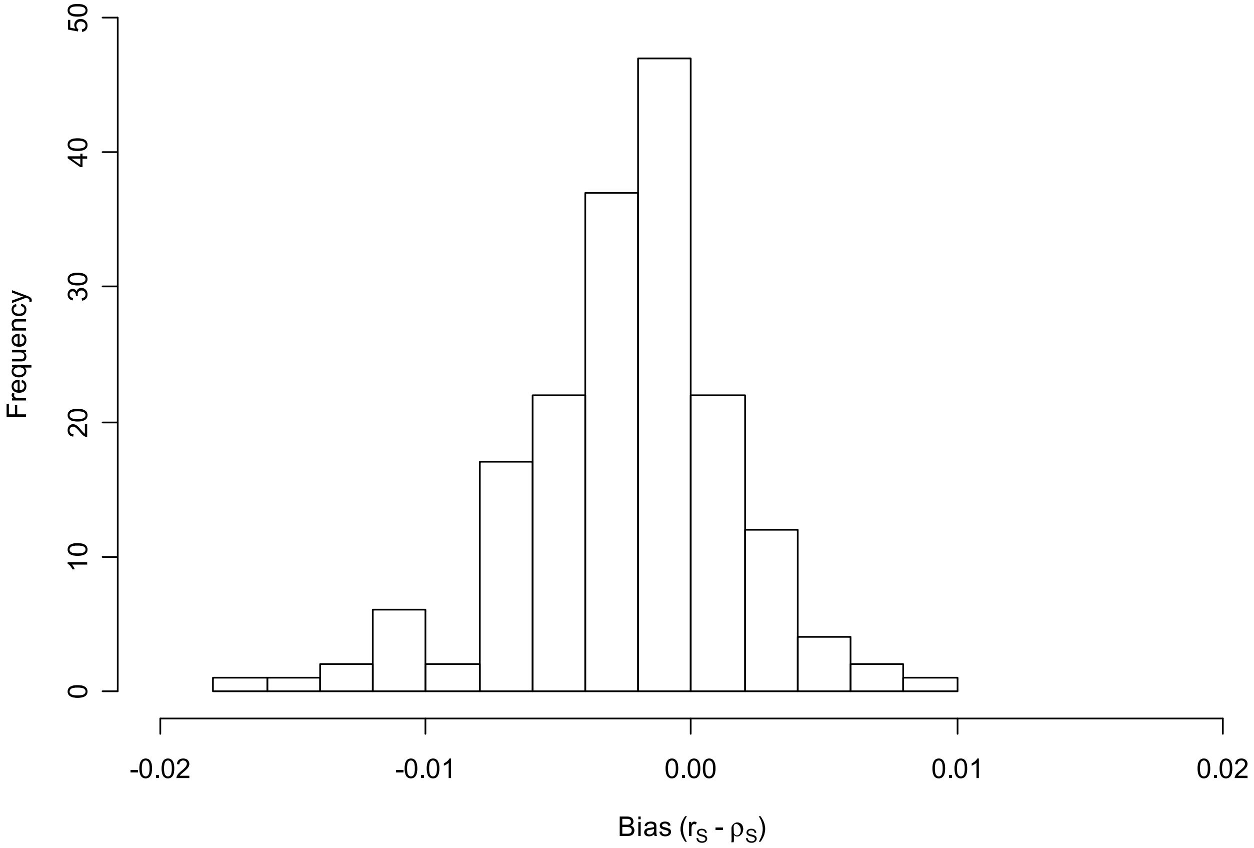


Figure 8: Histogram showing the bias in r_S as an estimator of ρ_S for all 176 cells of the study.



Under many conditions, both analytic and bootstrap methods provided CIs whose coverage approximated the nominal level of .95 well. These conditions included small values of ρ_S (between .00 and .30), moderate to large sample sizes (at least 50 cases), and symmetric (unimodal or uniform) marginal distributions. At larger values of ρ_S, the analytic methods tended to underestimate sampling error, yielding CIs that were too narrow and provided coverage less than the nominal level. This occurred for all marginal distributions studied, but the deviance was much greater for asymmetric than for symmetric distributions, and greater for unimodal than uniform distributions among those that were symmetric. Generally speaking, the BCA bootstrap method was robust across all values of ρ_S and each type of marginal distribution. To the extent that this method showed evidence of an Achilles' heel, it was the sometimes erratic coverage in the smallest samples studied (N = 25). Nonetheless, in many conditions with N = 25 and in nearly all conditions with $N \geq 50$, the BCA bootstrap method yielded CIs whose coverage was as good as or better than that of the analytic methods. At large values of ρ_S, this difference was substantial.

Although the study design spanned a broad array of data conditions - including several kinds of marginal distributions, sample sizes ranging from 25 to 200, and rank correlations ranging from .00 to .90 in ordinal data sets with relatively small numbers of categories - a number of issues remain to be clarified by future research. First, contingency tables of only two sizes were studied. Using Woods' (2007) investigation as a launching pad, the design included variables with either four or five categories crossed in 5 × 5 or 4 × 5 tables. With the exception of the symmetric, unimodal populations, 4 × 5 tables led to poorer coverage than 5 × 5 tables. Because there are only two table sizes, it is impossible to determine whether this effect is due to the variables' unequal numbers of categories or due to the inclusion of a variable with fewer categories. Teasing apart these possibilities would require independently manipulating the number of categories for each variable and the equality vs. inequality of these numbers across variable pairs.

The use of only two table sizes also prohibits the generalization of results to either smaller or larger tables. At one extreme, it is possible to calculate r_S for two dichotomous variables. However, there are many other measures of association available for the analysis of 2 × 2 tables, each of which was developed to address a specific type of research question (for an overview, see Kraemer, Kazdin, Offord, Kessler, Jensen, & Kupfer, 1999). It seems unlikely that one would select r_S as the most appropriate measure for a 2 × 2 table, but there remain table sizes between 2 × 2 and 4 × 5 that merit further study.

Because the analytic methods studied here involve ad hoc adjustments to a technique developed for use with bivariate normal data, using them with increasingly small table sizes - which necessitate deviations from bivariate normality - is likely to lead to less satisfactory results. Bootstrap methods may be especially well-suited to these conditions, and this possibility should be studied. At the other extreme, ordinal data with increasingly large numbers of categories would approximate continuous distributions. As table sizes increase, it becomes possible for data to approximate bivariate normality more closely, and the difference in coverage between analytic and bootstrap CIs probably will depend on distributional forms. The present study suggests that the bootstrap holds important advantages with asymmetric distributions; whether or not this generalizes to larger table sizes should be studied.

Also worthy of investigation is the possibility that bootstrap methods might yield CIs for ρ_S with even better coverage if a larger number of bootstrap samples is used. In the present study, B = 2,000 bootstrap samples per replication sample were generated and analyzed both because this value is recommended in the bootstrap literature (e.g., DiCiccio & Efron, 1996; Efron & Tibshirani, 1993) and has been used in similar simulation studies (e.g., Chan & Chan, 2004) and because available computing resources made a value this large feasible in the context of the study design. Even though the BCA bootstrap method performed fairly well in an absolute sense, and as good as or better than the analytic alternatives under most conditions, there remains room for improvement. For example, across the 176 data conditions studied here, coverage for the bootstrap CIs was within the α = .05 control limits of the nominal coverage level only 90% of the time, not 95% of the time.

When using nonparametric bootstrap techniques such as the percentile or BCA methods, which locate the limits of CIs by indexing positions within an empirical sampling distribution, it is important to attain sufficient precision in the tails of this distribution. A larger value of B would help to flesh out these tails. Moreover, it should improve the estimates of the median bias (z_0) and acceleration (a) parameters that are used to adjust the positions for locating the lower and upper limits of the CI. Whereas z_0 may change relatively little with increasing B, a is akin to a skewness statistic and its sampling error is not trivial; larger values of B should be especially useful in obtaining better estimates of a. All of this takes on greater importance if one wishes to construct CIs with even higher confidence levels than the usual .95, which was used exclusively in this study. For example, using the percentile bootstrap method by locating the values that define the middle 99% of an empirical sampling distribution requires a very large value of B to stabilize its tails, which are defined by only .5% of bootstrap samples apiece (e.g., 10 samples in each tail for B = 2,000).

Even though there are fruitful areas for follow-up research and no method of constructing CIs for ρ_S can guarantee that the observed coverage will equal the nominal level under all data conditions, researchers who would like to use r_S to measure the association between two variables can be advised to calculate and report CIs. With at least a moderate sample size (e.g., $N \geq 50$), the bootstrap BCA method with B = 2,000 appears to provide good coverage levels for any ρ_S from .00 to .90, even with as few as 4 or 5 ordered categories. If N is at least 25, the smallest value studied here, the analytic methods usually provided satisfactory coverage levels when ρ_S was not too large. For asymmetric distributions, coverage was good until ρ_S reached .50, and for symmetric distributions (unimodal or uniform), coverage was good until ρ_S reached .70. The only situations in which one would be well-advised to refrain from constructing CIs for ordinal data like those studied here are for small samples in which one's data are distributed asymmetrically and produce large values of r_S. Of course, conditions such as these would be extremely challenging for any correlational analysis - whether it involves testing H_0 or constructing a CI, using r_S or another measure of association - and it may be preferable to refrain from drawing strong conclusions from such data unless and until a method can be developed that handles them satisfactorily.

Notes

[1] The coefficient r_S is sometimes referred to as "Spearman's rho," which can be ambiguous in that Greek letters often are reserved for values calculated in populations rather than samples. In the present article, r_S will be used to denote the sample estimate of ρ_S, the population value of Spearman's rank correlation.

[2] Results for r_S published in Woods (2007) are superseded by those in a correction (Woods, 2008).

[3] Lee and Rodgers (1998) distinguished univariate and bivariate resampling for bootstrap applications with correlation coefficients. Whereas univariate resampling was found to be more useful for tests of statistical significance, it yields samples in which the marginal distributions reproduce those in the original data but the variables are uncorrelated (save for sampling error). As Lee and Rodgers note, bivariate resampling is required to construct CIs because this preserves not only the marginal distributions, but also the correlation in the original data. Thus, bivariate resampling was used exclusively for analyses presented in this paper.

[4] Categories were recoded to consecutive natural numbers. In the original populations used by Woods (2007), the coding of some variables began at 0 and others at 1, and some variables had frequencies of 0 at intermediate category numbers (e.g., scores of 0, 1, 2, and 4 occurred, with no scores of 3). Because this recoding preserved scores' rank order, it did not affect results.

[5] For the data condition with $\rho_S = .90$ and a 4 × 5 contingency table with symmetric marginal frequency distributions, ρ_S in the finite population of 100,000 cases was .8713. As in all other conditions, CI coverage was evaluated against the correlation observed in the finite population, not the correlation specified in the design, so the failure to generate a finite population with a .90 correlation should not bias the coverage results.

References

American Psychological Association. (2009). *Publication manual* (6th ed.). Washington, DC: Author.

Bonnett, D. G., & Wright, T. A. (2000). Sample size requirements for estimating Pearson, Kendall, and Spearman correlations. *Psychometrika*, *65,* 23-28.

Caruso, J. C., & Cliff, N. (1997). Empirical size, coverage, and power of confidence intervals for Spearman's rho. *Educational and Psychological Measurement*, *57,* 637-654.

Chan, W., & Chan, D. W. L. (2004). Bootstrap standard error and confidence intervals for the correlation corrected for range restriction: A simulation study. *Psychological Methods, 9,* 369-385.

Cliff, N. (1996). *Ordinal methods for behavioral data analysis.* Mahwah, NJ: Lawrence Erlbaum Associates.

DiCiccio, T. J., & Efron, B. (1996). Bootstrap confidence intervals. *Statistical Science, 11,* 189-228.

Edgington, E. S. (1987). *Randomization tests.* New York: Marcel Dekker.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. Boca Raton, FL: CRC Press.

Fieller, E. C., Hartley, H. O., & Pearson, E. S. (1957). Tests for rank correlation coefficients: I. *Biometrika*, *44,* 470-481.

Gonzalez, R., & Nelson, T. O. (1996). Measuring ordinal association in situations that contain tied scores. *Psychological Bulletin*, *119,* 159-165.

Higgins, J. J. (2004). *An introduction to modern nonparametric statistics*. Pacific Grove, CA: Brooks/Cole.

Kline, R. B. (2005). *Beyond significance testing: Reforming data analysis methods in behavioral research*. Washington, DC: American Psychological Association.

Kraemer, H. C., Kazdin, A. E., Offord, D. R., Kessler, R. C., Jensen, P. S., & Kupfer, D. J. (1999). Measuring the potency of risk factors for clinical or policy significance. *Psychological Methods*, *4,* 257-271.

Lee, W.-C., & Rodgers, J. L. (1998). Bootstrapping correlation coefficients using univariate and bivariate sampling. *Psychological Methods*, *3,* 91-103.

Ruscio, J., & Kaczetow, W. (2008). Simulating multivariate nonnormal data using an iterative approach. *Multivariate Behavioral Research*, *43*, 355-381.

Spearman, C. S. (1904). The proof and measurement of association between two things. *American Journal of Psychology*, *15,* 72-101.

Ruscio, J., Ruscio, A. M., & Meron, M. (2007). Applying the bootstrap to taxometric analysis: Generating empirical sampling distributions to help interpret results. *Multivariate Behavioral Research*, *44,* 349-386.

Stevens, S. S. (1946, June 7). On the theory of scales of measurement. *Science*, *103,* 677-680.

Wilcox, R. R. (2003). *Applying contemporary statistical techniques.* San Diego, CA: Academic Press.

Wilkinson, L., and the APA Task Force on Statistical Inference (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54,* 594-604.

Woods, C. M. (2007). Confidence intervals for gamma-family measures of ordinal association. *Psychological Methods*, *12,* 185-204.

Woods, C. M. (2008). Correction to Woods (2007). *Psychological Methods*, *13*, 72-73.

Zar, J. H. (1972). Significance testing of the Spearman rank correlation coefficient. *Journal of the American Statistical Association*, *67,* 578.

Journal of Modern Applied Statistical Methods
November, 2008, Vol. 7, No. 2, 435-445

1538 – 9472/08/$95.00

Two Dimension Marginal Distributions of Crossing Time and Renewal Numbers Related to Two-Stage Erlang Processes

Mir Ghulam Hyder Talpur
Ball State University

Iffat Zamir
University of Sindh,
Jamshoro, Pakistan

M. Masoom Ali
Ball State University

The two dimensional marginal transform, probability density and cumulative probability distribution functions for the random variables T_{ξ_N} (time taken by servers during vacations), ξ_N (number of vacations taken by servers) and η_N (number of customers or units arriving in the system) are derived by taking combinations of these random variables. One random variable is controlled at one time to determine the effect of the other two random variables simultaneously.

Key Words: Two dimensional marginal distribution, Erlang processes, Markov processes, renewal processes.

Introduction

Biggins and Cannings (1987) found that a Markov renewal process $\{(X_n, T_n): n \geq 0\}$ might have two constituents, and that $\{X_n : n \geq 0\}$ is a homogenous Markov chain where $(T_{n+1} - T_n)$ is the sojourn time in $X_n(T_0 = 0)$. Thus, X_n could be the state entered at T_n and left at T_{n+1}, assuming that $\{X_n : n \geq 0\}$ and $\{T_{n+1} - T_n \geq 0\}$ are independent, and the distribution of $(T_{n+1} - T_n)$ is dependent on $\{X_n : n \geq 0\}$ through X_n and X_{n+1} only (otherwise not dependent on *n*). It is assumed that sojourn time is always strictly positive. When the initial state is *i*, which is $X_0 = i$, the return to state *i* is an ordinary renewal process, and the visit to $j \neq i$ forms a delayed renewal process (the delay being the elapsed time until the first visit to *j*). Thus, Cinlar (1969) proved the theory of Markov renewal processes which generalizes those of renewal processes and the Markov chain and is a blend of the two. Biggins and Cannings (1987) applied the Markov Chain $\{X_n : n \geq 0\}$ to a finite state space assuming it to be the case in what followed, so that all introduced matrices are finite. In addition, the time T_n is integer-valued for transforms used for generating functions (with argument z) and Laplace transforms. They showed that the Markov renewal process theory provided a useful framework for the Markov chain model with wider applicability to the occurrence of sequences in the Markov chains, specifically on type one counters. Results are applied to problems regarding the reliability of the consecutive k-out-of-n:F system (Koutras & Papastavirdis, 1993; Godbole, 1993; Fu & Koutras,1994). The geometric distribution of order k was one of the simplest waiting time distributions. Several waiting time problems have been studied in more general situations (Ebneshahrshoob & Sobel, 1990; Kreos, 1992; Aki, 1992; Aki & Hirano, 1989, 1993, 1999; Mohanty, 1994). A class of waiting time problems was proposed by Ebneshahrshoob and Sobel (1990), who obtained the probability

Mir Ghulam Hyder Talpur is a Visiting Scholar in the Department of Mathematical Sciences. Email: ghtalpur@usindh.edu.pk. Iffat Zamir is in the Department of Statistics. Email: izamirmalik@yahoo.com. M. Masoom Ali is in the Department of Mathematical Sciences. Email: mali@bsu.edu.

generating functions (PGF) of waiting time distributions for a run of 0 of length r and a run of 1 of length k. Ling (1990) studied the distribution of waiting time for first time occurrence among E's when X's are independently and identically distributed (IID) and finite valued random variables, and all k's had the same value. Aki and Hirano (1993) obtained the PGF's of the distributions of the sooner or later waiting time for the same event as Ebneshahrshoob and Sobel. Talpur and Shi (1994) found the two dimension marginal distributions of crossing time and renewal numbers related with two Poisson processes using probability arguments, and constructing an absorbing Markov process. In this article, the same technique is extended for the case of the two stage Erlang process.

Methodology

An extensive literature review has shown that renewal processes are widely used in reliability theory and in models of queuing theory. Both theories are based on counting processes, and situations where the differences between two or more counting processes examined are common. Stochastic processes can be helpful in analyzing such situations. Kroese (1992) showed the difference process of the two counting processes as

$$D(t) = N_1(t) - N_2(t),$$

where $N_1(t)$ and $N_2(t)$ are two counting processes associated with the corresponding renewal sequences of $\{X_i\ \}$ and $\{Y_j\}$. The problem considered here is extended from the work of Kroese (1992) and Talpur & Shi (1994) and is based upon the renewal sequence of two variables $\{X_i\}$ and $\{Y_j\}$ as shown in Figure 1.
Let

$$\xi_N = \min_{n\to\infty}\{n / T_n \geq S_N\},$$

where ξ_N is a random variable and N is a constant.

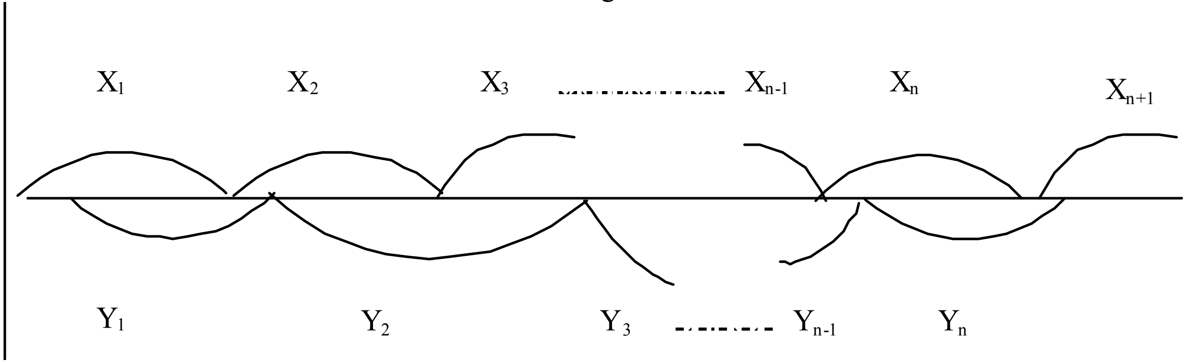
$$S_0 = 0,\ S_N = X_1 + X_2 + \ldots + X_n$$

$$T_0 = 0,\ T_n = Y_1 + Y_2 + \ldots + Y_n$$

$$T_{\xi_N} = \sum_{j=1}^{\xi_N} Y_j\ ,$$

X represents the inter arrival, and Y is the number of vacations performed by the server. Both variables are discrete and have renewal processes at each occurrence. The level of absorption is achieved at the n^th^ arrival of X_n; after the n^th^ arrival, the n^th^ vacation Y_n of the server occurs. The difference between the times at which the n^th^ vacation occurred and the n^th^ customer arrived is the crossing time of the server. The probability generating function, probability density function, and cumulative probability distribution function for the two dimensional marginal distribution for the three random variables T_{ξ_N} (time taken by servers during vacations), ξ_N (number of vacations taken by servers), and η_N (number of customers or units arriving in the system) are thus obtained.

Figure 1



Assumptions

Let N be a constant and X_i and Y_j be two sequences of random variables. Assume that X_i ($i = 1, 2, 3, \ldots$) is independently and identically distributed (IID) with a finite mean, λ^{-1}, and that Y_j ($j = 1,2,3,\ldots$) is IID with a finite mean μ^{-1}. Assume also that $N_1(t)$ is the Erlang process associated with X_i, in which the distribution of X_i is the 2-stage Erlang distribution, and $N_2(t)$ is the Erlang process associated with Y_j, in which the distribution of Y_j is the 2-stage Erlang distribution. In addition, assume that X_i and Y_j are mutually independent.

Absorbing Markov process and absorbing time distribution

Consider a Markov process $\{X(t), t \geq 0\}$ on the state space $E = (0,1,2,\ldots)$. If E_0 and E_1 are two non-null subsets of E and they satisfy the cases

$$E_0 \cup E_1 = E$$

and

$$E_0 \cap E_1 = \varnothing,$$

then, E_0 and E_1 are called a partition of E. If E_0 is the absorbing state set and E_1 is the transient state set, and α_E is the initial condition, the absorbing Markov process (AMP) is constructed to analyze the problem. Consider the AMP $\{N_1(t), N_2(t), I(t), J(t)\}$, in which $N_1(t)$ and $N_2(t)$ are the counting processes associated with X_i and Y_j respectively. I(t) and J(t) represent the phases of X_i and Y_j at time t respectively, and its state space is

$$E = \{(i,k,j,l),(i',j')/ i,j = 0,1,\ldots;\ k,l = 1,2;\\ i' = N', N'+1',\ldots;\ j' = 1',2',\ldots;\}$$

where (i',j') are absorbing states. The transition of states is illustrated in Figure 2. Let

$$P_{ij}(k,l,t) = p\{N_1(t) = i, N_2(t) = j, I(t) = k, J(t) = l\}$$

and

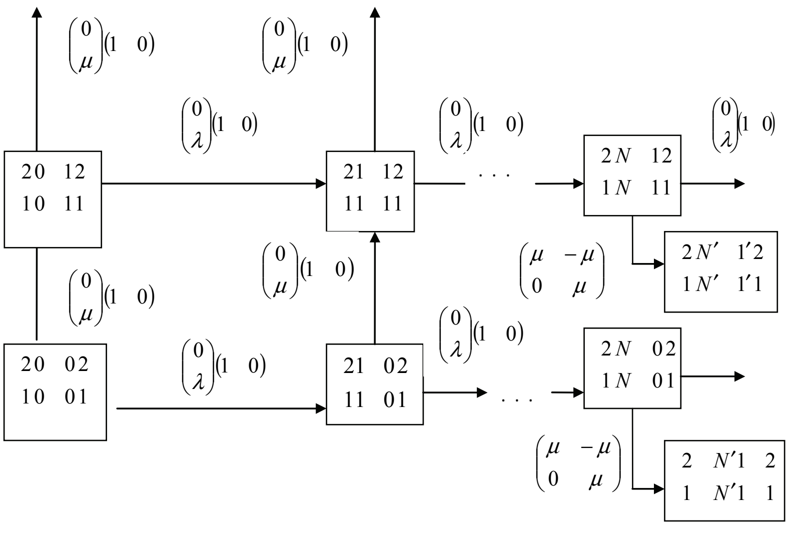
$$P_{ij}(t) = \\ \left[p_{ij}(1,1,t),\ldots p_{ij}(1,n,t),\ldots p_{ij}(m,1,t),\ldots p_{ij}(m,n,t)\right].$$

From the transition rate diagram, the systems differential equations are as follows

$$P'_{ij}(t) = \\ p_{ij}(t)\left\{-\left\{\begin{pmatrix}\lambda & -\lambda\\ 0 & \lambda\end{pmatrix} + \begin{pmatrix}\mu & -\mu\\ 0 & \mu\end{pmatrix}\right\}\right\} \\ + p_{i-1,j}(t)\begin{pmatrix}0 & 0\\ \lambda & 0\end{pmatrix} + p_{ij-1}(t)\begin{pmatrix}0\\ \mu\end{pmatrix}\begin{pmatrix}1 & 0\end{pmatrix}$$

where $i = 0,1,\ldots N-1;, j = 0,1,2,\ldots;$. (2.1)

Figure 2: Transition Rate Diagram


$\begin{pmatrix}0\\ \mu\end{pmatrix}(1\ \ 0)$
$\begin{pmatrix}0\\ \lambda\end{pmatrix}(1\ \ 0)$
$\begin{pmatrix}\mu & -\mu\\ 0 & \mu\end{pmatrix}$
20 12
10 11
21 12
11 11
2N 12
1N 11
2N′ 1′2
1N′ 1′1
20 02
10 01
21 02
11 01
2N 02
1N 01
2 N′1 2
1 N′1 1
. . .

and

$$P'_{ij}(t) = p_{ij}(t)\left\{-\left\{\begin{pmatrix}\lambda & -\lambda\\ 0 & \lambda\end{pmatrix}+\begin{pmatrix}\mu & -\mu\\ 0 & \mu\end{pmatrix}\right\}\right\} + p_{i-1j}(t)\begin{pmatrix}0 & 0\\ \lambda & 0\end{pmatrix},$$

where $i = N, N+1, ...;, j = 0,1,...$ (2.2)

Using these differential equations, Talpur and Iffat (2007) obtained the joint distribution for three random variables. The two dimension marginal distributions for the same problem were also obtained in this study.

Two dimensional marginal probability distribution function for T_{ξ_N}, ξ_N

The number of arriving customers is fixed in order to observe the effect of time taken by server vacations and the number of vacations taken. The two dimensional marginal probability generating function (probability transform function), two dimensional marginal probability density function, and the two dimensional marginal cumulative probability distribution function for random variables T_{ξ_N}, ξ_N are computed under the following theorems.

Theorem 3.1

The two dimensional marginal probability generating function of the two random variables T_{ξ_N}, ξ_N is given by:

$$f^*(s,u) = u(1 \quad 0)\left\{\begin{pmatrix}s+\lambda+\mu & -(\lambda+\mu)\\ -u\mu & s+\lambda+\mu\end{pmatrix}^{-1}\begin{pmatrix}0 & 0\\ \lambda & 0\end{pmatrix}\right\}^{N}\begin{pmatrix}s+\lambda+\mu & -(\lambda+\mu)\\ -\lambda & s+\lambda+\mu\end{pmatrix}^{-1}\begin{pmatrix}\mu\\ \mu\end{pmatrix}.$$

Proof 3.1

The two dimensional marginal probability generating function of two random variables T_{ξ_N}, ξ_N is calculated from the joint probability generating function of three random variables T_{ξ_N}, ξ_N and η_N (Talpur & Iffat, 2007):

$$f^*(s,u,z) = u(1 \quad 0)\left\{\begin{pmatrix}s+\lambda+\mu & -(\lambda+\mu)\\ -u\mu & s+\lambda+\mu\end{pmatrix}^{-1}\begin{pmatrix}0 & 0\\ \lambda z & 0\end{pmatrix}\right\}^{N}\begin{pmatrix}s+\lambda+\mu & -(\lambda+\mu)\\ -\lambda z & s+\lambda+\mu\end{pmatrix}^{-1}\begin{pmatrix}\mu\\ \mu\end{pmatrix}$$

If z is close to 1, then the two dimensional marginal probability generating function is:

$$f^*(s,u) = u(1 \quad 0)\left\{\begin{pmatrix}s+\lambda+\mu & -(\lambda+\mu)\\ -u\mu & s+\lambda+\mu\end{pmatrix}^{-1}\begin{pmatrix}0 & 0\\ \lambda & 0\end{pmatrix}\right\}^{N}\begin{pmatrix}s+\lambda+\mu & -(\lambda+\mu)\\ -\lambda & s+\lambda+\mu\end{pmatrix}^{-1}\begin{pmatrix}\mu\\ \mu\end{pmatrix}. \qquad (3.1)$$

Theorem 3.2

The two dimensional marginal probability density function of two random variables T_{ξ_N}, ξ_N is given by:

$$p\{T_{\xi_N} \le t, \xi_N = j\} = \sum_{i=N}^{\infty}\binom{N+j-2}{j-1}\left[\begin{aligned}&\lambda^i\mu^j(\lambda+\mu)^{j+i-1}\frac{t^{2j+2i-2}}{(2j+2i-2)!}e^{-(\lambda+\mu)t}\\ &+\lambda^i\mu^j(\lambda+\mu)^{j+i}\frac{t^{2j+2i-1}}{(2j+2i-1)!}e^{-(\lambda+\mu)t}\end{aligned}\right]dt$$

Proof 3.2

As used by Talpur and Shi (1994), the following equation can be obtained by definition of the z and L transform

$$f^*(s,u)=$$
$$\sum_{j=1}^{\infty}\int_{0}^{\infty}\exp(-st)dp\left\{T_{\xi_N}\le t,\xi_N=j\right\}u^j$$

Inserting the value from equation (3.1) results in:

$$f^*(s,u)=$$
$$u\begin{pmatrix}1 & 0\end{pmatrix}\left\{\begin{pmatrix} s+\lambda+\mu & -(\lambda+\mu)\\ -u\mu & s+\lambda+\mu\end{pmatrix}^{-1}\begin{pmatrix}0 & 0\\ \lambda & 0\end{pmatrix}\right\}^N$$
$$\begin{pmatrix} s+\lambda+\mu & -(\lambda+\mu)\\ -\lambda & s+\lambda+\mu\end{pmatrix}^{-1}\begin{pmatrix}\mu\\ \mu\end{pmatrix}$$

Let $a=s+\lambda+\mu$ and applying the rule of power series as in Pipes and Harwil (1970) results in:

$$f^*(s,u)\;=$$
$$\frac{u}{a}\left\{\frac{\lambda}{a}\sum_{k=0}^{\infty}\left(\frac{\lambda+\mu}{a}\right)^{k+1}\left(\frac{u\mu}{a}\right)^{k}\right\}^N$$
$$\mu\sum_{l=0}^{\infty}\left(\frac{\lambda+\mu}{a}\right)^{l}\left(\frac{\lambda}{a}\right)^{l}\left\{1+\frac{\lambda+\mu}{a}\right\}$$

Applying the negative binomial distribution and simplifying the series (Hogg & Craig, 2006), and setting $j=k+1$ and $i=N+l$, results in:

$$f^*(s,u)=$$
$$\sum_{j=1}^{\infty}\sum_{i=N}^{\infty}\begin{pmatrix}N+j-1\\ j-1\end{pmatrix}\left(\frac{u}{s+\lambda+\mu}\right)^{j}\left(\frac{\lambda}{s+\lambda+\mu}\right)^{i}$$
$$\left\{\left(\frac{\lambda+\mu}{s+\lambda+\mu}\right)^{j+i-1}+\left(\frac{\lambda+\mu}{s+\lambda+\mu}\right)^{j+i}\right\}u^{k+1}$$

After comparing the coefficient of u^j and taking the inverse of the Laplace transform, the two dimensional probability density function for the two random variables time taken by vacation of servers with respect to number of vacations is obtained as follows:

$$p\left\{T_{\xi_N}\le t,\xi_N=j\right\}=$$
$$\sum_{i=N}^{\infty}\begin{pmatrix}N+j-2\\ j-1\end{pmatrix}\lambda^i\mu^j$$
$$\begin{bmatrix}(\lambda+\mu)^{j+i-1}\dfrac{t^{2j+2i-2}}{(2j+2i-2)!}\\ +(\lambda+\mu)^{j+i}\dfrac{t^{2j+2i-1}}{(2j+2i-1)!}\end{bmatrix}e^{-(\lambda+\mu)t}$$

(3.2)

Theorem 3.3

The two dimensional marginal cumulative probability distribution function of two random variables T_{ξ_N},ξ_N is given by:

$$p\left\{T_{\xi_N}\le t,\xi_N=j\right\}=$$
$$\sum_{i=N}^{\infty}\begin{pmatrix}N+j-2\\ j-1\end{pmatrix}\left(\frac{\lambda}{\lambda+\mu}\right)^{i}\left(\frac{\mu}{\lambda+\mu}\right)^{j}$$
$$\left[\sum_{r=0}^{2i+2j-1}\frac{[(\lambda+\mu)t]^r}{r!}+\sum_{r=1}^{2i+2j}\frac{[(\lambda+\mu)t]^r}{r!}\right]e^{-(\lambda+\mu)t}$$

Proof 3.3

The two dimensional marginal cumulative probability distribution function T_{ξ_N},ξ_N, is obtained by integrating the probability density function (Medhi, 1982).

$$p\left\{T_{\xi_N}\le t,\xi_N=j\right\}$$
$$=\int_{0}^{\infty}dp\left\{T_{\xi_N}\le t,\xi_N=j\right\}dt$$
$$=\int_{0}^{\infty}\sum_{i=N}^{\infty}\begin{pmatrix}N+j-2\\ j-1\end{pmatrix}\lambda^i\mu^j$$
$$\begin{bmatrix}(\lambda+\mu)^{j+i-1}\dfrac{t^{2j+2i-2}}{(2j+2i-2)!}\\ +(\lambda+\mu)^{j+i}\dfrac{t^{2j+2i-1}}{(2j+2i-1)!}\end{bmatrix}e^{-(\lambda+\mu)t}\,dt$$

Integration by parts is applied to obtain the cumulative probability distribution function for the length of vacations taken by servers with the number of vacations taken by servers:

$$p\left\{T_{\xi_N} \le t, \xi_N = j\right\} = F(t,u)$$
$$= \sum_{i=N}^{\infty} \binom{N+j-2}{j-1} \left(\frac{\lambda}{\lambda+\mu}\right)^i \left(\frac{\mu}{\lambda+\mu}\right)^j$$
$$\left[\sum_{r=0}^{2i+2j-1} \frac{[(\lambda+\mu)t]^r}{r!} + \sum_{r=1}^{2i+2j} \frac{[(\lambda+\mu)t]^r}{r!}\right] e^{-(\lambda+\mu)t} \tag{3.3}$$

Two dimensional marginal probability distribution functions for T_{ξ_N}, η_N

The effect of the number of vacations taken by servers combined with the numbers of customers arriving is studied by controlling the number of vacations taken by the servers. The two dimensional marginal probability transform function (probability generating function), two dimensional probability density function, and two dimensional marginal cumulative probability distribution function are obtained. The time taken by number of vacations by servers with the number of arriving units are represented by random variables T_{ξ_N}, η_N.

Theorem 4.1

The two dimensional marginal probability generating function (probability transform function) for random variables T_{ξ_N}, η_N is:

$$f^*(s,z) =$$
$$\begin{pmatrix} 1 & 0 \end{pmatrix} \left\{ \begin{pmatrix} s+\lambda+\mu & -(\lambda+\mu) \\ -\mu & s+\lambda+\mu \end{pmatrix}^{-1} \begin{pmatrix} 0 & 0 \\ \lambda z & 0 \end{pmatrix} \right\}^N$$
$$\begin{pmatrix} s+\lambda+\mu & -(\lambda+\mu) \\ -\lambda z & s+\lambda+\mu \end{pmatrix}^{-1} \begin{pmatrix} \mu \\ \mu \end{pmatrix}$$

Proof 4.1

The two dimension marginal probability generating function for the two random variables, T_{ξ_N}, η_N, is calculated from the joint probability generating function (joint probability transform function) for the three random variables T_{ξ_N}, ξ_N and η_N as given by Talpur and Iffat (2007):

$$f^*(s,u,z) =$$
$$u\begin{pmatrix} 1 & 0 \end{pmatrix} \left\{ \begin{pmatrix} s+\lambda+\mu & -(\lambda+\mu) \\ -u\mu & s+\lambda+\mu \end{pmatrix}^{-1} \begin{pmatrix} 0 & 0 \\ \lambda z & 0 \end{pmatrix} \right\}^N$$
$$\begin{pmatrix} s+\lambda+\mu & -(\lambda+\mu) \\ -\lambda z & s+\lambda+\mu \end{pmatrix}^{-1} \begin{pmatrix} \mu \\ \mu \end{pmatrix}$$

Let u be close to 1^- for controlling the effect of the random variable ξ_N. Then we get the two dimensional marginal probability generating function of the two random variables T_{ξ_N}, η_N as

$$f^*(s,z) =$$
$$\begin{pmatrix} 1 & 0 \end{pmatrix} \left\{ \begin{pmatrix} s+\lambda+\mu & -(\lambda+\mu) \\ -\mu & s+\lambda+\mu \end{pmatrix}^{-1} \begin{pmatrix} 0 & 0 \\ \lambda z & 0 \end{pmatrix} \right\}^N$$
$$\begin{pmatrix} s+\lambda+\mu & -(\lambda+\mu) \\ -\lambda z & s+\lambda+\mu \end{pmatrix}^{-1} \begin{pmatrix} \mu \\ \mu \end{pmatrix} \tag{4.1}$$

Theorem 4.2

The two dimensional marginal probability density function for random variables T_{ξ_N}, η_N is:

$$f(t,z) =$$
$$\sum_{j=0}^{\infty} \binom{N+j-1}{j-1} \lambda^i \mu^j$$
$$\begin{bmatrix} (\lambda+\mu)^{j+i} \dfrac{t^{2j+2i-1}}{(2j+2i-1)!} \\ +(\lambda+\mu)^{j+i-1} \dfrac{t^{2j+2i}}{(2j+2i)!} \end{bmatrix} e^{-(\lambda+\mu)t} dt$$

Proof 4.2

The following equation as given by Talpur and Shi (1994) can be expressed by the definition of z and L transform as:

$$f^*(s,z) = \sum_{i=N}^{\infty}\int_0^{\infty}\exp(-st)dp\left\{T_{\xi_N}\le t,\eta_N=i\right\}z^i$$

$$=\begin{pmatrix}1 & 0\end{pmatrix}\left\{\begin{pmatrix}s+\lambda+\mu & -(\lambda+\mu)\\ -\mu & s+\lambda+\mu\end{pmatrix}^{-1}\begin{pmatrix}0 & 0\\ \lambda z & 0\end{pmatrix}\right\}^N \begin{pmatrix}s+\lambda+\mu & -(\lambda+\mu)\\ -\lambda z & s+\lambda+\mu\end{pmatrix}^{-1}\begin{pmatrix}\mu\\ \mu\end{pmatrix}.$$

Let a=$s+\lambda+\mu$, the following results after algebraic manipulation:

$$f^*(s,z) = \frac{\sum_{j=1}^{\infty}\sum_{i=N}^{\infty}\binom{N+j-2}{j-1}\left(\frac{\mu}{a}\right)^j\left(\frac{\lambda}{a}\right)^i}{\left\{\left(\frac{\lambda+\mu}{a}\right)^{j+i}+\left(\frac{\lambda+\mu}{a}\right)^{j+i+1}\right\}z^i}.$$

After substituting the value of *a*, comparing the coefficient of z^i, and taking the inverse of the Laplace transform, the two dimensional marginal probability density function for the variables time taken by servers vacations with respect to the number of customers arriving, T_{ξ_N},η_N, is established by:

$$f(t,z)= \sum_{j=1}^{\infty}\binom{N+j-2}{j-1}\lambda^i\mu^j\begin{bmatrix}\left(\lambda+\mu\right)^{j+i}\dfrac{t^{2j+2i-1}}{\left(2j+2i-1\right)!}\\ +\left(\lambda+\mu\right)^{j+i+1}\dfrac{t^{2j+2i}}{\left(2j+2i\right)!}\end{bmatrix}e^{-(\lambda+\mu)t} \qquad (4.2)$$

Theorem 4.3

The two dimensional marginal probability distribution function of random variables T_{ξ_N},η_N is:

$$p\left\{T_{\xi_N}\le t,\eta\ _N=i\right\}=$$

$$F(t,z)=\sum_{j=1}^{\infty}\binom{N+j-2}{j-1}\left(\frac{\lambda}{\lambda+\mu}\right)^i\left(\frac{\mu}{\lambda+\mu}\right)^j$$

$$\left[\sum_{r=0}^{2j+2i-1}\frac{\left[\left(\lambda+\mu\right)t\right]^r}{r!}+\sum_{r=1}^{2j+2i}\frac{\left[\left(\lambda+\mu\right)t\right]^r}{r!}\right]e^{-(\lambda+\mu)t} \qquad (4.3)$$

Proof 4.3

The two dimensional cumulative probability distribution function for two random variables, T_{ξ_N},η_N, is obtained by integrating the two dimension marginal probability density function for the same random variables.

$$p\left\{T_{\xi_N}\le t,\eta\ _N=i\right\}=$$

$$\sum_{j=1}^{\infty}\binom{N+j-2}{j-1}\lambda^i\mu^j \int_0^{\infty}\begin{bmatrix}\left(\lambda+\mu\right)^{j+i}\dfrac{t^{2j+2i-1}}{\left(2j+2i-1\right)!}\\ +\left(\lambda+\mu\right)^{j+i+1}\dfrac{t^{2j+2i}}{\left(2j+2i\right)!}\end{bmatrix}e^{-(\lambda+\mu)t}\,dt$$

After algebraic manipulation and applying the integration by parts the proof is obvious.

Two dimensional marginal distribution functions for ξ_N,η_N

The effect of the number of vacations taken by servers on the number of customer arrivals was studied by controlling the time taken with the number of vacations made by servers. The two dimensional marginal probability transform function (probability generating function), two dimensional probability density function and two dimensional marginal cumulative probability distribution function for the number of server vacations with number of arriving customers, as represented by random variables ξ_N,η_N, are now calculated.

Theorem 5.1

The two dimensional marginal probability generating function (probability transform function) for the two random variables ξ_N,η_N is:

$$f'(u,z) = u\begin{pmatrix}1 & 0\end{pmatrix} \left\{\begin{pmatrix}\lambda+\mu & -(\lambda+\mu) \\ -u\mu & \lambda+\mu\end{pmatrix}^{-1}\begin{pmatrix}0 & 0 \\ \lambda z & 0\end{pmatrix}\right\}^{N} \begin{pmatrix}\lambda+\mu & -(\lambda+\mu) \\ -\lambda z & \lambda+\mu\end{pmatrix}^{-1}\begin{pmatrix}\mu \\ \mu\end{pmatrix}.$$

The two dimensional probability density function for random variables ξ_N, η_N is:

$$f(u,z) = \binom{N+j-2}{j-1} 2\left(\frac{\mu}{\lambda+\mu}\right)^{j}\left(\frac{\lambda}{\lambda+\mu}\right)^{i}.$$

The two dimensional marginal cumulative probability distribution function for the random variables ξ_N, η_N is:

$$p\{\xi_N = j, \eta_N = i\} = \sum_{j=1}^{\infty}\sum_{i=N}^{\infty} 2\binom{N+j-2}{j-1}\left(\frac{\mu}{\lambda+\mu}\right)^{j}\left(\frac{\lambda}{\lambda+\mu}\right)^{i}.$$

Proof 5.1

The two dimensional marginal probability generating function (probability transform function) for the two random variables ξ_N, η_N is obtained from the joint probability transform function of the three random variables as:

$$f^{*}(s,u,z) = u\begin{pmatrix}1 & 0\end{pmatrix} \left\{\begin{pmatrix}s+\lambda+\mu & -(\lambda+\mu) \\ -u\mu & s+\lambda+\mu\end{pmatrix}^{-1}\begin{pmatrix}0 & 0 \\ \lambda z & 0\end{pmatrix}\right\}^{N} \begin{pmatrix}s+\lambda+\mu & -(\lambda+\mu) \\ -\lambda z & s+\lambda+\mu\end{pmatrix}^{-1}\begin{pmatrix}\mu \\ \mu\end{pmatrix}$$

Let s be close to 0^{+} to control the effect of time by the number of vacations taken by the number of servers. The two dimensional marginal probability generating function for random variables ξ_N, η_N is obtained using:

$$f'(u,z) = u\begin{pmatrix}1 & 0\end{pmatrix} \left\{\begin{pmatrix}\lambda+\mu & -(\lambda+\mu) \\ -u\mu & \lambda+\mu\end{pmatrix}^{-1}\begin{pmatrix}0 & 0 \\ \lambda z & 0\end{pmatrix}\right\}^{N} \begin{pmatrix}\lambda+\mu & -(\lambda+\mu) \\ -\lambda z & \lambda+\mu\end{pmatrix}^{-1}\begin{pmatrix}\mu \\ \mu\end{pmatrix} \tag{5.1a}$$

The definition of the z transform is expressed by the following equation (Talpur & Shi, 1994):

$$f(u,z) = \sum_{i=N}^{\infty} p\{\xi_N = j, \eta_N = i\} u^{j} z^{i} = u\begin{pmatrix}1 & 0\end{pmatrix} \left\{\begin{pmatrix}\lambda+\mu & -(\lambda+\mu) \\ -u\mu & \lambda+\mu\end{pmatrix}^{-1}\begin{pmatrix}0 & 0 \\ \lambda z & 0\end{pmatrix}\right\}^{N} \begin{pmatrix}\lambda+\mu & -(\lambda+\mu) \\ -\lambda z & \lambda+\mu\end{pmatrix}^{-1}\begin{pmatrix}\mu \\ \mu\end{pmatrix}$$

Using the same process as in theorem 2.2 results in:

$$f(u,z) = \sum_{j=1}^{\infty}\sum_{i=N}^{\infty}\binom{N+j-2}{j-1} 2\left(\frac{\mu}{\lambda+\mu}\right)^{j}\left(\frac{\lambda}{\lambda+\mu}\right)^{i} u^{j} z^{i},$$

and comparing the coefficients u and z, the following proof is obtained:

$$f(u,z) = \binom{N+j-2}{j-1} 2\left(\frac{\mu}{\lambda+\mu}\right)^{j}\left(\frac{\lambda}{\lambda+\mu}\right)^{i} \tag{5.1b}$$

Two dimensional marginal cumulative probability distribution functions for two random variables, ξ_N and η_N, was obtained by summing their density function and the number of vacations made by servers with the number of arriving units.

Results

As shown in Table 1, the two dimensional marginal probability distributions of random variables involving the crossing time spent for the number of vacations taken by servers (T_{ξ_N}) followed by the number of vacations by the service channels (ξ_N) shows a two stage Erlang distribution for the probability density function for achieving the absorption state. The cumulative probability distribution function is found to be a Gamma distribution.

The two dimensional marginal probability distribution for random variables involving the crossing time spent for the number of vacations taken by servers (T_{ξ_N}) with a reasonable number of arriving customers (η_N) for achieving the absorption state is shown in Table 2. The two variables show a two stage Erlang distribution for the probability density function. The Gamma distribution is satisfied for the cumulative probability distribution function.

The probability density function for two random variables is expressed as a negative binomial distribution. The cumulative probability distribution function also satisfied the negative binomial distribution. As Medhi (1982) expressed, if the parameter λ (intensity function) of a Poisson process is a random variable with Gamma distribution, then the mixed Poisson distribution is Negative binomial.

Table 1: Two Dimensional marginal probability distributions of random variables $T_{\xi_{N,}}\xi_N$

Transform function $f^*(s,u)$	$u(1\quad 0)\left\{\begin{pmatrix} s+\lambda+\mu & -(\lambda+\mu) \\ -u\mu & s+\lambda+\mu \end{pmatrix}^{-1}\begin{pmatrix} 0 & 0 \\ \lambda & 0 \end{pmatrix}\right\}^N$ $\begin{pmatrix} s+\lambda+\mu & -(\lambda+\mu) \\ -z\lambda & s+\lambda+\mu \end{pmatrix}^{-1}\begin{pmatrix} \mu \\ \mu \end{pmatrix}$
Probability density function $f(t,u)=$ $\{T_{\xi_N} \le t, \xi_N = j\}$	$\sum_{i=N}^{\infty}\binom{N+j-2}{j-1}\lambda^i\mu^j$ $\left[(\lambda+\mu)^{j+i-1}\frac{t^{2j+2i-2}}{(2j+2i-2)!}+(\lambda+\mu)^{j+i}\frac{t^{2j+2i-1}}{(2j+2i-1)!}\right]e^{-(\lambda+\mu)t}$
Probability distribution function $\mathrm{p}\{T_{\xi_N} \le t, \xi_N = j\}$	$\sum_{i=N}^{\infty}\binom{N+j-2}{j-1}\left(\frac{\lambda}{\lambda+\mu}\right)^i\left(\frac{\mu}{\lambda+\mu}\right)^j\left[\sum_{r=0}^{2j+2i-1}\frac{[(\lambda+\mu)t]^r}{r!}e^{-(\lambda+\mu)t}+\right.$ $\left.\sum_{r=0}^{2j+2i}\frac{[(\lambda+\mu)t]^r}{r!}e^{-(\lambda+\mu)t}\right]$

Table 2: Two Dimensional marginal probability distributions of random variables $T_{\xi_{N,}}\eta_N$

Transform function $f^*(s,z)$	$(1\quad 0)\left\{\begin{pmatrix} s+\lambda+\mu & -(\lambda+\mu)\\ -\mu & s+\lambda+\mu \end{pmatrix}^{-1}\begin{pmatrix} 0 & 0\\ \lambda z & 0 \end{pmatrix}\right\}^{N}$ $\begin{pmatrix} s+\lambda+\mu & -(\lambda+\mu)\\ -z\lambda & s+\lambda+\mu \end{pmatrix}^{-1}\begin{pmatrix} \mu\\ \mu \end{pmatrix}$
Probability density function $f(t,z)=$ $\{T_{\xi_N} \le t, \eta_N = i\}$	$\sum_{j=1}^{\infty}\binom{N+j-2}{j-1}\lambda^{i}\mu^{j}$ $\left[(\lambda+\mu)^{j+i}\frac{t^{2j+2i-1}}{(2j+2i-1)!}+(\lambda+\mu)^{j+i+1}\frac{t^{2j+2i}}{(2j+2i)!}\right]e^{-(\lambda+\mu)t}$
Probability distribution function $\mathrm{p}\{T_{\xi_N} \le t, \eta_N = i\}$	$\sum_{j=1}^{\infty}\binom{N+j-2}{j-1}\left(\frac{\lambda}{\lambda+\mu}\right)^{i}\left(\frac{\mu}{\lambda+\mu}\right)^{j}$ $\left[\sum_{r=0}^{2j+2i-1}\frac{[(\lambda+\mu)t]^{r}}{r!}e^{-(\lambda+\mu)t}\right.$ $\left.+\sum_{r=0}^{2j+2i}\frac{[(\lambda+\mu)t]^{r}}{r!}e^{-(\lambda+\mu)t}\right]$

Table 3: Two dimensional marginal probability distribution functions of random variables ξ_N, η_N

Transform function $f^*(u,z)$	$(1\quad 0)\left\{\begin{pmatrix} \lambda+\mu & -(\lambda+\mu)\\ -u\mu & \lambda+\mu \end{pmatrix}^{-1}\begin{pmatrix} 0 & 0\\ \lambda z & 0 \end{pmatrix}\right\}^{N}$ $\begin{pmatrix} \lambda+\mu & -(\lambda+\mu)\\ -z\lambda & \lambda+\mu \end{pmatrix}^{-1}\begin{pmatrix} \mu\\ \mu \end{pmatrix}$
Probability density function $f(u,z)$	$\binom{N+j-2}{j-1}2\left(\frac{\mu}{\lambda+\mu}\right)^{j}\left(\frac{\lambda}{\lambda+\mu}\right)^{i}$
Probability distribution function F(u, z) = $\mathrm{p}\{T_{\xi_N} \le t, \eta_N = i\}$	$\sum_{j=1}^{\infty}\sum_{i=N}^{\infty}2\binom{N+j-2}{j-1}\left(\frac{\lambda}{\lambda+\mu}\right)^{i}\left(\frac{\mu}{\lambda+\mu}\right)^{j}$

References

Aki, S. (1992). Waiting time problems for a sequence of discrete random variables. *Annals of the Institute of Statistical Mathematics, 44*, 363-378.

Aki, S., & Hirano, K. (1993). In K. Matisita, et al. (Ed.), Discrete distributions related to succession queue in two-state Markov chain. *Statistical Science and Data Analysis,* 467-474. VSP: Amsterdam.

Aki, S., & Hirano, K. (1999). Sooner and later waiting time problems for urns in Markov dependent bivariate trials. *Annals of the Institute of Statistical Mathematics, 51*, 17-29.

Biggins, J. D., & Cannings, C. (1987). Markov renewal processes, counters and repeated sequence in Markov chain. *Advances in Applied Probability, 19*, 521-545.

Cinlar, E. (1969). Markov renewal theory. *Advances in Applied Probability, 1*, 123-187.

Ebneshahrashoob, M., & Sobel, M. (1990). Sooner and later waiting time problems for Bernoulli trials: Frequency and run quotas. *Statistics and Probability Letters, 9*, 5-11.

Fu, J. C., & Koutras, M. V. (1994). Distribution of runs: A Markov chain approach. *Journal of the American Statistical Association, 89*, 1050-1058.

Godbole, A. P. (1993). Approximate reliabilities of m-consecutive-k-out-of-n: failure systems. *Statistica Sinica, 3*, 321-327

Hogg, R.V., & Craig, A. T. (1995). *Introduction to Mathematical Statistics.* Macmilan Publishing Co., Inc.: New York, NY.

Koutras, M. V., & Papastavridis, S. G. (1993). On the number of runs and related statistics. *Statistica Sinica. 3*, 277-294.

Kroese, D. P. (1992). The difference of two renewal processes level crossing and infimum. *Stochastic Model, 8*, 221-243.

Ling, K. L. (1988). On binomial distributions of order k. *Statistics and Probability Letters, 6*, 247-250.

Medhi, J. (1982). *Stochastic processes.* Wiley Eastern Ltd.: India.

Mohanty, S. G. (1994). Success runs of length k in Markov dependent trials. *Annals of the Institute of Statistical Mathematics, 46*, 777-796.

Pipes, A. L., & Harwil, R. L. (1970). *Applied Mathematics for Engineers and Physicsts, International Student Edition.* McGraw-Hill Book Inc.: Tokyo.

Talpur, G. H., & Shi, D. H. (1994). Some joint distribution of crossing time and renewal numbers related with two difference processes. *Journal of Shanghai University of Science and Technology, 17*, 245-256.

Talpur, G. H., & Iffat, Z. (2007). *The joint distribution of crossing time and renewal numbers related with two-stage Erlang processes*. May 2007. Conference in honor of Prof. M. Masoom Ali: Department of Mathematical Sciences. Ball State University.

Journal of Modern Applied Statistical Methods
November 2008, Vol. 7, No. 2, 446-453

1538 – 9472/08/$95.00

Size-Biased Generalized Negative Binomial Distribution

Khurshid Ahmad Mir
Govt.Degree College Bemina,
Srinagar (J&K) India.

A size biased generalized negative binomial distribution (SBGNBD) is defined and a recurrence relationship for the moments of SBGNBD is established. The Bayes' estimator for a parametric function of one parameter when two other parameters of a known size-biased generalized negative binomial distribution is derived. Prior information on one parameter is given by a beta distribution and the parameters in the prior distribution are assigned by computer using Monte Carlo and R-software.

Key words: Generalized negative binomial distribution, size-biased generalized negative binomial distribution, zero-truncated generalized negative binomial distribution; size biased negative binomial distribution, goodness of fit, Bayes' estimation.

Introduction

Jain and Consul (1971) first defined generalized negative binomial distribution (GNBD), and it was subsequently obtained by Consul and Shenton (1972, 1975) as a particular family of the Lagrangian distribution. The parameter space of the distribution was further modified by Consul and Gupta (1995). The probability function of the GNBD is given by

$$P_1(X=x)=\frac{m}{m+\beta x}\binom{m+\beta x}{\mathrm{x}}\alpha^x(1-\alpha)^{m+\beta x-x};$$
$$x=0,1,2..... \tag{1.1}$$

where

$$0<\alpha<1, m>0 \text{ and } \left|\alpha\beta\right|<1.$$

The probability model (1.1) reduces to the binomial distribution when β = 0, and to the negative binomial distribution when β = 1. It also resembles the Poisson distribution at β=½ because, for this value of β, the mean and variance are approximately equal. Jain and Consul (1971) obtained the first four non-central moments by using a recurrence relation and Shoukri (1980) obtained a recurrence relation among the central moments. The model (1.1) has many important applications in various fields of study and is useful in queuing theory and branching processes. Famoye and Consul (1989) considered a stochastic model for the GNBD and gave some other interesting applications of this model. The moments about the origin of the model (1.1) are given as:

$$\mu_1' = \frac{\mathrm{m}\alpha}{(1-\alpha\beta)} \tag{1.2}$$

$$\mu_2' = \frac{(\mathrm{m}\alpha)^2}{(1-\alpha\beta)^2}+\frac{\mathrm{m}\alpha(1-\alpha)}{(1-\alpha\beta)^3} \tag{1.3}$$

$$\mu_3' = \frac{(m\alpha)^3}{(1-\alpha\beta)^3}+\frac{3(m\alpha)^2(1-\alpha)}{(1-\alpha\beta)^4}$$
$$+\frac{m\alpha(1-\alpha)}{(1-\alpha\beta)^5}\left[1-2\alpha+\alpha\beta(2-\alpha)\right]$$

Khurshid Ahmad Mir is a Senior Assistant Professor in the Department of Statistics. Email: him at mirkhurshid_ahmad@yahoo.co.in or khrshdmir@yahoo.com. Mailing Address: Khaiwan Narwara Srinagar, Kashmir (India) Postal Code: 190002.

(1.4)

$$\mu_4' = \frac{(m\alpha)^4}{(1-\alpha\beta)^4} + \frac{6(m\alpha)^3(1-\alpha)}{(1-\alpha\beta)^5}$$
$$+\frac{(m\alpha)^2(1-\alpha)\left[7-11\alpha+4\alpha\beta(2-\alpha)\right]}{(1-\alpha\beta)^6}$$
$$+\frac{m\alpha(1-\alpha)}{(1-\alpha\beta)^7}+\begin{bmatrix}1-6\alpha+6\alpha^2\\ +2\alpha\beta\left(4-9\alpha+4\alpha^2\right)\\ +\alpha^2\beta^2\left(6-6\alpha+\alpha^2\right)\end{bmatrix}$$

(1.5)

and variance

$$\mu_2 = \frac{m\alpha(1-\alpha)}{(1-\alpha\beta)^3}.$$

(1.6)

Jain and Consul (1971) discussed the method of moments of estimation, and Gupta (1972, 1975) and Hassan (1995) obtained maximum likelihood estimations. Jani (1977), Kumar and Consul (1980), and Consul and Famoye (1989) studied the minimum variance unbiased estimation of GNBD, while Islam and Consul (1986) examined its Bayesian method of estimation. Recently, Consul and Famoye (1980) and Famoye (1997) discussed these methods in brief with respect to the model (1.1). Estimation techniques in the case of GNBD are not simple, all involve computation and can become tedious and time intensive.

The weighted distributions arise when observations generated from a stochastic process are not given an equal chance of being recorded, but instead are recorded according to some weight function. When the weight function depends on the lengths of the units of interest, the resulting distribution is called length biased. More generally, when the sampling mechanism selects units with probability proportional to some measure of the unit size, the resulting distribution is called size-biased. Such distributions arise, for example, in life length studies (see Blumenthal, 1967; Consul, 1989; Gupta 1975, 1976, 1979, 1984; Gupta & Tripathi, 1987, 1992; Schaeffer, 1972).

Size-biased generalized negative binomial distribution (SBGNBD) taking the weights of the probabilities as the variate values, are defined in this study. The moments of size-biased GNBD are also obtained. As far as estimation the parameters of a size-biased generalized negative binomial distribution (SBGNBD) is concerned, no method seems to have evolved to date, thus a Bayes' estimator of size-biased generalized negative binomial distribution is presented. A computer program in R-software has been developed to ease computations while estimating the parameters for data. A goodness of fit test is employed to test the program's improvement over the Bayes' estimator of the zero truncated generalized negative binomial distribution (ZTGNBD) and of the size biased negative binomial distribution (SBNBD).

The Truncated Generalized Negative Binomial Distribution

Jain and Consul's (1997) generalized negative binomial distribution (1.1) can be truncated at x = 0. The probability function of the zero-truncated GNBD is given by:

$$P_2\left(X=x\right)=\frac{\dfrac{m}{m+\beta x}\dbinom{m+\beta x}{x}\alpha^x\left(1-\alpha\right)^{m+\beta x-x}}{\left[1-\left(1-\alpha\right)^m\right]},$$

x =1, 2…..

(2.1)

where $0 < \alpha < 1, m > 0\ and\ |\alpha\beta| \leq 1$.

Bansal and Ganji (1997) obtained the Bayes' estimator of zero-truncated generalized negative binomial distribution (2.1). Famoye and Consul (1993) defined a truncated GNBD using (1.1); they obtained an estimator of its parameters by using different estimation methods.

Methodology

A size-biased generalized negative binomial distribution (SBGNBD) - a particular case of the weighted generalized negative binomial - taking weights as the variate value is defined and moments of SBGNBD are obtained.

Using (1.1) and (1.2), results in the following:

$$\sum_{x=0}^{\infty} x \cdot P_1(X=x) = \frac{m\alpha}{(1-\alpha\beta)}$$

thus,

$$\sum_{x=1}^{\infty} P_3(X=x) = 1$$

represents a probability distribution. This gives the size-biased generalized negative binomial distribution (SBGNBD) as:

$$P_3(X=x)$$

$$= (1-\alpha\beta)\binom{m+\beta x-1}{x-1}\ \alpha^{x-1}\ (1-\alpha)^{m+\beta x-x}\ ;$$

$$x = 1, 2, \ldots,\ \text{where } 0<\alpha<1,\ m>0,\ |\alpha\beta|<1$$

(3.1)

Putting $\beta = 0$ and $\beta = 1$, results in size-biased binomial (SBB) and size-biased negative binomial (SBNB) distributions.

Moments of SBGNBD

The r^th moment, $\mu_r'(s)$, about origin of the size-biased GNBD (3.1) can be defined as:

$$\mu_r'(s) = \sum_{x=1}^{\infty} x^r \cdot P_3(X=x)\ ;\ r = 1, 2, 3, \ldots$$

(3.2)

$\mu_0'(s)$ =1, and for r ≥ 1, and

$$\mu_r'(s) = \frac{1-\alpha\beta}{m\alpha}\sum_{x=0}^{\infty} x^{r+1} P_1(X=x)$$

$$\mu_r'(s) = \frac{1-\alpha\beta}{m\alpha}\mu_{r+1}'$$

(3.3)

where μ_{r+1}' is the (r + 1)th moment about the origin of (1.1). The first three moments of (3.1) about the origin using relations from (1.2) to (1.5) in (3.2) can be obtained by:

$$\mu_1'(s) = \frac{1-\alpha\beta}{m\alpha}\mu_2'$$

$$\mu_1'(s) = \frac{m\alpha}{1-\alpha\beta} + \frac{1-\alpha}{(1-\alpha\beta)^2}$$

(3.4)

which is the mean of (3.1). Similarly, for r = 2 in (3.2) using relation (1.4):

$$\mu_2'(s) =$$

$$\frac{(m\alpha)^2}{(1-\alpha\beta)^2} + \frac{3m\alpha(1-\alpha)}{(1-\alpha\beta)^3} + \frac{(1-\alpha)}{(1-\alpha\beta)^4}\left[1-2\alpha+\alpha\beta(2-\alpha)\right]$$

(3.5)

Using relation (1.5) for r = 3 in (3.2) results in:

$$\mu_3'(s) =$$

$$\frac{(m\alpha)^3}{(1-\alpha\beta)^3} + \frac{6(m\alpha)^2(1-\alpha)}{(1-\alpha\beta)^4} + \frac{m\alpha(1-\alpha)}{(1-\alpha\beta)^5}\left[7-11\alpha-4\alpha\beta(2-\alpha)\right]$$

(3.6)

The variance $\mu_2(s)$ of (3.1) using (3.3) and (3.4) is obtained by:

$$\mu_2(s) = \frac{m\alpha(1-\alpha)}{(1-\alpha\beta)^3} + \frac{\alpha(1-\alpha)}{(1-\alpha\beta)^4}\left[\beta(2-\alpha)-1\right]$$

(3.7)

The higher moments of (3.1) about the origin can also be obtained similarly by using (3.2).

Bayes' Estimation in Size-biased Generalized Negative Binomial Distribution

The likelihood function of SBGNBD (3.1) is:

$$L(\underline{x}\,|\,\alpha,\beta) =$$

$$(1-\alpha\beta)^n \prod_{i=1}^{n}\binom{m+\beta x_i-1}{x_i-1}\alpha^{\sum\limits_{i=}^{n} x_i - n}\ (1-\alpha)^{mn+\beta\sum\limits_{i=1}^{n} x_i - \sum\limits_{i=1}^{n} x_i}$$

$$= K\left(1-\alpha\beta\right)^{n}\alpha^{y-n}\left(1-\alpha\right)^{mn+\beta y-y} \quad (4.1)$$

where

$$y=\sum_{i=1}^{n}x_i \text{ and } K=\prod_{i=1}^{n}\binom{m+\beta x_i-1}{x_i-1}.$$

Because $0<\alpha<1$, it is assumed that prior information about α came from the beta distribution. Thus,

$$f(\alpha)=\frac{\alpha^{a-1}(1-\alpha)^{b-1}}{B(a,b)};\ 0<\alpha<1,\ a>0,\ b>0.$$

(4.2)

Using Bayes' Theorem, the posterior distribution of *α* from (4.1) and (4.2) can be written as:

$$p(\alpha\mid y)=\frac{(1-\alpha\beta)^{n}\alpha^{y+a-n-1}(1-\alpha)^{mn+\beta y-y+b-1}}{\int_{0}^{1}(1-\alpha\beta)^{n}\alpha^{y+a-n-1}(1-\alpha)^{mn+\beta y-y+b-1}d\alpha}.$$

(4.3)

Under square error loss function the Bayes'estimator of parametric function α^z is the posterior mean given as

$$\hat{\alpha}^{z}=\int_{0}^{1}\alpha^{z}\,p(\alpha\mid y)d\alpha$$

$$=\frac{\int_{0}^{1}(1-\alpha\beta)^{n}\ \alpha^{y+a-n+z}(1-\alpha)^{mn+\beta y-y+b-1}}{\int_{0}^{1}(1-\alpha\beta)^{n}\alpha^{y+a-n-1}(1-\alpha)^{mn+\beta y-y+b-1}d\alpha}.$$

(4.4)

where

$$\int_{0}^{1}(1-\alpha\beta)^{n}\ \alpha^{y+a-n+z}(1-\alpha)^{mn+\beta y-y+b-1}d\alpha=$$

$$\frac{\Gamma(y+a-n+z)\ \Gamma(\beta y+mn+b-y)\ {}^{2}F_{1}\left[-n,y+a-n+z,\beta y+mn+a+b-n+1,\beta\right]}{\Gamma(\beta y+mn+a+b-n+z)}$$

(4.5)

and

$$\int_{0}^{1}(1-\alpha\beta)^{n}\ \alpha^{y+a-n-1}(1-\alpha)^{mn+\beta y-y+b-1}d\alpha=$$

$$\frac{\Gamma(y+a-n)\ \Gamma(\beta y+mn+b-y)\ {}^{2}F_{1}\left[-n,y+a-n,\beta y+mn+a+b-n,\beta\right]}{\Gamma(\beta y+mn+a+b-n)}$$

(4.6)

Using relations (4.5) and (4.6) in (3.4), the Bayes' estimator of α^z becomes:

$$\hat{\alpha}^{z}=$$

$$\frac{\Gamma(y+a-n+z)\ \Gamma(\beta\mathrm{y}+\mathrm{mn}+\mathrm{a}+\mathrm{b\text{-}n})\ {}^{2}F_{1}\left[-n,y+a-n+z,\beta y+mn+a+b-n+1,\beta\right]}{\Gamma(y+a-n)\ \Gamma(\beta y+mn+a+b-n+z)\ {}^{2}F_{1}\left[-n,y+a-n,\beta y+mn+a+b-n,\beta\right]}$$

Similarly, the Bayes' estimator of the parametric function $(1-\alpha)^z$ can also be obtained as:

$$(1-\alpha)^{z}=\frac{\int_{0}^{1}(1-\alpha\beta)^{n}\ \alpha^{y+a-n-1}(1-\alpha)^{mn+\beta y-y+b-1+z}}{\int_{0}^{1}(1-\alpha\beta)^{n}\alpha^{y+a-n-1}(1-\alpha)^{mn+\beta y-y+b-1}d\alpha}$$

(4.8)

where

$$\int_{0}^{1}(1-\alpha\beta)^{n}\ \alpha^{y+a-n-1}(1-\alpha)^{mn+\beta y-y+b-1+z}d\alpha=$$

$$\frac{\Gamma(y+a-n)\;\Gamma(\beta y+mn+b-y+z)\;{}^{2}F_{1}\left[-n,y+a-n,\beta y+mn+a+b-n+z,\beta\right]}{\Gamma(\beta y+mn+a+b-n)}. \tag{4.9}$$

Using the values from (4.9) and (4.6) in (4.8), the Bayes' estimator of the parametric function $(1-\alpha)^z$ can be obtained as

$$(1-\alpha)^z = \frac{\Gamma(\beta\mathrm{y}+\mathrm{mn}+\mathrm{b}\text{-}\mathrm{y}+\mathrm{z})\;\Gamma(\beta\mathrm{y}+\mathrm{mn}+a+\mathrm{b}\text{-}\mathrm{n})\;{}^{2}F_{1}\left[-n,y+a-n,\beta y+mn+a+b-n+z,\beta\right]}{\Gamma(\beta y+mn+a+b-n+z)\;\Gamma(\beta y+mn+b-y)\;{}^{2}F_{1}\left[-n,y+a-n,\beta y+mn+a+b-n,\beta\right]} \tag{4.10}$$

The Bayes' estimator for some parametric functions $\phi(\alpha)$ and for particular models of SBGNBD are shown in Tables 4.1 and 4.2.

Conclusion

A computer program in R-Software was developed to ease computations while estimating the parameters for data. The expected frequencies and Chi-square obtained are shown in tables 5.1, 5.2 and 5.3. Assuming that the parameter α is unknown and that it has a beta distribution with parameters *a* and *b*, the Bayes' relative frequencies are estimated by using the estimator of (2.1) and (3.1). Since no other information is provided about the values of *a* and *b,* except that they are both positive and real, a range of values from 1 to 50 were considered for *a* and *b*, and the values of (2.1) and (3.1) were computed. Three sets of simulated values were obtained with the help of R-software: one each for the parameter combination (α=0.5, β=0.3, a=b=1), (α=0.6, β=0.5, a=b=2) and (α=0.6, β=0.7, a=b=3). We noted that the estimated Bayes' frequencies were quite close to the simulated sample frequencies when *a* and *b* were equal and that the variation in the Bayes' frequencies was very little as the equal values of *a* and *b* increased. The graph also reveals that the simulated frequencies and the estimated Bayes' frequencies are very close to each other for almost all values of X.

References

Bansal, A. K., & Ganji, M. (1997). Bayes' estimation in a decapitated generalized negative binomial distribution and some of its applications. *Journal of Information and optimization* Sciences, *18*(1), 189-198.

Blumenthal, S. (1967). Proportional sampling in life length studies. *Technometrics*, *9*, 205-218.

Consul, P. C., & Famoye, F. (1989). Minimum variance unbiased estimation for the Lagrange power series distributions. *Statistics*, *20*, 401-415.

Consul, P. C. (1975). Some new characterization of discrete Lagrangian distributions. In *Statistical distributions in scientific work, characterizations and applications*, G. P. Patil, Kotz, S.,& Ord, J. K. *(Eds.)*, 279-290. Dordrecht: Reidel.

Consul, P. C., & Famoye, F. (1995). On the generalized negative binomial distribution. *Communication in Statistics, Theory and Methods, 24*(2), 495-472.

Consul, P. C. & Gupta, R. C. (1980). The generalized negative binomial distribution and its characterization by zero regression. *SIAM Journal of Applied Mathematics*, *39*(2), 231-237.

Consul, P. C., & Shenton, L. R. (1972). Use of Lagrange expansion for generating discrete generalized probability distribution. *SIAM Journal of Applied Mathematics, 23*(2), 239-248.

Consul, P. C., & Shenton, L. R. (1975). On the probabilistic structure and properties of discrete Lagrangian distributions. In *Statistical distributions in scientific work, characterizations and applications*, G. P. Patil, Kotz, S.,& Ord, J. K. *(Eds.)*, 41-48. Dordrecht: Reidel.

Famoye, F. (1997). Parameter estimation of generalized negative binomial distribution, *Communications in Statistics-Simulation and Computation*, *26*, 269-279.

Famoye, F., & Consul, P. C. (1989). A stochastic urn model for the generalized negative binomial distribution. *Statistics*, *20*, 607-613.

Famoye, F., & Consul, P. C. (1993). The truncated generalized negative binomial distribution, *Journal of Applied Statistical Science*, *1*(2), 141-157.

Gupta, R. C. (1974). Modified power series distribution and some of its applications. *Sankhya: The Indian Journal of Statistics*, *36B*(3), 288-298.

Gupta, R. C. (1975). Some characterizations of discrete distributions. *Communication in Statistics, Theory and Methods*, *5*(1), 45-48.

Gupta, R. C. (1976). Some characterizations of distributions by properties of their forward and backward recurrence times in a renewal process. *Scandinavian Journal of Statistics*, *3*, 215-216.

Gupta, R. C. (1979). Waiting time paradox and size-biased sampling. *Communication in Statistics, Theory and* Methods, *8*, 601-607.

Gupta, R. C. (1984). Some characterization of renewal densities with emphasis in reliability. *Mathematische Operations-forschung und Statistics*, *15*, 571-579.

Gupta, R. C., & Tripathi, R. C. (1987). A comparison between the ordinary and the length biased modified power series distributions with applications. *Communication in Statistics, Theory and Methods*, *16*(4), 1195-1206.

Gupta, R. C., & Tripathi, R. C. (1992). Statistical inference based on the length-biased data for modified power series distributions. *Communication in Statistics, Theory and Methods,* 21(2), 519-537.

Hassan, A (1995). Problems of estimation in Lagrangian probability distribution. Ph.D Thesis, *Patna University, Patna.*Islam, M. N., & Consul, P. C. (1986). Bayesian estimation in generalized negative binomial distribution, *Biometrics*, *28*, 250-256.

Jain, G. C., & Consul, P. C. (1971). A generalized negative binomial distribution. *SIAM Journal of Applied Mathematics*, *21*(4), 501-513.

Jani, P. N. (1977). Minimum variance unbiased estimate for some left-truncated modified power series distributions, *Sankhya, B3*(39), 258-278.

Kumar, A., & Consul, P. C. (1980). Minimum variance unbiased estimation for modified power series distribution. *Communication in Statistics, Theory and Methods, 9*(12), 1261-1275.

Meegama, S. A. (1980). Socio-economic determinants of infant and child mortality in Sri Lanka. *An analysis of post war experience international statistical institute (World Fertility Survey) Netherlands.*

Scheaffer, R. L. (1972). Size biased sampling. *Technometrics, 14*, 635-644.

Shoukri, M. M. (1980). Estimation of generalized distributions. Unpublished Ph.D. thesis, *University of Calgary, Calgary, Canada.*

Singh, S. N., & Yadav, R. C. (1971). Trends in rural out-migration at household level. *Rural Demography*, *8*, 53-61.

Table 4.1: Bayes' Estimators of SBGNBD

Parametric Function $\phi(\alpha)$	Bayes' Estimator of SBGNBD
α	$\frac{(y+a-n)^{2}F_{1}[-n, y+a-n+1, b+mn+\beta y-n+a+1, \beta]}{(\beta y+mn+a+b-n)^{2}F_{1}[-n, y+a-n, b+mn+\beta y-n+a, \beta]}$
$(1-\alpha)$	$\frac{(\beta \mathrm{y}+\mathrm{mn}+\mathrm{b\text{-}y})^{2}F_{1}[-n, y+a-n, \beta y+mn+a+b-n+1, \beta]}{(\beta y+mn+a+b-n)^{2}F_{1}[-n, y+a-n, \beta y+mn+a+b-n, \beta]}$

Table 4.2: Bayes' $\hat{\alpha}$ Estimators

β	Distribution	Bayes' Estimator $\hat{\alpha}$
1	SBNBD	$\frac{y+a-n}{y+mn+a+b}$
0	SBBD	$\frac{y+a-n}{mn+a+b-n}$

Table 5.1: Number of mothers (f_x) in Sri Lanka having at least one neonatal death according to number of neonatal deaths (x) Meegama (1980) (a=b=2, m=5, β=0.3)

x	f_x	Expected Frequency		
		BSBNBD	BZTGNBD	BSBGNBD
1	567	545.25	549.22	547.45
2	135	154.67	153.03	150.47
3	28	27.31	29.65	29.41
4	11	16.61	12.69	15.65
5	5	2.16	1.41	3.02
Total	746	746	746	746
Estimates $\hat{\alpha}$		0.48	0.49	0.51
χ^2		3.7953	3.0477	2.738

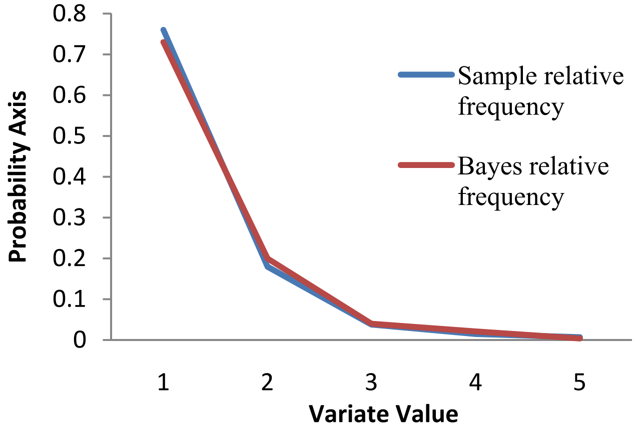
Table 5.2: Number of workers (f_x) having at least one accident according to number of accidents (x) (a=b=2, m=7, β=0.5)

x	f_x	Expected Frequency		
		BSBNBD	ZTGNBD	BSBGNBD
1	2039	2033.32	2031.45	2033.45
2	312	325.33	322.78	320.15
3	35	29.28	32.98	33.26
4	3	1.95	2.56	2.89
5	1	0.12	0.23	0.25
Total	2,390	2,390	2,390	2,390
Estimates $\hat{\alpha}$		0.465	0.493	0.503
χ^2		2.428	0.68	0.4077

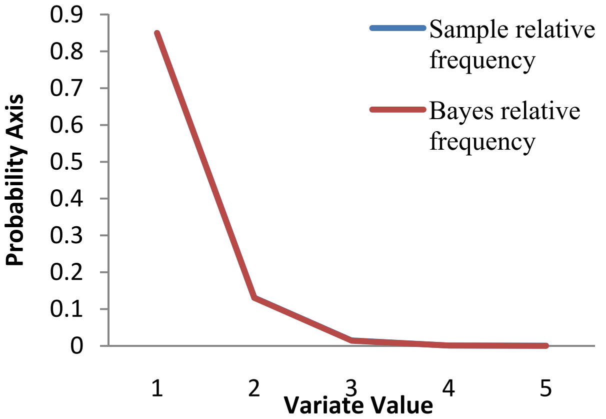
Table 5.3: Number of households (f_x) having at least one migrant according to number of migrants (x) Singh and Yadav (1980) (a=b=2, m=9, β=0.7)

x	f_x	Expected Frequency		
		BSBNBD	BZTGNBD	BSBGNBD
1	375	370.87	368.37	371.81
2	143	156.29	155.79	151.49
3	49	48.42	49.12	50.21
4	17	11.44	13.24	12.51
5	2	2.40	3.01	2.89
6	2	0.47	0.33	0.73
7	1	0.09	0.11	0.30
8	1	0.02	0.03	0.06
Total	590	590	590	590
Estimates $\hat{\alpha}$		0.475	0.489	0.493
χ^2		6.9458	4.06227	3.16908

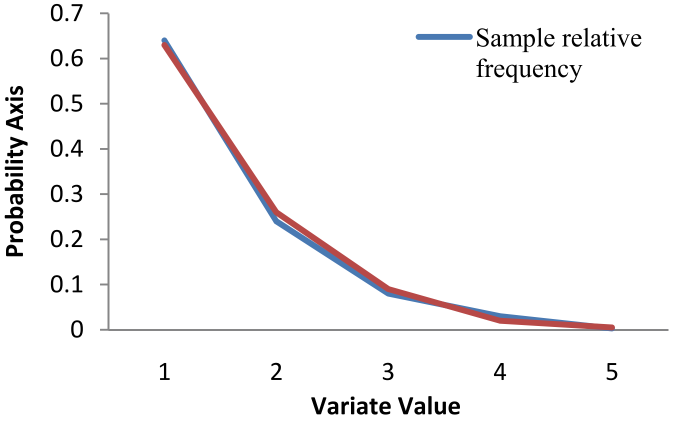
Graph 1: Sample Relative Frequency and Bayes' Relative Frequency for a=b=2, m=5, β=0.3



Graph 2: Sample Relative Frequency and Bayes' Relative Frequency for a=b=2, m=7, β=0.5



Graph 3: Sample Relative Frequency and Bayes' Relative Frequency for a=b=2, m=9, β=0.7



Journal of Modern Applied Statistical Methods
November, 2008, Vol. 7, No. 2, 454-466

1538 – 9472/08/$95.00

Non-Parametric Quantile Selection for Extreme Distributions

Wan Zawiah Wan Zin
Universiti Kebangsaan Malaysia

Abdul Aziz Jemain
Universiti Kebangsaan Malaysia

The objective is to select the best non-parametric quantile estimation method for extreme distributions. This serves as a starting point for further research in quantile application such as in parameter estimation using LQ-moments method. Thirteen methods of non-parametric quantile estimation were applied on six types of extreme distributions and their efficiencies compared. Monte Carlo methods were used to generate the results, which showed that the method of Weighted Kernel estimator of Type 1 was more efficient than the other methods in many cases.

Keywords: Order statistics, sample quantiles, kernel quantile estimators, weighted kernel quantile estimators, HD quantile, weighted HD quantile, LQ-moments, IMSE.

Introduction

In model fitting, one of the key steps is finding the accurate estimates of parameters based on the data in-hand. Several well-known methods include the maximum-likelihood method (ML), method of moments (MOM) and Probability Weighted Moments (PWM). An extension of PWM, termed L-moments, was introduced by Sillitto (1951) for increased accuracy and ease of use of PWM-based analysis.

Mudholkar & Hutson (1998) introduced LQ-moments, an extension of L-moments that was found to be more robust. LQ-moments are constructed using a series of robust linear location measures in place of expectations of order statistics in the L-moments. The r-th LQ-moment, ξ_r of X is defined as:

$$\xi_r = r^{-1}\sum(-1)^k \binom{r-1}{k} \tau_{p,\alpha}(X_{r-k:r}), \quad r = 1,2,.... \tag{1}$$

where $0 \le \alpha \le 1/2$, $0 \le p \le 1/2$, and

$$\begin{aligned}
&\tau_{p,\alpha}(X_{r-k:r}) \\
&= pQ_{X_{r-k:r}}(\alpha) + (1-2p)Q_{X_{r-k:r}}(1/2) \\
&+ pQ_{X_{r-k:r}}(1-\alpha) \\
&= pQ_X\left(B^{-1}_{r-k:r}(\alpha)\right) + (1-2p)Q_X\left(B^{-1}_{r-k:r}\left(\frac{1}{2}\right)\right) \\
&+ pQ_X\left(B^{-1}_{r-k:r}(1-\alpha)\right)
\end{aligned} \tag{2}$$

is the linear combination of symmetric quantiles of the order statistics $X_{r-k:r}$ with $Q_X(\cdot) = F_X^{-1}(\cdot)$ as the quantile function of the random variable X, and $B^{-1}_{r-k:r}(\alpha)$ denotes the corresponding α-th quantile of a beta random variable with parameters $r-k$ and $k+1$. From (2) it can be concluded that proper selection of quantile estimators is crucial to obtain the most accurate parameter estimation based on LQ-moments. As there are many non-parametric quantile estimation methods available, selection is based on statistical ground to propose the most efficient method in many cases.

Wan Zawiah is a Lecturer at the School of Mathematical Sciences. Her research interest is in extreme value analysis and extreme distribution analysis with application in rainfall data. Email: w_zawiah@ukm.my. Abdul Aziz Jemain is a Statistics Professor at the School of Mathematical Sciences. His research interests cover the area of multivariate analysis, demography and multi-criteria decision making. Email: azizj@ukm.my.

Methodology

The quantile function estimators

Let $X_1, X_2, \ldots.., X_n$ be independent and identically distributed random variables with common continuous distribution function (cdf) $F(x), x \in \Re$. Let $X_{1:n} \le X_{2:n} \le \ldots \le X_{n:n}$ denote the corresponding order statistics. The population quantile function, $Q(u)$ of a distribution is defined as:

$$Q(u) = \inf\{x : F(x) \ge u\}, \qquad 0 < u < 1. \quad (3)$$

A traditional estimator of Q(u) is the *u*-th sample quantile given by

$$SQ_u = X_{([nu]+1):n} \quad (4)$$

where $[nu]$ denotes the integral part of *nu* (David, 2003). However, this estimator suffers a drawback in efficiency, caused by the variability of individual order statistics (Huang, 2001). In their article on LQ-moments, Mudholkar & Hutson (1998) employed the linear interpolation-based quantile (LIQ) estimator, defined as,

$$\hat{Q}_X(u) = (1-\varepsilon)X_{[n'u]:n} + \varepsilon X_{[n'u]+1:n}, \quad (5)$$

where $\varepsilon = n'u - [n'u]$ and $n' = n+1$. This is the simplest estimator, and is available in most statistical software packages. It was used as the base for efficiency study in this research.

To overcome the drawback in efficiency of (4), many authors use *L* statistics to reduce the variability. A popular class of kernel quantile estimators has been applied for improving the efficiency of sample quantiles, using an appropriate weight function to average over the order statistics (Sheather & Marron, 1990). Parzen (1979) provided the formula

$$KQ_u = \sum_{i=1}^{n} \left(\int_{(i-1)/n}^{i/n} K_h(t-u)dt \right) X_{i:n}, \quad (6)$$

where *K* is a density function symmetric about 0, $h \to 0$ as $n \to \infty$ and $K_h(\bullet) = \left(\frac{1}{h}\right) K\left(\frac{\bullet}{h}\right)$.

Using the classical empirical distribution function S_n, given by

$$S_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \le x), x \in \Re, \quad (7)$$

where I_A is the indicator function of set *A*, the following are various approximation forms of KQ_u which are often used for practical reasons:

$$KQ_{u.1} = \sum_{i=1}^{n} \left(n^{-1} K_h \left(\frac{i}{n} - u \right) \right) X_{i:n}, \quad (8)$$

$$KQ_{u.2} = \sum_{i=1}^{n} \left(n^{-1} K_h \left(\frac{i - \frac{1}{2}}{n} - u \right) \right) X_{i:n}, \quad (9)$$

$$KQ_{u.3} = \sum_{i=1}^{n} \left(n^{-1} K_h \left(\frac{i}{n+1} - u \right) \right) X_{i:n}, \quad (10)$$

$$KQ_{u.4} = \frac{\sum_{i=1}^{n} \left(K_h \left(\frac{i - \frac{1}{2}}{n} - u \right) \right) X_{i:n}}{\sum_{i=1}^{n} \left(K_h \left(\frac{i - \frac{1}{2}}{n} - u \right) \right)} \quad (11)$$

Huang & Brill (1995, 1996) introduced a level crossing empirical distribution function

$$F_n(x) = \sum_{i=1}^{n} I(X_{i:n} \le x) w_{i,n}, \quad (12)$$

where the data point weights are

$$w_{i,n} = \begin{cases} \frac{1}{2}\left(1 - \frac{n-2}{\sqrt{n(n-1)}}\right), & i = 1, n \\ \frac{1}{\sqrt{n(n-1)}}, & i = 2,3,\ldots,n-1 \end{cases} \quad (13)$$

From (12) and (13), they obtained the following level crossing *u*-th sample quantile to estimate *Q(u),* namely,

$$SQ_{u(lc)} = X_{([b]+2)}, \quad (14)$$

where

$$b = \sqrt{n(n-1)}\left(u - \frac{1}{2}\left(1 - \frac{n-2}{\sqrt{n(n-1)}} \right) \right) \quad (15)$$

Huang & Brill (1999) then introduced the level crossing *u*-th sample kernel quantile given by,

$$WKQ_{u(lc)} = \sum_{i=1}^{n} \left(\int_{q_{i-1,n}}^{q_{i,n}} K_h(t-u)dt \right) X_{i:n}, \quad (16)$$

where $q_{i,n} = \sum_{j=1}^{i} w_{j,n}$ and $w_{i,n}$ is given in (13).

The approximation forms of $WKQ_{u(lc)}$ corresponding to (8)-(11) are as below:

$$WKQ_{u.1(lc)} = \sum_{i=1}^{n} \left(n^{-1} K_h \left(\sum_{j=1}^{i} w_{j,n} - u \right) \right) X_{i:n}, \quad (17)$$

$$WKQ_{u.2(lc)} = \sum_{i=1}^{n} \left(n^{-1} K_h \left(\sum_{j=1}^{i-1} w_{j,n} + \frac{1}{2} w_{i,n} - u \right) \right) X_{i:n}, \quad (18)$$

$$WKQ_{u.3(lc)} = \sum_{i=1}^{n} \left(n^{-1} K_h \left(\sum_{j=1}^{i} w_{j,n} \frac{n}{n+1} - u \right) \right) X_{i:n}, \quad (19)$$

$$WKQ_{u.4(lc)} = \frac{\sum_{i=1}^{n} \left(K_h \left(\sum_{j=1}^{i-1} w_{j,n} + \frac{1}{2} w_{i,n} - u \right) \right) X_{i:n}}{\sum_{i=1}^{n} \left(K_h \left(\sum_{j=1}^{i-1} w_{j,n} + \frac{1}{2} w_{i,n} - u \right) \right)}. \quad (20)$$

In the study, Huang & Brill investigated the relative efficiency of the *u*-th sample level crossing quantile, $SQ_{u(lc)}$ in (14) relative to the *u*-th sample quantile SQ_u in (4) and the relative efficiency of the level crossing quantile estimator $KQ_{u(lc)}$ in (16) relative to the ordinary kernel quantile estimator KQ_u in (6). From both theoretical and computational points of view, they showed that the proposed level crossing estimations were more efficient in many cases, especially for the tails of the distribution and for small sample sizes. Their simulation used the exponential and three types of generalized lambda distribution with small sample sizes (n=10 and n=20).

The selection of kernel or bandwidth of the kernel estimators has always been a sensitive problem. To overcome this, Harrell & Davis (1982) proposed an *L*-quantile estimator of *Q(u),* defined by,

$$HD_u = \sum_{i=1}^{n} \left(\int_{\frac{(i-1)}{n}}^{\frac{i}{n}} \frac{1}{\beta(a,b)} u^{a-1}(1-u)^{b-1} du \right) X_{i:n}, \quad (21)$$

where $a = (n+1)u$, $b=(n+1)v$, $v=1-u$ and $\beta(a,b)$ is the beta function with parameters *a* and *b*.

Huang (2001) proposed a level-crossing HD quantile estimator based on (12) and (21) as follows:

$$WHD_u = \sum_{i=1}^{n} \left(\int_{q_{i-1,n}}^{q_{i,n}} \frac{1}{\beta\{a,b\}} u^{a-1}(1-u)^{b-1} dy \right) X_{i:n}, \quad (22)$$

where $q_{i,n} = \sum_{j=1}^{i} w_{j,n}$, $q_{0:n} = 0$ and $w_{i,n}$ is given in (13).

Similar to previous research, Huang investigated the relative efficiency of the level crossing quantile estimator $HD_{u(lc)}$ in (22) relative to the ordinary quantile estimator HD_u in (21). From both theoretical and computational points of view, the result proved that the proposed level crossing estimations are more efficient in many cases, especially for the tails of the distribution and for small sample sizes. In their simulation, the exponential and three types of generalized lambda distribution with small sample sizes (n=10 and 20) were used.

Thirteen quantile estimation methods are used: (4), (5), (8) - (11), (14), (17) - (20), (21) and (22). An efficiency study is conducted based on integrated mean square error (IMSE) to determine the most efficient quantile estimation methods for several extreme values distributions. LIQ is used as the base because it is the simplest, is easily available in most statistical packages and is most often used quantile estimation method. The relative efficiency results are compared; the method with the lowest IMSE relative efficiency was considered the best and was recommended.

Extreme values distributions

In this research, six common extreme-values distributions were investigated, namely, the Generalized Extreme Value (GEV), Generalized Pareto Distribution (GPD), Generalized Logistic Distribution (GLD), the three-parameter Lognormal (LN3) and Pearson (PE3) distributions and the five-parameter extreme events such as extreme rainfall and flood.

Table 1 provides the list of the extreme value distributions, their corresponding quantile functions and the associated parameters to be tested. The parameters are ε, the position parameter, α, the scale parameter and κ is the Wakeby distribution (WAK5). These six

Table 1: Extreme Value Distributions

<table>
<tr><th rowspan="2">Distribution</th><th rowspan="2">Quantile Function, Q(u)</th><th colspan="5">Parameters</th></tr>
<tr><th>ε</th><th>α</th><th colspan="3">κ</th></tr>
<tr><td>1. Generalized Extreme Value (GEV)</td><td>$\varepsilon + \frac{\alpha}{\kappa}\left[1-\left(-\ln u\right)^{\kappa}\right]$</td><td>0</td><td>1</td><td colspan="3">-0.3, -0.2, -0.1, 0.1, 0.2, 0.3</td></tr>
<tr><td>2. Generalized Pareto Distribution (GPD)</td><td>$\varepsilon + \frac{\alpha}{\kappa}\left[1-\left(1-u\right)^{\kappa}\right]$</td><td>0</td><td>1</td><td colspan="3">-0.3, -0.2, -0.1, 0.1, 0.2, 0.3</td></tr>
<tr><td>3. Generalized Logistic Distribution (GLD)</td><td>$\varepsilon + \frac{\alpha}{\kappa}\left[1-\left\{\left(\frac{1-u}{u}\right)\right\}^{\kappa}\right]$</td><td>0</td><td>1</td><td colspan="3">-0.1, -0.2, -0.3, -0.4, -0.5, -0.6</td></tr>
<tr><td>4. The three-parameter Lognormal distribution (LN3)</td><td>$\varepsilon + \frac{\alpha}{\kappa}\left[1-e^{-\kappa Z}\right]$</td><td>0</td><td>1</td><td colspan="3">0.2(0.2)1.2</td></tr>
<tr><td>5. The three-parameter Pearson distribution (PE3)</td><td>$\frac{2}{\gamma}\left(1+\frac{\gamma z_u}{6}-\frac{\gamma^2}{36}\right)^3-\frac{2}{\gamma}$
(by Wilson-Hilferty transformation and Z_u is the u-th quantile of the standard normal distribution)</td><td>0</td><td>1</td><td colspan="3">1, 2, 3, 4, 6, 8</td></tr>
<tr><td rowspan="7">6. The five-parameter Wakeby distribution (WAK5)</td><td rowspan="7">$\varepsilon + \alpha\left[1-(1-u)^{\beta}\right]$
$-\gamma\left[1-(1-u)^{-\delta}\right]$</td><td rowspan="7">0</td><td rowspan="7">1</td><td>β</td><td>γ</td><td>δ</td></tr>
<tr><td>16</td><td>4</td><td>0.2</td></tr>
<tr><td>7.5</td><td>5</td><td>0.12</td></tr>
<tr><td>1</td><td>5</td><td>0.12</td></tr>
<tr><td>16</td><td>10</td><td>0.04</td></tr>
<tr><td>1</td><td>10</td><td>0.04</td></tr>
<tr><td>2.5</td><td>10</td><td>0.02</td></tr>
</table>

distributions are commonly applied in regional frequency analysis to model many situations of shape parameter unless stated otherwise. The distributions are studied at various shape parameters, κ while fixing the position, ε and scale parameters, α at 0 and 1 respectively, except for Wakeby distribution. The parameters selected were based on previous studies (e.g. for Wakeby) the parameters were proposed by Landwher, et al. (1980). Ani & Aziz (2007) studied and compared the efficiency of (5), (17) and (22) quantile estimators based on this distribution. They performed simulation on GEV based on LQ-moments and the results showed that $WKQ_{u.1(lc)}$ (17) was the most efficient quantile estimator.

Simulation Study

Several Monte Carlo simulation experiments were conducted to determine the best quantile estimators corresponding to different extreme values distributions. The data with small sample sizes, *n*=10(5)30 were generated from respective distribution quantile functions at various values of *u* = 0.01, 0.25, 0.33, 0.50, 0.66, 0.75, 0.90 and replicated (*m*) 5,000 times each.

For the kernel and weighted kernel quantile estimators, the Gaussian Kernel was used $K(u)=(2\pi)^{-\frac{1}{2}}\exp\left(-\frac{1}{2}u^2\right)$ with an optimal bandwidth $h_{opt}=\left(\frac{uv}{n}\right)^{\frac{1}{2}}$ where *v=1-u*, as proposed by Sheather & Marron (1990).

The expected values obtained from the quantile estimators, $\hat{Q}_i(u)$ were compared with the distribution actual (population) *u*-th quantile value, $Q(u)$, that is bias

$$\text{Bias} = \frac{1}{m}\sum_{i=1}^{m}\left(\hat{Q}_i(u)-Q(u)\right).$$

From this value the mean square value was calculated

$$\text{MSE} = \frac{1}{m}\sum_{i=1}^{m}\left(\hat{Q}_i(u)-Q(u)\right)^2 \quad (4.2)$$

along with the integrated mean square errors (IMSE), which is defined as the sum of Mean Square Error across all defined *u* values. The IMSE from all other methods was divided by the IMSE from LIQ to gain the relative efficiency. The estimator which gave the lowest relative IMSE was selected as the best estimator.

The computational results comparing various quantile estimation methods for various distributions are shown in Tables 2-7 for the six extreme distributions respectively. Note that bold font indicates the smallest IMSE value; the most efficient at each respective group.

Results

Table 2 shows the relative efficiency values for six types of Generalized Extreme Value (GEV) distribution. The selection of best quantile estimation method changes when the shape parameter changes from negative to positive.

Table 2 also shows that when the shape parameter is negative, as in GEV types 1,2 and 3, the method suggested was the Weighted Kernel Quantile estimator of Type 1, $WKQ_{u.1(lc)}$, as in (17). However, when the shape parameter is positive, as in GEV types 4, 5 and 6, the most efficient method was the Kernel Quantile estimator Type 4 as in (11) for GEV types 5 and 6. The result is similar in the case of *n*=10 for GEV type 4, but for this type, the more efficient estimator was the Weighted Kernel Quantile estimator Type 4, $WKQ_{u.4(lc)}$, as in (20) followed by the Kernel Quantile estimator Type 4. Hence, we suggest that – in the case of GEV distribution – when analyzing data which is lower-bounded (κ<0), as in most hydrological data, the best estimator would be Weighted Kernel Quantile estimator Type 1, $WKQ_{u.1(lc)}$, and for data that is upper-bounded (κ>0), the Kernel Quantile estimator Type 4, $KQ_{u.4}$, would be the best choice.

The IMSE relative efficiency for six types of Generalized Pareto distribution (GPD) is shown in Table 3. Similar to the GEV case, the selection of best quantile estimation method changes when the shape parameter changes from negative to positive. From Table 3, in almost all cases, the best estimator was the Weighted Kernel Quantile estimator Type 1, $WKQ_{u.1(lc)}$,

Table 2: Generalized Extreme Value (GEV) Distribution

Parameters: Position, ε = 0; Scale, α = 1; Shape, κ = -0.3 (GEV1)

n	SQP1	SQP2	KQ1	KQ2	KQ3	KQ4	WKQ1	WKQ2	WKQ3	WKQ4	HDQ	WHDQ
10	1.430	0.764	0.378	0.487	0.546	0.635	**0.373**	0.481	0.534	0.466	0.838	0.703
15	1.183	1.176	0.338	0.453	0.529	0.588	**0.333**	0.446	0.528	0.431	0.894	0.695
20	1.102	**0.158**	0.171	0.245	0.294	0.316	0.168	0.236	0.305	0.229	0.530	0.389
25	0.200	0.200	0.186	0.273	0.354	0.346	**0.182**	0.257	0.367	0.252	0.602	0.418
30	0.422	0.361	0.279	0.395	0.526	0.485	**0.270**	0.366	0.547	0.363	0.770	0.529

Parameters: Position, ε = 0; Scale, α = 1; Shape, κ = -0.2 (GEV2)

n	SQP1	SQP2	KQ1	KQ2	KQ3	KQ4	WKQ1	WKQ2	WKQ3	WKQ4	HDQ	WHDQ
10	1.286	0.810	0.475	0.559	0.667	0.658	**0.457**	0.538	0.635	0.524	0.811	0.705
15	1.133	1.121	0.455	0.542	0.653	0.636	**0.437**	0.524	0.628	0.508	0.885	0.714
20	1.100	0.289	0.262	0.320	0.379	0.384	**0.255**	0.312	0.377	0.303	0.564	0.439
25	0.377	0.377	0.308	0.375	0.455	0.444	**0.299**	0.362	0.460	0.355	0.647	0.489
30	0.586	0.514	0.366	0.448	0.550	0.527	**0.355**	0.426	0.564	0.421	0.743	0.558

Parameters: Position, ε = 0; Scale, α = 1; Shape, κ = -0.1 (GEV3)

n	SQP1	SQP2	KQ1	KQ2	KQ3	KQ4	WKQ1	WKQ2	WKQ3	WKQ4	HDQ	WHDQ
10	1.055	0.700	0.522	0.570	0.741	0.588	**0.486**	0.528	0.688	0.516	0.686	0.616
15	1.143	1.128	0.595	0.650	0.799	0.692	**0.562**	0.621	0.737	0.599	0.874	0.746
20	1.101	0.465	0.394	0.432	0.515	0.469	**0.377**	0.418	0.486	0.404	0.608	0.504
25	0.536	0.536	0.434	0.478	0.562	0.525	**0.418**	0.463	0.545	0.451	0.678	0.553
30	0.720	0.643	0.493	0.541	0.633	0.595	**0.478**	0.524	0.625	0.514	0.753	0.613

Parameters: Position, ε = 0; Scale, α = 1; Shape, κ = 0.1 (GEV4)

n	SQP1	SQP2	KQ1	KQ2	KQ3	KQ4	WKQ1	WKQ2	WKQ3	WKQ4	HDQ	WHDQ
10	0.797	0.691	0.659	0.645	0.889	**0.560**	0.590	0.587	0.793	**0.560**	0.584	0.567
15	1.067	1.051	0.879	0.865	1.112	0.767	0.804	0.820	0.958	**0.758**	0.820	0.778
20	1.074	0.806	0.692	0.686	0.833	0.631	0.645	0.664	0.723	**0.615**	0.688	0.640
25	0.836	0.836	0.724	0.716	0.825	0.681	0.690	0.707	0.736	**0.663**	0.740	0.688
30	0.915	0.876	0.724	0.718	0.805	0.705	0.699	0.715	0.742	**0.679**	0.770	0.711

Parameters: Position, ε = 0; Scale, α = 1; Shape, κ =0. 2 (GEV5)

n	SQP1	SQP2	KQ1	KQ2	KQ3	KQ4	WKQ1	WKQ2	WKQ3	WKQ4	HDQ	WHDQ
10	0.734	0.690	0.710	0.666	0.921	**0.555**	0.628	0.615	0.812	0.575	0.562	0.558
15	1.033	1.022	0.957	0.915	1.182	**0.779**	0.867	0.886	1.005	0.798	0.800	0.781
20	1.041	0.931	0.815	0.785	0.954	**0.694**	0.757	0.780	0.814	0.703	0.721	0.696
25	0.924	0.924	0.916	0.794	0.862	**0.734**	0.790	0.810	0.812	0.741	0.764	0.736
30	0.975	0.949	0.821	0.795	0.888	**0.749**	0.790	0.804	0.796	0.750	0.782	0.751

Parameters: Position, ε = 0; Scale, α = 1; Shape, κ = 0.3 (GEV6)

n	SQP1	SQP2	KQ1	KQ2	KQ3	KQ4	WKQ1	WKQ2	WKQ3	WKQ4	HDQ	WHDQ
10	0.702	0.704	0.757	0.682	0.938	**0.564**	0.664	0.649	0.821	0.595	0.565	0.566
15	1.004	1.006	1.005	0.930	1.198	**0.779**	0.905	0.930	1.013	0.815	0.788	0.780
20	0.979	0.969	0.868	0.817	0.993	**0.712**	0.802	0.834	0.840	0.737	0.729	0.714
25	0.951	0.951	0.880	0.839	0.967	**0.754**	0.832	0.860	0.836	0.777	0.774	0.756
30	0.979	0.973	0.863	0.828	0.923	**0.766**	0.828	0.848	0.817	0.783	0.788	0.768

Table 3: Generalized Pareto Distribution (GPD)

Parameters: Position, ε = 0; Scale, α = 1; Shape, κ = -0.3 (GPD1)

n	SQP1	SQP2	KQ1	KQ2	KQ3	KQ4	WKQ1	WKQ2	WKQ3	WKQ4	HDQ	WHDQ
10	1.437	0.719	0.331	0.457	0.521	0.603	**0.325**	0.438	0.508	0.429	0.819	0.677
15	1.215	1.209	0.305	0.434	0.506	0.576	**0.301**	0.421	0.505	0.409	0.903	0.695
20	1.103	**0.145**	0.165	0.250	0.316	0.325	0.162	0.236	0.323	0.232	0.556	0.397
25	0.169	0.169	0.166	0.257	0.339	0.328	**0.163**	0.239	0.351	0.236	0.594	0.406
30	0.390	0.331	0.257	0.378	0.513	0.470	**0.249**	0.347	0.534	0.346	0.771	0.519

Parameters: Position, ε = 0; Scale, α = 1; Shape, κ = -0.2 (GPD2)

n	SQP1	SQP2	KQ1	KQ2	KQ3	KQ4	WKQ1	WKQ2	WKQ3	WKQ4	HDQ	WHDQ
10	1.250	0.721	0.404	0.510	0.636	0.604	**0.386**	0.468	0.603	0.469	0.764	0.653
15	1.241	1.230	0.399	0.510	0.612	0.622	**0.385**	0.480	0.587	0.475	0.893	0.717
20	1.103	0.273	0.248	0.315	0.386	0.383	**0.239**	0.298	0.381	0.296	0.573	0.437
25	0.337	0.337	0.265	0.341	0.422	0.416	**0.257**	0.322	0.429	0.321	0.632	0.467
30	0.547	0.462	0.339	0.433	0.549	0.519	**0.327**	0.403	0.563	0.404	0.755	0.551

Parameters: Position, ε = 0; Scale, α = 1; Shape, κ = -0.1 (GPD3)

n	SQP1	SQP2	KQ1	KQ2	KQ3	KQ4	WKQ1	WKQ2	WKQ3	WKQ4	HDQ	WHDQ
10	1.049	0.675	0.458	0.551	0.749	0.570	**0.425**	0.470	0.689	0.484	0.677	0.598
15	1.154	1.136	0.522	0.619	0.786	0.674	**0.492**	0.553	0.718	0.559	0.884	0.735
20	1.102	0.391	0.333	0.390	0.476	0.435	**0.317**	0.358	0.448	0.360	0.593	0.477
25	0.465	0.465	0.370	0.427	0.517	0.482	**0.356**	0.397	0.500	0.400	0.658	0.517
30	0.660	0.566	0.420	0.490	0.593	0.557	**0.405**	0.455	0.588	0.459	0.742	0.580

Parameters: Position, ε = 0; Scale, α = 1; Shape, κ = 0.1 (GPD4)

n	SQP1	SQP2	KQ1	KQ2	KQ3	KQ4	WKQ1	WKQ2	WKQ3	WKQ4	HDQ	WHDQ
10	0.760	0.603	0.498	0.593	0.897	0.499	0.445	**0.442**	0.788	0.475	0.538	0.507
15	1.055	1.032	0.734	0.835	1.161	0.727	0.666	**0.651**	0.964	0.683	0.812	0.744
20	1.106	0.708	0.580	0.634	0.814	0.577	0.539	**0.523**	0.678	0.540	0.655	0.591
25	0.754	0.754	0.623	0.669	0.818	0.637	0.588	**0.576**	0.702	0.590	0.721	0.645
30	0.891	0.812	0.645	0.681	0.801	0.672	0.618	**0.607**	0.715	0.618	0.757	0.676

Parameters: Position, ε = 0; Scale, α = 1; Shape, κ = 0.2 (GPD5)

n	SQP1	SQP2	KQ1	KQ2	KQ3	KQ4	WKQ1	WKQ2	WKQ3	WKQ4	HDQ	WHDQ
10	0.656	0.571	0.497	0.601	0.948	0.462	0.436	**0.417**	0.815	0.458	0.478	0.464
15	0.984	0.958	0.800	0.918	1.326	0.727	0.715	**0.669**	1.062	0.714	0.762	0.728
20	1.109	0.864	0.716	0.787	1.047	0.663	0.659	**0.618**	0.829	0.644	0.704	0.665
25	0.882	0.882	0.770	0.818	1.014	0.714	0.724	**0.675**	0.821	0.695	0.755	0.715
30	0.975	0.907	0.762	0.795	0.943	0.726	0.728	**0.685**	0.794	0.700	0.772	0.727

Parameters: Position, ε = 0; Scale, α = 1; Shape, κ = 0.3 (GPD6)

n	SQP1	SQP2	KQ1	KQ2	KQ3	KQ4	WKQ1	WKQ2	WKQ3	WKQ4	HDQ	WHDQ
10	0.584	0.553	0.503	0.615	0.992	0.441	0.437	**0.405**	0.841	0.454	0.441	0.441
15	0.936	0.912	0.842	0.986	1.472	0.722	0.746	**0.677**	1.140	0.733	0.724	0.717
20	1.114	1.036	0.865	0.962	1.315	0.738	0.788	**0.710**	0.998	0.749	0.743	0.734
25	0.993	0.993	0.897	0.962	1.223	0.780	0.838	**0.759**	0.942	0.786	0.787	0.777
30	1.045	1.005	0.868	0.908	1.093	0.778	0.827	**0.758**	0.876	0.777	0.789	0.777

as in (2.15) and in all cases for positive shape parameter, the most efficient estimator was Weighted Kernel Quantile estimator Type 2 $WKQ_{u.2(lc)}$, as in (18). Hence, it is suggested that in the case of GPD, when analyzing data which is lower-bounded (κ<0), as in most hydrological data, the best estimator would be the Weighted Kernel Quantile estimator Type 1, $WKQ_{u.1(lc)}$, and for data that is upper-bounded (κ>0), the Weighted Kernel Quantile estimator Type 2 $WKQ_{u.2(lc)}$, would be the best choice.

For Generalized Logistic Distribution (GLD), the IMSE relative efficiency values point to several selections, as show in Table 4. Compared to the other two previous distributions, no one obvious estimator can be considered the most efficient for all types of GLD included in this study. The frequently quoted choices are Weighted Kernel Quantile estimator Type 1, ($WKQ_{u.1(lc)}$), SQP1 and 2, and Kernel Quantile estimator Type 4, ($KQ_{u.4}$). However, using Weighted Kernel Quantile estimator Type 1, ($WKQ_{u.1(lc)}$), is recommended since this estimator is frequently quoted as the most efficient compared to the others, and for ease of further analysis in its future application.

The IMSE relative efficiency values for Lognormal Type 3 (LN3) distributions are displayed in Table 5. This table shows that, for LN3 Types 1 and 2, the suggested estimator is the Weighted Kernel Quantile estimator Type 4, $WKQ_{u.4(lc)}$, for LN3 types 3, 4 and 5 it was the Weighted Kernel Quantile estimator Type 1, $WKQ_{u.1(lc)}$, and there was no obvious choice for LN3 Type 6. Hence, $WKQ_{u.1(lc)}$ is recommended for this distribution because it is the best estimator for 3 types of LN3 (LN3 Types 3, 4 and 5) in this study, however, further analysis of the IMSE relative efficiency values for LN3 Types 2 and 6 showed that this method gave the second smallest IMSE.

Table 6 shows the IMSE relative efficiency values for the Pearson Type 3 (PE3) distribution. In general, for PE3 Types 1 and 2, the recommended estimator was $WKQ_{u.4(lc)}$ for PE3 Types 3 and 4, was $WKQ_{u.1(lc)}$ for PE3 Type 5, and for Type 6 was $WKQ_{u.2(lc)}$. Because only one type of estimator from the thirteen choices available needs to be chosen, although the simulation results showed three different methods, $WKQ_{u.1(lc)}$ is recommended to use in other distributions. Another possible alternative would be to use $WKQ_{u.4(lc)}$ and $WKQ_{u.2(lc)}$ as quantile estimators.

Finally, Table 7 shows the IMSE relative efficiency values for the Wakeby Type 5 (WAK5) distribution. Although the most efficient quantile estimator for WAK5 Types 1 and 2 was $WKQ_{u.1(lc)}$, the $WKQ_{u.2(lc)}$ is often recommended for WAK5 Types 3, 4, 5, and 6. Hence, for WAK5, $WKQ_{u.2(lc)}$ is recommended as the quantile estimation method, with $WKQ_{u.1(lc)}$ as another alternative.

Conclusion

Table 8 summarizes the two most efficient quantile estimation methods (in sequence) with respect to the six extreme distributions.

Table 8: The Top Two Most Efficient Quantile Estimation Methods

Distribution	Most Efficient	2nd Most Efficient
GEV	$WKQ_{u.1(lc)}$	$WKQ_{u.4(lc)}$
GPD	$WKQ_{u.1(lc)}$	$WKQ_{u.2(lc)}$
GLD	$WKQ_{u.1(lc)}$	$WKQ_{u.4(lc)}$
LN3	$WKQ_{u.1(lc)}$	$KQ_{u.1(lc)}$
PE3	$WKQ_{u.1(lc)}$	$WKQ_{u.4(lc)}$
WAK5	$WKQ_{u.2(lc)}$	$WKQ_{u.1(lc)}$

The IMSE relative efficiency of level crossing estimators was compared to the ordinary quantile estimator and the number of times the result showed that the level crossing estimators are better than the ordinary quantile

Table 4: Generalized Logistic Distribution (GLD)

Parameters: Position, ε = 0; Scale, α = 1; Shape, κ =-0.1 (GLD1)

n	SQP1	SQP2	KQ1	KQ2	KQ3	KQ4	WKQ1	WKQ2	WKQ3	WKQ4	HDQ	WHDQ
10	0.852	0.772	0.882	0.784	0.939	**0.667**	0.808	0.815	0.869	0.692	0.676	0.674
15	1.064	1.073	1.113	1.011	1.128	**0.856**	1.047	1.078	1.039	0.899	0.877	0.870
20	1.155	1.084	0.977	0.897	0.945	**0.762**	0.943	0.970	0.879	0.815	0.768	0.775
25	1.088	1.088	1.072	0.991	0.995	0.853	1.058	1.076	0.940	0.921	**0.846**	0.870
30	1.117	1.210	1.138	1.051	1.016	0.912	1.143	1.150	0.974	0.997	**0.896**	0.937

Parameters: Position, ε = 0; Scale, α = 1; Shape, κ =-0.2 (GLD2)

n	SQP1	SQP2	KQ1	KQ2	KQ3	KQ4	WKQ1	WKQ2	WKQ3	WKQ4	HDQ	WHDQ
10	1.027	0.797	0.753	0.720	0.839	0.668	0.702	0.722	0.790	**0.647**	0.728	0.692
15	1.109	1.114	0.893	0.856	0.953	0.778	0.849	0.869	0.892	**0.773**	0.876	0.819
20	1.125	0.800	0.718	0.684	0.731	**0.616**	0.697	0.703	0.689	0.628	0.691	0.646
25	0.843	0.843	0.823	0.778	0.805	**0.701**	0.812	0.805	0.766	0.728	0.772	0.732
30	0.944	1.020	0.937	0.876	0.873	**0.795**	0.941	0.918	0.844	0.838	0.853	0.828

Parameters: Position, ε = 0; Scale, α = 1; Shape, κ =-0.3 (GLD3)

n	SQP1	SQP2	KQ1	KQ2	KQ3	KQ4	WKQ1	WKQ2	WKQ3	WKQ4	HDQ	WHDQ
10	1.150	0.672	0.570	0.594	0.681	0.599	**0.541**	0.577	0.653	0.541	0.716	0.644
15	1.132	1.135	0.644	0.676	0.754	0.683	0.621	0.663	0.723	**0.620**	0.884	0.760
20	1.105	0.506	0.461	0.476	0.523	0.472	0.450	0.470	0.506	**0.441**	0.616	0.521
25	0.554	0.554	0.536	0.543	0.594	0.529	0.530	0.538	0.577	**0.508**	0.688	0.577
30	0.688	0.738	0.651	0.652	0.704	0.636	0.651	0.652	0.691	**0.620**	0.813	0.685

Parameters: Position, ε = 0; Scale, αa = 1; Shape, κ =-0.4(GLD4)

n	SQP1	SQP2	KQ1	KQ2	KQ3	KQ4	WKQ1	WKQ2	WKQ3	WKQ4	HDQ	WHDQ
10	1.468	0.893	0.479	0.579	0.609	0.694	**0.464**	0.559	0.595	0.533	0.894	0.769
15	1.130	1.131	0.378	0.485	0.559	0.565	**0.369**	0.459	0.550	0.439	0.887	0.687
20	1.104	**0.200**	0.213	0.271	0.320	0.309	0.210	0.256	0.318	0.246	0.517	0.388
25	0.196	**0.196**	0.224	0.294	0.372	0.329	0.221	0.273	0.370	0.265	0.591	0.415
30	**0.329**	0.341	0.343	0.428	0.547	0.465	0.342	0.401	0.547	0.391	0.804	0.547

Parameters: Position, ε = 0; Scale, α = 1; Shape, κ =-0.5 (GLD5)

n	SQP1	SQP2	KQ1	KQ2	KQ3	KQ4	WKQ1	WKQ2	WKQ3	WKQ4	HDQ	WHDQ
10	1.616	0.891	0.374	0.513	0.510	0.683	**0.370**	0.499	0.507	0.477	0.931	0.777
15	1.297	1.297	0.248	0.390	0.436	0.514	**0.246**	0.364	0.436	0.348	0.872	0.666
20	1.105	**0.060**	0.121	0.217	0.293	0.274	0.120	0.194	0.295	0.188	0.538	0.362
25	0.163	**0.163**	0.196	0.288	0.387	0.337	0.194	0.262	0.391	0.257	0.612	0.411
30	**0.299**	0.306	0.338	0.532	0.831	0.590	0.335	0.468	0.834	0.465	0.907	0.551

Parameters: Position, ε = 0; Scale, α = 1; Shape, κ =-0.6 (GLD6)

n	SQP1	SQP2	KQ1	KQ2	KQ3	KQ4	WKQ1	WKQ2	WKQ3	WKQ4	HDQ	WHDQ
10	1.513	0.510	0.198	0.334	0.340	0.474	**0.197**	0.316	0.340	0.302	0.740	0.583
15	1.257	1.257	0.208	0.362	0.415	0.491	**0.207**	0.333	0.417	0.319	0.864	0.647
20	1.106	**0.065**	0.116	0.205	0.272	0.262	0.115	0.184	0.275	0.179	0.514	0.351
25	0.074	**0.074**	0.119	0.218	0.312	0.271	0.118	0.192	0.316	0.188	0.570	0.366
30	0.126	**0.124**	0.187	0.344	0.536	0.408	0.185	0.297	0.540	0.294	0.831	0.488

Table 5: Log-Normal Type 3 Distribution (LN3)

Parameters: Position, ε = 0; Scale, α = 1; Shape, κ =-0.2 (LN3_1)

n	SQP1	SQP2	KQ1	KQ2	KQ3	KQ4	WKQ1	WKQ2	WKQ3	WKQ4	HDQ	WHDQ
10	0.846	0.730	0.761	0.658	0.876	0.591	0.674	0.692	0.783	**0.588**	0.619	0.599
15	1.067	1.055	0.934	0.842	1.054	0.764	0.845	0.907	0.929	**0.752**	0.826	0.779
20	1.077	0.801	0.723	0.669	0.795	0.627	0.670	0.724	0.710	**0.611**	0.686	0.637
25	0.796	0.796	0.719	0.683	0.781	0.663	0.681	0.735	0.718	**0.639**	0.733	0.674
30	0.896	0.844	0.716	0.693	0.774	0.690	0.689	0.738	0.730	**0.659**	0.764	0.697

Parameters: Position, ε = 0; Scale, α = 1; Shape, κ =-0.4 (LN3_2)

n	SQP1	SQP2	KQ1	KQ2	KQ3	KQ4	WKQ1	WKQ2	WKQ3	WKQ4	HDQ	WHDQ
10	0.980	0.729	0.641	0.614	0.794	0.598	0.585	0.618	0.729	**0.555**	0.664	0.617
15	1.106	1.092	0.762	0.745	0.916	0.732	0.705	0.758	0.832	**0.676**	0.851	0.763
20	1.100	0.602	0.527	0.525	0.620	0.531	0.498	0.538	0.573	**0.485**	0.638	0.556
25	0.641	0.641	0.544	0.553	0.638	0.577	0.521	0.568	0.608	**0.521**	0.697	0.598
30	0.786	0.713	0.570	0.590	0.676	0.625	**0.552**	0.598	0.658	0.560	0.752	0.640

Parameters: Position, ε = 0; Scale, α = 1; Shape, κ =-0.6 (LN3_3)

n	SQP1	SQP2	KQ1	KQ2	KQ3	KQ4	WKQ1	WKQ2	WKQ3	WKQ4	HDQ	WHDQ
10	1.163	0.764	0.545	0.577	0.709	0.631	**0.515**	0.573	0.670	0.537	0.748	0.665
15	1.141	1.128	0.568	0.613	0.732	0.682	**0.542**	0.615	0.695	0.576	0.883	0.741
20	1.103	0.446	0.384	0.418	0.493	0.465	**0.368**	0.416	0.477	0.394	0.611	0.500
25	0.509	0.509	0.414	0.457	0.540	0.509	**0.400**	0.449	0.531	0.432	0.671	0.540
30	0.693	0.609	0.461	0.513	0.604	0.577	**0.448**	0.502	0.608	0.487	0.748	0.598

Parameters: Position, ε = 0; Scale, α = 1; Shape, κ =-0.8 (LN3_4)

n	SQP1	SQP2	KQ1	KQ2	KQ3	KQ4	WKQ1	WKQ2	WKQ3	WKQ4	HDQ	WHDQ
10	1.313	0.763	0.457	0.532	0.625	0.640	**0.444**	0.527	0.604	0.506	0.804	0.688
15	1.193	1.182	0.439	0.523	0.607	0.635	**0.427**	0.520	0.596	0.498	0.891	0.719
20	1.105	0.284	0.303	0.316	0.414	0.377	**0.274**	0.301	0.394	0.294	0.559	0.431
25	0.372	0.372	0.291	0.359	0.436	0.433	**0.283**	0.348	0.447	0.340	0.639	0.479
30	0.590	0.505	0.364	0.447	0.552	0.528	**0.352**	0.425	0.569	0.419	0.745	0.557

Parameters: Position, ε = 0; Scale, α = 1; Shape, κ =-1.0 (LN3_5)

n	SQP1	SQP2	KQ1	KQ2	KQ3	KQ4	WKQ1	WKQ2	WKQ3	WKQ4	HDQ	WHDQ
10	1.502	0.859	0.404	0.521	0.561	0.692	**0.402**	0.522	0.555	0.502	0.906	0.761
15	1.208	1.200	0.338	0.451	0.514	0.592	**0.335**	0.449	0.519	0.432	0.897	0.699
20	1.102	0.206	0.203	0.274	0.326	0.350	**0.200**	0.268	0.337	0.260	0.552	0.413
25	0.287	0.287	0.231	0.313	0.395	0.389	**0.225**	0.297	0.409	0.292	0.623	0.446
30	0.444	0.376	0.275	0.387	0.510	0.474	**0.266**	0.357	0.531	0.354	0.755	0.521

Parameters: Position, ε = 0; Scale, α = 1; Shape, κ =-1.2 (LN3_6)

n	SQP1	SQP2	KQ1	KQ2	KQ3	KQ4	WKQ1	WKQ2	WKQ3	WKQ4	HDQ	WHDQ
10	1.600	0.878	**0.363**	0.504	0.524	0.706	0.367	0.508	0.526	0.488	0.948	0.783
15	1.237	1.232	**0.265**	0.401	0.451	0.552	0.266	0.393	0.462	0.378	0.893	0.682
20	1.104	**0.123**	0.145	0.225	0.278	0.300	0.143	0.215	0.289	0.209	0.522	0.376
25	0.196	0.196	0.175	0.271	0.356	0.350	**0.171**	0.254	0.373	0.249	0.610	0.419
30	0.347	0.294	0.233	0.364	0.503	0.454	**0.226**	0.331	0.525	0.328	0.772	0.510

Table 6: Pearson Type 3 Distribution (PE3)

Parameters: Position, ε = 0; Scale, α = 1; Shape, κ = 1 (PE3_1)

n	SQP1	SQP2	KQ1	KQ2	KQ3	KQ4	WKQ1	WKQ2	WKQ3	WKQ4	HDQ	WHDQ
10	0.916	0.741	0.747	0.647	0.844	0.599	0.668	0.697	0.763	**0.579**	0.642	0.611
15	1.086	1.072	0.873	0.788	0.973	0.745	0.797	0.862	0.874	**0.710**	0.833	0.768
20	1.103	0.727	0.659	0.611	0.720	0.592	0.615	0.667	0.658	**0.561**	0.669	0.607
25	0.735	0.735	0.662	0.632	0.723	0.630	0.629	0.682	0.678	**0.591**	0.718	0.644
30	0.863	0.808	0.683	0.665	0.748	0.672	0.658	0.705	0.716	**0.628**	0.763	0.681

Parameters: Position, ε = 0; Scale, α = 1; Shape, κ = 2 (PE3_2)

n	SQP1	SQP2	KQ1	KQ2	KQ3	KQ4	WKQ1	WKQ2	WKQ3	WKQ4	HDQ	WHDQ
10	1.119	0.788	0.634	0.617	0.747	0.654	0.599	0.646	0.707	**0.580**	0.748	0.678
15	1.120	1.099	0.673	0.672	0.794	0.715	0.639	0.697	0.752	**0.630**	0.873	0.755
20	1.106	0.535	0.464	0.469	0.543	0.503	0.445	0.482	0.522	**0.443**	0.625	0.529
25	0.582	0.582	0.474	0.493	0.565	0.542	**0.460**	0.503	0.560	0.470	0.681	0.566
30	0.754	0.667	0.510	0.541	0.628	0.593	**0.495**	0.539	0.626	0.511	0.738	0.608

Parameters: Position, ε = 0; Scale, α = 1; Shape, κ = 3 (PE3_3)

n	SQP1	SQP2	KQ1	KQ2	KQ3	KQ4	WKQ1	WKQ2	WKQ3	WKQ4	HDQ	WHDQ
10	1.286	0.804	0.539	0.576	0.648	0.685	**0.534**	0.601	0.637	0.568	0.826	0.721
15	1.160	1.141	0.530	0.582	0.660	0.684	**0.520**	0.592	0.653	0.565	0.896	0.743
20	1.099	0.383	0.321	0.362	0.414	0.431	**0.315**	0.364	0.419	0.351	0.589	0.470
25	0.473	0.473	0.367	0.418	0.491	0.487	**0.359**	0.409	0.500	0.400	0.661	0.518
30	0.666	0.571	0.406	0.471	0.563	0.546	**0.396**	0.452	0.577	0.445	0.733	0.570

Parameters: Position, ε = 0; Scale, α = 1; Shape, κ = 4 (PE3_4)

n	SQP1	SQP2	KQ1	KQ2	KQ3	KQ4	WKQ1	WKQ2	WKQ3	WKQ4	HDQ	WHDQ
10	1.398	0.819	**0.463**	0.541	0.575	0.698	0.469	0.556	0.577	0.545	0.883	0.750
15	1.181	1.167	**0.420**	0.502	0.555	0.644	0.421	0.508	0.567	0.499	0.904	0.721
20	1.062	0.285	0.251	0.304	0.351	0.382	**0.248**	0.299	0.363	0.297	0.564	0.433
25	0.396	0.396	0.298	0.364	0.436	0.445	**0.293**	0.348	0.452	0.348	0.645	0.485
30	0.596	0.509	0.358	0.439	0.542	0.524	**0.349**	0.409	0.559	0.412	0.741	0.552

Parameters: Position, ε = 0; Scale, α = 1; Shape, κ = 6 (PE3_5)

n	SQP1	SQP2	KQ1	KQ2	KQ3	KQ4	WKQ1	WKQ2	WKQ3	WKQ4	HDQ	WHDQ
10	1.129	0.751	0.415	0.463	0.477	0.650	0.417	**0.400**	0.479	0.529	0.888	0.728
15	1.193	1.191	0.387	0.428	0.476	0.585	0.383	**0.366**	0.478	0.452	0.898	0.693
20	0.631	**0.192**	0.247	0.264	0.316	0.346	0.237	0.218	0.314	0.266	0.556	0.409
25	0.275	0.275	0.293	0.304	0.387	0.388	0.273	**0.247**	0.378	0.296	0.626	0.444
30	0.433	0.391	0.383	0.383	0.512	0.474	0.349	**0.302**	0.490	0.363	0.748	0.517

Parameters: Position, ε = 0; Scale, α = 1; Shape, κ = 8 (PE3_6)

n	SQP1	SQP2	KQ1	KQ2	KQ3	KQ4	WKQ1	WKQ2	WKQ3	WKQ4	HDQ	WHDQ
10	0.870	0.802	0.568	0.515	0.566	0.674	0.532	**0.442**	0.540	0.614	0.875	0.735
15	1.184	1.184	0.530	0.481	0.548	0.617	0.501	**0.413**	0.523	0.533	0.894	0.708
20	0.357	0.257	0.335	0.296	0.352	0.372	0.312	**0.253**	0.331	0.315	0.562	0.425
25	0.337	0.337	0.386	0.333	0.412	0.414	0.355	**0.273**	0.382	0.347	0.632	0.462
30	0.461	0.455	0.509	0.426	0.545	0.513	0.458	**0.351**	0.495	0.436	0.751	0.543

Table 7: Wakeby 5-parameter Distribution (WAK5)

Parameters: Position, ε = 0; Scale, α = 1 and γ = 4; Shape, β = 16 and δ = 0.20 (WAK5_1)

n	SQP1	SQP2	KQ1	KQ2	KQ3	KQ4	WKQ1	WKQ2	WKQ3	WKQ4	HDQ	WHDQ
10	1.097	0.631	0.375	0.490	0.692	0.532	**0.342**	0.407	0.622	0.422	0.671	0.577
15	1.157	1.150	0.416	0.543	0.709	0.626	**0.392**	0.476	0.640	0.491	0.895	0.716
20	1.107	0.312	0.266	0.337	0.424	0.395	**0.254**	0.303	0.395	0.315	0.576	0.447
25	0.369	0.369	0.291	0.366	0.455	0.433	**0.281**	0.333	0.441	0.349	0.639	0.482
30	0.555	0.500	0.354	0.444	0.549	0.526	**0.345**	0.406	0.549	0.426	0.755	0.562

Parameters: Position, ε = 0; Scale, α = 1 and γ = 5; Shape, β = 7.5 and δ = 0.12 (WAK5_2)

n	SQP1	SQP2	KQ1	KQ2	KQ3	KQ4	WKQ1	WKQ2	WKQ3	WKQ4	HDQ	WHDQ
10	0.868	0.587	0.394	0.540	0.816	0.503	**0.353**	0.372	0.711	0.452	0.577	0.526
15	1.111	1.108	0.563	0.738	1.003	0.721	**0.521**	0.538	0.840	0.653	0.878	0.770
20	1.159	0.555	0.398	0.500	0.618	0.516	**0.383**	0.394	0.531	0.472	0.640	0.550
25	0.622	0.622	0.443	0.540	0.623	0.569	**0.435**	0.446	0.560	0.525	0.700	0.601
30	0.768	0.734	0.509	0.610	0.677	0.655	**0.509**	0.525	0.639	0.607	0.791	0.679

Parameters: Position, ε = 0; Scale, α = 1 and g = 5; Shape, β = 1 and δ = 0.12 (WAK5_3)

n	SQP1	SQP2	KQ1	KQ2	KQ3	KQ4	WKQ1	WKQ2	WKQ3	WKQ4	HDQ	WHDQ
10	0.782	0.636	0.532	0.667	0.940	0.540	0.463	**0.457**	0.832	0.528	0.573	0.552
15	1.082	1.082	0.757	0.892	1.135	0.751	0.685	**0.658**	0.965	0.731	0.846	0.789
20	1.128	0.803	0.670	0.748	0.888	0.624	0.622	**0.578**	0.749	0.627	0.694	0.653
25	0.840	0.840	0.771	0.828	0.927	0.709	0.736	**0.679**	0.798	0.720	0.777	0.736
30	0.926	0.980	0.862	0.898	0.963	0.785	0.839	**0.771**	0.850	0.804	0.852	0.814

Parameters: Position, ε = 0; Scale, α = 1 and γ = 10; Shape, β = 16 and δ = 0.04 WAK5_4)

n	SQP1	SQP2	KQ1	KQ2	KQ3	KQ4	WKQ1	WKQ2	WKQ3	WKQ4	HDQ	WHDQ
10	0.583	0.450	0.384	0.504	0.939	0.385	**0.311**	0.313	0.750	0.360	0.427	0.396
15	1.011	1.004	0.654	0.842	1.370	0.705	0.557	**0.534**	1.005	0.664	0.810	0.731
20	1.146	0.668	0.485	0.602	0.855	0.564	0.440	**0.418**	0.629	0.534	0.655	0.584
25	0.766	0.766	0.531	0.642	0.820	0.638	0.503	**0.475**	0.632	0.613	0.726	0.652
30	0.910	0.868	0.580	0.697	0.821	0.715	0.566	**0.533**	0.672	0.695	0.790	0.723

Parameters: Position, ε = 0; Scale, α = 1 and γ = 10; Shape, β = 1 and δ = 0.04 (WAK5_5)

n	SQP1	SQP2	KQ1	KQ2	KQ3	KQ4	WKQ1	WKQ2	WKQ3	WKQ4	HDQ	WHDQ
10	0.606	0.594	0.557	0.732	1.061	0.515	0.468	**0.436**	0.916	0.543	0.501	0.515
15	0.905	0.909	0.967	1.172	1.531	0.828	0.848	**0.755**	1.244	0.888	0.787	0.824
20	1.161	1.339	1.139	1.302	1.560	0.954	1.040	0.918	1.252	1.032	**0.900**	0.955
25	1.258	1.258	1.230	1.342	1.496	1.017	1.157	1.021	1.216	1.108	**0.957**	1.027
30	1.192	1.313	1.244	1.311	1.378	1.034	1.203	1.068	1.151	1.130	**0.976**	1.055

Parameters: Position, ε = 0; Scale, α = 1 and γ = 10; Shape, β = 2.5 and δ = 0.02 (WAK5_6)

n	SQP1	SQP2	KQ1	KQ2	KQ3	KQ4	WKQ1	WKQ2	WKQ3	WKQ4	HDQ	WHDQ
10	0.565	0.568	0.432	0.692	1.064	0.484	0.374	**0.333**	0.881	0.520	0.453	0.482
15	0.860	0.878	0.800	1.171	1.603	0.831	0.717	**0.637**	1.229	0.904	0.763	0.826
20	1.386	1.513	1.041	1.408	1.732	1.042	0.971	**0.876**	1.312	1.135	0.955	1.047
25	1.365	1.365	1.111	1.415	1.594	1.092	1.071	**0.980**	1.237	1.187	1.003	1.108
30	1.293	1.387	1.114	1.355	1.427	1.083	1.102	1.020	1.149	1.174	**1.004**	1.111

estimators was calculated; these findings are summarized in the Table 9.

Table 9: IMSE Relative Efficiency when Comparing Between Level Crossing and Ordinary Quantile Estimators

$\frac{WKQ_{u.1(lc)}}{KQ_{u.1}}$	$\frac{WKQ_{u.2(lc)}}{KQ_{u.2}}$	$\frac{WKQ_{u.3(lc)}}{KQ_{u.3}}$	$\frac{WKQ_{u.4(lc)}}{KQ_{u.4}}$	$\frac{WHD_{u}}{HD_{u}}$
95%	82%	83%	80%	92%

Hence, it can be concluded that the level crossing estimators are better than the ordinary quantile estimators as shown in our analysis most of the time.

Analysis on the most efficient method among the ordinary quantile estimators family showed that the $KQ_{u.1}$ quantile estimation method is the most efficient.

References

Ani, S., & Jemain, A. A. (2007). LQ-moment: Application to the generalized extreme value. *Journal of Applied Sciences*, *7(1)*, 115-120.

Ben-Zvi, A., & Azmon, B. (1997). Joint use of L-moment diagram and goodness-of-fit test: A case study of diverse series, *Journal of Hydrology*, *198*, (*1-4*, Nov.). p 245-259.

David, H. A., & Nagaraja, H. N. (2003). *Order statistics*. (3rd ed.) NY: Wiley.

Harrel, F. E., & Davis, C. E. (1982). A new distribution-free quantile estimator. *Biometrika*, *69(3)*, 635-640.

Hosking, J .R. M., & Wallis, J. R. (1987). Parameter and quantile estimation for the generalized Pareto distribution. *Technometrics*, *29*, 339–349.

Hosking, J. R. M., & J. R. Wallis. (1997). *Regional frequency analysis*. Cambridge: Cambridge University Press.

Huang, M. L., & Brill, P. H. (1995). A new weighted density estimation method. *American Statistical Association Proceedings of the Statistical Computing Section*, p. 125-130.

Huang, M. L., & Brill, P. H. (1996). Empirical distribution based on level crossings. *Working paper Series*, *No. W96-01*, University of Windsor, Faculty of Business Administration.

Huang, M. L., & Brill, P. H. (1999). A level crossing quantile estimation method. *Statistics and Probability Letters*, *45*, 111-119.

Huang, M. L. (2001). On a distribution-free quantile estimator. *Computational Statistics and Data Analysis*, *37*, 477-486.

Landwehr, J. M., Matalas, N. C., & Wallis, J. R. (1979b). Estimation of parameters and quantiles of Wakeby distributions. *Water Resources Research*, *15*, 1361–1379

Mudholkar, G. S., Hutson, A. D. (1998). LQ-Moments:Analogs of L-Moments. *Journal of Statistical Planning and Inference*, *71*, 191-208.

Parzen, E. (1979). Nonparametric statistical data modeling, *Journal of the American Statistical Association*, *74*, 105–121.

Rao, A. R., & K. H. Hamed. (2000). *Flood Frequency Analysis*. Boca Raton, London, New York, Washington, D.C: CRC Press.

Shabri, A., & Jemain, A. A. (2007). LQ-moment: Application to the generalized extreme value. *Journal of Applied Sciences*, *7(1)*, 115-120.

Sheather, S. J., & Marron, J. S. (1990). Kernel quantile estimators. *Journal of American Statistical Association*, *85*, 410-416.

Shorack, G. R., &Wellner, J. S. (1986). *Empirical processes with applications to statistics*. NY: Wiley.

Sillitto G. P. (1951). Interrelations between certain linear systematic statistics of samples from any continuous population. *Biometrika*, *38*, 377-382.

Vogel, R. M., & Fennessey, N. M. (1993). L-moment diagrams should replace product-moment diagrams. *Water Resources Research*, *29*(*6*), 1745-1752.

Journal of Modern Applied Statistical Methods
November, 2008, Vol. 7, No. 2, 467-470

Frequency Domain Modeling with Piecewise Constant Spectra

Erhard Reschenhofer
University of Vienna, Austria

Using piecewise constant functions as models for the spectral density of the differenced log real U.S. GDP it was found that these models have the capacity to compete with the spectral densities implied by ARMA models. According to AIC and BIC the piecewise constant spectral densities are superior to ARMA.

Key words: Spectral analysis, piecewise constant spectra, ARMA spectra, aggregate output.

Introduction

Univariate ARMA models are used in empirical economics as simple, purely statistical models for properly transformed macroeconomic time series (such as the first differences of the logs of the real GDP), and for the description of the serial correlation in the errors of more complex models such as linear or nonlinear multivariate regression models. A typical example of the first type is the study by Campbell & Mankiw (1987) who used ARMA(p,q) models with $p \leq 3$ and $q \leq 3$ to investigate the long-run behavior of aggregate output. The persistence of output shocks can be measured by the cumulative impulse response or, equivalently, by the value of the spectral density at frequency zero, however, two drawbacks exist. The first is that the model parameters must be estimated by numerical optimization routines, which depend heavily on the starting values and can easily get stuck at local optima (e.g., Hauser, et al., 1999). The second is the extreme sensitivity of inference to the order of the ARMA representation (e.g., Christiano & Eichenbaum, 1990).

Recently, interest has shifted from univariate to multivariate modeling (e.g., Blanchard & Quah, 1989; Pesaran, et al., 1993; Pesaran & Shin, . However, a multivariate approach based on economic theory and the information contained in a much larger data set is not necessarily better than a simple univariate time series model, because both the estimation and the identification of multivariate models is many orders of magnitude more difficult. But even in situations where multivariate models outperform univariate models, the latter are often used as benchmarks for the former (see, e.g., Schumacher & Dreger, 2004). Thus, univariate ARMA models still have an important role to play. This article proposes competitive alternatives to ARMA models for the purpose of estimating the spectral densities of macroeconomic time series.

Methodology

The following piecewise constant functions are proposed:

$$g_r(\omega) = a(b_1 1_{[\alpha_0,\alpha_1)} + b_2 1_{[\alpha_1,\alpha_2)} + \ldots + b_{r-1} 1_{[\alpha_{r-2},\alpha_{r-1})} + 1_{[\alpha_{r-1},\alpha_r]}),\ \omega \in [0,\pi], \quad (1)$$

where $r \geq 2$ and $0=\alpha_0<\alpha_1<\ldots<\alpha_r=\pi$, for the approximation of the spectral densities of macroeconomic time series. There are 2(r-1)+1 parameters that must be estimated, a, $b_1, \ldots, b_{r-1}$, $\alpha_1, \ldots, \alpha_{r-1}$. An obvious choice for the first parameter is:

Erhard Reschenhofer is Associate Professor in the Department of Statistics and Decision Support Systems. Email: erhard.reschenhofer@univie.ac.at.

$$a=\frac{I_1+...+I_m}{b_1 s_1+b_2(s_2-s_1)+...+b_{r-1}(s_{r-1}-s_{r-2})+(m-s_{r-1})}, \qquad (2)$$

where s_j is the largest integer such that:

$$\frac{2\pi s_j}{n}<\alpha_j. \qquad (3)$$

The parameters $b_1, \ldots, b_{r-1}, \ldots, s_1, \ldots, s_{r-1}$ can be found by maximizing the Whittle likelihood

$$\prod_{k=1}^{m}\frac{1}{g_r(\omega_k)}\exp(-\frac{I_k}{g_r(\omega_k)}), \qquad (4)$$

or, equivalently,

$$-\sum_{k=1}^{n}\log(g_r(\omega_k))-\sum_{k=1}^{m}\frac{I_k}{g_r(\omega_k)}, \qquad (5)$$

where

$$\omega_k=\frac{2\pi k}{n}, \ k=1,\ldots,m. \qquad (6)$$

The parameters $\alpha_1, \ldots, \alpha_{r-1}$ can be obtained from $s_1, \ldots, s_{r-1}$ via

$$\alpha_j=\frac{2\pi s_j}{n}+\frac{\pi}{n}. \qquad (7)$$

To demonstrate the usefulness of this approach, the seasonally adjusted quarterly real U.S. GDP from 1947.1 to 2007.1 was downloaded from FRED® (Federal Reserve Economic Data) and the spectral density of the first differences of the log GDP was approximated by the piecewise constant functions g_j, j=2,3,4.

Results

Figure 1 compares the three piecewise constant spectral densities with the best three ARMA spectral densities selected by BIC. One of these three ARMA models, namely the ARMA(3,2) model, is the best ARMA model according to AIC. Apart from the ARMA models of order (2,3) and (3,3), whose spectral densities are very similar to that of the ARMA(3,2) model, all other ARMA models ($p\leq8$ & $q=0$, $p=0$ & $q\leq8$, $1\leq p, q\leq3$) have much higher AIC values than the ARMA(3,2) model. To facilitate the comparison between the piecewise constant spectral densities g_2, g_3, and g_4, and the ARMA spectral densities slightly modified AIC and BIC values (AIC* and BIC*) obtained from the Whittle likelihood were used. Among the top models both according to AIC* and BIC* (see Tables 1 and 2) are g_2, g_3, and g_4. Overall, g_2 has the smallest BIC* value and g_4 has the smallest AIC* value.

Table 1: AIC values (obtained from the Whittle likelihood) for piecewise constant spectral densities g(r) & ARMA(p,q) spectral densities, respectively, fitted to the differenced log real U.S. GDP

	1	2	3	4	5	6	7	8
g(r)		-2456.8	-2457.5	-2461.6				
AR(p)	-2447.3	-2446.9	-2447.7	-2448.6	-2448.1	-2446.2	-2444.2	-2442.4
MA(q)	-2440.5	-2448.2	-2447.8	-2445.9	-2447.7	-2445.7	-2443.8	-2445.2
ARMA(1,q)	-2446.2	-2447.2	-2445.8					
ARMA(2,q)	-2445.8	-2450.7	-2457.6					
ARMA(3,q)	-2449.1	-2459.2	-2457.1					

Table 2: BIC values (obtained from the Whittle likelihood) for piecewise constant spectral densities g(r) & ARMA(p,q) spectral densities, respectively, fitted to the differenced log real U.S. GDP

	1	2	3	4	5	6	7	8
g(r)		-2446.3	-2440.1	-2437.3				
AR(p)	-2440.3	-2436.5	-2433.7	-2431.2	-2427.3	-2421.9	-2416.4	-2411.1
MA(q)	-2433.5	-2437.8	-2433.9	-2428.5	-2426.8	-2421.3	-2416.0	-2413.9
ARMA(1,q)	-2435.8	-2433.3	-2428.4					
ARMA(2,q)	-2431.9	-2433.3	-2436.8					
ARMA(3,q)	-2431.7	-2438.3	-2432.8					

Conclusion

The results obtained show that piecewise constant spectral densities are extremely useful tools for the spectral analysis of macroeconomic time series and can outperform the more sophisticated ARMA spectral densities. This finding is striking given that twenty-five ARMA spectral densities were tried but only three piecewise constant spectral densities. It may also serve as a severe warning not to over-interpret certain characteristics of estimated ARMA spectral densities such as a decline or incline near frequency zero.

References

Blanchard, J. O., & Quah, D. (1989). The dynamic effects of aggregate supply and demand disturbances. *American Economic Review*, *79*, 655-673.

Campbell, J. Y., & Mankiw, N. G. (1987). Are output fluctuations transitory? *The Quarterly Journal of Economics*, *102*, 857-880.

Christiano, L. J., & Eichenbaum, M. (1990). Unit roots in real GNP: Do we know, and do we care? *Carnegie-Rochester Conference Series on Public Policy*, *32*, 7-61.

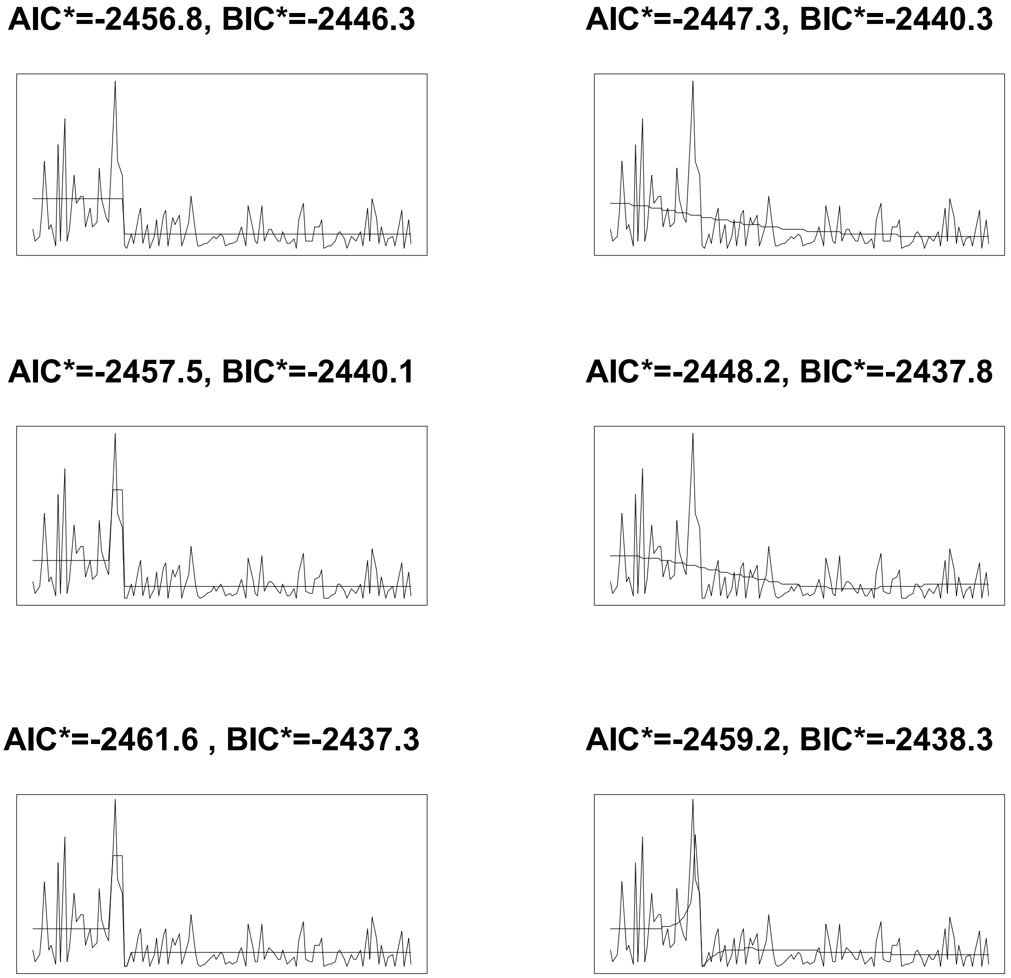
Hauser, M. A., Pötscher, B. M., & Reschenhofer, E. (1999). Measuring persistence in aggregate output: ARMA models, fractionally integrated ARMA models and nonparametric procedures. *Empirical Economics*, *24*, 243-269.

Pesaran, H. M., Pierse R. G., & Lee, K. C. (1993). Persistence, co-integration and aggregation: A disaggregate analysis of output fluctuations in the U.S. economy. *Journal of Econometrics*, *56*, 57-88.

Pesaran, H. M., & Shin, Y. (1996). Cointegration and speed of convergence to Equilibrium. *Journal of Econometrics*, *71*, 117-143.

Schumacher, C., & Dreger, C. (2004). Estimating large-scale factor models for economic activity in Germany: Do they outperform simpler models? *Journal of Economics and Statistics 224*, 732-750.

Figure 1: Periodogram of differenced log GDP together with piecewise constant spectral densities (with two, three, and four pieces) & ARMA spectral densities AR(1), MA(2), ARMA(3,2)



AIC*=-2456.8, BIC*=-2446.3

AIC*=-2447.3, BIC*=-2440.3

AIC*=-2457.5, BIC*=-2440.1

AIC*=-2448.2, BIC*=-2437.8

AIC*=-2461.6 , BIC*=-2437.3

AIC*=-2459.2, BIC*=-2438.3

Journal of Modern Applied Statistical Methods
November, 2008, Vol. 7, No. 2, 471-477

Multi-Group Confirmatory Factor Analysis for Testing Measurement Invariance in Mixed Item Format Data

Kim H. Koh
Nanyang Technological University
Singapore

Bruno D. Zumbo
University of British Columbia
Canada

This simulation study investigated the empirical Type I error rates of using the maximum likelihood estimation method and Pearson covariance matrix for multi-group confirmatory factor analysis (MGCFA) of full and strong measurement invariance hypotheses with mixed item format data that are ordinal in nature. The results indicate that mixed item formats and sample size combinations do not result in inflated empirical Type I error rates for rejecting the true measurement invariance hypotheses. Therefore, although the common methods are in a sense sub-optimal, they don't lead to researchers claiming that measures are functioning differently across groups – i.e., a lack of measurement invariance.

Key words: Multi-Group Confirmatory Factor Analysis, Measurement Invariance, Binary and Ordinal Items.

Introduction

Multi-group confirmatory maximum likelihood factor analysis has become the most commonly used scale-level technique to evaluate measurement invariance/ equivalence of a test across different groups (e.g., gender, language), over different mediums of administration (e.g., web-based versus paper-and-pencil testing), or across accommodated and non-accommodated conditions. Measurement invariance is tenable when the relations between observed variables and latent construct(s) are identical across relevant groups. In particular, individuals with the same standing on a latent variable but sampled from different subpopulations should have the same expected observed score on a test of that variable (Horn and McArdle, 1992). The common understanding in the research literature is that without measurement invariance, observed means (or latent means) are not directly comparable (Drasgow & Kanfer, 1985).

Mixed item format data are often found in educational measurement wherein many classroom and large-scale assessments in use today are blended instruments that include a mixture of multiple-choice and constructed-response items. Typically, multiple-choice items are dichotomously scored and constructed-response items are polytomously (partial-credit) scored. These two types of scores are on an ordinal scale. Two commonly encountered, and interrelated, problems associated with ordinal scale are measurement scale coarseness and multivariate nonnormality. Measurement scale coarseness is caused by a crude classification of the latent variables to ordinal scales with small numbers of response categories. Because of the discrete nature of ordinal scales, the distributions of the response data obtained from dichotomous and polytomous items are not conducive to multivariate normality.

Ideally, data derived from an ordinal scale should be analyzed using estimation methods that are designed for use with such data. Weighted Least Squares (WLS, Jöreskog

Kim H. Koh is Assistant Professor, Centre for Research in Pedagogy and Practice, National Institute of Education. Email: khkoh@nie.edu.sg. Bruno D. Zumbo is Professor of Measurement, Evaluation and Research Methodology, as well as member of the Department of Statistics and the Institute of Applied Mathematics. Email him at: bruno.zumbo@ubc.ca. An earlier version of this article was presented at the 2007 American Educational Research Association (AERA) conference.

& Sörbom, 1996), Asymptotic Distribution Free (ADF, Browne, 1984), or Robust Maximum Likelihood estimation of model parameters using the polychoric correlation and asymptotic covariance matrix is theoretically sound for MGCFA with ordinal and mixed item format data. Practitioners, however, seldom use these methods. The implicit reasoning appears to be two-fold: (a) there is lack of awareness of these relatively new methods, and (b) these new methods are understood to require large sample sizes; larger than ones found in many research settings, and are, generally, not computationally viable with tests or measures involving more than 25 items[1].

Consequently, the ordinal-scaled data are often treated as if they were continuous and analyzed with the normal theory Maximum Likelihood (ML) estimation method and Pearson covariance matrix. The purpose, therefore, of this study was to investigate the statistical properties of the maximum likelihood factor analysis of a Pearson covariance matrix for testing measurement invariance hypotheses in MGCFA with mixed item format data. Specifically, the study examined the effects of mixed item formats and sample size combinations on the Type I error rates of ML-based chi-square difference tests for two commonly investigated measurement invariance hypotheses, namely strong and full invariance.

To be clear, we are not advocating using a Pearson covariance matrix for testing measurement invariance with mixed item formats, but rather we are interested in investigating: (a) what happens to the Type I error rates for those researchers who continue to choose to use these sub-optimal methods, and (b) the empirical Type I error rate of the extant research literature that used these sub-optimal methods (before the more optimal ones were widely available) for measurement invariance. We are also not advocating for the exclusive use of hypothesis testing in this context. Our aim is to reflect common research and applied measurement practice (both in terms of the methods used and the type of data) and hence to document the Type I error rates that one would find in these applied settings. This matter of keeping an eye on everyday research practice will come up again in the Methods Section when we describe the various hypothesis tests we are investigating.

Theoretical Framework

The fundamental idea underlying the measurement models in MGCFA is the use of a set of observable variables (i.e., items) to represent the latent variable(s). When the ordinal-scaled items are used as proxies for the latent continuous variable(s), the assumptions of interval measurement scale and multivariate normality are violated. Measurement errors induced by a crude categorization of the latent continuous variables can lead to the violations of the covariance structure. Because the Pearson covariance is attenuated in the ordinal variables, the covariance structure model may not hold for the observed variables. Therefore, ML estimation based on the distorted sample covariance matrix is likely to be biased.

When ordinal data are used with the ML estimation method and Pearson covariance matrix in single-group confirmatory factor

[1] The WLS/ADF estimation method requires relatively large sample sizes (i.e., at least 2,000-5,000 observations per group, Browne, 1984) to alleviate problems due to convergence or improper solutions and is not a viable method for models with a large number of items. Also, diagonally weighted least squares with the corresponding asymptotic covariance matrix and the polychoric (or tetrachoric) covariance matrix is limited due to the fact that no more than 25 items can be used due to the excessive computer memory demands with the so-called weight matrix, i.e., asymptotic covariance matrix of the vectorized elements of the observed covariance matrix. With p variables there are L elements in the same covariance matrix, and the weight matrix is of order LxL, where L=(p(p+1))/2. Therefore, as an example, for a model that has 20 items, the weight matrix would contain 22,155 distinct elements and for 25 items the weight matrix would contain 52,975 distinct elements. Likewise, the Satorra-Bentler corrected chi-square in LISREL and Muthen's estimation method for ordered categorical data in the software Mplus are also limited by the large number of items that are found in large-scale educational measurement. Therefore, most applied research in MGCFA has ordinal or mixed item format data with small sample sizes and large numbers of items, therefore these computational and statistical restrictions prevent many applied researchers from using the WLS/ADF estimation method.

analysis, the chi-square goodness of fit statistic is inflated due to departures from multivariate normality in the observed variables, albeit negligible bias is found in the model parameter estimates (e.g., Hutchinson & Olmos, 1998; Muthén & Kaplan, 1992; Potthast 1993; Rigdon & Ferguson, 1991). Hence, using the ML chi-square statistic as a formal test statistic of model-data fit under the conditions of multivariate nonnormality leads to an inflated Type I error rate for rejecting a true model.

Methods

Simulation data focused on the situation wherein one has a test with a mixture of dichotomously and polytomously scored items. The design variables were three conditions of mixed item formats and six sample size combinations, resulting in a 3 × 6 factorial design with 18 cells in our simulation experimental design. Within each cell, 100 replications were generated.

A 30 item test was simulated with mixed item formats that were varied according to the proportions of dichotomous and polytomous items as follows:

A. 67% (20) dichotomous items and 33% (10) polytomous items (3 scale points),

B. 50% (15) dichotomous items and 50% (15) polytomous items (3 scale points), and

C. 33% (10) dichotomous items and 67% (20) polytomous items (3 scale points).

These item format proportions reflect the real achievement assessment data found in educational testing contexts such as the Trends in International Mathematics and Science Study (TIMSS) and the National Assessment of Educational Progress (NAEP). Given that most of the achievement data, when partial scores are allotted, use 3-category polytomous items, the polytomous items in the simulation were limited to item responses with 3 scale points.

The sample size combinations consisted of equal and unequal sample sizes for the two groups: 200 vs. 200; 500 vs. 500; 800 vs. 800; 200 vs. 500; 200 vs. 800; and 500 vs. 800. These were the typical sample sizes across two groups used with the ML estimation method and Pearson covariance matrix in MGCFA applied research.

Simulation Procedure

For unidimensional dichotomous items, the item responses were generated from the three-parameter logistic (3PL) item response theory model (Birnbaum, 1968),

$$P_i(\theta) = c_i + \frac{(1-c_i)}{1+\exp[-1.7a_i(\theta-b_i)]},$$

where a_i, b_i and c_i are the item i discrimination, difficulty, and guessing parameters, respectively. The $P_i(\theta)$ denotes the probability of answering correctly to item i by a randomly selected examinee with ability θ. The 3PL item parameters a, b, and c of each of the 20 dichotomous items were real item parameter estimates taken from the 1999 *TIMSS Mathematics Achievement Test*.

Using a random number generator to produce numbers uniformly distributed on the interval [0,1], the probabilities were converted to either 0s or 1s to reflect examinee item scores. When the random number selected was less than or equal to $P_i(\theta)$, a 1 was assigned to an examinee for item i, and a 0 otherwise (Hambleton & Rovinelli, 1986).

For the polytomously scored items, the generalized partial credit model (GPCM)(Muraki, 1992) was used to generate unidimensional polytomous item responses, which were categorized into r_i+1 ordered score categories (0, 1, …, r_i) for i-th item. The model states that the probability of getting item score $U_j=q$ for a randomly sampled examinee with ability θ to the i-th item is given by

$$P_{i,q}(\theta) = \Pr ob(U_i = q|\theta) =$$

$$\frac{\exp[\Sigma_{v=0}^{q} 1.7a_i(\theta-b_i+d_{iv})]}{\Sigma_{j=0}^{ri} \exp[\Sigma_{v=0}^{j} 1.7a_i(\theta-b_i+d_{iv})]},$$

$$q = 0,1,\ldots, r_i,$$

where a_i is the slope parameter of item i; b_i is the location parameter of item i; and d_{iv} are a set of threshold parameters of item i with associated constrains d_{i0}= 0 and

$\sum_{v=1}^{r_i} d_{iv} =$ 0 (Muraki, 1992). A total of 20 polytomous item parameters (*a*s, *b*s, *d*s) were obtained from the TIMSS data.

The approach described by González-Romá, Hernández & Gómez-Benito (2002) was used to generate ordered polytomous items. For each examinee, a latent trait estimate θ was generated from a standard normal distribution, *N*(0,1). The GPCM probabilities were summed across categories to create a cumulative probability for each score level, and then the probability of responding above category k [$P_k^*(\theta)$] was computed. For each simulated item and examinee a single random number (*u*) was randomly sampled from a uniform distribution over the interval [0,1], and the item scores were assigned as follows:

$$\text{k} = 3 \text{ if } P_2^*(\theta) \geq u$$
$$\text{k} = 2 \text{ if } P_2^*(\theta) < u \leq P_1^*(\theta)$$
$$\text{k} = 1 \text{ if } P_1^*(\theta) < u\,.$$

Two population data were simulated with equivalent parameters to represent measurement invariance. The population data consisted of 20 dichotomous and 20 polytomous items. Data sets with different proportions of dichotomous and polytomous items were then created by a random selection of the items from the first two population data. As can be seen in Table 1, the item response distributions across groups for each of the mixed item format conditions were only slightly negatively skewed.

Table 1: Mean Skewness of the Mixed Item Format Population Data

Mixtures of Item Formats	Mean Skewness
67% Dichotomous and 33% Polytomous Items	-0.39
50% Dichotomous and 50% Polytomous Items	-0.44
33% Dichotomous and 67% Polytomous Items	-0.40

Testing for Measurement Invariance Hypotheses

Three MGCFA nested models were used for the testing of the strong and full measurement invariance hypotheses. Model 1 served as a baseline model where no parameters were constrained between groups. The baseline model was properly specified and hence model misspecification was not a condition in the study. The first chi-square value was obtained from the baseline model for comparison with more constrained models. In Model 2 (i.e., strong measurement invariance model), the number of factors and factor loadings were constrained to be equal across groups. The number of factors, factor loadings, and error variances were constrained to equality across groups in Model 3 (i.e., full measurement invariance model). The tenability of an invariance hypothesis is determined by the statistical significance of the chi-square difference test between two nested models. A non-significant chi-square difference test statistic (e.g., baseline model versus full measurement invariance model) indicates that the full measurement invariance hypothesis is tenable.

It should be noted that, with an eye toward reflecting what goes on in research practice, we did not test for the equality of intercepts -- and hence we did not use a mean and covariance structure (MACS) model (Wu, Li, & Zumbo, 2007). That is, even though there has been periodic advocacy for testing for equality of intercepts it has been largely neglected in applied measurement practice. A thorough review of empirical tests of measurement invariance in applied psychology by Vandenberg and Lance (2000) revealed that although 99% of the studies that they had reviewed investigated loading invariance, only 12% investigated intercept equality and 49% investigated residual variance equality. Therefore by not using the MACS model and not testing intercepts we are not advocating that one ignore intercept equality but rather we are aiming to reflect common research practice. In short, we want our empirical Type I error rates from our simulation study to reflect those error rates in the research literature and in practice.

Estimation Method

The MGCFA was conducted by using the Pearson product moment covariance matrices along with the normal theory ML estimation method in the LISREL 8.53.

Dependent Variables

For each combination of the conditions, MGCFA was conducted for testing the two hypotheses of measurement invariance. Effects of mixed item formats and sample size combinations on the tests of hypotheses of measurement invariance were analyzed through the mean rejection rates of the true models (Type I error rates).

Results

A quality check on the simulated data was conducted by testing the full and strong measurement invariance hypotheses at the population level for each mixed item format combination. As can be seen in Table 2, the differences in chi-squares between models, that is, baseline vs. full invariance, and baseline vs. strong invariance are not statistically significant at the alpha level of .05. The results indicate that the factor structure of the artificial achievement test is invariant across groups. Thus, any sample data drawn from the population data are expected to yield equivalent factor structures for the two groups in the MGCFA framework.

The results in Table 3 show that the empirical rejection rates of the ML chi-square difference test have the nominal alpha (.05) that fall within their two-tailed confidence interval (at a Bonferroni corrected confidence interval of 99%) for the full and strong measurement invariance hypotheses across mixed item formats and sample size combinations. This indicates that mixed item formats and sample size combinations do not affect the empirical Type I error rates of the ML chi-square difference tests in the hypotheses testing of full and strong measurement invariance. Keep in mind that the item response distributions across groups are not very skewed.

Conclusion

The findings of the current study suggest that the practice of using multi-group confirmatory maximum likelihood factor analysis of a Pearson covariance matrix to test measurement invariance hypotheses with mixed item format data does not lead to inflated chi-square difference test statistics. These findings are certainly welcome news for someone reading and reviewing the extant research literature and research reports. However, although these are positive findings, we encourage researchers to use methods that treat the data as ordinal (e.g., polychoric matrices or perhaps full-information methods) and to test for the equality of intercepts. Our results lead us to conclude that although common practice is, in a sense, sub-optimal it at least is not leading to a tendency to over-claim differences in measurement scales across groups – i.e., an inflated Type I error rate.

[The reference list can be found after the subsequent tables.]

Table 2: Maximum Likelihood Chi-square Goodness-of-Fit Statistics between Models

Mixed Item Format	*Model*	*Chi-square Difference Statistic*	*P*
67% Dichotomous Items 33% Polytomous Items (20:10)	Baseline vs. Full Invariance	$\Delta\chi^2 = 32$, $\Delta df = 60$	1.00
	Baseline vs. Strong Invariance	$\Delta\chi^2 = 21$, $\Delta df = 30$	.89
50% Dichotomous Items 50% Polytomous Items (15:15)	Baseline vs. Full Invariance	$\Delta\chi^2 = 38$, $\Delta df = 60$	.99
	Baseline vs. Strong Invariance	$\Delta\chi^2 = 23$, $\Delta df = 30$	.82
33% Dichotomous Items 67% Polytomous Items (10:20)	Baseline vs. Full Invariance	$\Delta\chi^2 = 39$, $\Delta df = 60$	.98
	Baseline vs. Strong Invariance	$\Delta\chi^2 = 23$, $\Delta df = 30$	.82

Note: Numbers of dichomotous and polytomous items are in parentheses.

Table 3: Empirical Type I Error Rates of ML Chi-square Difference Test for the Full and Strong Measurement Invariance Hypotheses Across Mixed Item Formats and Sample Size Combinations

Sample Sizes (n1: n2)	*Hypothesis*	*Mixed Item Formats*		
		67% Dichotomous 33% Polytomous	*50% Dichotomous 50% Polytomous*	*33% Dichotomous 67% Polytomous*
200 : 200	FI	.01	.02	.01
	SI	.00	.00	.00
500 : 500	FI	.00	.01	.00
	SI	.02	.01	.02
800 : 800	FI	.00	.01	.00
	SI	.01	.01	.00
200 : 500	FI	.00	.03	.00
	SI	.02	.00	.01
200 : 800	FI	.00	.03	.00
	SI	.00	.02	.00
500 : 800	FI	.00	.02	.02
	SI	.01	.01	.01

Note: Those empirical Type I error rates that have the nominal alpha (.05) outside of their two-tailed confidence interval (at a Bonferroni corrected confidence interval of 99%) would be in **bold font**. FI and SI denote Full and Strong Measurement Invariance Hypotheses, respectively.

References

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord and M. R. Novick (Eds.), *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Browne, M.W. (1984). Asymptotically distribution-free methods for the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology*, *37*, 62-83.

Drasgow, F., & Kanfer, R. (1985). Equivalence of psychological measurement in heterogeneous populations. *Journal of Applied Psychology*, *70*, 662-680.

González-Romá, V., Hernández, A., & Gómez-Benito, J. (2002). *An evaluation of the multiple-group mean and covariance structure analysis model for detecting differential item functioning in graded response items*. Paper presented at the International Test Commission (ITC) Conference on Computer-Based Testing and the Internet. Winchester, UK.

Hambleton, R. K., & Rovinelli, R. J. (1986). Assessing the dimensionality of a set of test items. *Applied Psychological Measurement, 10*, 287-302.

Horn, J. L., & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research, 18*, 117-144.

Hutchinson, S. R., & Olmos, A. (1998). Behavior of descriptive fit indexes in confirmatory factor analyses using ordered categorical data. *Structural Equation Modeling, 5,* 344-364.

Jöreskog, K. & Sörbom, D. (1996). *LISREL 8: User's reference guide*. Chicago, IL: Scientific Software International.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*, 159-176.

Muthén, B., & Kaplan, D. (1992). A comparison of some methodologies for the factor analysis for non-normal Likert variables: A note on the size of the model. *British Journal of Mathematical and Statistical Psychology, 45*, 19-30.

Potthast, M. J. (1993). Confirmatory factor analysis of ordered categorical variables with large models. *British Journal of Mathematical and Statistical Psychology, 46*, 273-286.

Rigdon, E. E., & Ferguson, Jr. (1991). The performance of the polychoric correlation coefficient and selected fitting functions in confirmatory factor analysis with ordinal data. *Journal of Marketing Research, Vol. XXVIII*, 491-497.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the MI literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3*, 4-69.

Wu. A. D., Li, Z., & Zumbo, B. D. (2007). Decoding the meaning of factorial invariance and updating the practice of multi-group confirmatory factor analysis: A demonstration with TIMSS data. *Practical Assessment, Research and Evaluation, 12*(3), 1-26.

Journal of Modern Applied Statistical Methods
November, 2008, Vol. 7, No. 2, 478-487

1538 – 9472/08/$95.00

An Optimum Allocation with a Family of Estimators Using Auxiliary Information in Sample Survey

Gajendra K. Vishwakarma Housila P. Singh
Vikram University, India

The problem of obtaining optimum allocation using auxiliary information in stratified random sampling. An optimum allocation with a family of estimators is obtained and its efficiency is compared with that of Neyman allocation based on Srivastava (1971) class of estimators and the optimum allocation suggested by Zaidi et al., (1989). It is shown that the proposed allocation is better in the sense having smaller variance compared to other optimum allocation.

Key words: Auxiliary variate, study variate, variance, optimum allocation, stratified random sampling.

Introduction

When a population contains heterogeneity among units in terms of value, survey users are advised to form several homogeneous groups, and the sampling design is known as stratified sampling. All designs, other than these, are generated as a further modification of simple random sampling and stratified sampling. Stratification is one of the most widely used techniques in sample survey design due to its dual purposes of providing samples that are representative of major sub-groups of the population and increasing the precision of estimators. It is also well established that the auxiliary information may lead to more efficient estimators: ratio, product and regression methods of estimation are examples in this context. This article suggests a class of estimators using auxiliary information in stratified random sampling and discusses its properties.

Let *y* be the study variate and *x* be the auxiliary variate, let the population $U=\left(U_1,U_2,U_3,...,U_N\right)$ of size N be divided into *L* stratum, and let N_h and n_h be the total number of units and sample size respectively in h^{th} stratum, such that $\sum_{h=1}^{L}N_h=N$ and $\sum_{h=1}^{L}n_h=n$. Next, let $\left(y_{hj},\,x_{hj}\right)$ be the pair of values according to the variate under study *y* and the auxiliary variate *x* respectively for j^{th}-unit $\left(j=1,2,3,...,N_h\right)$ in the h^{th} sample of size n_h selected by simple random sampling from the h^{th} stratum $\left(j=1,2,3,...,N_h;\,h=1,2,3,...,L\right)$. For simplicity, assume that N_h is large enough compared to n_h so that $f_h=\frac{n_h}{N_h}\approx 0$. Denote

$$\bar{Y}=\sum_{h=1}^{L}W_h\bar{Y}_h,\ \bar{X}=\sum_{h=1}^{L}W_h\bar{X}_h,$$

$$\bar{Y}_h=\frac{1}{N_h}\sum_{j=1}^{N_h}y_{hj},\ \ \bar{X}_h=\frac{1}{N_h}\sum_{j=1}^{N_h}x_{hj}$$

$$\bar{y}_h=\frac{1}{n_h}\sum_{j=1}^{n_h}y_{hj},\ \ \bar{x}_h=\frac{1}{n_h}\sum_{j=1}^{n_h}x_{hj},$$

$$\bar{y}_{st}=\sum_{h=1}^{L}W_h\bar{y}_h,\ \bar{x}_{st}=\sum_{h=1}^{L}W_h\bar{x}_h,$$

Address correspondence to Gajendra K. Vishwakarma, School of Studies in Statistics, Vikram University, Ujjain - 456010, M.P., India. E-mail: vishwagk@rediffmail.com.

$$W_h = \frac{N_h}{N},\ S_{yh}^2 = \frac{1}{(N_h - 1)} \sum_{j=1}^{N_h} \left(y_{hj} - \overline{Y}_h\right)^2,$$

$$S_{xh}^2 = \frac{1}{(N_h - 1)} \sum_{j=1}^{N_h} \left(x_{hj} - \overline{X}_h\right)^2,$$

$$S_{xyh} = \frac{1}{(N_h - 1)} \sum_{j=1}^{N_h} \left(y_{hj} - \overline{Y}_h\right)\left(x_{hj} - \overline{X}_h\right)$$

$$s_{yh}^2 = \frac{1}{(n_h - 1)} \sum_{j=1}^{n_h} \left(y_{hj} - \overline{y}_h\right)^2,$$

$$s_{xh}^2 = \frac{1}{(n_h - 1)} \sum_{j=1}^{n_h} \left(x_{hj} - \overline{x}_h\right)^2,$$

$$s_{xyh} = \frac{1}{(n_h - 1)} \sum_{j=1}^{n_h} \left(x_{hj} - \overline{x}_h\right)\left(y_{hj} - \overline{y}_h\right)$$

$$\mu_{rsh} = \frac{1}{N_h} \sum_{j=1}^{N_h} \left(y_{hj} - \overline{Y}_h\right)^r \left(x_{hj} - \overline{X}_h\right)^s,$$

$$\rho_h = \frac{S_{xyh}}{S_{yh}\, S_{xh}},\ r_h = \frac{s_{xyh}}{s_{yh}\, s_{xh}}$$

$$C_{yh}^2 = \frac{S_{yh}^2}{\overline{Y}_h^2},\ C_{xh}^2 = \frac{S_{xh}^2}{\overline{X}_h^2},\ \lambda_{rsh} = \frac{\mu_{rsh}}{\left(\mu_{20h}^r\, \mu_{02h}^s\right)^{1/2}}$$

$$a_h = \frac{\overline{x}_h}{\overline{X}_h}, b_h = \frac{s_{xh}^2}{S_{xh}^2} \text{ and } c_h = \frac{r_h}{\rho_h}.$$

Writing,

$$e_{0h} = \frac{\overline{y}_h}{\overline{Y}_h} - 1 = (t_h - 1),$$

$$e_{1h} = \frac{\overline{x}_h}{\overline{X}_h} - 1 = (a_h - 1),$$

$$\eta_{0h} = \frac{s_{yh}^2}{S_{yh}^2} - 1,\ \eta_{1h} = \frac{s_{xh}^2}{S_{xh}^2} - 1 = (b_h - 1)$$

and $\delta_h = \dfrac{r_h}{\rho_h} - 1 = (c_h - 1)$

results in,

$$E(e_{0h}) = E(e_{1h}) = E(\eta_{0h}) = E(\eta_{1h}) = 0,$$

$$E(\delta_h) = \frac{1}{n_h}[3\rho_h(\lambda_{40h} + \lambda_{04h}) - 4(\lambda_{31h} + \lambda_{13h}) + 2\rho_h\lambda_{22h}]8\rho_h,$$

$$E\left(e_{0h}^2\right) = \frac{1}{n_h} C_{yh}^2,\ E\left(e_{1h}^2\right) = \frac{1}{n_h} C_{xh}^2,$$

$$E\left(\eta_{0h}^2\right) = \frac{1}{n_h}(\lambda_{40h} - 1), E\left(\eta_{1h}^2\right) = \frac{1}{n_h}(\lambda_{04h} - 1),$$

$$E\left(\delta_h^2\right) = \frac{1}{n_h}[D_h], E(e_{0h}e_{1h}) = \frac{\lambda_{30h}}{n_h} C_{yh},$$

$$E(e_{0h}\eta_{1h}) = \frac{\lambda_{12h}}{n_h} C_{yh},\ E(e_{1h}\eta_{0h}) = \frac{\lambda_{21h}}{n_h} C_{xh},$$

$$E(e_{1h}\eta_{1h}) = \frac{\lambda_{03h}}{n_h} C_{xh},\ E(e_{0h}\delta_h) = \frac{A_{0h}}{n_h} C_{yh},$$

$$E(e_{1h}\delta_h) = \frac{A_{1h}}{n_h} C_{xh},\ E(\eta_{0h}\delta_h) = \frac{B_{0h}}{n_h},$$

$$E(\eta_{1h}\delta_h) = \frac{B_{1h}}{n_h}, E(\eta_{0h}\eta_{1h}) = \frac{1}{n_h}(\lambda_{22h} - 1),$$

where,

$$\begin{aligned} D_h &= [\rho_h^2(\lambda_{40h} + \lambda_{04h}) \\ &\quad - 4\rho_h(\lambda_{31h} + \lambda_{13h}) \\ &\quad + 2\left(2 + \rho_h^2\right)\lambda_{22h}]/4\rho_h^2 \\ A_{0h} &= [2\lambda_{21h} - \rho_h(\lambda_{12h} + \lambda_{30h})]/2\rho_h \\ A_{1h} &= [2\lambda_{12h} - \rho_h(\lambda_{21h} + \lambda_{03h})]/2\rho_h \\ B_{0h} &= [2\lambda_{31h} - \rho_h(\lambda_{40h} + \lambda_{22h})]/2\rho_h \\ B_{1h} &= [2\lambda_{13h} - \rho_h(\lambda_{04h} + \lambda_{22h})]/2\rho_h. \end{aligned}$$

Using this background and following Srivastava (1971) a family of estimators of population mean $\overline{Y}$ may be defined as

$$\hat{\overline{Y}}_q = \sum_{h=1}^{L} W_h \overline{y}_h q_h(a_h), \qquad (1)$$

where $q_h(.)$ is a function of (a_h) such that $q_h(1) = 1$ and satisfies certain regularity conditions similar to those given by Srivastava (1971).

To the first degree of approximation, the variance of $\hat{\overline{Y}}_q$ is given by

$$V\left(\hat{\overline{Y}}_q\right) = \sum_{h=1}^{L} W_h^2\, \overline{Y}_h^2\, \frac{1}{n_h}\left[C_{yh}^2 + C_{xh}^2\, q_{h1}^2(1) + 2\rho_h C_{xh} C_{yh} q_{h1}(1)\right] \qquad (1.2)$$

which is minimized for

$$q_{h1}(1) = -\rho_h \frac{C_{yh}}{C_{xh}} \qquad (1.3)$$

Thus, the resulting minimum variance of $\hat{\bar{\bar{Y}}}_q$ is given by

$$\min . V\left(\hat{\bar{\bar{Y}}}_q\right) = \sum_{h=1}^{L} W_h^2 \frac{1}{n_h} S_{yh}^2 (1-\rho_h^2) \qquad (1.4)$$

Following Srivasrava and Jhajj (1981), Zaidi et. al. (1989) suggested a class of estimators of population mean $\bar{Y}$ as

$$\hat{\bar{\bar{Y}}}_t = \sum_{h=1}^{L} W_h \bar{y}_h t_h (a_h, b_h) \qquad (1.5)$$

where t_h (.) is a function of $(a_h,\, b_h)$ such that $t_h(1,\,1) = 1$, which satisfies certain regularity conditions similar to those given by Srivastava and Jhajj (1981).

To the first degree of approximation the variance of $\hat{\bar{\bar{Y}}}_t$ is given by

$$\begin{aligned} V\left(\hat{\bar{\bar{Y}}}_t\right) = \sum_{h=1}^{L} W_h^2\, \bar{Y}_h^2 \frac{1}{n_h} \big[& C_{yh}^2 + C_{xh}^2\, t_{h1}^2(1,\,1) \\ & + (\lambda_{04h} - 1) t_{h2}^2 (1,\,1) \\ & + 2\rho_h C_{xh} C_{yh} t_{h1}(1,\,1) \\ & + 2\lambda_{12h} C_{yh} t_{h2}(1,\,1) \\ & + 2 C_{xh} \lambda_{03h} t_{h1}(1,\,1)\, t_{h2}(1,\,1) \big] \end{aligned} \qquad (1.6)$$

which is minimized for

$$\left.\begin{aligned} t_{h1}(1,\,1) &= \frac{C_{yh}\left[\lambda_{12h}\lambda_{03h} - \rho_h(\lambda_{04h}-1)\right]}{C_{xh}\left[\lambda_{04h} - \lambda_{03h}^2 - 1\right]} \\ t_{h2}(1,\,1) &= \frac{C_{yh}\left[\rho_h\lambda_{03h} - \lambda_{12h}\right]}{\left[\lambda_{04h} - \lambda_{03h}^2 - 1\right]} \end{aligned}\right\} \qquad (1.7)$$

and the minimum variance of $\hat{\bar{\bar{Y}}}_t$ is given by

$$\min . V\left(\hat{\bar{\bar{Y}}}_t\right) = \sum_{h=1}^{L} W_h^2 \frac{S_{yh}^2}{n_h} \left[(1-\rho_h^2) - \frac{(\rho_h \lambda_{03h} - \lambda_{12h})^2}{(\lambda_{04h} - \lambda_{03h}^2 - 1)} \right] \qquad (1.8)$$

The crux of this article is to suggest an optimum allocation with a family of estimators considered by Srivastava and Jhajj (1983) and compares its efficiency with that of Neyman allocation and others. It is seen that the proposed allocation is better in the sense of having lesser variance than other.

The Suggested Family of Estimators

Whatever the sample chosen, let (a_h, b_h, c_h) assume values in a bounded closed convex subset, R of the three dimensional real space containing the point $(1,\,1,\,1)$. Let $g_h(a_h,\, b_h,\, c_h)$ be the function of a_h, b_h and c_h, such that $g_h(1,\,1,\,1) = 1$, and satisfies the following conditions:

1. In R, the function $g_h(a_h,\, b_h,\, c_h)$ is continuous and bounded.
2. The first and second partial derivatives of $g_h(a_h,\, b_h,\, c_h)$ exist and are continuous and bounded.

Define a family of estimators for population mean $\bar{Y}$ as

$$\hat{\bar{\bar{Y}}}_g = \sum_{h=1}^{L} W_h \bar{y}_h g_h (a_h, b_h, c_h) \qquad (2.1)$$

Expanding $g_h(a_h,\, b_h,\, c_h)$ about the point $(1,\,1,\,1)$ in a second order Taylor's series and noting that the second partial derivatives of *g* are bounded. We have

$$E\left(\hat{\bar{\bar{Y}}}_g\right) = \bar{Y} + 0\left(n^{-1}\right),$$

so that bias of $\hat{\bar{\bar{Y}}}_g$ is of the order of n^{-1}. Thus, to the first degree of approximation the variance of $\hat{\bar{\bar{Y}}}_g$ is given by

$$V\left(\hat{\bar{\bar{Y}}}_g\right) = E\left(\hat{\bar{\bar{Y}}}_g - \bar{Y}\right)^2$$

$$\begin{aligned}= \sum_{h=1}^{L} W_h^2\, \overline{Y}_h^2 \frac{1}{n_h}\Big[C_{yh}^2 + C_{xh}^2\, g_{h1}^2(1,\,1,\,1) \\ + (\lambda_{04h}-1) g_{h2}^2(1,\,1,\,1) \\ + D_h\, g_{h3}^2(1,\,1,\,1) \\ + 2\rho_h C_{xh} C_{yh} g_{h1}(1,\,1,\,1) \\ + 2C_{yh}\, \lambda_{12h} g_{h2}(1,\,1,\,1) \\ + 2C_{yh} A_{0h} g_{h3}(1,\,1,\,1) \\ + 2C_{xh}\lambda_{03h} g_{h1}(1,\,1,\,1) g_{h2}(1,\,1,\,1) \\ + 2C_{xh} A_{1h} g_{h1}(1,\,1,\,1) g_{h3}(1,\,1,\,1) \\ + 2B_{1h} g_{h2}(1,\,1,\,1) g_{h3}(1,\,1,\,1)\Big] \qquad (2.2)\end{aligned}$$

where $g_{h1}(1,\,1,\,1)$, $g_{h2}(1,\,1,\,1)$ and $g_{h3}(1,\,1,\,1)$ denote the first order partial derivates of $g_h(a_h,\, b_h,\, c_h)$ at the point $(1,\,1,\,1)$. Differentiating (2.2) partially with respect to $g_{h1}(.)$, $g_{h2}(.)$ and $g_{h3}(.)$, and equating them to zero the following equations

$$\begin{bmatrix} C_{xh}^2 & C_{xh}\lambda_{03h} & C_{xh}A_{1h} \\ C_{xh}\lambda_{03h} & (\lambda_{04h}-1) & B_{1h} \\ C_{xh}A_{1h} & B_{1h} & D_h \end{bmatrix} \begin{bmatrix} g_{h1}(.) \\ g_{h2}(.) \\ g_{h3}(.) \end{bmatrix} = -C_{hy} \begin{bmatrix} \rho_h C_{xh} \\ \lambda_{12h} \\ A_{0h} \end{bmatrix} \qquad (2.3)$$

Solving (2.3), the optimum values of $g_{h1}(.)$, $g_{h2}(.)$ and $g_{h3}(.)$ were obtained respectively as

$$\begin{aligned} g_{1h}(1,\,1,\,1) = \frac{C_{yh}}{K_h C_{xh}} \big[\{\lambda_{12h}\lambda_{03h} - \rho_h(\lambda_{04h}-1)\} D_h \\ + \{(\lambda_{04h}-1)A_{1h} - \lambda_{03h}B_{1h}\} A_{0h} \\ - (\lambda_{12h}A_{1h} - \rho_h B_{h1}) B_{1h}\big] \end{aligned}$$

$$\begin{aligned} g_{h2}(1,\,1,\,1) = \frac{C_{yh}}{K_h} \big[(\rho_h\lambda_{03h} - \lambda_{12h}) D_h \\ + (\lambda_{03h}A_{1h} - B_{1h}) A_{0h} \\ + (\lambda_{12h}A_{1h} - \rho_h B_{h1}) A_{1h}\big] \end{aligned}$$

$$\begin{aligned} g_{h3}(1,\,1,\,1) = \frac{C_{yh}}{K_h} \big[\{\rho_h(\lambda_{04h}-1) - \lambda_{12h}\lambda_{03h}\} A_{h1} \\ - (\lambda_{04h} - \lambda_{03h}^2 - 1) A_{0h} \\ - (\rho_h\lambda_{03h} - \lambda_{1h2}) B_{1h}\big] \end{aligned}$$

where,

$$\begin{aligned} K_h = \big[(\lambda_{04h} - \lambda_{03h}^2 - 1) D_h - (\lambda_{04h}-1) A_{1h}^2 \\ + 2\lambda_{03h} A_{1h} B_{1h} - B_{1h}^2\big] \end{aligned}$$

Thus, the minimum variance of $\left(\hat{\overline{Y}}_g\right)$ is given by

$$\begin{aligned} \min . V\left(\hat{\overline{Y}}_g\right) = \sum_{h=1}^{l} W_h^2 \frac{S_{yh}^2}{n_h} \Big[(1-\rho_h^2) \\ - \frac{(\rho_h\lambda_{03h} - \lambda_{12h})^2}{(\lambda_{04h} - \lambda_{03h}^2 - 1)} \\ - \frac{\{G_h\}^2}{K_h(\lambda_{04h} - \lambda_{03h}^2 - 1)}\Big] \qquad (2.4) \end{aligned}$$

where,

$$\begin{aligned} G_h = (\lambda_{12h}\lambda_{03h} - \rho_h\lambda_{04h} + \rho_h) A_{1h} \\ + (\lambda_{04h} - \lambda_{03h}^2 - 1) A_{0h} + (\rho_h\lambda_{03h} - \lambda_{12h}) B_{1h} \end{aligned}$$

In (2.4), the first term on the right hand side gives the minimum asymptotic variance of the family when only $\overline{X}_h$ is used, and the first two terms give the minimum asymptotic variance when both $\overline{X}_h$ and S_{hx}^2 are used. The third term gives the reduction in asymptotic variance when ρ_h is also used along with $\overline{X}_h$ and S_{hx}^2.

Efficiency Comparisons

It is known that the variance of usual unbiased estimators in stratified sampling under SRSWOR is

$$V(\overline{y}_{st}) = \sum_{h=1}^{L} W_h^2 \frac{S_{yh}^2}{n_h} \qquad (3.1)$$

From (1.4) and (3.1) the following results

$$V(\overline{y}_{st}) - \min . V\left(\hat{\overline{Y}}_q\right) = \sum_{h=1}^{L} W_h^2 \rho_h^2 \frac{S_{yh}^2}{n_h} \geq 0 \qquad (3.2)$$

which, in turn, yields the inequality

$$\min . V\left(\hat{\overline{Y}}_q\right) \leq V(\overline{y}_{st}) \qquad (3.3)$$

From (1.4) and (1.8)

$$\min.V\left(\hat{\bar{Y}}_q\right) - \min.V\left(\hat{\bar{Y}}_t\right)$$

$$= \sum_{h=1}^{L} W_h^2 \frac{S_{yh}^2}{n_h} \frac{\{\rho_h \lambda_{03h} - \lambda_{12h}\}^2}{K_h(\lambda_{04h} - \lambda_{03h}^2 - 1)} \geq 0 \qquad (3.4)$$

which gives the inequality

$$\min.V\left(\hat{\bar{Y}}_t\right) \leq \min.V\left(\hat{\bar{Y}}_q\right) \qquad (3.5)$$

Further from (1.8) and (2.4)

$$\min.V\left(\hat{\bar{Y}}_t\right) - \min.V\left(\hat{\bar{Y}}_g\right)$$

$$= \sum_{h=1}^{L} W_h^2 \frac{S_{yh}^2}{n_h} \frac{\{G_h\}^2}{K_h(\lambda_{04h} - \lambda_{03h}^2 - 1)} \geq 0 \qquad (3.6)$$

which gives the inequality

$$\min.V\left(\hat{\bar{Y}}_g\right) \leq \min.V\left(\hat{\bar{Y}}_t\right) \qquad (3.7)$$

Thus from (3.3), (3.5) and (3.7) we have

$$\min.V\left(\hat{\bar{Y}}_g\right) \leq \min.V\left(\hat{\bar{Y}}_t\right) \leq \min.V\left(\hat{\bar{Y}}_q\right) \leq V(\bar{y}_{st}) \qquad (3.8)$$

It follows from (3.8) that the proposed estimator $\hat{\bar{Y}}_g$ is better than $\bar{y}_{st}$, $\hat{\bar{Y}}_q$ and $\hat{\bar{Y}}_t$ at its optimum conditions.

Optimum Allocation

The variance of $\bar{y}_{st}$ under the Neyman allocation

$$n_h = n \frac{W_h S_{yh}}{\sum_{h=1}^{L} W_h S_{yh}} \qquad (4.1)$$

$$V(\bar{y}_{st})_N = \frac{1}{n}\left(\sum_{h=1}^{L} W_h S_{yh}\right)^2 \qquad (4.2)$$

To minimize $\min.V\left(\hat{\bar{Y}}_q\right)$, $\min.V\left(\hat{\bar{Y}}_t\right)$ and $\min.V\left(\hat{\bar{Y}}_g\right)$, consider the cost function

$$C^* = C_0 + \sum_{h=1}^{L} C_h n_h\,, \qquad (4.3)$$

where C_0 and C_h are the overhead cost and cost per unit within h^{th} stratum respectively, for the given cost restriction

$$C_1 n_1 + C_2 n_2 + ... + C_L n_L = C^* - C_0 \qquad (4.4)$$

Using Lagrange's method of multipliers, the optimum allocation in order to minimize $\min.V\left(\hat{\bar{Y}}_q\right)$, $\min.V\left(\hat{\bar{Y}}_t\right)$ and $\min.V\left(\hat{\bar{Y}}_g\right)$ respectively is

$$n_h = n \frac{W_h S_{yh}(1-\rho_h^2)^{1/2} \Big/ \sqrt{C_h}}{\sum_{h=1}^{L} W_h S_{yh}(1-\rho_h^2)^{1/2} \Big/ \sqrt{C_h}} \qquad (4.5)$$

$$n_h = n \frac{W_h S_{yh}\left\{(1-\rho_h^2) - \frac{(\rho_h\lambda_{03h} - \lambda_{12h})^2}{(\lambda_{04h} - \lambda_{03h}^2 - 1)}\right\}^{1/2} \Big/ \sqrt{C_h}}{\sum_{h=1}^{L} W_h S_{yh}\left\{(1-\rho_h^2) - \frac{(\rho_h\lambda_{03h} - \lambda_{12h})^2}{(\lambda_{04h} - \lambda_{03h}^2 - 1)}\right\}^{1/2} \Big/ \sqrt{C_h}} \qquad (4.6)$$

and

$$n_h = n \frac{W_h S_{yh}\left\{(1-\rho_h^2) - \frac{(\rho_h\lambda_{03h} - \lambda_{12h})^2}{(\lambda_{04h} - \lambda_{03h}^2 - 1)} - \frac{\{G_h\}^2}{(\lambda_{04h} - \lambda_{03h}^2 - 1)K_h}\right\}^{1/2} \Big/ \sqrt{C_h}}{\sum_{h=1}^{L} W_h S_{yh}\left\{(1-\rho_h^2) - \frac{(\rho_h\lambda_{03h} - \lambda_{12h})^2}{(\lambda_{04h} - \lambda_{03h}^2 - 1)} - \frac{\{G_h\}^2}{(\lambda_{04h} - \lambda_{03h}^2 - 1)K_h}\right\}^{1/2} \Big/ \sqrt{C_h}} \qquad (4.7)$$

In particular, if $C_h = C$ for the given cost function $C^* = C_0 + nC$, the optimum allocation (4.5), (4.6) and (4.7) respectively reduce to

$$n_h = n \frac{W_h S_{yh}(1-\rho_h^2)^{1/2}}{\sum_{h=1}^{L} W_h S_{yh}(1-\rho_h^2)^{1/2}} \qquad (4.8)$$

$$n_h = n\frac{W_h S_{yh}\left\{(1-\rho_h^2) - \dfrac{(\rho_h\lambda_{03h}-\lambda_{12h})^2}{(\lambda_{04h}-\lambda_{03h}^2-1)}\right\}^{1/2}}{\sum_{h=1}^{L} W_h S_{yh}\left\{(1-\rho_h^2) - \dfrac{(\rho_h\lambda_{03h}-\lambda_{12h})^2}{(\lambda_{04h}-\lambda_{03h}^2-1)}\right\}^{1/2}} \tag{4.9}$$

and

$$n_h = n\frac{W_h S_{hy}\left\{(1-\rho_h^2) - \dfrac{(\rho_h\lambda_{03h}-\lambda_{12h})^2}{(\lambda_{04h}-\lambda_{03h}^2-1)} - \dfrac{\{G_h\}^2}{(\lambda_{04h}-\lambda_{03h}^2-1)K_h}\right\}^{1/2}}{\sum_{h=1}^{L} W_h S_{hy}\left\{(1-\rho_h^2) - \dfrac{(\rho_h\lambda_{03h}-\lambda_{12h})^2}{(\lambda_{04h}-\lambda_{03h}^2-1)} - \dfrac{\{G_h\}^2}{(\lambda_{04h}-\lambda_{03h}^2-1)K_h}\right\}^{1/2}} \tag{4.10}$$

Substituting the values of n_h from (4.8), (4.9) and (4.10) respectively in (1.4), (1.8) and (2.4) the resulting variances of $\hat{\bar{Y}}_q$, $\hat{\bar{Y}}_t$ and $\hat{\bar{Y}}_g$ are

$$\min.V\left(\hat{\bar{Y}}_q\right)_O = \frac{1}{n}\left[\sum_{h=1}^{L} W_h S_{yh}(1-\rho_h^2)^{1/2}\right]^2 \tag{4.11}$$

$$\min.V\left(\hat{\bar{Y}}_t\right)_O = \frac{1}{n}\left[\sum_{h=1}^{L} W_h S_{yh}\{(1-\rho_h^2)\} - \frac{(\rho_h\lambda_{03h}-\lambda_{12h})^2}{(\lambda_{04h}-\lambda_{03h}^2-1)}\}^{1/2}\right]^2 \tag{4.12}$$

and

$$\min.V\left(\hat{\bar{Y}}_g\right)_O = \frac{1}{n}\left[\sum_{h=1}^{L} W_h S_{yh}\{(1-\rho_h^2) - \frac{(\rho_h\lambda_{03h}-\lambda_{12h})^2}{(\lambda_{04h}-\lambda_{03h}^2-1)} - \frac{\{G_h\}^2}{(\lambda_{04h}-\lambda_{03h}^2-1)K_h}\}^{1/2}\right]^2 \tag{4.13}$$

From (4.2), (4.11), (4.12) and (4.13) it can be easily proved that

$$\min.V\left(\hat{\bar{Y}}_g\right)_O \le \min.V\left(\hat{\bar{Y}}_t\right)_O \le \min.V\left(\hat{\bar{Y}}_q\right)_O \le V(\bar{y}_{st})_N, \tag{4.14}$$

which clearly indicates that the proposed optimum allocation is better than Neyman allocation $(\bar{y}_{st})$ and the optimum allocation based on Srivastava (1971) family of estimators and the optimum allocation envisaged by Zaidi et al., (1989) in the sense of having smaller variance.

Empirical Study

The performance of various families of estimators of the population mean $\bar{Y}$ through six natural population data sets has been illustrated.

To examine the performance of the estimators $\hat{\bar{Y}}_q$, $\hat{\bar{Y}}_t$ and $\hat{\bar{Y}}_g$ with respect to $\bar{y}_{st}$ under optimum allocation we have computed the percent relative efficiencies of *t* with respect to $\bar{y}_{st}$ using the formula,

$$PRE(t,\, \bar{y}_{st}) = \frac{V(\bar{y}_{st})_N}{\min V(t)_O} \times 100,$$

where t = $\hat{\bar{Y}}_q$, $\hat{\bar{Y}}_t$, $\hat{\bar{Y}}_g$; results are presented in Table 5.1.

Conclusion

Table 5.1 clearly indicates that the proposed family of estimator $\hat{\bar{Y}}_g$ is more efficient than the usual unbiased estimator $\bar{y}_{st}$, $\hat{\bar{Y}}_q$ and the Zaidi, et al. (1989) estimator, $\hat{\bar{Y}}_t$. Thus the proposed family of estimator $\hat{\bar{Y}}_g$ would be preferred over $\bar{y}_{st}$, $\hat{\bar{Y}}_q$ and $\hat{\bar{Y}}_t$.

Table 5.1: Percent Relative Efficiencies of $\hat{\bar{Y}}_q$, $\hat{\bar{Y}}_t$, and $\hat{\bar{Y}}_g$ with respect to $\bar{y}_{st}$

Population	$PRE\left(\hat{\bar{Y}}_q, \bar{y}_{st}\right)$	$PRE\left(\hat{\bar{Y}}_t, \bar{y}_{st}\right)$	$PRE\left(\hat{\bar{Y}}_g, \bar{y}_{st}\right)$
I	872.12	879.51	2308.29
II	351.30	367.04	690.30
III	420.66	496.89	571.88
IV	856.61	984.67	1746.53
V	615.88	727.70	1003.45
VI	147.64	242.84	362.15

Population I: Singh and Chaudhary (1986, p. 162)
y: total number of trees, x: area under orchards in ha.

$N = 25$, $L = 3$, $N_1 = 6$, $N_2 = 8$, $N_3 = 11$

Stratum	Values of parameters for h^{th} stratum					
No.	S_{yh}	ρ_h	λ_{12h}	λ_{21h}	λ_{03h}	λ_{30h}
1	273.45103	0.9215191	-0.2276668	-0.071714	-0.2400887	0.138323
2	509.03212	0.9737715	1.6980145	1.6304126	1.7646005	1.576411
3	256.6819	0.8826909	1.0289035	0.8472329	1.2344161	0.5897102

Stratum	Values of parameters for h^{th} stratum (continued)				
No.	λ_{22h}	λ_{04h}	λ_{40h}	λ_{13h}	λ_{31h}
1	1.2773905	1.3483853	1.5310737	1.239425	1.3741684
2	4.4920977	4.7537207	4.2700966	4.6186087	4.3727487
3	3.264646	4.3492128	2.684855	3.7646968	2.8334168

For illustration take $n = 10$, $n_1 = 3$, $n_2 = 3$, $n_3 = 4$

Population II: Singh and Mangat (1996, p. 194)
y: pocket money, x: annual income

$N = 27$, $L = 3$, $N_1 = 4$, $N_2 = 10$, $N_3 = 13$

Stratum	Values of parameters for h^{th} stratum					
No.	S_{yh}	ρ_h	λ_{12h}	λ_{21h}	λ_{03h}	λ_{30h}
1	225.46249	0.9527907	0.9817665	0.9616631	0.9637509	0.906753
2	108.14085	0.8074107	0.1045162	0.0851702	0.0745106	-0.0097243
3	98.871841	0.7621946	-0.1720774	-0.0129786	-0.0879664	-0.1103153

Stratum	Values of parameters for h^{th} stratum (continued)				
No.	λ_{22h}	λ_{04h}	λ_{40h}	λ_{13h}	λ_{31h}
1	2.1256188	2.1872063	2.1224402	2.1470526	2.1142848
2	1.4455092	1.7719919	2.1393301	1.484715	1.5986642
3	1.6145628	1.9933334	1.5608654	1.6582907	1.3338932

For illustration take $n = 10$, $n_1 = 2$, $n_2 = 4$, $n_3 = 5$

Population III: Singh and Mangat (1996, p. 207)
y: no. refrigerators sold in current year, x: no. refrigerators sold last summer

$N = 42$, $L = 4$, $N_1 = 14$, $N_2 = 9$, $N_3 = 12$, $N_4 = 7$

Stratum	Values of parameters for h^{th} stratum					
No.	S_{yh}	ρ_h	λ_{12h}	λ_{21h}	λ_{03h}	λ_{30h}
1	12.911576	0.7929927	-0.019159	0.3665704	-0.3717353	0.8009986
2	13.201431	0.8697081	0.4460543	0.402637	0.4681387	0.3062423
3	15.05344	0.9191256	-0.1618712	-0.2565663	-0.128619	-0.4344209
4	13.062123	0.9055795	0.2273419	-0.0915551	0.5905558	-0.3916206

Stratum	Values of parameters for h^{th} stratum (continued)				
No.	λ_{22h}	λ_{04h}	λ_{40h}	λ_{13h}	λ_{31h}
1	1.8121436	2.2006301	3.3060221	1.7701281	2.263858
2	1.5135141	2.2975185	1.6129147	1.7937746	1.4355898
3	1.928372	1.9632339	2.7733335	1.815768	2.2420385
4	1.7822884	2.4742281	1.9126016	2.0034381	1.7549122

For illustration take $n = 16$, $n_1 = 5$, $n_2 = 3$, $n_3 = 5$, $n_4 = 3$

Population IV: Singh and Mangat (1996, p. 212)
y: leaf area for newly developed strain of wheat, x: weight of leaves

$N = 39$, $L = 3$, $N_1 = 12$, $N_2 = 13$, $N_3 = 14$

Stratum	Values of parameters for h^{th} stratum					
No.	S_{yh}	ρ_h	λ_{12h}	λ_{21h}	λ_{03h}	λ_{30h}
1	6.3362112	0.9202367	0.429305	0.5097853	0.23599	0.5031633
2	5.5075918	0.9154022	0.9960984	0.815551	1.0341649	0.5847596
3	6.7413528	0.9668189	0.2057622	0.2971175	0.083846	0.3360654

Stratum	Values of parameters for h^{th} stratum (continued)				
No.	λ_{22h}	λ_{04h}	λ_{40h}	λ_{13h}	λ_{31h}
1	1.9123464	2.2748233	1.9394547	2.0257975	1.879711
2	2.970998	3.436904	2.9819269	3.0966741	2.9303901
3	2.5134376	2.8955496	2.3448986	2.6759523	2.3988602

For illustration take $n = 14$, $n_1 = 4$, $n_2 = 5$, $n_3 = 5$

Population V: Singh and Mangat (1996, p. 218)
y: juice quantity, x: weight of cane

$N = 25$, $L = 3$, $N_1 = 6$, $N_2 = 12$, $N_3 = 7$

Stratum	Values of parameters for h^{th} stratum					
No.	S_{yh}	ρ_h	λ_{12h}	λ_{21h}	λ_{03h}	λ_{30h}
1	8.9442719	0.9455626	0.576173	0.6492226	0.4598407	0.688919
2	15.05042	0.948196	0.9857208	0.9738854	0.9465183	0.9187277
3	10.965313	0.7532234	1.0354011	0.8915649	0.8581802	0.727283

Stratum	Values of parameters for h^{th} stratum (continued)				
No.	λ_{22h}	λ_{04h}	λ_{40h}	λ_{13h}	λ_{31h}
1	2.2641624	2.2865633	2.3437501	2.2586791	2.2886912
2	3.379509	3.2689734	3.792407	3.2777466	3.5484598
3	2.3117711	3.1306353	2.3294286	2.487514	2.2170337

For illustration take $n = 10$, $n_1 = 3$, $n_2 = 4$, $n_3 = 3$

Population VI: Singh and Mangat (1996, p. 219)
y: total number of milch cows 1993, x: total number of milch cows 1990
$N = 24$, $L = 3$, $N_1 = 7$, $N_2 = 12$, $N_3 = 5$

Stratum	Values of parameters for h^{th} stratum					
No.	S_{yh}	ρ_h	λ_{12h}	λ_{21h}	λ_{03h}	λ_{30h}
1	4.197505	0.7654592	-0.4418403	-0.4494459	0.0382842	-0.324885
2	4.0778411	0.4066542	-0.2762718	-0.2448949	0.1507925	-0.6181979
3	3.6469165	0.4945774	-0.8119799	-0.2847418	-0.569229	-0.0912794

Stratum	Values of parameters for h^{th} stratum (continued)				
No.	λ_{22h}	λ_{04h}	λ_{40h}	λ_{13h}	λ_{31h}
1	1.1348072	1.8497596	1.6555367	1.0929828	1.3169373
2	0.5695984	2.312027	2.7509735	0.8349021	0.6748404
3	1.3461457	1.8333916	1.5925434	1.1123488	1.0704605

For illustration take $n = 10$, $n_1 = 3$, $n_2 = 5$, $n_3 = 2$

References

Singh, D., & Chaudhary, F. S. (1986). *Theory and analysis of sample survey designs*. New Delhi, India: Wiley Eastern Ltd.

Singh, R., & Mangat, N. S. (1996). *Element of survey sampling.* London, England: Kluwer Academic Publishers.

Srivastava, S. K. (1971). A generalized estimator for the mean of a finite population using multi-auxiliary information. *Journal of the American Statistical Association*, *66*, 404-407.

Srivastava, S. K., & Jhajj, H. S. (1981). A class of estimators of the population mean in survey sampling using auxiliary information. *Biometrika*, *68*, 341-343.

Srivastava, S. K., & Jhajj, H. S. (1983). Class of estimators of mean and variance using auxiliary information when correlation coefficient is known. *Biometrical Journal*, *24*(4), 401-409.

Zaidi, S., Shanker, U., & Singh, R. K. (1989). An optimum allocation with a class of estimators using auxilary information. *Journal of Statistical Research*, *23*, 1-4.

Journal of Modern Applied Statistical Methods
November, 2008, Vol. 7, No. 2, 488-500

Analyzing Incomplete Categorical Data: Revisiting Maximum Likelihood Estimation (Mle) Procedure

Hoo Ling Ping
The University of Nottingham,
Malaysia

M. Ataharul Islam
University of Dhaka,
Bangladesh

Incomplete data poses formidable difficulties in the application of statistical techniques and requires special procedures to handle. The most common ways to solve this problem are by ignoring, truncating, censoring or collapsing those data, but these may lead to inappropriate conclusions because those data might contain important information. Most of the research for estimating cell probabilities involving incomplete categorical data is based on the EM algorithm. A likelihood approach is employed for estimating cell probabilities for missing values and makes comparisons between maximum likelihood estimation (MLE) and the EM algorithm. The MLE can provide almost the same estimates as that of the EM algorithm without any loss of properties. Results are compared for different distributional assumptions. Using clinical trial results from a group of 59 epileptics, results from the application of MLE and EM algorithm are compared and the advantages of MLE are highlighted.

Key words: Incomplete categorical data, maximum likelihood estimation (MLA), EM algorithm, multinomial distribution, binomial distribution, Poisson distribution, Newton-Raphson method.

Introduction

Incomplete data is referred to as data in which entries are missing, were a prior zero or are undetermined (Fienberg, 1980). Incomplete data is one of the main obstacles to researchers; this is especially true in the case of incomplete categorical data. The most common ways to solve this problem are by ignoring, truncating, censoring or collapsing those data; however, such procedures may lead to confusion and/or inappropriate conclusions because those data might contain important information.

Little & Rubin (1987) defined the missing data mechanisms as *ignorable missing data mechanism* and *non-ignorable missing data mechanism.* The ignorable missing data mechanism involves process missing completely at random (MCAR) and missing at random (MAR). When the missingness is independent of both unobserved and observed data, the non-response process is named as MCAR. However, if the missingness is independent of the unobserved measurement conditionally on the observed data, the non-response process is called MAR. Non-ignorable missing data mechanisms involve *informative* process. When the process is neither MCAR nor MAR, then the process is termed informative. This article considers the missing data mechanism as a non-ignorable missing data mechanism.

The problem of estimation for incomplete contingency tables under the quasi-independence model was examined by Fienberg (1970), who used the maximum likelihood estimation (MLE) procedure. Similarly, MLE for the Poisson and Multinomial sampling distributions for the incomplete contingency tables in the presence of missing row and missing column data were considered by Chen & Fienberg (1974). Chen & Fienberg (1976) extended their works which focused on cross-classifications containing some totally mixed up cell frequencies with multinomial sampling. In the following year, Dempster, Laird & Rubin

Hoo Ling Ping is Assistant Professor in the Department of Applied Mathematics, Faculty of Engineering. Email: lphoo_04@yahoo.com. M. Ataharul Islam is Professor in the Department of Statistics, Biostatistics and Informatics. Email: mataharul@yahoo.com.

presented MLE of incomplete data and named the algorithm EM since each iteration of the algorithm involved expectation (E) and maximization (M) steps. This method has been used extensively by other researchers especially for incomplete categorical data. Among others, Fuchs (1982), Nordheim (1984), Fay (1986), Baker & Laird (1988), and Philips (1993) have used the EM algorithm for analyzing incomplete categorical data. The EM algorithm was used to improve the convergence of the EM by incorporating the Newton-Rapson approach by Baker (1994) and Galecki & Molenberghs (2001). The EM algorithm is well developed (Lauritzen, 1995) to exploit the computational scheme of Lauritzen & Spiegelhalter (1988) to perform the E-step of EM algorithm to find MLE in hierarchical log-linear models and recursive models for contingency tables with missing data. Molenberghs & Goetchebeur (1997) presented a simple expression of the observed data log-likelihood for the EM algorithm. Tang, et. al. (2007) also found that the EM algorithm is the most widely used approach for finding the maximum likelihood estimate for incomplete data situations, but it lacks the direct provision of a measure of precision for the estimators and has a slow rate of convergence.

Because the EM algorithm was introduced, the MLE procedure was largely ignored by researchers until 1985. Stasny (1985) used MLE to process the model based on data from a Current Population Survey, and also used a Labor Force Survey to estimate gross flow data. Most recently, Lyles & Allen (2003) proposed MLE with multinomial likelihood, properly accounting for missing data and assuming that the probability of missing exposure depends on true exposure.

In this article, not only is the missing row or missing column data redistributed, but also both row and column missing data for multinomial sampling by extending the works of Chen & Fienberg (1974). Both row and column missing data are also investigated for the EM algorithm which has not been studied before. The MLE method for Poisson and Binomial sampling distributions was also examined as an extension of the works of Chen & Fienberg (1974). The binomial distribution can be considered a special case of the Multinomial distribution. The same sampling patterns for the EM algorithm are considered here. The Newton-Raphson method was adopted in the MLE procedure to make convergence faster. Results of the MLE and EM algorithm are compared and the advantages of MLE are highlighted.

This article is organized as follows: data taken from Diggle, Liang & Zeger (1994) is described, followed by the formulation of the MLE and EM algorithms. Finally, results are discussed, testing independence is presented and conclusions are put forth.

Methodology

The data considered herein was referred from Diggle, Liang & Zeger (1994) based on a clinical trial of 59 epileptics. For each patient, the number of epileptic seizures was recorded during an eight week baseline period. Patients were then randomized to either a treatment group with anti-epileptic drug progabide (0) or to a placebo group (1) and the number of seizures was recorded in four consecutive two-week intervals. Table 1 shows the 2x2 artificial incomplete contingency table; rows refer to the treatment and columns refer to the results of the last treatment for the patient. The result of treatment is recorded as Y.

Maximum likelihood estimation (MLE), Poisson and multinomial distribution

Chen & Fienberg (1974) considered the MLE for incomplete contingency tables when missing row and column data existed. Their works are extended by considering incomplete contingency tables where either row or column, or both row and column are missing.

Let the fully cross-classified count for the $(i, j)^{th}$ cell of an r x c contingency table be x_{ij}, R_i (i = 1, 2, …, r) is the count of the partially classified individuals corresponding to the i^{th} row, C_j (j = 1, 2, …, c) is the count of the partially classified individuals corresponding to the j^{th} column, and D is the count of missing in both row and column. (See Figure 1.) Therefore the total sample size is:

$$N=\sum_{ij} x_{ij}+\sum_{i} R_i+\sum_{j} C_j+D=x_{++}+R_{+}+C_{+}+D. \quad (1)$$

Table1: Incomplete data

a) No missing on treatment and Y

		Y		Total
		≤ 5	> 5	
Treatment	0	13	7	20
	1	12	7	19
Total		25	14	39

b) Missing Y, treatment, Y and treatment

		Missing Y
Treatment	0	3
	1	7
Total		10

	Y		Total
Missing Treatment	Yes	No	
	2	2	4

	Missing Y	Total
Missing Treatment	6	
Total		6

Figure 1: Illustration for complete observed and incomplete data

a) Complete observed data

	Row				Total
Column	x_{11}	x_{12}	…	x_{1c}	x_{1+}
	x_{21}	x_{22}	…	x_{2c}	$_{2+}$
	⋮	⋮	⋱	⋮	⋮
	x_{r1}	x_{r2}	…	x_{rc}	x_{c+}
Total	x_{+1}	x_{+2}		x_{+c}	x

b) Incomplete units

	Missing column
Row	R_{1+}
	R_{2+}
	⋮
	R_{r+}
Total	R

	Column				Total
Missing row	C_{+1}	C_{+2}	…	C_{+c}	C

	Missing column	Total
Missing row	D	
Total		D

When the original sampling scheme is Poisson with an expected value m_{ij} for the (i, j)th cell, parameters associated with the cells (illustrated in Table 2) where $\lambda_{1(i)}$, $\lambda_{2(j)}$ and λ are referred to the probabilities of losing its row, column, and both row and column identity, respectively. The cell probability of multinomial sampling for a completely classified (i, j)th cell is π_{ij} and $\sum_i \sum_j \pi_{ij} = 1$. By replacing π_{ij} with m_{ij}, the likelihood function for Poisson is

$$\exp\left(-\sum_i \sum_j m_{ij}\right) \prod_i \prod_j \left[\left(1-\lambda_{1(i)}-\lambda_{2(j)}-\lambda\right) m_{ij}\right]^{x_{ij}}$$
$$\prod_i \left(\lambda_{1(i)} m_{i+}\right)^{R_i} \prod_j \left(\lambda_{2(j)} m_{+j}\right)^{C_i} \left(\lambda m_{++}\right)^{D} \tag{2}$$

where

$m_{i+}=\sum_i m_{ij}$, $m_{+j}=\sum_j m_{ij}$, and $m_{++}=\sum_i \sum_j m_{ij}$

for all i and j.

Equation (2) is a product of functions f_1 and f_2 defined as follows

$$f_1 = \frac{\left[\prod_i \prod_j \left(1-\lambda_{1(i)}-\lambda_{2(j)}-\lambda\right)^{x_{ij}}\right]}{\left[\prod_i \lambda_{1(i)}^{R_i}\right]\left[\prod_j \lambda_{2(j)}^{C_j}\right]\left[\lambda^{D}\right]} \tag{3}$$

and

$$f_2 = \frac{\exp\left(-\sum_i \sum_j m_{ij}\right)\left[\prod_i \prod_j m_{ij}^{x_{ij}}\right]}{\left[\prod_i m_{i+}^{R_i}\right]\left[\prod_j m_{+j}^{C_j}\right]\left[m_{++}^{D}\right]}. \tag{4}$$

when considering the unrestricted log linear model where

$$\log m_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}. \tag{5}$$

Therefore,

$$\log f_2 = \begin{matrix} -\sum_i \sum_j m_{ij} + \sum_i \sum_j x_{ij} \log m_{ij} + \\ \sum_i R_i \log m_{i+} + \sum_j C_j \log m_{+j} + D \log m_{++} \end{matrix}$$
$$= \begin{matrix} -\sum_i \sum_j e^{\mu+\alpha_i+\beta_j+\gamma_{ij}} + \sum_i \sum_j x_{ij} \\ \left(\mu+\alpha_i+\beta_j+\gamma_{ij}\right) + \sum_i R_i \left(\mu+\alpha_i+\beta_+ + \gamma_{i+}\right) \end{matrix}$$
$$+\sum_j C_j \left(\mu+\alpha_+ + \beta_j + \gamma_{i+}\right)$$
$$+D\left(\mu+\alpha_+ + \beta_+ + \gamma_{++}\right) \tag{6}$$

Differentiating (6) with μ, α_i, β_j and γ_{ij},t, results in:

$$\frac{\partial \log f_2}{\partial \mu} = \begin{matrix} -\sum_i \sum_j e^{\mu+\alpha_i+\beta_j+\gamma_{ij}} \\ +\sum_i \sum_j x_{ij} + \sum_i R_i + \sum_j C_j + D \end{matrix}$$
$$= -\sum_i \sum_j m_{ij} + \sum_i \sum_j x_{ij} + \sum_i R_i + \sum_j C_j + D$$
$$= - m_{++} + x_{++} + R_+ + C_+ + D$$

$$\frac{\partial \log f_2}{\partial \alpha_i} = - m_{i+} + x_{i+} + R_i + \sum_j C_j \left(\frac{m_{ij}}{m_{+j}}\right) + D\left(\frac{m_{i+}}{m_{++}}\right)$$
$$\frac{\partial \log f_2}{\partial \beta_j} = - m_{+j} + x_{+j} + \sum_i R_i \left(\frac{m_{ij}}{m_{i+}}\right) + C_j + D\left(\frac{m_{+j}}{m_{++}}\right)$$
$$\frac{\partial \log f_2}{\partial \gamma_{ij}} = m_{ij} + x_{ij} + R_i \left(\frac{m_{ij}}{m_{i+}}\right) + C_j \left(\frac{m_{ij}}{m_{+j}}\right) + D\left(\frac{m_{ij}}{m_{++}}\right) \tag{7}$$

When (7) is equal to 0,

$$\hat{m}_{ij} = x_{ij} + R_i \left(\frac{\hat{m}_{ij}}{m_{i+}}\right) + C_j \left(\frac{\hat{m}_{ij}}{m_{+j}}\right) + D\left(\frac{\hat{m}_{ij}}{m_{++}}\right). \tag{8}$$

As from Chen & Fienberg (1974), (8) is not able to be solved in closed from; initial estimates of the $\{\hat{m}_{ij}\}$ will be considered as

$$m_{ij}^{(0)} = \left(\frac{x_{ij}}{x_{++}}\right) N. \tag{9}$$

On the first iteration, from (8)

$$\hat{m}_{ij}^{(1)} = x_{ij} + R_i \left(\frac{\hat{m}_{ij}^{(0)}}{m_{i+}^{(0)}}\right) + C_j \left(\frac{\hat{m}_{ij}^{(0)}}{m_{+j}^{(0)}}\right) + D\left(\frac{\hat{m}_{ij}^{(0)}}{m_{++}^{(0)}}\right), \tag{10}$$

therefore on (k+1)th iteration,

$$\hat{m}_{ij}^{(k+1)} = x_{ij} + R_i \left(\frac{\hat{m}_{ij}^{(k)}}{m_{i+}^{(k)}}\right) + C_j \left(\frac{\hat{m}_{ij}^{(k)}}{m_{+j}^{(k)}}\right) + D\left(\frac{\hat{m}_{ij}^{(k)}}{m_{++}^{(k)}}\right) \tag{11}$$

When k $\rightarrow \infty$, $\left|\hat{m}_{ij}^{(k+1)} - \hat{m}_{ij}^{(k)}\right| \le \varepsilon$.

Table 2: Underlying probabilities for a 2x2 table

	Fully Classified Table		Row Supplemental Margin
	$\left(1-\lambda_{1(1)}-\lambda_{2(1)}-\lambda\right)\pi_{11}$	$\left(1-\lambda_{1(1)}-\lambda_{2(2)}-\lambda\right)\pi_{12}$	$\lambda_{1(1)}\pi_{1+}$
	$\left(1-\lambda_{1(2)}-\lambda_{2(1)}-\lambda\right)\pi_{21}$	$\left(1-\lambda_{1(2)}-\lambda_{2(2)}-\lambda\right)\pi_{22}$	$\lambda_{1(2)}\pi_{2+}$
			Missing row and column
Column Supplemental Margin	$\lambda_{2(1)}\pi_{+1}$	$\lambda_{2(2)}\pi_{+2}$	$\lambda\pi_{++}$

Since $\pi_{ij} = \dfrac{x_{ij}}{x_{++}}$ for complete classified multinomial sampling, from (9) results,

$$\hat{\pi}_{ij}^{(0)} = \frac{\hat{m}_{ij}^{(0)}}{N},$$

and on the (k+1)th iteration, $\hat{\pi}_{ij}^{(k+1)} = \dfrac{\hat{m}_{ij}^{(k+1)}}{N}$.

Poisson and binomial distributions

Now consider the complete contingency tables where there exist missing column. The fully cross-classified count for the (i, j) cell of an r x 2 contingency table is x_{ij}, and R_i (i = 1, 2, …, r) is the count of the partially classified individuals corresponding to the ith row. Therefore the total sample size is

$$N = \sum_{ij} x_{ij} + \sum_i R_i$$
$$= x_{++} + R_{+} \quad (12)$$

When the original sampling scheme is Poisson with expected value m_{ij} for the (i, j) cell, then the parameters associated with the cells (illustrated in Table 3) where $\lambda_{1(i)}$, is referred to the probabilities of losing its row.

The cell probability of Binomial sampling for complete classified of (i, j) cell is π_{ij} and $\sum_i \sum_j \pi_{ij} = 1$. By replacing π_{ij} with m_{ij}, the likelihood function for Poisson is

$$\exp\left(-\sum_i \sum_j m_{ij}\right)$$
$$\prod_i \prod_j \left[\left(1-\lambda_{1(i)}\right) m_{ij}\right]^{x_{ij}} \prod_i \left(\lambda_{1(i)} m_{i+}\right)^{R_i},$$
$$(13)$$

where $m_{i+} = \sum_i m_{ij}$ for all i and j.

Equation (13) is a product of a function

$$f_1 = \left[\prod_i \prod_j \left(1-\lambda_{1(i)}\right)^{x_{ij}}\right]\left[\prod_i \lambda_{1(i)}^{R_i}\right] \quad (14)$$

and

$$f_2 = \exp\left(-\sum_i \sum_j m_{ij}\right)\left[\prod_i \prod_j m_{ij}^{x_{ij}}\right]\left[\prod_i m_{i+}^{R_i}\right]. \quad (15)$$

When considering the unrestricted log linear model where

$$\log m_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}, \quad (16)$$

then,

$$\begin{aligned}\log f_2 = & -\sum_i\sum_j m_{ij} + \sum_i\sum_j x_{ij}\log m_{ij} \\ & + \sum_i R_i \log m_{i+} \\ & -\sum_i\sum_j e^{\mu+\alpha_i+\beta_j+\gamma_{ij}} + \sum_i\sum_j x_{ij} \\ = & \left(\mu+\alpha_i+\beta_j+\gamma_{ij}\right) \quad . \\ & +\sum_i R_i\left(\mu+\alpha_i+\beta_+ +\gamma_{i+}\right)\end{aligned} \tag{17}$$

Differentiating (17) with μ, α_i, β_j and γ_{ij}, results in

$$\begin{aligned}\frac{\partial \log f_2}{\partial \mu} &= -\sum_i\sum_j e^{\mu+\alpha_i+\beta_j+\gamma_{ij}} + \sum_i\sum_j x_{ij} + \sum_i R_i \\ &= -\sum_i\sum_j m_{ij} + \sum_i\sum_j x_{ij} + \sum_i R_i \\ &= -\mathrm{m}_{++} + \mathrm{x}_{++} + \mathrm{R}_+\end{aligned}$$

$$\frac{\partial \log f_2}{\partial \alpha_i} = -\mathrm{m}_{i+} + \mathrm{x}_{i+} + \mathrm{R}_i$$

$$\frac{\partial \log f_2}{\partial \beta_j} = -\mathrm{m}_{+j} + \mathrm{x}_{+j} + \sum_i R_i\left(\frac{m_{ij}}{m_{+j}}\right)$$

$$\frac{\partial \log f_2}{\partial \gamma_{ij}} = -\mathrm{m}_{ij} + \mathrm{x}_{ij} + R_i\left(\frac{m_{ij}}{m_{i+}}\right) \tag{18}$$

and, when (18) is equal to 0

$$\hat{m}_{ij} = x_{ij} + R_i\left(\frac{\hat{m}_{ij}}{m_{i+}}\right). \tag{19}$$

Initial estimates of the $\{\hat{m}_{ij}\}$ were considered as

$$m_{ij}^{(0)} = \left(\frac{x_{ij}}{x_{i+}}\right)m_{i+}. \tag{20}$$

where $m_{i+} = x_{i+} + R_i$.

On the first iteration, from (19),

$$\hat{m}_{ij}^{(1)} = x_{ij} + R_i\left(\frac{\hat{m}_{ij}^{(0)}}{m_{i+}}\right). \tag{21}$$

So, on the (k+1)th iteration,

$$\hat{m}_{ij}^{(k+1)} = x_{ij} + R_i\left(\frac{\hat{m}_{ij}^{(k)}}{m_{i+}}\right), \tag{22}$$

when k $\to \infty$, $\left|\hat{m}_{ij}^{(k+1)} - \hat{m}_{ij}^{(k)}\right| \le \varepsilon$.

If an underlying Binomial sampling scheme is assumed, then

$$p_{ij} = \frac{x_{ij}}{x_{i+}}.$$

Therefore, from (20),

$$\hat{p}_{ij}^{(0)} = \hat{m}_{ij}^{(0)} \Big/ m_{i+}$$

and

$$\hat{p}^{(0)} = \sum_i\sum_j \hat{m}_{ij}^{(0)} \Big/ N.$$

On the (k+1)th iteration,

$$\hat{p}_{ij}^{(k+1)} = \hat{m}_{ij}^{(k+1)} \Big/ m_{i+}$$

and

$$\hat{p}^{(k+1)} = \sum_i\sum_j \hat{m}_{ij}^{(k+1)} \Big/ N.$$

Table 3: Underlying probabilities for a 2x2 table

Fully Classified Table		Row Supplemental Margin
$\left(1-\lambda_{1(1)}\right)\pi_{11}$	$\left(1-\lambda_{1(1)}\right)\pi_{12}$	$\lambda_{1(1)}\pi_{1+}$
$\left(1-\lambda_{1(2)}\right)\pi_{21}$	$\left(1-\lambda_{1(2)}\right)\pi_{22}$	$\lambda_{1(2)}\pi_{2+}$

Formulation of Newton-Raphson

From Le (1992), the iterative solution for a parameter estimation on (k+1)th iteration will be considered as

$$\hat{\theta}^{(k+1)} = \hat{\theta}^{(k)} + \Delta\hat{\theta}, \tag{23}$$

where θ is the parameter and

$$\Delta\hat{\theta} = -\left(\frac{d\ln L}{d\theta}\right) \Big/ \left(\frac{d^2 \ln L}{d\theta^2}\right).$$

Differentiating (7) with γ_{ij} and equal to 0, then results in

$$m_{ij} = R_{i+}\left[\frac{m_{ij}}{m_{i+}} - \left(\frac{m_{ij}}{m_{i+}}\right)^2\right] + C_{+j}\left[\frac{m_{ij}}{m_{+j}} - \left(\frac{m_{ij}}{m_{+j}}\right)^2\right] + D\left[\frac{m_{ij}}{m_{++}} - \left(\frac{m_{ij}}{m_{++}}\right)^2\right] \tag{24}$$

To avoid the confusion of m_{ij} for (7) and (24), let $m1_{ij}$ and $m2_{ij}$ for (7) and (24), respectively.

For application of the Newton-Raphson method in the two-way incomplete contingency table, consider

$$\hat{m}1_{ij}^{(1)} = x_{ij} + R_{i+}\left(\frac{\hat{m}_{ij}^{(0)}}{m_{i+}^{(0)}}\right) + C_{+j}\left(\frac{\hat{m}_{ij}^{(0)}}{m_{+j}^{(0)}}\right) + D\left(\frac{\hat{m}_{ij}^{(0)}}{m_{++}^{(0)}}\right) \tag{25}$$

$$m2_{ij}^{(1)} = R_{i+}\left[\frac{m1_{ij}^{(0)}}{m1_{i+}^{(0)}} - \left(\frac{m1_{ij}^{(0)}}{m1_{i+}^{(0)}}\right)^2\right] + C_{+j}\left[\frac{m1_{ij}^{(0)}}{m1_{+j}^{(0)}} - \left(\frac{m1_{ij}^{(0)}}{m1_{+j}^{(0)}}\right)^2\right]. + D\left[\frac{m1_{ij}^{(0)}}{m1_{++}^{(0)}} - \left(\frac{m1_{ij}^{(0)}}{m1_{++}^{(0)}}\right)^2\right] \tag{26}$$

where $\hat{m}_{ij}^{(0)}$ is the same with (9) and

$$\hat{m}_{ij}^{(1)} = m1_{ij}^{(0)} - \frac{m1_{ij}^{(0)}}{m2_{ij}^{(0)}}. \tag{27}$$

On the (k+1)th iteration,

$$\hat{m}1_{ij}^{(k)} = x_{ij} + R_{i+}\left(\frac{\hat{m}_{ij}^{(k-1)}}{m_{i+}^{(k-1)}}\right) + C_{+j}\left(\frac{\hat{m}_{ij}^{(k-1)}}{m_{+j}^{(k-1)}}\right) + D\left(\frac{\hat{m}_{ij}^{(k-1)}}{m_{++}^{(k-1)}}\right). \tag{28}$$

$$m2_{ij}^{(k)} = R_{i+}\left[\frac{m1_{ij}^{(k-1)}}{m1_{i+}^{(k-1)}} - \left(\frac{m1_{ij}^{(k-1)}}{m1_{i+}^{(k-1)}}\right)^2\right] + C_{+j}\left[\frac{m1_{ij}^{(k-1)}}{m1_{+j}^{(k-1)}} - \left(\frac{m1_{ij}^{(k-1)}}{m1_{+j}^{(k-1)}}\right)^2\right] + D\left[\frac{m1_{ij}^{(k-1)}}{m1_{++}^{(k-1)}} - \left(\frac{m1_{ij}^{(k-1)}}{m1_{++}^{(k-1)}}\right)^2\right] \tag{29}$$

$$\hat{m}_{ij}^{(k+1)} = m1_{ij}^{(k)} - \frac{m1_{ij}^{(k)}}{m2_{ij}^{(k)}}. \tag{30}$$

For an accelerated convergence, these equations were employed to obtain the maximum likelihood estimators.

The EM algorithm: Formulation of the EM algorithm for contingency table

The EM approach for incomplete categorical data on the basis of Multinomial, Binomial and Poisson assumptions is now investigated.

Multinomial Distributions

For Multinomial distributions, the complete data log likelihood is

$$\log L_c(\pi_i) = \sum_{i=1}^{n-1} (x_i+z_i)\log \pi_i + (x_n+z_n) \log(1-\pi_1-\pi_2-\ldots-\pi_{n_{-1}}), \tag{31}$$

where unobservable or missing data are referred to as $z_i = (z_1, z_2, \ldots, z_n)^T$ and $z_i = r_i+c_i+d_i$ with r_i being missing column data, c_i missing row data, and d_i both row and column missing data on cell ith. Differentiating (31) with respect to π_i, results in

$$\hat{\pi}_i = \frac{x_i + z_i}{x_n + z_n}\pi_n. \tag{32}$$

Since $\sum_{i=1}^{n}\pi_i$ = 1, therefore from (32),

$$\hat{\pi}_i = \frac{x_i + z_i}{N} \qquad (33)$$

where $\sum_{i=1}^{n}\left(x_i + z_i\right) = \text{N}$.

The E- and M-values on the first iteration for cell (i, j) were considered as follows.

E-step:

$$m_{ij}^{(1)} = \text{x}_{\text{ij}} + \text{R}_{\text{i+}}\left(\frac{\pi_{ij}}{\pi_{i+}}\right) + \text{C}_{\text{+j}}\left(\frac{\pi_{ij}}{\pi_{+j}}\right) + \text{D}\pi_{\text{ij}}$$

where $m_{ij}^{(1)}$ is the expected of cell (i, j) on the first iteration.

M-step:

$$\hat{\pi}_{ij}^{(1)} = m_{ij}^{(1)} \big/ N,$$

where $\pi_{ij}^{(1)}$ is the probability for cell (i, j).

On the (k+1)th iteration, the E- and M-steps were defined as follows:

E-step:

$$m_{ij}^{(k+1)} = \text{x}_{\text{ij}} + \text{R}_{\text{i+}}\left(\frac{\hat{\pi}_{ij}^{(k)}}{\hat{\pi}_{i+}^{(k)}}\right) + \text{C}_{\text{+j}}\left(\frac{\hat{\pi}_{ij}^{(k)}}{\hat{\pi}_{+i}^{(k)}}\right) + \text{D}\pi_{\text{ij}}$$

M-step:

$$\hat{\pi}_{ij}^{(k+1)} = m_{ij}^{(k+1)} \big/ N.$$

The E- and M-steps were alternated and repeated until

$$\left|\hat{\pi}_{ij}^{(k+1)} - \hat{\pi}_{ij}^{(k)}\right| = \left|\frac{m_{ij}^{(k+1)}}{N} - \frac{m_{ij}^{(k)}}{N}\right|$$
$$= \left|\frac{m_{ij}^{(k+1)} - m_{ij}^{(k)}}{N}\right| \le \varepsilon.$$

Therefore, when $\text{k} \longrightarrow \infty$, $\lim_{k\to\infty}\left|\hat{\pi}_{ij}^{(k+1)} - \hat{\pi}_{ij}^{(k)}\right|$ = 0, and $\hat{\pi}_{ij}^{(k+1)} = \hat{\pi}_{ij}^{(k)} = \pi^{*}$.

Binomial distribution

For the binomial distribution, the complete-data log likelihood is

$$\log \text{L}_{\text{c}}(p_{i1}) = (\text{x}_{\text{i1}} + \text{z}_{\text{i1}})\log p_{i1} + (\text{x}_{\text{i2}} + \text{z}_{\text{i2}})\log(1 - p_{i1}),$$

for i = 1, …, n, and z_i is referred to as unobservable or missing data on the ith row where $\text{z}_{\text{i1}} + \text{z}_{\text{i2}} = \text{z}_\text{i}$. Differentiating with respect to p_{i1} results in

$$\hat{p}_{i1} = \frac{x_{i1} + z_{i1}}{x_{i1} + x_{i2} + z_i}. \qquad (34)$$

From (34), if all rows are summed, the following is obtained

$$\sum_{i=1}^{I}\hat{p}_{i1} = \frac{\sum_{i=1}^{I} r_{i1} + z_{i1}}{\sum_{i=1}^{I} r_{i1} + r_{i2} + z_i}.$$

Since $\sum_{i=1}^{I} r_{i1} + r_{i2} + z_i = N$, total sample,

$$\hat{p}_{+j} = \frac{\sum_{i=1}^{n} x_{ij}^{(1)}}{N}.$$

The E- and M-values on the first iteration for cell (i, j) were considered as follows:

E-step:

$$m_{ij}^{(1)} = \text{x}_{\text{ij}} + \text{R}_{\text{i+}}\text{p}_{\text{ij}}$$

where $m_{ij}^{(1)}$ is the expected value of cell (i, j) on the first iteration and $\text{p}_{\text{ij}} = \text{x}_{\text{ij}}/\text{x}_{\text{i+}}$.

M-step:

$$\hat{p}_{ij}^{(1)} = m_{ij}^{(1)} \Big/ \left(m_{i1}^{(1)} + m_{i2}^{(1)}\right),$$

and

$$\hat{p}_{+j}^{(1)} = \frac{\sum_{i=1}^{n} m_{ij}^{(1)}}{N}$$

On the (k+1)th iteration, the E- and M-steps were defined as follows:

E-step:

$$m_{ij}^{(k+1)} = \mathrm{x}_{\mathrm{ij}} + \mathrm{R}_{\mathrm{i+}}\, p_{ij}^{(k)}.$$

M-step:

$$\hat{p}_{ij}^{(k+1)} = m_{ij}^{(k+1)} \Big/ \left(m_{i1}^{(k+1)} + m_{i2}^{(k+1)}\right),$$

and

$$\hat{p}_{+j}^{(k+1)} = \frac{\sum_{i=1}^{n} m_{ij}^{(k+1)}}{N}.$$

The E- and M-steps were alternated and repeated until

$$\left|\hat{p}_{ij}^{(k+1)} - \hat{p}_{ij}^{(k)}\right| = \left|\frac{m_{ij}^{(k+1)}}{m_{i1}^{(k+1)} + m_{i2}^{(k+1)}} - \frac{m_{ij}^{(k)}}{m_{i1}^{(k)} + m_{i2}^{(k)}}\right| \le \varepsilon,$$

and

$$\left|\hat{p}_{+j}^{(k+1)} - \hat{p}_{+j}^{(k)}\right| = \left|\frac{\sum_{i=1}^{n} m_{ij}^{(k+1)}}{N} - \frac{\sum_{i=1}^{n} m_{ij}^{(k)}}{N}\right|$$

$$= \left|\frac{\sum_{i=1}^{n} m_{ij}^{(k+1)} - \sum_{i=1}^{n} m_{ij}^{(k)}}{N}\right| \le \varepsilon.$$

Therefore, when k $\rightarrow \infty$, $\lim_{k\to\infty} \left|\hat{p}_{ij}^{(k+1)} - \hat{p}_{ij}^{(k)}\right| = 0$ and $\lim_{k\to\infty} \left|\hat{p}_{+j}^{(k+1)} - \hat{p}_{+j}^{(k)}\right| = 0.$

Poisson distribution

For the Poisson distribution, the complete-data log likelihood is

Log $L_c(y; \theta_i)$ =

$$\sum_{i=1}^{n} \left[(x_i + z_i)\log(\theta_i) - \theta_i - \log(x_i + z_i)! \right] \tag{35}$$

where $z_1 + z_2 + \ldots + z_n$ is referred to as unobservable or missing data. By differentiating (35) with respect to θ_i,

$$\hat{\theta}_i = x_i + z_i. \tag{36}$$

Referring to Figure 1, the E- and M-values on the first iteration for the cell (i, j) was considered as:

E-step:

$$z_{ij}^{(1)} = R_{ij}^{(1)} + C_{ij}^{(1)} + D_{ij}^{(1)},$$

where

$$R_{ij}^{(1)} = \mathrm{R}_{\mathrm{i+}}\left(\frac{x_{ij}}{x_{i+}}\right),\ C_{ij}^{(1)} = \mathrm{C}_{\mathrm{+j}}\left(\frac{x_{ij}}{x_{+j}}\right), \text{ and}$$

$$D_{ij}^{(1)} = \mathrm{D}\left(\frac{x_{ij}}{x}\right).$$

M-step:

$$\hat{\theta}_{ij}^{(1)} = x_{ij} + z_{ij}^{(1)}.$$

On the (k+1)th iteration, the E- and M-steps were defined as follows:

E-step:

$$z_{ij}^{(k+1)} = R_{ij}^{(k+1)} + C_{ij}^{(k+1)} + D_{ij}^{(k+1)},$$

where

$$R_{ij}^{(k+1)} = \mathrm{R}_{\mathrm{i+}}\left(\frac{\theta_{ij}^{k}}{\theta_{i+}^{k}}\right),\ C_{ij}^{(k+1)} = \mathrm{C}_{\mathrm{+j}}\left(\frac{\theta_{ij}^{k}}{\theta_{+j}^{k}}\right), \text{ and}$$

$$D_{ij}^{(k+1)} = \mathrm{D}\left(\frac{\theta_{ij}^{k}}{N}\right),$$

and N is total sample.

M-step:

$$\hat{\theta}_{ij}^{(1)} = x_{ij} + z_{ij}^{(1)}.$$

The E- and M-steps were alternated and repeated until

$$\left|\hat{\theta}_{ij}^{(k+1)} - \hat{\theta}_{ij}^{(k)}\right| = \left|\left(x_{ij} + z_{ij}^{(k+1)}\right) - \left(x_{ij} + z_{ij}^{(k)}\right)\right|$$

$$= \left|z_{ij}^{(k+1)} - z_{ij}^{(k)}\right| \le \varepsilon$$

Therefore, when k $\rightarrow \infty$

$$\lim_{k\to\infty} \left|\hat{\theta}_{ij}^{(k+1)} - \hat{\theta}_{ij}^{(k)}\right| = 0,$$

and it may be said that $\hat{\theta}_{ij}^{(k+1)} = \hat{\theta}_{ij}^{(k)} = \theta^{*}$.

Results

The results of MLE, adopting Newton-Raphson in MLE and the M-step of the EM algorithm for the Poisson distribution are presented in Tables 4, 5 and 6 respectively.

Table 4 MLE for Poisson distribution

	Cells			
Iteration	(1,1)	(1,2)	(2,1)	(2,2)
1	17.99	10.13	19.23	11.66
2	17.72	10.05	19.35	11.89
3	17.67	10.03	19.35	11.96
4	17.66	10.02	19.34	11.98
5	17.67	10.02	19.33	11.99
6	17.67	10.01	19.33	11.99
7	17.67	10.01	19.33	11.99

Table 5: Adopting Newton-Raphson in MLE for Poisson distribution

	Cells			
Iteration	(1,1)	(1,2)	(2,1)	(2,2)
1	17.92	9.7	19.62	11.76
2	17.68	9.38	19.87	12.06
3	17.66	9.34	19.78	12.22
4	17.65	9.34	19.77	12.24
5	17.65	9.34	19.77	12.24

Table 6: M-step for Poisson distribution

	Cells			
Iteration	(1,1)	(1,2)	(2,1)	(2,2)
1	17.99	10.13	19.22	11.66
2	17.72	10.04	19.35	11.9
3	17.67	10.02	19.35	11.96
4	17.67	10.02	19.34	11.98
5	17.67	10.01	19.33	11.98
6	17.67	10.01	19.33	11.99
7	17.67	10.01	19.33	11.99

The results of MLE, adopting Newton-Raphson in MLE and the M-step of the EM algorithm for the Multinomial distribution are presented in Tables 7, 8 and 9 respectively. The results of MLE and the M-step for the Binomial distribution are presented in Tables 10, 11 and 12 respectively.

Based upon results, both the MLE and the EM algorithms converge on the 7th iteration (see Tables 4 and 6), and both methods give the same results. However, by adopting the Newton-Raphson in the MLE, the results on the 5th iteration were obtained (see Table 5). Although it seems that the EM algorithm was converging the same as the MLE, the EM algorithm involves two calculation steps on each iteration.

In other words, the EM takes longer to compute the results compared with the MLE. After adopting the Newton-Raphson in the MLE, it was able to give faster convergence without as much deviance in the results as the EM algorithm. Tables 7 and 8 were obtained by considering the last iteration of Tables 4 and 5 respectively. The results were also the same for the Multinomial distribution for the MLE and the EM algorithm. By comparing the results of Table 8 with the last iteration of Table 9, it is observed that the results are not much different. However Table 11 was obtained by considering the last iteration of Table 10. Results shown in Tables 11 and 12 were the same as those obtained for the Binomial distribution.

Testing independence

For two-way contingency tables, the null hypothesis of statistical independence is $H_0 : \pi_{ij} = \pi_{i+} \pi_{+j}$ for all i and j. The likelihood-ratio statistic, G^2 is asymptotically equivalent to χ^2 when $n \to \infty$ with d.f. = (r – 1)(c – 1) where r is the number of rows and c is the number of columns in the contingency table.

According to Schafer (1997), $G^2 = 2\left[\ell(\hat{\pi} \mid Y_{obs}) - \ell(\tilde{\pi} \mid Y_{obs})\right]$, where $\ell(\hat{\pi} \mid Y_{obs})$ is the unrestricted ML estimate ($\hat{\pi}$) and $\ell(\tilde{\pi} \mid Y_{obs})$ is the restricted ML estimate ($\tilde{\pi}$). Thus, $\ell(\pi \mid Y_{obs})$ is considered as:

$$\ell(\pi \mid Y_{obs}) = \ell_A(\pi \mid Y_{obs}) + \ell_B(\pi \mid Y_{obs}) + \ell_C(\pi \mid Y_{obs}) + \ell_D(\pi \mid Y_{obs}).$$

For the Multinomial and Poisson distributions with the MLE and EM algorithm,

$$G^2 = 2\left[\ell(\hat{\pi} \mid Y_{obs}) - \ell(\tilde{\pi} \mid Y_{obs})\right].$$

For the Binomial distribution, H_0: $p = p_{i1}$, therefore

$$G^2 = 2[\ \ell(\hat{p} \mid Y_{obs}) - \ell(\tilde{p} \mid Y_{obs})\ ],$$

where $\ell(\hat{p} \mid Y_{obs})$ is the unrestricted ML estimate of $\hat{p}$ and $\ell(\tilde{p} \mid Y_{obs})$ is the restricted ML estimate of $\tilde{p}$. For both the MLE and the EM algorithms $\ell(p \mid Y_{obs})$ is considered as:

$$\ell(p \mid Y_{obs}) = \ell_A(p \mid Y_{obs}) + \ell_B(p \mid Y_{obs}).$$

Therefore, adopting the Newton-Raphson in the MLE and EM algorithms for Multinomial and Poisson distributions, G^2=0.02. However, for the Binomial distribution, G^2=0.01. From these results, it may be concluded that treatment type is independent of the results of treatment for the Multinomial and Poisson distributions, and the number of seizure pain which is less than five is the same for treatment 0 and 1 for the Binomial distribution.

Table 7: MLE for the Multinomial distribution

Cells			
(1,1)	(1,2)	(2,1)	(2,2)
0.2992	0.1697	0.3276	0.2032

Table 8: Adopting Newton-Raphson in MLE for the Multinomial distribution

Cells			
(1,1)	(1,2)	(2,1)	(2,2)
0.2992	0.1583	0.3351	0.2075

Table 9: M-step for Multinomial distribution

Iteration	Cells			
	(1,1)	(1,2)	(2,1)	(2,2)
1	0.3049	0.1716	0.3259	0.1976
2	0.3003	0.1702	0.3278	0.2017
3	0.2995	0.1698	0.3278	0.2027
4	0.2993	0.1698	0.3278	0.2031
5	0.2992	0.1697	0.3276	0.2031
6	0.2992	0.1697	0.3276	0.2032
7	0.2992	0.1697	0.3276	0.2032

Table 10: MLE for the Poisson distribution

Iteration	Cells			
	(1,1)	(1,2)	(2,1)	(2,2)
1	14.95	8.05	16.42	9.58
2	14.95	8.05	16.42	9.58

Table 11: MLE for the Binomial distribution

Cells					
(1,1)	(2,1)	(+, 1)	(1, 2)	(2,2)	(+, 2)
0.65	0.6315	0.6402	0.35	0.3685	0.3598

Table 12: M-step for the Binomial distribution

Iteration	Cells					
	(1,1)	(2,1)	(+, 1)	(1, 2)	(2,2)	(+, 2)
1	0.65	0.6315	0.6402	0.35	0.3685	0.3598
2	0.65	0.6315	0.6402	0.35	0.3685	0.3598

Conclusion

The EM algorithm is more complicated than the MLE, because the EM algorithm involves the E-(expectation) and M-(maximization) steps. This makes the calculations more complicated and also increases the amount of time required to calculate results as compared with the MLE, which is more straightforward for estimating cell probabilities in cases of incomplete categorical data. For example when consider a contingency table with a Poisson sampling scheme, for MLE, the expected value is obtained as in (11) by considering the previous iteration of the expected value. However, for the EM algorithm, before calculating the expected value in the M-step, the E-step - which involves the estimation of initial cell probability first – must first be considered. For the Binomial sampling scheme, the convergence for estimation of p_{i1} and p can be obtained when first considering Poisson sampling employing the MLE procedure. Again, if the EM algorithm is considered, the E-step is required first in order to obtain an initial estimate for p_{ij}. Similar explanations may be given for Multinomial sampling cases where, if MLE is considered, the Poisson sampling must be addressed before using the last iteration to obtain $\hat{\pi}_{ij}$. The EM algorithm, however, requires step by step convergence starting from the initial value for $\hat{\pi}_{ij}$ before convergence is achieved.

The MLE can better perform by adopting the Newton-Raphson method, because this method helps to accelerate the convergence. When the MLE is adopted with that of Newton-Raphson, as a convergence method, it is clear that the MLE and the EM algorithm are two different kinds of algorithms. The MLE algorithm provides a direct way to maximize the final expected value, while the EM algorithm involves expectation before the maximization; however, the EM algorithm demonstrates the distribution of missing values at each step until convergence on the basis of the marginal probabilities.

The MLE is much simpler than the EM algorithm when one is interested simply in final results. If interest lies in understanding the distribution of missing values in more detail, the EM algorithm is the better choice.

References

Baker, S. G. (1994). Composite linear models for incomplete multinomial data. *Statist. Med., 13*, 609-622.

Baker, S. G., & Laird, N. M. (1988). Regression analysis for categorical variables with outcome subject to nonignorable nonresponse. *Journal of the American Statistical Association, 83*, 62-69.

Bishop, Y. M. M., Fienberg, S. E., & Holland, P. W. (1975). *Discrete Multivariate Analysis:Theory and Practice*. MIT Press: Cambridge, MA.

Chen, T., & Fienberg, S. E. (1974). Two-dimensional contingency tables with both completely and partially cross-classified data. *Biometrics, 30*, 629-642.

Chen, T., & Fienberg, S. E. (1976). The analysis of contingency tables with incompletely classified data. *Biometrics, 32*, 133-144.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. Roy. Statist. Soc. B, 39*, 1-38.

Diggle, P.J., Liang K. Y., & Zeger S. L. (1994) *Analysis of Longitudinal Data*. Oxford University Press: Oxford, England.

Fay, R. E. (1986). Causal models for patterns of nonresponse. *Journal of the American Statistical Association, 81*, 354-365.

Fienberg, S. E. (1970). Quasi-independence and maximum likelihood estimation in incomplete contingency tables. *Journal of the American Statistical Association*, *65*, 1610-1616.

Fienberg, S. E. (1980). *The analysis of cross-classified categorical data.* MIT Press: Cambridge, MA.

Fuchs, C. (1982). Maximum likelihood estimation and model selection in contingency tables with missing data. *Journal of the American Statistical Association*, *77*, 270-278.

Galecki, A. T., Have, T. R. T. & Molenberghs, G. (2001). A simple and fast alternative to the EM algorithm for incomplete categorical data and latent class models. *Comp. Stat. & Data Analysis*, *35*, 265-281.

Lauritzen, S. L. (1995). The EM algorithm for graphical association models with missing data. *Comp. Stat. & Data Analysis*, *19*, 191-201.

Lauritzen, S. L., & Spiegelhalter, D. J. (1988). Local computations with probabilities on graphical structures and their application to expert systems (with discussion). *Journal of the Royal Statistical Society B*, *50*, 157-224.

Le, C. T. (1992). *Fundamentals of Biostatistical Inference.* Marcel Dekker, Inc.: New York, NY.

Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data*. John Wiley: New York, NY.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data*, *2nd Ed.* John Wiley: New Jersey.

Lyles, R. H., & Allen, A. S. (2003). Missing data in the 2x2 table: patterns and likelihood-based analysis for cross-sectional studies with supplemental sampling Stats. *Med., 22*, 517-534.

McLachlan, G. J., & Krishnan, T. (1997). *The EM algorithm and extensions*. John Wiley: New York, NY.

Molenberghs, G., & Goetchebeur, E. (1997). Simple fitting algorithm for incomplete categorical data. *Journal of the Royal Statistical Society B*, *59*, 401-414.

Nordheim, E. V. (1984). Inference from nonrandomly missing categorical data: an example from a genetic study on Turner's syndrome. *Journal of the American Statistical Association*, *79*, 772-780.

Phillips, M. J. (1993). Contingency tables with missing data, *The Statistician*, *42*, 9-18.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data.* Chapman and Hall/CRC: New York, NY.

Stasny, E. A. (1985). Modeling non-ignorable non-response in panel data. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 349-354.

Tang, M. L., Ng, K. W., Tian, G. L. & Tan, M. (2007). On improved EM algorithm and confidence interval construction for incomplete rxc tables. *Computational Statistics and Data Analysis*, *51*: 2919-2933.

Journal of Modern Applied Statistical Methods
November, 2008, Vol. 7, No. 2, 501-505

1538 – 9472/08/$95.00

Adaptive Estimation of Heteroscedastic Linear Regression Model Using Probability Weighted Moments

Faqir Muhammad
Allama Iqbal Open University

Muhammad Aslam
Bahauddin Zakariya University

G.R. Pasha
Bahauddin Zakariya University

An adaptive estimator is presented by using probability weighted moments as weights rather than conventional estimates of variances for unknown heteroscedastic errors while estimating a heteroscedastic linear regression model. Empirical studies of the data generated by simulations for normal, uniform, and logistically distributed error terms support our proposed estimator to be quite efficient, especially for small samples.

Key words: Adaptive estimator, estimated weighted least squares, heteroscedasticity, probability weighted moments.

Introduction

The basic version of linear regression model assumes homoscedasticity of error terms. If this assumption is not met then the regression disturbances whose variances are not constant across observations are heteroscedastic. In the presence of heteroscedasticity, the method of ordinary least squares (OLS) does not result in biased and inconsistent parameter estimates. However, OLS estimates are no longer best linear unbiased estimators (BLUE). That is, among all the unbiased estimators, OLS does not provide the estimate with the smallest variance. In addition, the standard errors of the estimates become biased and inconsistent when heteroscedasticity is present. This, in turn, leads to bias in test statistics and confidence intervals. Depending on the nature of the heteroscedasticity, significance tests can be too high or too low. These effects are not ignorable as earlier noted by Geary (1966), White (1980) and Pasha (1982), among many others.

When the form of heteroscedasticity is known, using weights to correct for heteroscedasticity is very simple by weighted least squares (WLS). If the form of heteroscedasticity is not known, the standard method of replication is used as given by Fuller and Rao (1978). In this approach, the unknown variance of each residual can be estimated first and these estimates can be used as weights in a second step and the resultant estimates are referred to as estimated weighted least squares (EWLS) estimates.

Pasha (1984) gave a comparison among EWLS and minimum norm quadratic unbiased estimator (MINQUE) and reported EWLS to be better than MINQU-based estimators for estimation of heteroscedastic linear regression model. Pasha and Ord (1994) presented two adaptive estimators, one based on overall test of heteroscedasticity and other on paired comparison procedures following the idea of Bancroft (1964) and Bancroft & Hans (1977). These estimators were also based on EWLS and the attractive performances of these adaptive estimators were reported for efficiency gain.

An adaptive estimator is presented in this article by using probability weighted moments (PWM) as weights for transforming

Faqir Muhammad is Faculty Dean and Professor in the Department of Statistics and Mathematics, Allama Iqbal Open University, Islamabad, Pakistan. E-mail: aioufsd@yahoo.com. Muhammad Aslam is Assistant Professor in the Department of Statistics, Bahauddin Zakariya University, Multan, Pakistan. E-mail: aslamasadi@bzu.edu.pk. G.R. Pasha is Professor in the Department of Statistics, Bahauddin Zakariya University, Multan, Pakistan E-mail: drpasha@bzu.edu.pk.

matrix rather than conventional estimates of unknown error variances as usually used in EWLS. Downton (1966) suggested a linear estimate of the standard deviation of the normal distribution as

$$S_p = \frac{2\sqrt{\pi}}{n(n-1)}\sum_{i=1}^{n}[i-0.5(n+1)]X_i.$$

Here X_i indicates ordered observations in a sample of size n. The estimate of the standard deviation using PWM is also a function of ordered observations as

$$S_{pw} = \frac{\sqrt{\pi}}{n}\sum_{i=1}^{n}[X_i - 2(1-\frac{i-0.5}{n})X_i].$$

The estimate of the mean is $\frac{\sum_{i=1}^{n} X_i}{n}$. The X_i's are the ordered observations and $(i - 0.5)/n$ is the empirical distribution function $F_n(X)$. Such estimator is also used by Muhammad et al. (1993). Greenwood (1979) explained the robustness of the PWM over the conventional moments to outliers by drawing more efficient inferences using PWM.

A heteroscedastic linear regression model and a usual EWLS estimator are given below. In addition, a new estimator based on probability weighted moments, denoted as PEWLS estimator, is presented. Finally, empirical results, an application for this approach and conclusions are put forth.

Methodology

Linear Regression Model with Heteroscedastic Errors and EWLS

Consider the following heteroscedastic linear regression model:

$$y_{ij} = x'_i\,\beta + u_{ij},\; i=1, 2, \ldots, k, j=1, 2, \ldots, n_i,\; \sum_{i}^{k} n_i = n \qquad (2.1)$$

where y_{ij} is the jth response at the ith design point x_i, x_i are known p-vectors, β is a p-vector of unknown parameters and u_{ij} are the mutually independent with $E\,(u_{ij}) = 0$ and $E\,(u_{ij}^2) = \sigma_i^2$, $j = 1, 2, \ldots, n_i$. The variances σ_i^2's are unknown and heteroscedastic. A matrix form of model (2.1) is

$$y = X\beta + u, \qquad (2.2)$$

where

$$y = (y_{11} \;.\;.\;.\; y_{1n_1} \;.\;.\;.\; y_{k1} \;.\;.\;.\; y_{kn_k})'_{n\times 1},$$

$$u = (u_{11} \;.\;.\;.\; u_{1n_1} \;.\;.\;.\; u_{k1} \;.\;.\;.\; u_{kn_k})'_{n\times 1},$$

and

$$X = (x_{11} \;.\;.\;.\; x_{1n_1} \;.\;.\;.\; x_{k1} \;.\;.\;.\; x_{kn_k})'_{n\times p},$$
$$x_{ij} = x_i, \quad j = 1, 2, \ldots, n_i,$$

with heteroscedastic error terms of covariance matrix Ω having typical ith diagonal elements σ_i^2.

The usual OLS estimator for β in (2.2) is

$$\hat{\beta}_{OLS} = (X'X)^{-1}X'y$$

Fuller and Rao (1978) presented EWLS estimator of β as

$$\hat{\beta}_{EWLS} = (X'\hat{\Omega}^{-1}X)^{-1}(X'\hat{\Omega}^{-1}y), \qquad (2.3)$$

where

$$\hat{\Omega} = \text{diag}\{\hat{\sigma}_1^2, \hat{\sigma}_2^2, \ldots, \hat{\sigma}_n^2\},$$

$$\hat{\sigma}_i^2 = n_i^{-1}\sum_{j=1}^{ni}(y_{ij} - x'_i\hat{\beta}_{OLS})^2.$$

PWM-based Adaptive Estimator (PEWLS)

Probability weighted moments are used as weights in transforming matrix $\hat{\Omega}$ in (2.3) and propose a new estimator as

$$\hat{\beta}_{PEWLS} = (X'\hat{\Phi}^{-1}X)^{-1}(X'\hat{\Phi}^{-1}y),$$

where

$$\hat{\Phi} = diag\{\hat{\phi}_1, \hat{\phi}_2, \cdots, \hat{\phi}_n\},$$

$$\hat{\phi}_i = \frac{\sqrt{\pi}}{n} \sum_{i=1}^{n} [Y_i - 2(1 - \frac{i - 0.5}{n})Y_i].$$

The estimate of the mean is $\frac{\sum_{i=1}^{n} Y_i}{n}$. The Y_i's are the ordered observations and $(i - 0.5)/n$ is the empirical distribution function $F_n(Y)$.

Results

A Monte Carlo study was performed on the model used by Jacquez, et al. (1968) among others in their numerical work.

$$y_{ij} = 1 + x_i + u_{ij}\ ; \quad i = 1, 2, 3, \ldots, k; \quad j = 1, 2, 3, \ldots, n_i \tag{4.1}$$

The u_{ij} are independently distributed with zero mean and variance σ_i^2. Different versions for the model (4.1) were used according to the following formations: n_i were set to be equal to m; m = 5, 10. k was chosen as k = 6, 8, 10. x_i were selected as ; for k = 6, x_i were (1, 2, 4, 7, 9, 10), for k = 8, x_i were (1, 2, 4, 5, 6, 7, 9, 10), and for k = 10, x_i were (1, 2, 3, 4, 5, 6, 7, 8, 9, 10). For each pair (m, k), two σ-pattern (data generating process: DGP) were chosen; DGP-I: $\sigma_i = (x_i +8)/9$, and DGP-II: $\sigma_i = (0.5\ x_i +1)/3$.

Different data sets are generated for each pair of (m, k) and σ-pattern for normal, uniform, and logistically distributed error terms. For each pair of (m, k) and σ-pattern, 2,000 simulations are run. On the basis of the generated data, in Table 4.1 and Table 4.2, the efficiency of the EWLS estimator relative to the PEWLS estimator for β, is compared as R.E = $MSE(\hat{\beta}_{PEWLS}) / MSE(\hat{\beta}_{EWLS})$.

The mean values of standard error of estimates of the regressions are compared by computing the ratios SE (PEWLS)/SE (EWLS). These ratios are shown in Table 4.3 and 4.4.

Table 4.1 shows the relative efficiencies under DGP-I. For normally distributed errors, PEWLS performs better than EWLS for all the pairs (m, k) in terms of efficiency. But for small samples (m = 5, k = 6), PEWLS is more efficient and the gain in efficiency reaches to 20% while comparing with that of EWLS. For m = 10, both estimators tend to become equal efficient as k increases from 6 to 10. For uniform and logistic errors, no substantial efficiency is observed while using PEWLS.

Table 4.2 (DGP-II) shows the same trend of efficiency as shown by Table 4.1 for all the tried error patterns. It is noted again that when m = 5 is fixed, the new proposed estimator shows more efficient behavior for small values of k, namely, for k = 6.

Table 4.3 and 4.4 show that the results of the adaptive estimator PEWLS are brightly encouraging with respect to the standard error of estimate for the fitted model even for all the selected pairs of (m, k) and the error patterns. For normal errors and small samples (m = 5), the results are quite impressive by using PWELS as compared to its competitor for all chosen k. The standard errors of estimates of the fitted model are about double for EWLS as compared to that of our proposed PEWLS (e.g., for k = 6, 8). Almost similar are the findings for the other tried error distributions so far. Same fashion of less standard error of estimates is observed for DGP-II in Table 4.4. These findings show that by using the proposed adaptive estimator, one can find better regression estimates as compared to that by using EWLS.

Application

To illustrate the computations of the proposed PEWLS estimators and to compare its performance with the EWLS, already available in the literature, take the example of compensation per employee ($) in Nondurable Manufacturing Industries of US Department of Commerce as quoted by Gujarati (2003, p. 392). This example is used to compare these findings in practical data with findings already available in the literature.

Table 5.1 reports the performance of OLS, EWLS and the proposed PEWLS estimators. First, OLS estimates are found and the presence of heteroscedasticity is noted by

Table 4.1: Relative Efficiency of PEWLS and EWLS Estimators of β (DGP-I)

k	Normal		Uniform		Logistic	
	$m = 5$	$m = 10$	$m = 5$	$m = 10$	$m = 5$	$m = 10$
6	0.8088	0.9779	0.9887	0.9344	0.9885	1.0000
8	0.8526	0.9839	0.9921	0.9617	0.9891	1.0051
10	0.9400	0.9899	0.9989	0.9625	0.9911	0.9656

Table 4.2: Relative Efficiency of PEWLS and EWLS Estimators of β (DGP-II)

k	Normal		Uniform		Logistic	
	$m = 5$	$m = 10$	$m = 5$	$m = 10$	$m = 5$	$m = 10$
6	0.8918	0.9915	1.0000	0.9268	0.9831	0.9943
8	0.9112	0.9652	0.9471	0.9915	0.9962	0.9952
10	1.0031	0.9705	1.0252	0.9966	0.9986	1.0000

Table 4.3: Ratios of Standard Error of Estimates of PEWLS and EWLS (DGP-I)

k	Normal		Uniform		Logistic	
	$m = 5$	$m = 10$	$m = 5$	$m = 10$	$m = 5$	$m = 10$
6	0.5721	0.9244	0.6592	0.9068	0.8470	0.9650
8	0.5944	0.9287	0.6720	0.9325	0.8169	0.9661
10	0.6515	0.9365	0.6263	0.9317	0.7474	0.9317

Table 4.3: Ratios of Standard Error of Estimates of PEWLS and EWLS (DGP-II)

k	Normal		Uniform		Logistic	
	$m = 5$	$m = 10$	$m = 5$	$m = 10$	$m = 5$	$m = 10$
6	0.6770	0.9477	0.7011	0.9616	0.6329	0.9899
8	0.6839	0.9234	0.6531	0.9577	0.7378	0.9965
10	0.6833	0.9307	0.7139	0.9603	0.6859	1.0011

using White's test (1980) with p-value 0.07. It is noted that the proposed estimator bear lower standard errors among all the remaining estimators presenting an adequate reliability for its adaptation. It is further noted that the proposed estimator give better R^2 and much improved standard errors of regression that confirms the adequacy of the fitted model. Similarly, the proposed adaptive estimator gives lowest Akaike Information Criteria (AIC) values that indicate the right specification of the weighting mechanism.

Conclusion

It was found that use of probability weighted moments as estimates of unknown heteroscedastic weights rather than conventional estimates of variances for unknown heteroscedastic errors while estimating a heteroscedastic linear regression model, makes more efficient estimations. This new formulation, considerably, contributes in reducing standard errors of estimates for fitted models. The gain in efficiency and the reduction

Table 5.1: Comparative Statistics

Estimators	Estimation of β_0			Estimation of β_1			R^2	S.E. of Regression	AIC
	$\hat{\beta}_0$	SE	*t*-statistic	$\hat{\beta}_1$	SE	*t*-statistic			
OLS	3417.70	81.04	42.17	148.81	14.40	10.33	0.9385	111.56	12.46
EWLS	3406.20	80.86	42.13	154.24	16.93	9.11	0.9645	126.54	12.71
PEWLS	3437.40	79.39	43.29	142.99	17.69	10.44	0.9842	103.87	12.31

in standard errors of estimates of regression model are appealing, especially, for small samples and thus make our new adaptation more attractive for many of practical situations of small samples.

References

Downton, F. (1966). Linear estimates with polynomial coefficients. *Biometrika 53*, 129.

Geary, R. C. (1966). A note on residual heretovariance and estimation efficiency in regression. *American Statistician, 20*, 30-31.

Greenwood, J. A., Landwehr, J. M., Matalas, N. C., and Wallis, J. R. (1979). Probability weighted moments: definition and relation to parameters of several distributions expressible in inverse form. *Water Resources Research, 15*, 1049-1054.

Gujarati, D. N. (2003). *Basic econometrics.* (4th ed.). NY: McGraw-Hill.

Horn, S. D., Horn, R. A., and Duncan, D. B. (1975). Estimating the heteroscedastic variances in linear models. *Journal of the American Statistical Association, 70*, 380-385.

Pasha, G. R. (1984). A comparative empirical study of WLS and MINQU-based estimators for small sample. *Karachi Journal of Mathematics,* 2, 31-42.

White, H. (1980). A heteroscedasticity-consistent covariance matrix estimator and direct test for heteroscedasticity. *Econometrica, 48*, 817-838.

Jacquez, J. A., Mathur, F. J., & Crawford, C. R. (1968). Linear regression with non-constant unknown error variances. *Biometrics, 24*, 607-627.

Muhammad, F., Ahmad, S., & Abdullah, M. (1993). Use of probability weighted moments in the analysis of means. *Biometrical Journal, 35*, 371-378.

Pasha, G. R. (1982). *Estimation methods for regression models with unequal error variances.* Unpublished Ph. D. dissertation, University of Warwick.

Pasha, G. R., & Ord, J. K. (1994). Adaptive estimators for heteroscedastic linear regression models, *Pakistan Journal of Statistics, 10*, 47-54.

Journal of Modern Applied Statistical Methods
November, 2008, Vol. 7, No. 2, 506-513

Variance Estimation in Time Series Regression Models

Samir Safi
The Islamic University of Gaza

The effect of variance estimation of regression coefficients when disturbances are serially correlated in time series regression models is studied. Variance estimation enters into confidence interval estimation, hypotheses testing, spectrum estimation, and expressions for the estimated standard error of prediction. Using computer simulations, the robustness of various estimators, including Estimated Generalized Least Squares (EGLS) was considered. The estimates of variance of the coefficient estimators produced by computer packages were considered. Models were generated with a second order auto-correlated error structure, considering the robustness of estimators based upon misspecified order. Ordinary Least Squares (OLS) (order zero) estimates outperformed first order EGLS. A full comparison of order zero and four estimators indicate that over specification is preferable to under specification.

Key words: Autoregressive models, auto-correlated, disturbances, ordinary least squares, generalized least squares.

Introduction

In the standard linear regression model,

$$y = X\,\beta + u\,, \qquad (1)$$

where y is the $(T \times 1)$ response variable; X is an $(T \times k)$ model matrix; β is a $(k \times 1)$ vector of unknown regression parameters; and u is a $(T \times 1)$ random vector of disturbances, it is well known that Ordinary Least Squares (OLS) yield unbiased, but inefficient estimates for the regression parameters with serially correlated disturbance structures. OLS regression estimates have larger sampling variances than the Generalized Least Squares (GLS) estimator which accounts for auto-correlated nature of disturbances.

An important consideration is the estimation of the standard errors of the estimators, because estimates of the variance enter into usual inference procedures such as prediction and confidence intervals, hypotheses testing, spectrum estimation, expressions for the estimated standard error of prediction, and other inferential procedures.

In practice, if using a statistical package to compute the OLS estimators the variance estimate produced would be based on $\sigma_u^2 \left(X'X\right)^{-1}$, which may be biased for the true variance $\sigma_u^2 \left(X'X\right)^{-1} X'\Sigma X \left(X'X\right)^{-1}$. For GLS estimation ($\Sigma$ known), on the other hand, the variance estimate is unbiased for the true variance of the GLS estimator. It is unclear, however, how the variance estimators for EGLS estimation behave. In order to investigate how well the variance estimators function in the different cases, the ratio of the variance of the OLS estimated variance to that for the estimated GLS estimators from the simulation results was computed.

The most commonly assumed process in both theoretical and empirical studies is the first-order autoregressive process, or AR(1), which can be represented in the autoregressive form as

$$u_t = \rho u_{t-1} + \varepsilon_t\,, \quad \varepsilon_t \sim \text{i.i.d. } N\left(0, \sigma_\varepsilon^2\right) \qquad (2)$$

where ρ is the first order autoregressive disturbance parameter. The second-order autoregressive process, or AR(2) error process, may be written

Samir Safi is Community Service and Continuing Education Dean. E-mail: samirsafi@gmail.com.

$$u_t = \phi_1 u_{t-1} + \phi_2 u_{t-2} + \varepsilon_t \qquad (3)$$

where ϕ_1 and ϕ_2 are the second-order autoregressive disturbance parameters.

Numerous articles describe the efficiency of the OLS coefficient estimator relative to the GLS estimator which takes this correlation into account. Safi & White (2006) have shown that, if the error structure is autoregressive and the dependent variable is non-stochastic and linear or quadratic, the OLS estimator performs nearly as well as its competitors. When faced with an unknown error structure, however, AR(4) may offer the best choice. Koreisha & Fang (2004) investigated the impact that the EIGLS correction may have on forecast performance. They found that, for predictive purposes, not much is gained in trying to identify the actual order and form of the auto-correlated disturbances or in using more complicated estimation methods such as GLS or MLE procedures which often require inversion of large matrices. Krämer & Marmol (2002) showed that OLS and GLS are asymptotically equivalent in the linear regression model with AR(p) disturbance and a wide range of trending independent variables, and that OLS based statistical inference is still meaningful after proper adjustment of the test statistics. Grenander & Rosenblatt (1957) gave necessary and sufficient conditions for X such that the OLS and GLS estimators have the same asymptotic covariance matrix. This class of X matrices includes polynomial and trigonometric polynomial functions of time.

In addition, it is known from Anderson's (1948) results that if the columns of observations on k independent variables are linearly dependent on a set of k eigen vectors of the variance matrix of the errors, then the efficiency of the OLS estimator will be identical with the GLS estimator for most values of the autocorrelation coefficient $|\rho| < 1$. By contrast, if this matrix is allowed to vary arbitrarily, the efficiency of the OLS relative to the GLS estimator with a known autocorrelation coefficient can approach zero. Good references of techniques for analysis in time series models are Anderson (1971) and Fuller (1996).

The GLS estimator based on an under parameterized AR(1) disturbance model structure with an estimated AR(1) coefficient denoted, EIGLS-AR(1) will have the highest variance estimation among the other estimators. For example, for some cases the variance estimation of EIGLS-AR(1) is at least more than six times higher than the OLS estimator. This indicates that EIGLS-AR(1) can be much less efficient than OLS.

This article is organized as follows: Simulation setup, definitions of the mean squared error of the variance for each of the regression coefficients, the bias and the variance of the estimated variance, and the ratio of the variance of the OLS estimated variance to that of four GLS estimators are introduced. Complete simulation results based on the variance of OLS and GLS estimated variance of each of the regression coefficients are shown and the ratio of variance estimation of OLS to that of GLS estimators for each of the regression coefficients is discussed. This simulation study was designed to compare the performance of different estimators and to characterize the effect of the design on the efficiency of OLS. Lastly, conclusions based on the comparison of the variance estimation of OLS and GLS on the regression coefficients is provided.

Methodology

The robustness of various estimators, including estimated generalized least squares (EGLS) was considered. These simulations examined the sensitivity of estimators to model misspecification.

The the ratios of the variances of the OLS estimator relative to four GLS estimated variances were compared: the GLS based on the correct disturbance model structure and known AR(2) coefficients denoted as GLS-AR(2); the GLS based on the correct disturbance model structure but with estimated AR(2) coefficients denoted as EGLS-AR(2); the GLS based on an under parameterized AR(1) disturbance model structure with an estimated AR(1) coefficient denoted as EIGLS-AR(1), and the GLS based on over parameterized AR(4) disturbance model structure with estimated AR(4) coefficients denoted as EIGLS-AR(4). AR(p) GLS corrections disturbances.

Three finite sample sizes (50, 100, and 200) and three non-stochastic design vectors of the independent variable were used; linear, quadratic, and exponential. A standard normal stochastic design vector of length 1,000 was generated, assuming the variance of the error term in AR(2) process was $\left(\sigma_{\varepsilon}^{2}=1\right)$. In addition, 1,000 observations were generated for each of the AR(2) error disturbances with four pairs of autoregressive coefficients: (.2,-.9), (.8,-.9), (.2,-.7), and (.2,-.1).

The regression coefficients β_0, and β_1 for an intercept and the slope were each chosen to equal one. Breusch (1980) has shown that for a fixed design, the distribution of $\frac{\hat{\beta}_{\text{EGLS}}-\beta}{\sigma_{u}^{2}}$ does not depend on the choice for β and σ_u^2, and the result holds even if the covariance matrix Σ is misspecified.

Definition 1

The simulation mean squared error ($\hat{\eta}_{\beta_j}$) of an estimated variance W, of the true variance (τ), is the function defined by $E_{\tau}\left(W-\tau\right)^{2}$. That is

$$\hat{\eta}_{\beta_j}=k^{-1}\sum_{i=1}^{k}\left(W_{ij}-\tau\right)^{2} \tag{4}$$

where j = 0,1, k is the number of simulations,

$$W_{ij}=\widehat{\text{Var}}\left(\beta_{ij}\right),\tau=\text{Var}_{T}\left(\beta_{j}\right).$$

An estimate with the smallest value in (4) indicates that it was the most efficient among other estimates.

Definition 2

The bias of an estimated variance (W), of the true variance (τ), is the difference between the expected value of W and τ. That is,

$$\hat{\delta}_{\beta_j}=E_{\tau}W-\tau \tag{5}$$

where

$$E_{\tau}W=k^{-1}\sum_{i=1}^{k}\widehat{\text{Var}}\left(\beta_{ij}\right).$$

An estimator whose bias is identically (in τ) equal to zero is called unbiased and satisfies $E_{\tau}W=\tau$ for all τ.

Note that $\tau=\text{Var}_{T}\left(\beta_{j}\right)$ is different for each case of the estimation procedure; since no known explicit formula exists for EGLS cases, this quantity is estimated from the simulation results in all cases.

Definition 3

The variance of the estimated variance (W), of the true variance (τ), is the difference between the estimated mean squared error ($\hat{\eta}_{\beta_j}$), and the bias of an estimated variance W, $\hat{\delta}_{\beta_j}$. That is,

$$\text{Var}\left(\widehat{\text{Var}}_{\beta_j}\right)=\hat{\eta}_{\beta_j}-\hat{\delta}_{\beta_j}^{2} \tag{6}$$

Definition 4

The ratio of the variance of the OLS estimated variance to that of GLS is

$$R_{\beta_j}=\frac{V_j}{V_{ji}} \tag{7}$$

where

$$V_j=\text{Var}\left(\widehat{\text{Var}}_{\beta_{j.\text{OLS}}}\right),$$
$$V_{ji}=\text{Var}\left(\widehat{\text{Var}}_{\beta_{j.\text{GLS}}}\right),$$
$$j=0,1,i=1,2,3,4$$

for four GLS estimates such that:

$$V_{j1}=\text{Var}\left(\widehat{\text{Var}}_{\beta_{j.\text{GLS}-\text{AR}(2)}}\right),$$
$$V_{j2}=\text{Var}\left(\widehat{\text{Var}}_{\beta_{j.\text{EGLS}-\text{AR}(2)}}\right),$$
$$V_{j3}=\text{Var}\left(\widehat{\text{Var}}_{\beta_{j.\text{EIGLS}-\text{AR}(1)}}\right),$$
$$V_{j4}=\text{Var}\left(\widehat{\text{Var}}_{\beta_{j.\text{EIGLS}-\text{AR}(4)}}\right).$$

A ratio (R_{β_j}), less than one indicates that the OLS estimate is more efficient than GLS, if R_{β_j} is close to one then the OLS estimate is nearly as efficient as GLS, otherwise, OLS performs poorly.

S-plus code was written to compute the ratio of the variance of the OLS estimated variance to that of GLS in (7) using the OLS and four GLS estimators.

Results

The simulation results based on the variances of OLS and GLS estimated variance of each of the regression coefficients using four GLS and OLS estimates are now discussed.

Tables (1) and (2) show the simulation results of the variances of OLS and four GLS estimated variance, $\text{Var}\left(\widehat{\text{Var}}_{\beta_0}\right)$ and $\text{Var}\left(\widehat{\text{Var}}_{\beta_1}\right)$ in (6), when the serially correlated disturbance is AR(2) process, under parameterized AR(1), and over parameterized AR(4) for linear design with all selected AR(2) coefficients and all sample sizes.

First, regardless of sample size, the selected autoregressive coefficients for all non-stochastic designs, OLS was more efficient than EIGLS-AR(1) in estimating both β_0 and β_1. This is shown in Table (1), when Φ = (.8,-.9) for a linear design with T=100, $\left[V_0, V_{03}\right]$= [7.9340E-04, 5.8309E-03] and $\left[V_1, V_{13}\right]$ = [8.0950E-04, 5.5626E-03]. For all cases EIGLS-AR(1) was the least efficient estimator. For example, when Φ = (.2,-.9) with T=200, V_{03} = 1.0822E-04 and V_{13} = 1.0868E-04.

Second, regardless of sample size and selected non-stochastic design, OLS was more efficient than GLS in estimating $\left(\beta_0, \beta_1\right)$ with Φ = (.2,-.1). For example, as shown in Table (2), with T=50, $\left[V_0, V_{01}\right]$ = [1.9062E-05, 2.5782E-05] and $\left[V_1, V_{11}\right]$= [1.9848E-05, 2.6844E-05]. Otherwise, the OLS estimator performed less efficiently than the GLS estimator. Furthermore, if Φ = (.2,-.1), OLS was more efficient than GLS estimates; EGLS- AR(2), and EIGLS-AR(4), for

Table 1: Panel (A) - Variances of OLS and GLS Estimators for Linear Design

		$(\Phi_1, \Phi_2) = (.2, -.9)$		$(\Phi_1, \Phi_2) = (.8, -.9)$	
Size	Estimator	V0	V1	V0	V1
50	VOLS	3.8994E-03	4.0602E-03	6.0748E-03	6.3253E-03
	VGLS AR(2)	1.9922E-06	2.4186E-06	1.3186E-05	1.5923E-05
	VEGLS AR(2)	2.9702E-06	3.5411E-06	2.0197E-05	2.3851E-05
	VEIGLS AR(1)	7.5176E-03	7.5862E-03	5.3587E-02	4.8020E-02
	VEIGLS AR(4)	2.1458E-05	1.9951E-05	1.2105E-04	1.1500E-04
100	VOLS	5.0757E-04	5.1788E-04	7.9340E-04	8.0950E-04
	VGLS AR(2)	2.3634E-07	2.6019E-07	1.5304E-06	1.6804E-06
	VEGLS AR(2)	3.0145E-07	3.2972E-07	2.2467E-06	2.4441E-06
	VEIGLS AR(1)	8.6461E-04	8.7130E-04	5.8309E-03	5.5626E-03
	VEIGLS AR(4)	1.6742E-06	1.7054E-06	9.0592E-06	9.2609E-06
200	VOLS	6.5781E-05	6.6444E-05	9.7015E-05	9.7993E-05
	VGLS AR(2)	3.2956E-08	3.4571E-08	1.5183E-07	1.5907E-07
	VEGLS AR(2)	4.0323E-08	4.2192E-08	2.3086E-07	2.4091E-07
	VEIGLS AR(1)	1.0822E-04	1.0868E-04	6.6333E-04	6.4879E-04
	VEIGLS AR(4)	1.7871E-07	1.8186E-07	1.0469E-06	1.0632E-06

all sample sizes for all design vectors. This is shown in Table (2). For a linear design with sample size T=100; the variances of the estimated variance of $\left(\beta_0, \beta_1\right)$ using OLS, EGLS-AR(2) and EIGLS-AR(4) were $\left[V_0, V_{02}, V_{04}\right]$ = [2.4432E-06, 1.9476E-05, 4.2741E-05] and $\left[V_1, V_{12}, V_{14}\right]$ = [2.4928E-06, 1.8614E-05, 3.6817E-05], respectively. Otherwise, GLS estimates were more efficient than OLS. The results for the other non-stochastic designs mimic the same behavior of the linear designs.

Table (3) shows the simulation results of the variances of OLS and GLS estimated variance for standardized normal stochastic design. OLS was more efficient than GLS estimators in estimating β_0 for all sample sizes with Φ = (.2,-.1). For example, when T=50, $\left[V_0, V_{01}, V_{02}, V_{04}\right]$= [2.0509E-05, 2.6708E-05, 2.2857E-04, 1.1503E-03].

For estimating the slope, β_1, OLS was nearly as efficient as GLS-AR(2), EGLS-AR(2), and EIGLS-AR(4) estimators for all sample sizes with AR(2) parametrization Φ = (.2,-.1). For example, when T=50, $\left[V_1, V_{11}, V_{12}, V_{14}\right]$= [4.5526E-05, 3.9870E-05, 3.8240E-05, 3.6470E-05]. Otherwise, OLS performed poorly. Second, the efficiency of OLS in estimating β_0 was more efficient than EIGLS-AR(1). For example, with AR(2) parametrization Φ = (.2,-.1) for T=50, $\left[V_0, V_{03}\right]$= [2.0509E-05,1.6262E-04]. However, the efficiency of OLS in estimating β_1 was nearly as efficient as EIGLS-AR(1), for example, with Φ = (.2,-.1) for T=50, $\left[V_1, V_{13}\right]$= [4.5526E-05, 4.0393E-05].

The simulation results based on the ratio of the variance of the estimated variance of OLS to that of GLS of each of the regression coefficients, R_β in (7) are now discussed. Tables (4) and (5) are presented for the linear design.

Table 2: Panel (B) - Variances of OLS and GLS Estimators for Linear Design

		(Φ_1, Φ_2) = (.2, -.7)		(Φ_1, Φ_2) = (.2,-.1)	
Size	Estimator	V0	V1	V0	V1
50	VOLS	1.7765E-04	1.8498E-04	1.9062E-05	1.9848E-05
	VGLS AR(2)	3.5205E-06	4.1751E-06	2.5782E-05	2.6844E-05
	VEGLS AR(2)	7.5677E-06	8.6003E-06	2.0664E-04	1.7487E-04
	VEIGLS AR(1)	4.2082E-04	4.1699E-04	1.5224E-04	1.3856E-04
	VEIGLS AR(4)	6.0168E-05	4.8976E-05	8.1543E-04	3.8291E-04
100	VOLS	2.4082E-05	2.4571E-05	2.4432E-06	2.4928E-06
	VGLS AR(2)	4.2622E-07	4.6385E-07	3.2092E-06	3.2743E-06
	VEGLS AR(2)	8.0958E-07	8.6773E-07	1.9476E-05	1.8614E-05
	VEIGLS AR(1)	5.1302E-05	5.1260E-05	1.9368E-05	1.8558E-05
	VEIGLS AR(4)	3.1549E-06	3.1480E-06	4.2741E-05	3.6817E-05
200	VOLS	2.8555E-06	2.8843E-06	3.0692E-07	3.1001E-07
	VGLS AR(2)	5.3668E-08	5.5979E-08	3.9671E-07	4.0070E-07
	VEGLS AR(2)	1.0279E-07	1.0652E-07	2.1704E-06	2.1293E-06
	VEIGLS AR(1)	5.7967E-06	5.7999E-06	2.2070E-06	2.1659E-06
	VEIGLS AR(4)	3.5784E-07	3.6114E-07	3.9794E-06	3.7836E-06

First, when the disturbance term is under parameterization, regardless of the sample size, the selected autoregressive coefficients, and for all the non-stochastic designs, OLS is more efficient than EIGLS-AR(1) in estimating both β_0 and β_1. For example, as shown in Table (4), when Φ = (.8,-.9) for the linear design with T=100, the ratio between V_0 and V_{03} for estimating the intercept, R_{β_0} is about 0.1361, and the ratio between V_1 and V_{13} for estimating the slope, R_{β_1} is about 0.1455. This result indicates that the variance of the OLS estimated variance would be around 0.1361 and 0.1455 times that of EIGLS-AR(1) for estimating the intercept and slope, respectively. This result shows that the variance estimation of EIGLS-AR(1) is at least more than six times higher than the OLS estimator. Moreover, for all cases EIGLS-AR(1) was the least efficient estimator.

Regardless of the example, shown in Table (5) with T=50, the ratio between V_0 and V_{01}, R_{β_0} = 0.7393 and the ratio between V_1 and V_{11}, R_{β_1} = 0.7394. Otherwise, the OLS estimator performed less efficiently than the GLS estimator.

When Φ = (.2,-.1), OLS was more efficient than GLS estimates; EGLS-AR(2), and EIGLS-AR(4), for all sample sizes for all design vectors. For example, as shown in Table (5), for the linear design with sample size T=100, the ratios between the estimated variance of (β_0, β_1) using OLS, EGLS-AR(2) and EIGLS-AR(4) were (0.1254,0.0572) and (0.1339,0.0677), respectively. Otherwise, OLS was less efficient than GLS estimates. The results for the other non-stochastic designs mimic the same behavior of the linear design.

Table 4: Panel (A) - Ratios of OLS and GLS Estimators for Linear Design

		$(\Phi_1, \Phi_2) = (.2, -.9)$		$(\Phi_1, \Phi_2) = (.8,-.9)$	
Size	Estimator	V0	V1	V0	V1
50	VOLS/VGLS2	1957.3616	1678.7466	460.7095	397.2509
	VOLS/VEGLS2	1312.8309	1146.5926	300.7824	265.2020
	VOLS/VEIGLS1	0.5187	0.5352	0.1134	0.1317
	VOLS/VEIGLS4	181.7205	203.5053	50.1835	55.0002
100	VOLS/VGLS2	2147.5934	1990.3878	518.4137	481.7247
	VOLS/VEGLS2	1683.7773	1570.6721	353.1306	331.2065
	VOLS/VEIGLS1	0.5871	0.5944	0.1361	0.1455
	VOLS/VEIGLS4	303.1820	303.6701	87.5786	87.4111
200	VOLS/VGLS2	1996.0370	1921.9448	638.9684	616.0478
	VOLS/VEGLS2	1631.3452	1574.7931	420.2377	406.7638
	VOLS/VEIGLS1	0.6079	0.6114	0.1463	0.1510
	VOLS/VEIGLS4	368.0923	365.3556	92.6731	92.1673

Table 5: Panel (B) - Ratios of OLS and GLS Estimators for Linear Design

		$(\Phi_1, \Phi_2) = (.2, -.7)$		$(\Phi_1, \Phi_2) = (.2,-.1)$	
Size	Estimator	V0	V1	V0	V1
50	VOLS/VGLS2	50.4616	44.3050	0.7393	0.7394
	VOLS/VEGLS2	23.4750	21.5081	0.0922	0.1135
	VOLS/VEIGLS1	0.4222	0.4436	0.1252	0.1432
	VOLS/VEIGLS4	2.9526	3.7769	0.0234	0.0518
100	VOLS/VGLS2	56.5013	52.9710	0.7613	0.7613
	VOLS/VEGLS2	29.7460	28.3162	0.1254	0.1339
	VOLS/VEIGLS1	0.4694	0.4793	0.1261	0.1343
	VOLS/VEIGLS4	7.6332	7.8053	0.0572	0.0677
200	VOLS/VGLS2	53.2076	51.5253	0.7737	0.7737
	VOLS/VEGLS2	27.7797	27.0774	0.1414	0.1456
	VOLS/VEIGLS1	0.4926	0.4973	0.1391	0.1431
	VOLS/VEIGLS4	7.9800	7.9868	0.0771	0.0819

Table (6) shows the ratio between the variance of OLS estimated variance and the variance of GLS estimates for all sample sizes for the standardized normal design.

First, with Φ = (.2,-.1) and all sample sizes, the ratio between the variance of OLS estimated variance and the variance of GLS estimates; GLS-AR(2), EGLS-AR(2), and EIGLS-AR(4) were significantly smaller than one for estimating an intercept. For example, when T=50, R_{β_0} = (0.7679, 0.0897, 0.0178). (See Table 6.) However, that ratio was slightly larger than the one for estimating the slope. For example, when T=50, R_{β_1} = (1.1419, 1.1905, 1.2483).

Second, regardless of sample size and AR(2) parametrization, the ratio between the variance of OLS estimated variance and the variance of EIGLS-AR(1) was significantly smaller than one for estimating an intercept. For example, with Φ = (.2,-.1) and T=50, R_{β_0} = 0.1261. However, the efficiency of OLS in estimating β_1 was nearly as efficient as EIGLS-AR(1). For example, with Φ = (.2,-.1) and T = 50, R_{β_1} =1.1271.

Conclusion

This study investigated the impact that variance estimators may have on inference based on the OLS estimator. The variance estimation is important because estimates of the variance enter into the usual inferential procedures such as confidence intervals, hypotheses testing, and spectrum estimation, as well as in expressions for the estimated standard error of prediction. The major finding is that, OLS (order zero) estimates outperform first order estimated generalized least squares, EIGLS-AR(1). In particular, the ratio of the variance estimation of the regression coefficients when the disturbance term is under parametrized, i.e. EIGLS-AR(1) has the highest ratio estimation among the other estimators. This indicates that EIGLS-AR(1) can be much less efficient than OLS.

Table 6: Ratios of OLS and GLS Estimators for Standardized Normal Design

		$(\Phi_1, \Phi_2) = (.2, -.7)$		$(\Phi_1, \Phi_2) = (.2,-.1)$	
Size	Estimator	V0	V1	V0	V1
50	VOLS/VGLS2	47.1201	12.0362	0.7679	1.1419
	VOLS/VEGLS2	27.7268	11.5177	0.0897	1.1905
	VOLS/VEIGLS1	0.5036	1.0508	0.1261	1.1271
	VOLS/VEIGLS4	6.1101	11.4395	0.0178	1.2483
100	VOLS/VGLS2	55.4232	13.1856	0.8045	1.1970
	VOLS/VEGLS2	31.4673	12.1070	0.1326	1.1990
	VOLS/VEIGLS1	0.5251	1.0636	0.1252	1.1595
	VOLS/VEIGLS4	8.9087	11.5746	0.0609	1.2367
200	VOLS/VGLS2	55.1432	13.6790	0.7756	1.1285
	VOLS/VEGLS2	31.4317	12.7201	0.1590	1.0819
	VOLS/VEIGLS1	0.5338	1.0553	0.1524	1.0836
	VOLS/VEIGLS4	10.0527	11.8398	0.0756	1.1029

References

Anderson, T. W. (1971). *The Statistical Analysis of Time Series*. John Wiley & Sons: New York, NY.

Anderson, T. W. (1948). On the theory of Testing Serial Correlation *Skandinavisk Aktuarietid skrift*, 31, 88-116.

Breusch, T. (1980). Useful invariance results for generalized regression models. *Journal of Econometrics, 13*, 327-340.

Fuller, W.A. (1996). *Introduction to Statistical Time Series, 2nd Edition*. John Wiley & Sons: New York, NY.

Grenander, U., & Rosenblatt, M. (1957). *Statistical Analysis of Stationary Time Series*. Wiley: New York, NY.

Koreisha, S. G., & Fang, Y. (2004). Forecasting with Serially Correlated Regression Models. *Journal of Statistical Computations and Simulation, 74*, 625-649

Krämer, W., & Marmol, F. (2002). OLS-Based Asymptotic Inference in Linear Regression Models with Trending Regressors and AR(p)-Disturbances. *Communications in Statistics: Theory and Methods, 31*, 261-270.

Safi, S., & White, A. (2006). The Efficiency of OLS in the Presence of Auto-Correlated Disturbances in Regression Models, *Journal of Modern Applied Statistical Methods*, *5,(1)*, 107-117.

Journal of Modern Applied Statistical Methods
May 2008, Vol. 7, No. 2, 514-525

Bootstrap Confidence Intervals and Coverage Probabilities of Regression Parameter Estimates Using Trimmed Elemental Estimation

Matthew Hall
Child Health Corporation
of America

Matthew S. Mayo
University of Kansas
Medical Center

Mayo and Gray introduced the leverage residual-weighted elemental (LRWE) classification of regression estimators and a new method of estimation called trimmed elemental estimation (TEE), showing the efficiency and robustness of TEE point estimates. Using bootstrap methods, properties of various trimmed elemental estimator interval estimates to allow for inference are examined and estimates with ordinary least squares (OLS) and least sum of absolute values (LAV) are compared. Confidence intervals and coverage probabilities for the estimators using a variety of error distributions, sample sizes, and number of parameters are examined. To reduce computational intensity, randomly selecting elemental subsets to calculate the parameter estimates were investigated. For the distributions considered, randomly selecting 50% of the elemental regressions led to highly accurate estimates.

Key words: Elemental subsets, elemental regression, robust regression, coverage probabilities.

Introduction

A popular method of finding a solution to the multiple linear regression model

$$Y = X\beta + \varepsilon \qquad (1.1)$$

is to make use of the ordinary least squares (OLS) solution:

$$\hat{\beta}_{OLS} = (X^tX)^{-1} X^t Y.$$

In this nomenclature, Y is a $n\times1$ vector of random observations, X is a $n \times p$ matrix of known constants, β is a $p \times 1$ vector of unknown parameters, and ε is a $n \times 1$ vector of random errors with $E(\varepsilon) = 0$ and $Var(\varepsilon) = \sigma^2I$. The OLS solution purposefully minimizes the sum of squared residuals

$$SSE(\hat{\beta}) = (Y - X\hat{\beta})^t (Y - X\hat{\beta}).$$

There are many reasons why this solution is desirable, such as ease of calculation and the well developed theory that supports it. However, the OLS solution is also known to be sensitive to outliers and/or violations of model assumptions.

Several attempts to develop solutions that are less sensitive to outliers have been developed. These include least absolute values (LAV) regression, which minimizes the sum of the absolute residuals, and L_p-norm regression, which minimizes the sum of the p^{th} powers of the absolute residuals. This article furthers the work of another method called the trimmed elemental estimator (TEE), first proposed by Mayo and Gray (1), that makes use of elemental subsets.

Elemental Subsets

In most cases when using model (1.1), n (the sample size) is much greater then p (the number of unknown parameters), and the system of equations becomes over-determined. However, in order to estimate

Matthew Hall is Senior Statistician, Department of Informatics. Email: matt.hall@chca.com. Matthew S. Mayo is Professor and Chair, Department of Biostatistics. Email him at MMAYO@kumc.edu.

$\beta_0, \beta_1, \beta_2, \ldots, \beta_p$, only $k = p+1$ observations are mathematically required. Thus, when solving the over-determined system, a choice must be made from infinitely many possible solutions in order to settle on a single regression line. One way to deal with this issue is to ignore the fact that only k observations are needed and to pool all n observations into a single system of k equations to solve: this is what OLS does. Alternatively, subsets of the data could be formed with exactly k observations, their corresponding fits found, and the best one taken: this is what LAV does. An even better method might be to take several of the fits in this scheme and use their combined information to settle on estimates. Mayo and Gray (1997) developed TEE for this purpose. Using either of these last two approaches makes use of elemental subsets and elemental regression.

An elemental subset of a data set is simply a subvector of the data. In the setting of model (1.1), a subvector $h = \{i_1, i_2, \ldots, i_p\}$ may be considered as a set of distinct indices from a set of n indices. X_h may be defined as the $p \times p$ submatrix of X containing the rows of X indexed by the subset h. Furthermore, Y_h can be defined as the corresponding $p \times 1$ subvector of Y. The solution to the elemental regression equation is given by:

$$\hat{\beta}_h = \left(\mathbf{X}_h^t \mathbf{X}_h\right)^{-1} \mathbf{X}_h^t \mathbf{Y}_h = \mathbf{X}_h^{-1} \mathbf{Y}_h .$$

With the advent of high speed computers, elemental regression has been revived from its forgotten past nearly 250 years ago. It was, in fact, a predecessor to least squares, introduced in 1755 by Boscovich. However, due to its computational intensity and the introduction of least squares, it fell out of favor with data analysts. The need for computational power is evident when considering even a small data set. For example, assume a sample size of 50 and the need to estimate three parameters. There are ${}_{50}C_3$ = 19,600 elemental subsets of the data that must be fit. This is clearly beyond human capability.

Renewed interest in elementals has occurred on many fronts. Going back to the early days of modern computers, Theil (1950) and Sen (1968) used elementals to develop simple linear regression estimators. On the diagnostics front, Rubin (1980), Hawkins (1993), and Welsch (1986) used elementals to detect outliers and perform other regression diagnostics. Rousseeuw and Bassett (1991) and Hawkins (1993) considered searching through the set of elemental regressions and selecting the optimal parameter estimates based on specified criteria. Hawkins further defined, for a specified fitting criterion, the best elemental estimator (BEE) as the optimal estimate over all elemental fits. Recently, Hawkins and Olive (2002) introduced the X-cluster algorithm as a form of elemental regression for large multiple regression datasets.

Mayo and Gray's (1997) contribution introduced regression estimators based on OLS in terms of elemental regression. Sheynin (1973) reported that Jacobi was the first to show that OLS can be viewed as a weighted average of elemental regressions:

$$\hat{\beta}_{OLS} = \frac{\sum_h \left|X_h^t X_h\right| \hat{\beta}_h}{\sum_h \left|X_h^t X_h\right|} = \sum_h \frac{\left|X_h^t X_h\right|}{\left|X^t X\right|} \hat{\beta}_h \qquad (1.2)$$

where h is the set of all possible elemental subsets and the single bars indicate determinates. Furthermore, the weights are defined as:

$$w_h = \frac{\left|\mathbf{X}_h^t \mathbf{X}_h\right|}{\left|\mathbf{X}^t \mathbf{X}\right|} .$$

Because these weights are between zero and one and must sum to one, OLS is a weighted average of the elemental regressions $\hat{\beta}_h$.

Mayo and Gray (1997) took this version of OLS and generalized it to a class of estimators which they called leverage-residual weighted elemental (LRWE) estimators of the form:

$$\hat{\beta}(\lambda, \rho) = \frac{\sum_h w[\lambda(h), \rho(h)] \hat{\beta}_h}{\sum_h w[\lambda(h), \rho(h)]} \qquad (1.3)$$

In this formulation, $\lambda(h)$ is a factor based on the leverage information for X_h, and $\rho(h)$ is a factor based on the degree of fit for the elemental regression h. The OLS version is observed (1.2) in this form where

$$\lambda(h) = \left|\mathbf{X}_h^t \mathbf{X}_h\right|, \ \rho(h) = 1 \text{ for all } h, \quad (1.4)$$

and

$$w[\lambda(h), \rho(h)] = \lambda(h)\rho(h).$$

OLS does not make use of the weight factor based on the degree of fit, $\rho(h)$. For this reason, in OLS, elemental regressions with extreme data points are weighted the same as those that behave normally. Thus, OLS can be easily influenced by the presence of outliers.

Trimmed Elemental Estimators

Instead of ignoring the goodness of fit of a regression to a set of elementals, $\rho(h)$ could be altered in the OLS formulation of (1.4). Mayo and Gray (1997) created what they called the trimmed elemental estimator (TEE) to trim out the elemental regressions that poorly fit the data or have extreme leverage. The benefit of such a strategy is to remove from consideration elemental regressions that are computed from outlying data, thus achieving a more robust regression. Using the same $\lambda(h)$ and $w[\lambda(h), \rho(h)]$ as in (1.4), they altered $\rho(h)$ to be the indicator function:

$$\rho(h) = \mathrm{I}\left(\sum_{i=1}^{n} |e_{hi}| \le (1-\alpha_p)100^{th} \text{ percentile of the } {}_nC_p \sum_{i=1}^{n} |e_{hi}| \text{ values}\right)$$

Here, α_p is a trimming constant between zero and one and $\sum_{i=1}^{n} |e_{hi}|$ is the sum of absolute residuals (SAE) resulting from the elemental estimate $\hat{\beta}_h$. Depending on the proportion of regressions one would like to remove from consideration as determined by their goodness of fit, α_p can be adjusted accordingly. Thus, many trimmed elemental regression estimators can be found and denoted by TEE(α_p).

Mayo and Gray (2001) used simulation results to show the robustness and efficiency properties of TEE(α_p) point estimates to normal and symmetric non-normal error distributions, a feature which OLS does not enjoy. Results showed that the TEE(α_p) offers high efficiency under normality and is very robust to non-normality. This article furthers their work by examining some bootstrap confidence intervals of the trimmed elemental estimator and their properties and reducing computational intensity through random selection of elemental regressions.

Methodology

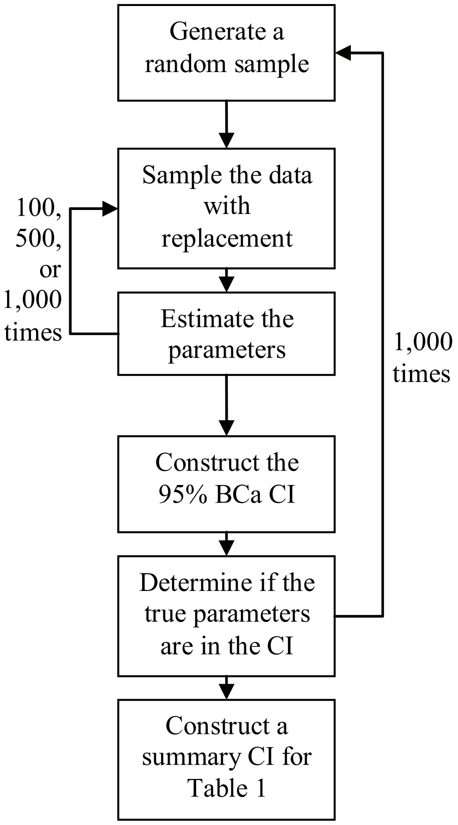
Simulation Design

Simulations were aimed at gaining a better understanding of the TEE(α_p) for inference by creating confidence intervals for the parameters and coverage probabilities under various scenarios. The objective was to compare these using the following methods: least absolute values (LAV), TEE(0.25), TEE(0.50), TEE(0.75), and OLS. Furthermore, a variety of error term distributions were assumed including: Normal, Laplace, Cauchy, 10% Contaminated Normal, and Student's t. These distributions were selected to provide a variety of weight in the tails of the distribution. In the simulations, Normal, Laplace, and t distribution parameter values had an error variance (σ^2) of 3.0. For the Normal distribution, standard normal variates were generated and multiplied by σ.

For Laplace, random variates from an exponential distribution were generated (mean = 1.0), randomly assigned a sign, and multiplied by $\sigma/2$. The Cauchy was the standard Cauchy distribution. For the 10% Contaminated Normal errors, standard normal variates were generated and-based on the value of a uniform random variate-were multiplied by either $\sqrt{5}\sigma$ (with probability 0.1) or σ (with probability 0.9). Finally, for the Student's t error distribution, three degrees of freedom were used in order for $\sigma^2 = 3$. The independent variable X was generated from a N(3,3) distribution.

In order to achieve the research goals, various quantities of 95% bias-corrected and accelerated (BC$_a$) bootstrap confidence intervals for OLS, LAV were calculated, and various trimmed elemental estimators and determined the number of times the true value of the parameter was in the intervals. Figure 1 shows the flowchart for the simulations.

Figure 1: Simulations flowchart.


Generate a random sample
Sample the data with replacement
100, 500, or 1,000 times
Estimate the parameters
1,000 times
Construct the 95% BCa CI
Determine if the true parameters are in the CI
Construct a summary CI for Table 1

The bootstrap is a well-developed approach to calculating approximate confidence intervals for parameter estimates when exact confidence intervals do not exist by repeatedly resampling the data with replacement. The BC_a method was introduced by Efron (1987) as an improvement to the bias-corrected (BC) method of Efron (1982) in order to provide confidence intervals for a wider class of problems. It constitutes a method for setting approximate confidence intervals for a parameter based on the percentiles of the bootstrap histogram, a bias correction, and an acceleration constant which measures how rapidly the standard error is changing on the normalized scale. For a complete review of various bootstrap confidence intervals including BC_a, see DiCiccio and Efron (1996). As a way of summarizing the BC_a confidence intervals, an overall 95% interval was calculated for each parameter. For this interval, the lower limit represents the value for which 2.5% of the lower boundaries of the BC_a confidence intervals are less than this value. Similarly, the upper limit represents the value for which 2.5% of the upper boundaries of the BC_a confidence intervals are greater than this value. All simulations were performed on a Dell 1.6GHz Pentium 4 computer with 1.0 GB of RAM using Digital FORTRAN 90.

In order to verify that the program was performing properly, the performance was tested using the two extreme methods under consideration: LAV, which takes only a single elemental regression, and OLS, which uses all of the elemental regressions. Comparing the parameter estimates ($p = 2$, $n = 25$) provided by the program for the three error distributions to the estimates provided by SAS© version 8e, agreement to five significant digits was obtained.

Results

In order to understand how the TEE(α $_p$) estimators would act under different situations, the following simulation scenarios were chosen:

a) a small sample size of 10 with two parameters;
b) a moderate sample size of 25 with three parameters;
c) a moderate sample size of 25 with two parameters; and
d) a large sample size of 100 with five parameters.

Sample sizes and number of parameters were chosen to limit computing time while allowing properties of the confidence intervals across a variety of scenarios to be ascertained. The results of simulations (c) and (d) are not presented here, they were performed to verify that the results did not change dramatically when the sample size and number of parameters was altered. The results of these simulations were very similar to the results discussed in greater detail below. Any exceptions are noted.

For these simulations, there were ${}_{10}C_2 = 45$, ${}_{25}C_3 = 2{,}300$, ${}_{25}C_2 = 300$, and, ${}_{100}C_5 = 75{,}287{,}520$ elemental subsets that had to be fit for each bootstrap sample, respectfully. For simulation (a), Table 1 shows the summary 95%

intervals for the BC$_a$ confidence intervals for β_1 using the method previously described. The smallest confidence interval in each scenario is highlighted. Figure 2 shows the coverage probabilities for the 1,000 BC$_a$ confidence interval created by the bootstrap (100, 500, or 1,000 samples) for β_1 from simulation (a). Similarly, Figure 3 shows the coverage probabilities for β_1 and β_2 from simulation (b).

From Table 1, it is evident that the summary intervals tend to tighten around the true values of the parameters as the number of bootstrap samples increase. As long as the error term is Normal or 10% Contaminated Normal, OLS does quite well. Furthermore, regarding the 1,000 bootstraps, it is apparent that OLS is difficult to distinguish from TEE(0.25) when the error is Normal, 10% Contaminated Normal, or Student's t. However, as expected, when the error term is either Cauchy or Laplace, OLS is clearly not the best choice. With a Cauchy error term, it appears that TEE(0.75) performs best for the slope regardless of sample size or the number of parameters (simulations (b) , (c), and (d) also showed TEE(0.75) to be superior). When the error follows the Laplace distribution, TEE(0.50) or TEE(0.25) seem to be the best (simulations (b), (c), and (d) showed TEE(0.50) to be slightly better than TEE(0.25)). In sum, it appears that TEE(0.50) performs very well for all of the error distributions considered. Although not shown, the results were very similar for the intercept in all four simulations with only slightly wider intervals. The parameter β_2 in simulation (b) had very similar results to those discussed above for β_1.

Figures 2 and 3 show how the different methods performed at covering the true values of the parameters with their 95% BC$_a$ confidence intervals for simulations (a) and (b), respectively. Although not shown in either figure, the confidence intervals for the intercept fail to include the true parameter more frequently than the slope confidence intervals. Nonetheless, the coverage probabilities for the intercept ranged from 0.90 to 0.97 for all simulations. Considering the 1,000 bootstrap samples (dashed lines) in the figures, the coverage probabilities for the error distributions studied ranges from 0.90 to 0.98. Thus, all of the methods captured the true values of the parameters quite well. However, regardless of the error distribution considered, TEE(0.50) appears to perform very consistently.

Furthermore, it is observed that either LAV or TEE(0.75) has the highest coverage probabilities, while OLS has the lowest for the error distributions under consideration. In fact, since the coverage probabilities were expected to be at 0.95, it is generally the case that LAV and TEE(0.75) performed above this level, TEE(0.50) and TEE(0.25) performed at this level, and OLS performed below this level. Hence, the coverage probability decreases as the trimming constant (α_p) decreases. The data from the other simulations were very similar and are not presented here. Once again, the coverage probabilities for β_2 in simulation (b) were similar to the probabilities for β_1 described above.

An objective in this article was to reduce the amount of necessary computations to achieve an acceptable estimate for the parameters using TEE(α_p). How this might be accomplished through random selection of elemental subsets as suggested by Hawkins (1993) for the BEE was investigated.

For simulation purposes, all of the elementals were first used to construct all of the elemental regressions $\hat{\beta}_h$. Specified proportions (30%, 50%, 70% or 90%) of these were then randomly selected in order to calculate parameter estimates through equation (1.3). This was performed with 10,000 data sets, and the median estimate was calculated for each error distribution at each percentage. The median was selected since it is a more robust measure of central tendency when compared to the mean. For β_1 when n=10 and p=2, the medians are displayed in Figure 4.

Using 50%, 70%, or 90% of the elemental regressions seems to provide accurate estimates for β_1 as long as the error distribution is one of those under consideration here. By selecting only 30% of the elemental regressions, the median estimates diverged further from the true value when compared to the other proportions, especially for the Normal, 10% Contaminated Normal, and the Student's.

Table 1: Summary intervals of 1,000 BC_a confidence intervals for β_1 when N=10, p=2. The true value is one.

	100 Bootstraps	500 Bootstraps	1000 Bootstraps
Normal			
LAV	(-2.645, 3.880)	(-2.023, 3.900)	(-1.930, 4.029)
TEE (0.75)	(-2.385, 3.537)	(-1.948, 3.706)	(-1.913, 3.952)
TEE (0.50)	(-2.032, 3.232)	(-1.620, 3.338)	(-1.525, 3.433)
TEE (0.25)	(-1.692, 2.950)	(-1.254, 3.065)	(-1.192, 3.111)
OLS	(-1.530, 2.838)	(-1.200, 2.994)	(-1.117, 3.106)
Cauchy			
LAV	(-29.597, 18.777)	(-31.469, 26.944)	(-25.958, 18.943)
TEE (0.75)	(-29.036, 16.883)	(-30.733, 26.192)	(-24.885, 18.313)
TEE (0.50)	(-31.722, 16.812)	(-28.499, 31.962)	(-29.893, 19.275)
TEE (0.25)	(-40.439, 27.037)	(-30.955, 31.913)	(-40.622, 24.040)
OLS	(-39.576, 31.148)	(-38.294, 38.800)	(-42.077, 22.391)
Laplace			
LAV	(-8.493, 7.962)	(-7.793, 8.521)	(-5.495, 7.960)
TEE (0.75)	(-8.335, 7.699)	(-7.340, 8.414)	(-5.157, 7.954)
TEE (0.50)	(-6.852, 6.901)	(-6.003, 7.533)	(-4.579, 6.931)
TEE (0.25)	(-6.895, 6.515)	(-5.709, 6.907)	(-4.794, 7.096)
OLS	(-7.371, 6.715)	(-5.719, 6.921)	(-4.974, 7.488)
Contam			
LAV	(-3.005, 4.390)	(-2.730, 4.278)	(-2.558, 4.666)
TEE (0.75)	(-2.876, 4.170)	(-2.685, 4.190)	(-2.528, 4.507)
TEE (0.50)	(-2.680, 3.965)	(-2.302, 6.644)	(-2.093, 4.187)
TEE (0.25)	(-2.517, 3.591)	(-1.948, 3.525)	(-1.635, 3.935)
OLS	(-2.470, 3.531)	(-1.807, 3.473)	(-1.672, 3.800)
T-distribution			
LAV	(-2.477, 3.895)	(-2.249, 3.555)	(-1.794, 4.330)
TEE (0.75)	(-2.554, 3.870)	(-2.161, 3.490)	(-1.842, 4.219)
TEE (0.50)	(-2.180, 3.537)	(-1.738, 3.164)	(-1.518, 3.894)
TEE (0.25)	(-1.808, 3.281)	(-1.482, 3.077)	(-1.288, 3.733)
OLS	(-1.746, 3.297)	(-1.447, 3.016)	(-1.280, 3.751)

Figure 2: Coverage probabilities of the 1,000 BC_a confidence intervals for β_1 when N=10 and p=2.

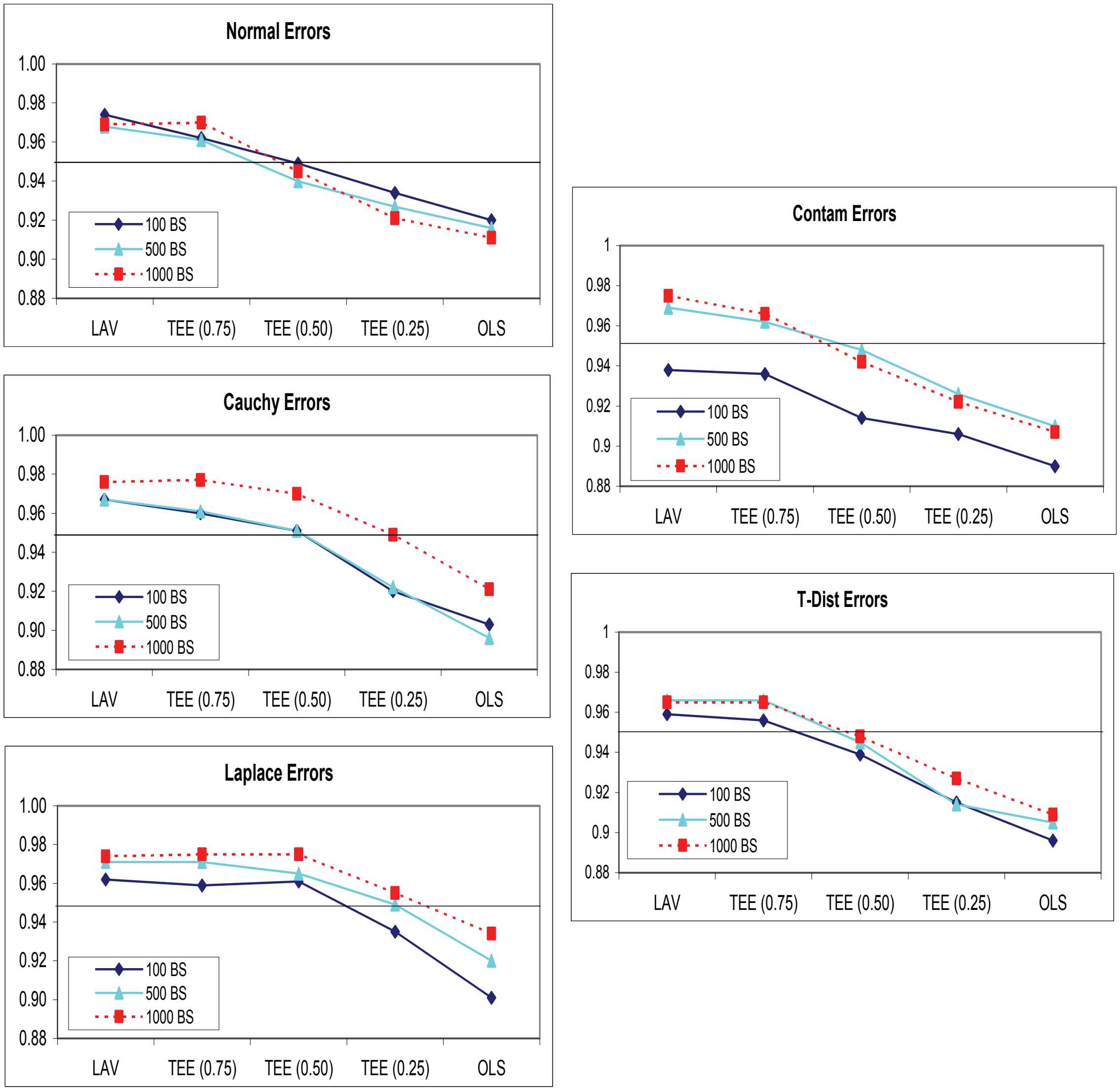
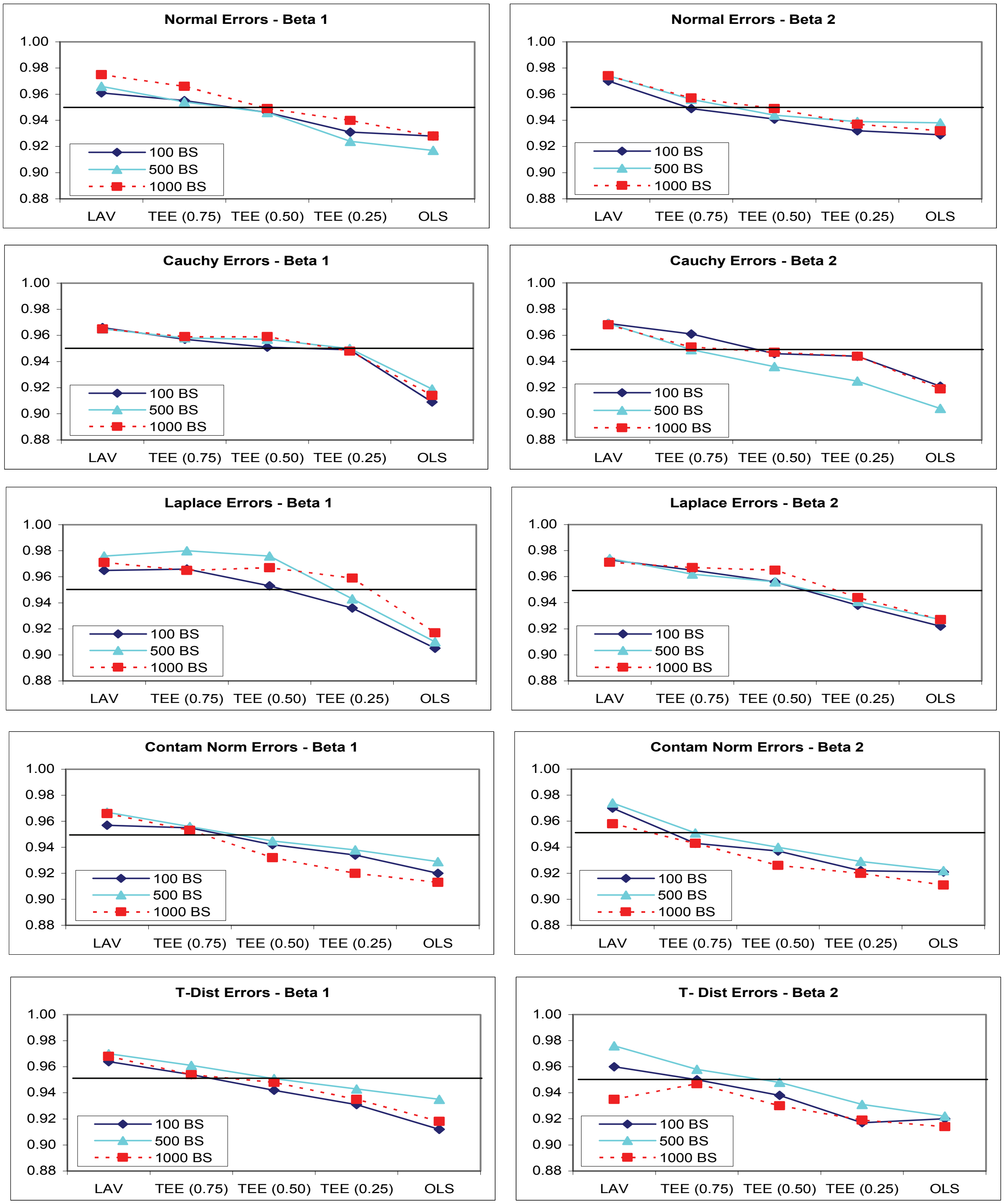


Figure 3: Coverage probabilities of the 1,000 BC$_a$ confidence intervals for β_1 (column 1) and β_2 (column 2) when N=25 and p=3.



Normal Errors - Beta 1

Normal Errors - Beta 2

Cauchy Errors - Beta 1

Cauchy Errors - Beta 2

Laplace Errors - Beta 1

Laplace Errors - Beta 2

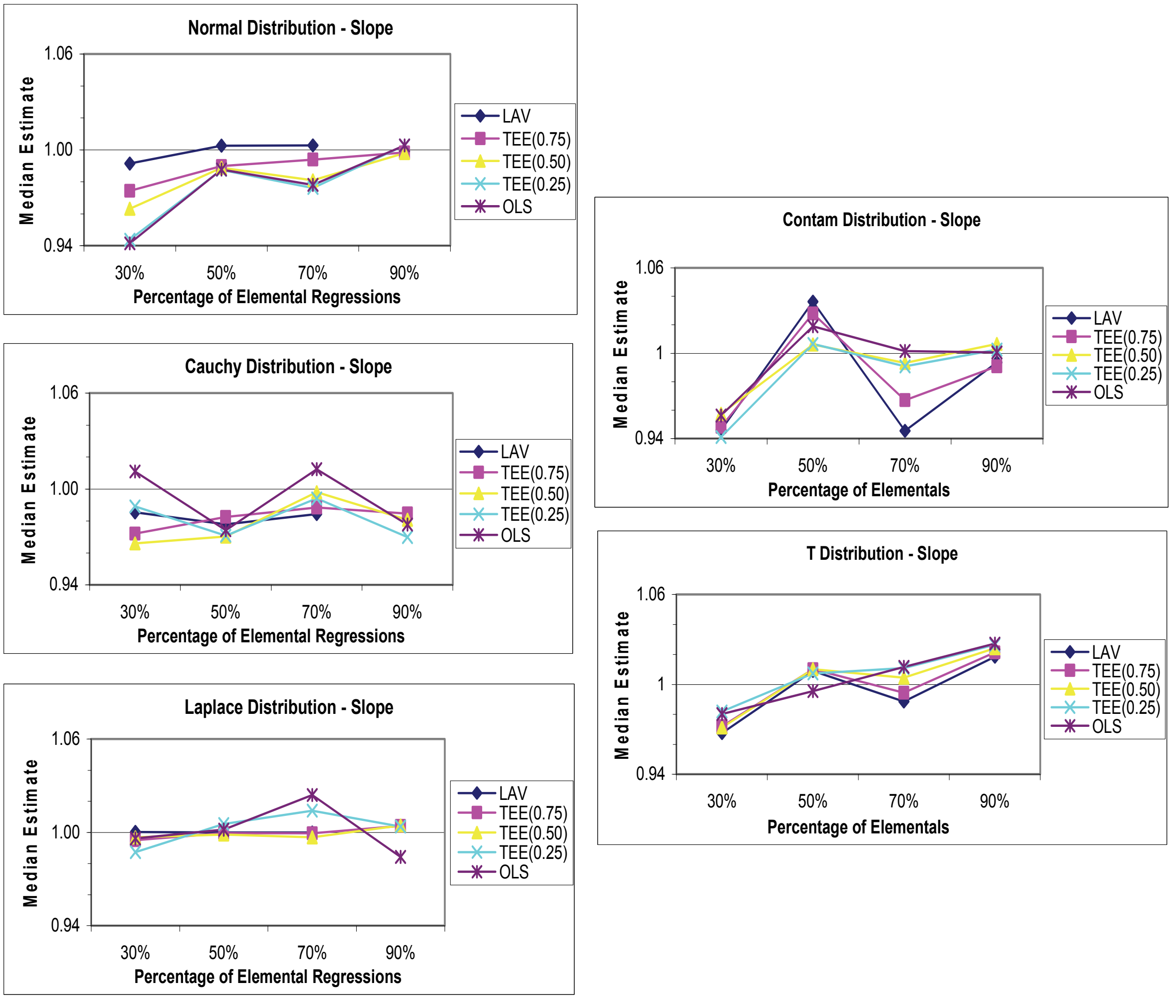
Contam Norm Errors - Beta 1

Contam Norm Errors - Beta 2

T-Dist Errors - Beta 1

T- Dist Errors - Beta 2

Figure 4: Median estimates for β_1 of 10,000 simulated data sets (N=10, p=2) using random selection of elemental regressions. The true value is one.



Thus, it appears that randomly selecting at least 50% of the elemental regressions is sufficient for producing accurate estimates. These results are similar for the intercept (data not shown) with the exception of using 50% of the elemental regressions with Laplace errors. In this situation, TEE(0.25) and OLS overestimated the intercept considerably. However, at 70%, the estimates behaved much more like those seen in Figure 4.

Figure 5 shows the coverage probabilities for the 95% BC_a confidence intervals using various quantities of bootstrap samples when n = 10, p = 2, and 50% of the elemental regressions are randomly selected. When similar simulations (n = 10, p = 2) are compared between Figure 5 (coverage probabilities when 50% of the elemental regressions are randomly selected) and Figure 2 (coverage probabilities without random selection), it is observed that results are quite similar. That is, while the coverage probabilities in Figure 5 are slightly higher than those in Figure 2, the trends seem similar. As was the case in Figure 2, generally speaking, LAV and TEE(0.75) over perform at the 95% level, TEE(0.50) and TEE(0.25) performed consistently at the 95% level, and OLS performed below the 95% level. Coverage probabilities from randomly selecting 70% and 90% of the elemental regressions produced similar results with the lines generally moving closer (as the percentage increased) to those observed in Figure 2.

Conclusion

The construction of BC_a confidence intervals for the trimmed elemental estimators have been demonstrated and their coverage probabilities have been. These are necessary extensions to Mayo and Grays original work and are additions to the development of TEE for inference purposes. In agreement with Mayo and Gray, this article showed that the trimmed elemental estimators are desirable in many situations. In fact, among those considered, they seem to be the clear choice when the error distribution is Cauchy or Laplace. Furthermore, for the Normal, 10% Contaminated Normal, or Student’s t error distributions, trimmed elemental estimators were found to be almost indistinguishable from OLS. In addition, TEE(0.50) performed consistently well in terms of estimation and coverage probabilities for all of the error distributions under consideration. It appears that a researcher could be fairly comfortable in choosing TEE(0.50), however knowledge of the process should guide this and utilization of traditional graphical procedures, such as residual and fitted value plots, might aid in determining the trimming constant. The TEE requires a large number of calculations as compared with OLS, therefore, it is desirable to use OLS when it is known that the assumptions for OLS are not violated and that there are no outliers present.

When data sets become larger and the number of parameters increases, increasing computational difficulties for LRWE estimators are present. Since there are ${}_nC_p$ elemental subsets that must be fit, ways must be found to decrease the number of computations. Hawkins (1993) suggested that using a random subsample of the elemental subsets would produce a good estimate for the best elemental estimator. This article examined such random subsamples to determine if this method is appropriate for reducing the number of calculations required for the trimmed elemental estimator. It was found that utilizing at least 50% of the elemental regressions generally provides good results as long as the error distribution is Normal, Cauchy, Laplace, 10% Contaminated Normal, or Student’s t. It was also observed that estimates tend to drift from the true value when random sampling falls to 30%.

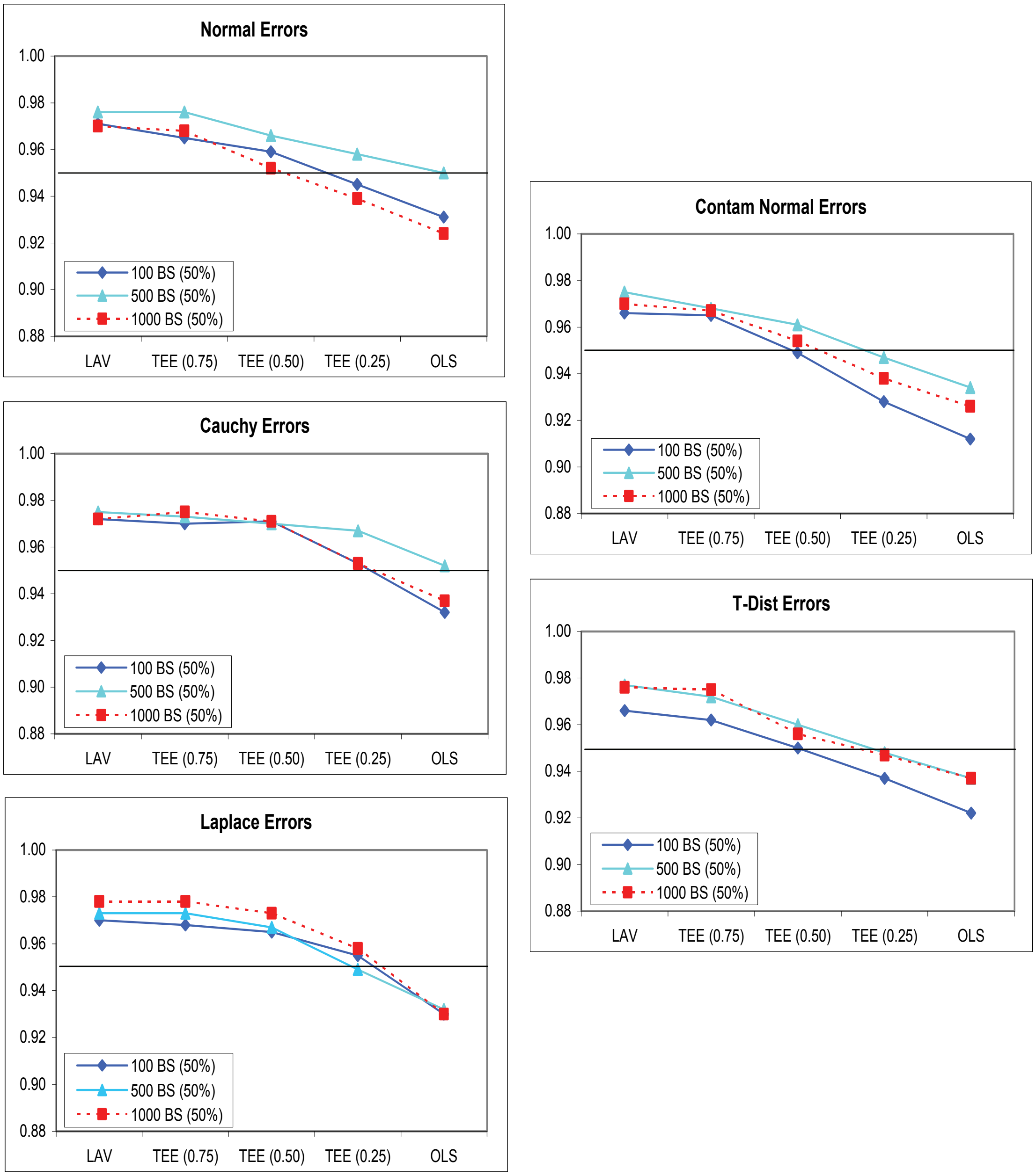
References

DiCiccio, T., & Efron, B. (1996). Bootstrap confidence intervals. *Statistical Science*, *11*, 189-212.

Efron, B. (1987). Better bootstrap confidence intervals. *Journal of the American Statistical Association*, *82*, 171-185.

Efron, B. (1982). The jackknife, the bootstrap, and other resampling plans. *Conference Board of the Mathematical Sciences*, *38,* Society for Industrial and Applied Mathematics – National Science Foundation.

Figure 5: Coverage probabilities of the 1,000 BC$_a$ confidence intervals for β_1 (N=10 and p=2) when randomly selecting 50% of the elemental regressions.



Efron, B., & Tibshirani, R. (1993). *An introduction to the bootstrap*. NY: Chapman and Hall.

Hawkins, D. (1993). The accuracy of elemental set approximations for regression. *Journal of American Statistical Association*, *88*, 580-589.

Hawkins, D., & Olive, D. (2002). Inconsistency of resampling algorithms for high-breakdown regression estimators and a new algorithm. *Journal of the American Statistical Association*, *97*, 136-148.

Mayo, M., & Gray, B. (1997). Elemental subsets: the building blocks of regression. *The American Statistician*, *51*, 122-129.

Mayo, M., & Gray, B. (2001). The robustness and efficiency of trimmed elemental estimation in regression analysis: a Monte Carlo simulation study. *Probabilistic Engineering Mechanics*, *16*, 323-330.

Rousseeuw, P., & Bassett, G. (1991). Robustness of the p-subset algorithm for regression with higher breakdown points. In W. Stahel (Ed.), *Robustness, Diagnostics, Computing and Graphics in* Statistics. NY: Springer.

Rubin, D. (1980). Composite points in weighted least squares regression. *Technometrics*, *22*, 343-348.

Sen, P. K. (1968). Estimates of the regression coefficient based on Kendall’s tau. *Journal of the American Statistical Association*, *63*, 1379-1389.

Sheynin, O.B. R.J. (1973). Boscovich’s work on probability. *Archive for History of Exact Sciences*, *9*, 306-324.

Theil, H. (1950). A rank invariant method of linear and polynomial regression analysis III. *Proceedings of the Royal Society A*, *53*, 1397-1412.

Welsch, R. (1986). Comment. *Statistical Science*, *1*, 403-405.

Journal of Modern Applied Statistical Methods
November, 2008, Vol. 7, No. 2, 526-534

1538 – 9472/08/$95.00

Least Squares Percentage Regression

Chris Tofallis
University of Hertfordshire
United Kingdom

In prediction, the percentage error is often felt to be more meaningful than the absolute error. We therefore extend the method of least squares to deal with percentage errors, for both simple and multiple regression. Exact expressions are derived for the coefficients, and we show how such models can be estimated using standard software. When the relative error is normally distributed, least squares percentage regression is shown to provide maximum likelihood estimates. The *multiplicative* error model is linked to least squares percentage regression in the same way that the standard additive error model is linked to ordinary least squares regression.

Key words: Regression, error measures, relative error, percentage regression, weighted least squares, multiplicative error, heteroscedasticity.

Introduction

When a regression model is used for prediction the size of the error is of interest. The magnitude of an error is not meaningful in isolation – it needs to be viewed in relation to the size of the observed or actual value. Percentage errors are often used for this purpose. Our definition of percentage error is 100 × (observed value − predicted value)/(observed value), as used in the fields of forecasting and time series analysis. In traditional least squares regression, an error of one unit is treated equally whether the dependent variable has a value of ten or a hundred, even though in percentage terms an error of one in ten would usually be considered more serious than an error of one in a hundred. In this article the method of least squares regression will be adapted to deal with percentage errors. There is a separate body of literature dealing with minimizing the mean absolute percentage error (MAPE), e.g. Narula & Wellington, 1977. This suffers from at least two deficiencies: (1) there is no formula for the coefficients (one must solve a linear programming problem to find them), and (2) the resulting parameter estimates may not be unique. The method presented in this article does not have these drawbacks.

It is important to highlight a difference between the above definition of relative error vs. (observed value − predicted value)/(predicted value). The latter was used by Book and Lao (1999) and Goldberg and Touw (2003). The question is: Should we compare the error with the actual observed value or with the value predicted from the model? The following may be one way of choosing. When dealing with a controlled scientific situation where the functional form of the underlying theoretical model is known, then any departures from the predictions may be due to measurement error; in this case, it may make sense to consider the error relative to the predicted value. If however, the ‘true’ underlying model or all its constituent variables are unknown then the ‘true’ value is also unknown and we recommend the approach taken here.

This is the usual situation in finance, economics, psychology and the other social sciences. For example, when forecasting the value of investments traded on the stock market it makes sense to relate prediction errors to the observed values. The same argument usually applies in the area of cost estimation. The people

Chris Tofallis is a Senior Lecturer in Operational Research at The Business School. He is a member of The Royal Statistical Society, and the Operational Research Society. Email: C.Tofallis@herts.ac.uk

paying the costs will find it more meaningful to assess the predictive ability of a cost-estimating relation (CER) using the error relative to what they actually paid, not relative to what the model predicted. Similarly, a prediction that a salary bonus would be $10k, but which actually turned out to be $5k corresponds to an error of 100% by the definition used in this article, whereas the other definition would rate this as only a 50% error in prediction.

The definition of relative error used here also has computational advantages over the other form. The minimization of the sum of squares of the other form cannot be solved exactly because the normal equations are a nonlinear system. Book and Lao (1999) noted numerical optimization techniques are usually necessary to find the coefficients; they pointed out that due to multiple local minima unreasonable solutions must be excluded, and that the is most plausible solution physically selected. Moreover, the resulting estimators are inconsistent. Goldberg and Touw (2003) explained the reason for this: “simply inflating the predictions in the denominator [of the relative error] will tend to deflate the percentage errors, at the expense of worsening the fit” (p. 62). This problem does not arise if the standard definition of relative error is used.

Before deriving the necessary equations for the coefficients, alternative approaches will be considered. Consider the simple case where a scatter plot of the data indicates that fitting a straight line (y = a + bx) is appropriate. One suggestion might be to use logarithms in the following way: regress ln(y) against x. The trouble with this is that the resulting model would ***not*** be a linear relationship between y and x, instead it would have ln(y) linearly related to x, and so y would be exponentially related to x. Although this does correspond to a straight line when the exponent is zero, the slope of the line is forced to be zero.

It is in fact a common misconception that regressing ln(y) is equivalent to minimizing the squared relative errors; it is approximately true only if all the errors are small, as then $\ln(\hat{y}/y) \approx (\hat{y}/y) - 1$. The difference in these regression models will be illustrated with a numerical example below.

Regressing ln(y) on ln(x), the fitted model is:

$$\ln(y) = A + B \ln(x),$$

hence,

$$y = \exp[A + B \ln(x)] = \exp(A)\, x^B,$$

which is a power law. For the case B = 1 this does correspond to a line, but it is forced to have a zero intercept and so passes through the origin.

Thus, both of these approaches involving log transformations are inadequate because they depart from a linear model in the original variables, which is our assumed starting point. Another suggestion might be to regress ln (y) on ln (a + bx). This is a non-linear problem requiring iterative computational procedures. By contrast, in the proposed approach exact expressions for the coefficients are available.

Derivation of Formulae for the Coefficients

An exact expression is now derived for the coefficients for percentage least squares regression. Let X be a matrix in which each column contains the data for one of the explanatory variables, and the first column contains the value 1 in each position. The aim is to obtain a coefficient b_i for each column variable, and the coefficient associated with the first column will be the constant.

The values of the dependent variable are contained in a column vector y, which is assumed strictly positive. The data in the ith row of the matrix is associated with the ith element of the y vector.

Traditionally, the sum of squared errors would be minimized, e^Te , where e denotes the vector of errors, y − Xb. (Superscript T denotes the transpose.) However, the primary interest is in the relative errors r (percentage error = 100 times relative error), so each error e_i needs to be divided by y_i , so $r_i = e_i / y_i$. Carrying out this division on the form y −Xb requires that the ith row of X be divided by y_i. This is achieved using the form r = Dy − DXb, where D is an *n by n* diagonal division matrix containing the value $1/y_i$ in the ith diagonal position and zeros elsewhere. D can be viewed as a matrix of weights.

Minimizing the sum of squares of relative errors $\sum r_i^2$, which, in vector notation, becomes

$$\begin{aligned} r^T r &= (Dy - DXb)^T (Dy - DXb) \\ &= (Dy)^T Dy \ - (Dy)^T DXb \\ &\quad - (DXb)^T Dy + b^T X^T D^2 Xb \end{aligned}$$

To find the minimum, differentiate this with respect to b and equate to zero:

$$-(Dy)^T Dx + X^T D^2 Xb = 0$$

This is the matrix equivalent of the normal equations of ordinary least squares regression. Notice that these equations have the great convenience of being linear in *b* and so can be easily solved.

Rearranging the previous equation:

$$X^T D^2 Xb = (DX)^T Dy,$$

and thus

$$\begin{aligned} b &= (X^T D^2 X)^{-1} (DX)^T Dy \\ &= (X^T D^2 X)^{-1} X^T D^2 y \qquad (1) \end{aligned}$$

It seems that this formula for the coefficients has not previously appeared as a solution for relative least squares.

If a spreadsheet is used for the calculations, the vector *b* can be computed directly using the matrix functions MINVERSE, MMULT (to multiply) and TRANSPOSE.

To satisfy the second order condition for a minimum, the second derivative of $r^T r$ with respect to b must be positive definite. This derivative equals $X^T D^2 X$ or $(DX)^T DX$. This square matrix will be positive definite if the columns of DX are linearly independent. Thus, the required *unique* minimum is obtained provided that no column of DX is expressible as a linear combination of the remaining columns.

If (1) is compared with the expression for ordinary least squares coefficients: $(X^T X)^{-1} X^T y$, observe that X has been replaced by DX, and y has been replaced by the vector Dy. Thus, D acts as a matrix of weights, as discussed further below.

In Ferreira et al. (2000)'s important article on relative least squares regression, expressions are derived for the coefficients, and also for their variance. They pointed out the connection between weighted least squares and relative least squares. Their formulae for the coefficients are in terms of ratios of determinants. These are less compact and less computationally convenient than the above formula (1), because a separate matrix has to be set up for each coefficient. A more practical computational method will be shown that can be applied using any standard software regression routine.

The consistency properties of relative least squares coefficients have been studied by Khoshgoftaar, et al. (1992). Using mild non-distributional assumptions such as independent error terms, a finite value for the expected measure of goodness of fit, and compact coefficient space, they prove that the coefficients are strongly consistent. That is, apart from a set of probability-measure zero, the coefficients will converge to the true values as the sample size increases.

Park and Stefanski (1998) also studied the best mean squared relative error prediction of y given x. Rather than provide formulae for coefficients, they assumed that some underlying distribution for y is given, and derive an expression for the predictor in terms of conditional inverse moments:

$$\hat{y} = E[y^{-1} \mid x] \,/\, E[y^{-2} \mid x].$$

They applied this using the lognormal and gamma distributions. They also showed that the mean squared relative prediction error is

$$\mathrm{var}\,(y^{-1} \mid x) \,/\, E[y^{-2} \mid x].$$

Observe that in their experience "engineers often think in terms of relative error" (p. 227), and that they were motivated to explore relative least squares by a consulting problem with environmental engineers, who "citing engineering and political reasons, were steadfast in their dissatisfaction with the usual prediction methods, that too frequently resulted in unacceptably large relative errors. They wanted a "simple, easily implemented, and generally applicable approach to predicting" (p. 228). Park and Shin (2005) applied this to stationary ARMA time series.

Returning to (1) for b and focusing on the simple straight-line case, it follows from the above that the slope for percentage regression is given by

$$b = \frac{\Sigma\frac{x}{y}\Sigma\frac{1}{y^2} - \Sigma\frac{1}{y}\Sigma\frac{x}{y^2}}{\Sigma\frac{1}{y^2}\Sigma\frac{x^2}{y^2} - (\Sigma\frac{x}{y^2})^2} \quad . \qquad (2)$$

(Note: all summations are from 1 to n, where n is the number of data points.) The intercept is given by

$$a = \frac{\Sigma\frac{1}{y} - b\Sigma\frac{x}{y^2}}{\Sigma\frac{1}{y^2}} \quad . \qquad (3)$$

The normal equation arising from differentiating r^Tr with respect to the intercept can be written in the form

$$\Sigma\frac{e}{y^2} = 0 \quad . \qquad (4)$$

This expression informs that the mean weighted error is zero if the weights are $1/y^2$. In vector terms this corresponds to $E[D^2 e] = 0$. From (3)

$$\Sigma\frac{1}{y_i} = a\Sigma\frac{1}{y_i^2} + b\Sigma\frac{x_i}{y_i^2}$$

it follows that there is a point through which the line will always pass (this would be the centroid of the data when using the ordinary least squares line). This is the point with coordinates given by

$$x = \frac{\Sigma\frac{x_i}{y_i^2}}{\Sigma\frac{1}{y_i^2}} \qquad y = \frac{\Sigma\frac{1}{y_i}}{\Sigma\frac{1}{y_i^2}}$$

Easy Computation by Transforming the Model Equation

Consider the model equation $y_i = a + bx_i + e_i$ and divide through by y_i , this yields

$$1 = \frac{a}{y_i} + b\frac{x_i}{y_i} + \frac{e_i}{y_i}. \qquad (5)$$

If ordinary least squares is used to regress the constant left hand side on the first two terms on the right, (notice there is now no constant term), then once again we are minimizing the sum of squared relative errors $\Sigma(e_i/y_i)^2$. Therefore, the same coefficients are derived, and the residuals will be the relative errors. This is a more convenient method of estimation, as even the Excel spreadsheet regression tool (part of the Analysis Toolpack) has the option to hold the constant to zero. Naturally, the above estimation approach carries over to the case of multiple explanatory variables.

The regression represented by (5) can be viewed as a novel form of weighted least squares with weights 1/y. Weighted least squares is a standard way of dealing with unequal variances (heteroscedasticity). In econometrics, for example, the heteroscedasticity problem has been dealt with by using weights which are a function of *one* of the explanatory variables and so some element of trial and error has been required to select this variable. (See, for example Greene, 2003, section 11.5). However, in this treatment it is not necessary to be concerned with choosing from the explanatory variables for the transformation, because the single dependent variable is used instead.

Saez and Rittmann (1992) carried out Monte Carlo investigations of relative least squares regression where the y-data does not have constant variance but does have constant *relative* variance. By using generated data they could compare estimated parameters with the known values from the generating model. They found that the 90% confidence regions for the coefficients were approximately centered on the true values, whereas this was not the case for ordinary least squares. The OLS confidence regions did not even always include the true values. The relative least squares confidence regions were also much smaller than those for

OLS. They concluded that relative least squares is superior to OLS for such heteroscedastic data.

Analysis of Relative Variance and Goodness of Fit

In ordinary least squares the disturbance term is orthogonal to each of the explanatory variables. From (5) the equivalent orthogonal relations for our weighted regression are:

$$\Sigma \frac{e_i}{y_i^{\,2}} = 0 \quad \text{and} \quad \Sigma \frac{e_i \; x_i}{y_i^{\,2}} = 0$$

The disturbance term is also orthogonal to the predicted dependent variable, which in this case corresponds to $\hat{y}_i/y_i$. Therefore

$$\Sigma \frac{e_i \; \hat{y}_i}{y_i^{\,2}} = 0 \quad \text{i.e. } \Sigma \frac{\hat{y}_i}{y_i}\left(1 - \frac{\hat{y}_i}{y_i}\right) = 0 \qquad (6)$$

Define the relative variance as:

$$\frac{1}{n}\sum \left(\frac{y - \bar{y}}{y}\right)^2$$

Ignoring the 1/n , this can be written as

$$\sum \frac{(\hat{y} - \bar{y} + y - \hat{y})^2}{y^2} =$$

$$\sum \frac{(\hat{y} - \bar{y})^2}{y^2} + \sum \frac{(y - \hat{y})^2}{y^2}$$

$$+ \sum \frac{(\hat{y} - \bar{y})(y - \hat{y})}{y^2}$$

The final term in the previous expression is zero as a consequence of the normal equations above.

Total relative variation =
Explained relative variation + Unexplained relative variation,

which is a decomposition of the relative variance.

A statistic can now be defined to measure the goodness of fit of our model, akin to r^2. The coefficient of relative determination is the ratio

$$\frac{\text{Explained relative variation}}{\text{Total relative variation}}$$

This ratio gives the proportion of the relative variation that is explained by the model. It will have a value in the range zero to one.

A Note on Measurement Scale

If all values of the dependent variable are re-scaled by multiplying by a positive constant, then the percentage errors remain unchanged. Consequently the resulting percentage least squares model will be equivalent to the original model, and it will provide equivalent predictions. For example if the y-variable is multiplied by 10 (e.g. due to conversion from centimeters to millimeters), then all coefficients in the fitted model equation will also be multiplied by 10.

If however, a constant is added to each value of the dependent variable then the percentage errors will not be the same as before. In this case the model fitted using percentage least squares will not be equivalent to the previously estimated model. The situation is exemplified when speaking of percentage changes in Fahrenheit temperature and percentage changes measured on the Celsius scale. The two are not the same because these scales do not share a common zero point. The dependent variable needs to be measured on a ratio scale when using percentage regression. This is because a percentage is not meaningful if one is permitted to shift the zero of the scale.

Maximum Likelihood

Is there a distribution for which the above estimators are maximum likelihood estimators? Consider the following multiplicative representation

$$y = X\beta u \qquad (7)$$

where u is multiplicative error factor, as opposed to an additive error term. Obviously, the expected value of u is desired to be unity, and thus the choice of the symbol *u*. E[y] = Xβ is desirable, so assume that the error factor is independent of the explanatory variables so that

E[y] = E[Xβ] E[u] = E[Xβ] = y

so that the estimate of the mean response will be unbiased.

Define $v_i = 1/ u_i$. Once there is an estimator *b* then the conditional estimate of the mean of y is ŷ = Xb , then

$$v_i = E[y_i]/y_i \qquad (8)$$

An error is indicated by this accuracy ratio differing from unity. Notice that $1 - v_i = r_i$, which is the relative error. Assume that the relative error is normally distributed with mean zero and constant variance (σ^2). This implies that v is normally distributed with mean value unity and constant variance (σ^2). [See the Appendix for the implications regarding the conditional distribution of y.] From (8), for any given x_i there is a one to one relationship between *y* and *v*. For a given data sample the likelihood function in terms of *v* is given by

$$\frac{1}{(2\pi\sigma^2)^{n/2}}\exp\left[-\frac{\sum_{i=1}^{n}(v_i-1)^2}{2\sigma^2}\right]$$

and in terms of y, the negative of the log likelihood becomes

$$\frac{\sum_{i=1}^{n}(\frac{X\beta}{y_i}-1)^2}{2\sigma^2}+\frac{n}{2}\ln\sigma^2+\frac{n}{2}\ln(2\pi) \qquad (9)$$

The summand is the square of the relative error, so it is now apparent that if the coefficient values are chosen to maximize the log likelihood, the same estimates for the coefficients as in (1) are obtained. The result is that when the relative error is normally distributed $N(1,\sigma^2)$ then the least squares percentage regression estimators are maximum likelihood estimators.

It is also possible to estimate σ^2 in the same way by differentiating the log likelihood with respect to σ^2 and setting the derivative to zero:

$$\frac{\sum_{i=1}^{n}(\frac{X\beta}{y_i}-1)^2}{2\sigma^4}-\frac{n}{2\sigma^2}=0$$

If the estimators are substituted for β, the following is obtained as the estimator for σ^2

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(\frac{Xb}{y_i}-1)^2$$

From (9) the log likelihood contains the sum of squared relative errors. If these are independent and identically distributed then for large *n*, the central limit theorem can be applied. This can be used to deduce a confidence interval for the coefficients.

Unbiasedness.

The estimator for β can be shown to be unbiased as follows. From (1)

$$\begin{aligned} E[b] &= E[(X^T D^2X)^{-1} X^T D^2 y] \\ &= E[(X^T D^2X)^{-1} X^T D^2 X \beta u] \\ &= E[\beta u] \quad . \end{aligned}$$

Assuming that the error factor is independent of β, we have: E[b] = E[β] E[u] = E[β] = β. Hence b is an unbiased estimator of β.

Example. The following table gives the sales figures from 18 different US industries, as well as the expenditure on research and development (millions of dollars). The sales variable has a wide range, and so it is likely that observations near the upper end will dominate over those at the lower end in positioning the regression line; this is because residuals for high sales are likely to be much larger. The correlation between the variables is 0.69 and a scatter plot shows evidence of heteroscedasticity.

Table 1. Sales and research & development expenditure in millions of dollars for 18 US industries.

Sales	R & D Expenses
6375	62.5
11626	92.9
14655	178.3
21869	258.4
26408	494.7
32406	1083
35108	1620.6
40295	421.7
70762	509.2
80553	6620.1
95294	3918.6
101314	1595.3
116141	6107.5
122316	4454.1
141650	3163.8
175026	13210.7
230614	1703.8
293543	9528.2

Source: Gujarati, 2003, page 424. Originally published in *Business Week* 1989.

If ordinary least squares is applied with sales as the dependent variable, the following model is obtained:

$$\text{Sales} = 43942 + 15.00\ \text{R\&D},$$

with p-values of 0.03 and 0.0015 for the intercept and slope respectively.

Consider the absolute percentage error (APE), defined as the residual expressed as a percentage of the observed value. The above model has a mean absolute percentage error (MAPE) of 105%, which is very poor. In fact three of the 18 industries have APEs exceeding 200%. The largest APEs occur for those industries which have low sales.

Some analysts advise taking logs of the dependent variable if one is interested in reducing percentage errors. If ordinary least squares is conducted, the following model is obtained:

$$\text{Ln(Sales)} = 10.341 + 0.000198\ \text{R\&D}$$

with p-values of 0.002 for the slope and essentially zero for the intercept. If the exponential is taken, it is possible to predict sales and calculate percentage errors. The MAPE is then 76%, which is an improvement. However, there are four industries with an APE exceeding 100%, three of these are at the lower end of the sales range.

Finally, consider our approach of minimizing the squared percentage residuals. The resulting model is found to be:

$$\text{Sales} = 8817 + 17.88\ \text{R\&D}$$

with p-values of 0.002 and 5×10^{-5} for the slope and intercept respectively.

The MAPE is now 38.5%. This is a large improvement as it is actually half of the percentage error from the log model. No residuals exceeded 100%, in fact the largest residual was 83%. The differences with the log model are worth emphasizing because it is a common misconception among statisticians that taking logs is equivalent to minimizing percentage errors. As mentioned in the introduction, this is true only in the limit as the residuals tend to zero.

Conclusion

Percentage error (relative to the observed value) is often felt to be more meaningful than the absolute error in isolation. The mean absolute percentage error (MAPE) is widely used in forecasting as a basis of model comparison, and regression models can be fitted which minimize this criterion. Unfortunately, no formula exists for the MAPE coefficients, and models for a given data set may not be unique. I have instead explored least squares regression based on the percentage error. I was able to derive exact expressions for the regression coefficients when the model is linear in these coefficients. Another advantage over MAPE is that this solution is unique.

The percentage errors are defined relative to the observed values. This is the standard definition of percentage error used in forecasting. When making predictions it usually makes more sense to relate the size of the error to the actual observation to measure its relative size. This is a departure from some of the existing literature on relative error least squares regression (e. g., Book & Lao (1999), & Goldberg & Touw, 2003), where the error relative to the *predicted* value has been used. The latter approach suffers on two counts. First, because the predicted values appear in the denominator of the fitting criterion, the latter value can be improved by inflating the predicted values, despite the fact that this worsens the fit (i.e., it gives biased estimates). Second, even for a linear model, estimation requires iteratively re-weighted least squares, which is computationally more demanding.

It has been shown that the proposed method is equivalent to a form of weighted least squares where, unusually, the weights depend on the dependent variable. This connection allowed us to develop a form which has great ease of computation. Indeed the models are attractive to the practitioner because they can easily be fitted using standard spreadsheet software. In comparing ordinary least squares with percentage least squares, the key difference is that the former ignores how large the residual is relative to the quantity being predicted, whereas the latter takes this into account. I believe that this method will be of use when the dependent variable has a wide range, as then the residuals at the upper end would dominate if ordinary least squares were used, unless the error variance is constant, which is often not the case in such situations.

It has also been shown that for a normally distributed multiplicative error model the least squares percentage estimators are maximum likelihood estimators. In short, the multiplicative error model is linked to least squares percentage regression in the same way that the standard additive error model is linked to ordinary least squares regression.

References

Book, S. A., & Lao, N. Y. (1999). Minimum-percentage-error regression under zero-bias constraints. *Proceedings of the 4th US Army Conference on Applied Statistics 1998*, US Army Research Lab report no. ARL-SR-84, pages 47-56.

Draper, N. R., & Smith, H. (1998). *Applied regression analysis*. 3rd edition. New York: Wiley.

Ferreira, J. M., Caramelo, L., & Chhabra, R. P. (2000). The use of relative residues in fitting experimental data: an example from fluid mechanics. *International Journal of Mathematical Education in Science and Technology*, 31(4), 545-552.

Goldberg, M. S., & Touw, A. E. (2003). *Statistical methods for learning curves and cost analysis*. Institute for Operations Research and the Management Sciences, Linthicum, MD.

Greene, W. H. (2003). *Econometric analysis*. 5th edition. New Jersey: Prentice Hall.

Gujarati, DN. (2003). *Basic econometrics*. 4th edition. NY: McGraw Hill

Khoshgoftaar, T. M., Bhattacharyya, B. B., & Richardson, G. D. (1992). Predicting software errors, during development, using nonlinear regression models: A comparative study. *IEEE Transactions on Reliability*, *41*(3) 390-395.

Narula, S. C., & Wellington, J. F. (1977). Prediction, linear regression and the minimum sum of relative errors. *Technometrics*, *19*(2), 185-190.

Park, H., & Shin, K-I. (2005). A shrinked forecast in stationary processes favouring percentage error. *Journal of Time Series Analysis*, *27*(1), 129-139.

Park, H., & Stefanski, L. A. (1998). Relative-error prediction. *Statistics and Probability Letters, 40*, 227-236.

Saez, P. B., & Rittmann, B. E. (1992). Model-parameter estimation using least squares. *Water Research*, *26*(6), 789-796.

Appendix: The distribution of y when the relative error is normally distributed

In deducing the maximum likelihood estimates, assume for a given x-value that the *relative* error ($r_i = 1 - \mu_y/y_i$) is normally distributed, N(0, σ^2). Consider the implication for the conditional distribution of y; from (8) $r_i = 1 - v_i = 1 - \mu_y/y_i$ and $v_i \sim N(1, \sigma^2)$. The conditional value of y should therefore follow the reciprocal normal distribution (not to be confused with the inverse normal). Specifically, use the change of variable rule to deduce the distribution of y_i for a given x_i (Greene, 2003, Appendix B6). This gives the following distributional form:

$$\frac{\mu_y}{y^2\sigma\sqrt{2\pi}}\exp\left[-\frac{\left(\frac{\mu_y}{y}-1\right)^2}{2\sigma^2}\right],$$

where σ is the standard deviation of the relative error, here assumed to have mean value unity. Figure 1 charts this density function for two values of σ.

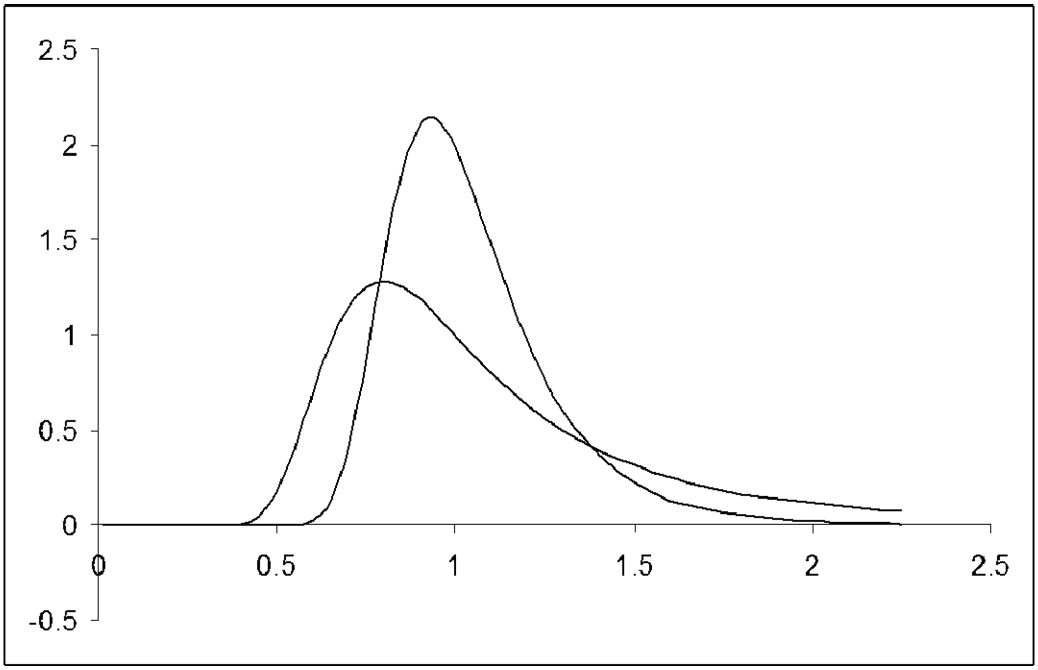


Figure 1. Probability density of y when the relative error is normally distributed with mean unity and σ = 20% (taller curve) and σ = 40% (shorter curve) .

Acknowledgements: I am grateful to Emeritus Professor MJR Healy, Dr. Paul Taylor, and Dr. Neil Spencer for sharing their statistical knowledge. Their comments on this work were most helpful. All errors remain my own responsibility.

Journal of Modern Applied Statistical Methods
November 2008, Vol. 7, No. 2, 535-545

Robust Predictive Inference for Multivariate Linear Models with Elliptically Contoured Distribution Using Bayesian, Classical and Structural Approaches

B. M. Golam Kibria
Florida International University

Predictive distributions of future response and future regression matrices under multivariate elliptically contoured distributions are discussed. Under the elliptically contoured response assumptions, these are identical to those obtained under matric normal or matric-t errors using structural, Bayesian with improper prior, or classical approaches. This gives inference robustness with respect to departure from the reference case of independent sampling from the matric normal or matric t to multivariate elliptically contoured distributions. The importance of the predictive distribution for skewed elliptical models is indicated; the elliptically contoured distribution, as well as matric t distribution, have significant applications in statistical practices.

Key words: Bayesian; Classical; Elliptically Contoured Distribution; Matric Normal; Matric-t; Multivariate Linear Model; Predictive Distribution; Robustness; Structural.

Introduction

The predictive inference for multivariate regression models has been researched extensively. For example, Guttman & Hougarrd (1985) considered the classical approach, Geisser (1965) and Zellner & Chetty (1965), Kowalski, et al. (1999), Thabane (2000), Thabane and Haq (2003), and Kibria, et al. (2002) considered the Bayesian method, Fraser and Haq (1969) considered the structural approach and Haq (1982) considered the structural relation of the model approach. The predictive distributions have been derived under assumptions of multivariate normal errors, but the assumption of normality and independency for error variables may not be appropriate in many practical situations, especially when the underlying distributions have heavier tails. For such cases, multivariate t-errors with liner models have been considered by several researchers, for example: Zellner (1976), Gnanadesikan (1977), Sutradhar and Ali (1989) and Kibria and Haq (1998, 1999a). In the case of the multivariate linear model, matric-t error has been considered by Kibria and Haq (2002) and Kibria (2006).

Using the structural relation of the model, Haq (1982) derived the predictive distribution for future responses under the matric normal distribution. He obtained the predictive distributions as matric-t with appropriate degrees of freedom. Kibria and Haq (2000) considered the predictive inference for future responses under the matric-t errors and obtained the predictive distribution as a matric-t with appropriate degrees of freedoms. Therefore, the distribution of a future response matrix is not affected by a change in the error distribution from matric normal to matric-t. The invariance of the predictive distribution for the future response matrix suggests that the predictive distribution would be invariant to a wide class of error distributions. A broader assumption is

B. M. Golam Kibria is an Associate Professor in the Department of Mathematics and Statistics at the Florida International University. He is the overseas managing editor of the *Journal of Statistical Research*, coordinating editor for the *Journal of Probability and Statistical Science*. He is an elected fellow of the *Royal Statistical Society* and the *International Statistical Institute.* Email him at: kibriag@fiu.edu.

considered here: that error terms have a multivariate elliptically contoured distribution. The elliptically contoured distribution includes various distributions: the multivariate normal, matric-t, multivariate Student's t, and multivariate Cauchy (see Ng 2000). The class of of normal distribution mixtures is a subclass of the elliptical distributions as well as the class of spherically symmetric distributions (Fang, et al., 1990).

Elliptically contoured distributions have been discussed extensively for traditional multivariate regression models by Anderson and Fang (1990), Fang and Li (1999), Kubokawa and Srivastava (2001), and Arellano-Valle, et al. (2006). This distribution has also been considered by Chib, et al. (1988), Kibria and Haq (1999b), Kibria (2003), and Kibria and Nadarajah (2006) in the context of predictive inference for linear regression models. Ng (2000) considered the model under the multivariate elliptically error contoured distribution using both Bayesian and classical approaches: he obtained the same predictive distribution with both approaches.

This article reviews predictive distributions for future response and future regression matrices under multivariate elliptically contoured error distributions. When the errors of model 1 are assumed to have an elliptically contoured distribution, the prediction distribution of future response and regression matrices are also obtained as matric-t distributions under structural relation, Bayesian, and classical approaches. The assumptions of normality and matric-t are robust to deviations in the direction of elliptical distributions as far as inferences about the future regression matrix and prediction is concerned. The distribution is said to be robust if it remains the same under violations of the normality assumption.

Methodology

Consider a set of n responses from the following multivariate linear model:

$$Y = \beta X + \Gamma E, \tag{1}$$

where Y is an $m \times n$ matrix of observed responses, β is an $m \times p$ matrix of regression parameters, X is a $p \times n$ $(n \geq p)$ known design matrix, Γ is an $m \times m$ matrix of scale parameter with $\Gamma\Gamma' = \Sigma$, where $|\Gamma| > 0$ and E is an $m \times n$ random error matrix. If it is assumed that E has a spherically contoured distribution with the probability density function:

$$f(E) \propto g\{tr(EE')\}, \tag{2}$$

(Anderson & Fang, 1990), where $g\{.\}$ is a non-negative function over $m \times m$ positive definite matrices such that $f(E)$ is a density function, then the response variable Y has an elliptically contoured distribution. Here E' denotes the transpose of the matrix E, and $tr(M)$ denotes the trace of the matrix M. To derive the prediction distribution,

$$B_E = EX(XX')^{-1} \tag{3}$$

and

$$S_E = (E - B_E X)(E - B_E X)'$$

are defined as the regression matrix of E on X and the sum of squares and product (SSP) matrix respectively. Consider C_E to be a non-singular matrix such that the error SSP matrix, S_E can be expressed as $C_E C'_E = S_E$, and the standardized residual matrix is:

$$W_E = C_E^{-1}(E - B_E X). \tag{4}$$

It follows from (4) that

$$E = B_E X + C_E W_E, \tag{5}$$

and, because $W_E W'_E = I_m$:

$$EE' = B_E XX'B'_E + C_E C_{E'}. \tag{6}$$

Considering a set of n_f future responses from the multivariate linear model defined in (1) as

$$Y_f = \beta X_f + \Gamma E_f, \qquad (7)$$

where Y_f and E_f are the $m \times n_f$ matrices of future responses and errors respectively, and X_f is an $p \times n_f$ ($n_f \geq p$) future design matrix. Assuming that E_f has the same distribution as E, then the joint distribution of E and E_f can be written as

$$f(E, E_f) \propto g\{tr(EE' + E_f E'_f)\}. \qquad (8)$$

Defining the quantities in (3) to (6) in terms of future errors as follows:

$$B_{E_f} = E_f X_f (X_f X'_f)^{-1} \qquad (9)$$

and

$$S_{E_f} = (E_f - B_{E_f} X_f)(E_f - B_{E_f} X_f)'$$

as the regression matrix of E_f on X_f and the sum of squares and product (SSP) matrix respectively. The standardized residual matrix and the future error matrix are respectively

$$W_{E_f} = C_{E_f}^{-1}(E_f - B_{E_f} X_f), \qquad (10)$$

and

$$E_f = B_{E_f} X_f + C_{E_f} W'_{E_f}. \qquad (11)$$

If $W_{E_f} W_{E_f}' = I_m$, then

$$E_f E_{f'} = B_{E_f} X_f X_{f'} B'_{E_f} + C_{E_f} C'_{E_f}, \qquad (12)$$

where $S_{E_f} = C_{E_f}'$ are the SSP matrix for future error variables.

Derivation of Predictive Distributions: The Structural Relation Approach

Following Fraser and Ng (1980), the joint density function of error statistics B_E, S_E, and E_f for given data (D) is obtained as

$$p(B_E, S_E, E_f \mid D) \propto$$
$$|S_E|^{\frac{n-m-p-1}{2}} g\left\{tr\left(B_E XX'B'_E + S_E + E_f E'_f\right)\right\}. \qquad (13)$$

To obtain the desired predictive distribution, the following transformation is made:

$$\begin{aligned} R &= S_E^{-\frac{1}{2}}(E_f - B_E X_f\} \\ U &= B_E \\ V &= V. \end{aligned} \qquad (14)$$

If the Jacobian of the transformation $J\{[E_f, B_E, S_E] \to [R, U, V]\}$ is equal to $|V|^{\frac{n_f}{2}}$, then the joint density of R, U, and V is

$$\begin{aligned} &p(R, U, V \mid D) \\ &\propto |V|^{\frac{n+n_f-m-p-1}{2}} g\left\{tr\left(UAU' + 2V^{\frac{1}{2}}RX'_f U' + V + V^{\frac{1}{2}}RR'V^{\frac{1}{2}}\right)\right\} \\ &\propto |V|^{\frac{n+n_f-m-p-1}{2}} g\left\{tr\left(tr(A^*) + tr(I_m + RHR')V\right)\right\}, \end{aligned} \qquad (15)$$

where

$A^* = (U + V^{\frac{1}{2}}RX'_f A^{-1})A(U + V^{\frac{1}{2}}RX'_f A^{-1})'$,
$H = (I_f - X_f A^{-1} X'_f)$, and $A = XX' + X_f X'_f$ is a symmetric matrix.

Following Ng (2000) in assuming that $I_m + RHR'$ is positive definite and Q is a non-singular matrix such that $Q'Q = I_m + RHR'$. The following transformation may be made:

$$\begin{aligned} Y &= QVQ' \\ Z &= U + V^{\frac{1}{2}}RX'_f A^{-1}, \end{aligned} \qquad (16)$$

the Jacobian of transformation is $|Q|^{-(m+1)}$, then the joint density function of R, Y and Z is as follows:

$$p(R,Y,Z\,|\,D) \;\propto\; |I_m + RHR'|^{-\frac{n+n_f-m}{2}}\,|Y|^{\frac{n+n_f-m-p-1}{2}}\,g\{tr(Y)+tr(ZAZ'\} \tag{17}$$

Integrating (17) with respect to Y and Z yields the density function of R as:

$$\begin{aligned} p(R\,|\,D) \;&\propto\; \iint p(R,Y,Z\,|\,D)\,dY\,dZ \\ &\propto\; |I_m + RHR|^{-\frac{n+n_f-p}{2}}\,. \end{aligned} \tag{18}$$

It may then be shown that:

$$\begin{aligned} R &= S_E^{-\frac{1}{2}}(E_f - B_E X_f\} \\ &= S_Y^{-\frac{1}{2}}(E_Y - B_Y X_f\}, \end{aligned} \tag{19}$$

where B_Y is the regression matrix of Y on X and $S_Y = (Y - B_Y)(Y - B_Y)'$ is the Wishart matrix. Thus, the prediction distribution of Y_f can be obtained from (18) and (19) as follows:

$$p(Y_f\,|\,D) \;\propto\; |I_m + S_Y^{-1}(Y_f - B_Y X_f)(I_{n_f} - X'_f A^{-1} X_f)(Y_f - B_Y X_f)'|^{\frac{n+n_f-m}{2}}, \tag{20}$$

which is a Matric-t density. The predictive distribution of the future responses for given data is an $m \times n_f$ dimensional matric-t distribution with $(n-p-m+1)$ degrees of freedom. The location parameter in the predictive density of Y_f is $B_Y X_f$ and the scale parameter matrix is $I_{n_f} - X'_f A^{-1} X_f$. This result coincides with that of Haq (1982), where he considered matric normal, and that of Kibria and Haq (2000) who considered the matric T error distribution. Thus, the predictive distribution of future responses are unaffected by departures from normality or dependent but uncorrelated assumptions to an elliptically contoured distribution. The shape parameter of the predictive distribution does not depend on the unknown parameter, instead, it depends on the sample observation and the dimension of the regression matrix.

Derivation of Predictive Distributions:

The Bayesian Approach

The density of $Y\,|\,\Sigma$ is given as

$$f(Y\,|\,\Sigma) \propto |\Sigma|^{-\frac{n}{2}}\, g\{tr(\Sigma^{-1}(Y-BX)(Y-BX)')\}, \tag{21}$$

Following Ng (2000), the Bayesian predictive distribution for future responses is obtained as follows. Suppose Y_f is an unobserved $m \times n_f$ of future observations, then the density function of (Y, Y_f) is given by:

$$\begin{aligned} &f(Y, Y_f\,|\,B,\Sigma) \propto \\ &|\Sigma|^{-\frac{n+n_f}{2}}\, g\{tr(\Sigma^{-1}[(Y-BX)(Y-BX)') \\ &+(Y_f - BX_f)(Y_f - BX_f)')]\}. \end{aligned} \tag{22}$$

The Bayesian predictive density of Y_f for given Y is defined as:

$$f(Y_f\,|\,Y) \propto \iint f(Y, Y_f\,|\,B,\Sigma)\,p(B,\Sigma^{-1})\,dB\,d\Sigma^{-1}, \tag{23}$$

where $p(B,\Sigma^{-1})$ is the non-informative prior density function of (B,Σ^{-1}) and is,

$$p(B,\Sigma^{-1}) \propto |\Sigma^{-1}|^{-\frac{m+1}{2}}\,. \tag{24}$$

The predictive density is obtained as

$$
\begin{aligned}
f(Y_f \mid Y) \quad &\propto \quad \iint |\Sigma|^{-\frac{n+n_f-m-1}{2}} \\
&\times \quad g\{tr(\Sigma^{-1}[(Y-BX)(Y-BX)') \qquad (25)\\
&+(Y_f - BX_f)(Y_f - BX_f)')]\} dBd\Sigma^{-1}.
\end{aligned}
$$

And the matrix expression in (25) can be rewritten as:

$$
\begin{aligned}
&(Y-BX)(Y-BX)')(Y_f - BX_f)(Y_f - BX_f)' = \\
&S_Y + (Y_f - \hat{B}X_f)H(Y_f - \hat{B}X_f)' \qquad (26)\\
&+(B-B^*)A(B-B^*)'
\end{aligned}
$$

where $B^* = (YX' + Y_f X_f)A^{-1}$. The matrices A and H are defined under equation (15). From the following transformation,

$$
\begin{aligned}
D &= \Sigma^{-\frac{1}{2}}(B-B^*) \\
G &= K\Sigma^{-1}K'
\end{aligned}
\qquad (27)
$$

where $KK' = S_Y + (Y_f - \hat{B}X_f)(Y_f - \hat{B}X_f)'$ and the Jacobian of the transformation $J[(B,\Sigma^{-1}) \rightarrow (D,G)$ is equal to $|G|^{-\frac{p}{2}} |K'K|^{-\frac{m-p+1}{2}}$, then (25) becomes

$$
\begin{aligned}
f(Y \mid Y_f) \quad &\propto \quad \iint |S_Y + (Y_f - \hat{B}X_f)H(Y_f - \hat{B}X_f)'|^{-\frac{n+n_f-k}{2}} \\
&|G|^{-\frac{n+n_f-m-p-1}{2}} \qquad g\{tr(G) + tr(DAD')+\}dDdG \\
&\propto \quad |I_m + S_Y^{-1}(Y_f - B_Y X_f)(I_{n_f} - X'_f A^{-1} X_f)(Y_f - B_Y X_f)'|^{\frac{n+n_f-m}{2}}. \qquad (28)
\end{aligned}
$$

Hence Y_f has a matric-t distribution with $n_f - m - p + 1$ degrees of freedom. Thus, the predictive distribution under the structural relation and the Bayesian approaches are the same.

Derivation of Predictive Distributions:
The Classical Approach

To obtain the predictive density of Y_f, it follows from Ng (2000) that $R = S_Y^{-\frac{1}{2}}(Y_f - \hat{B}X_f)$ is the studentized variable, and $S_Y^{-\frac{1}{2}}$ is the symmetric square root of S_Y^{-1}. Since R is invariant under the transformations $Y \rightarrow BX + CY$, $Y_f \rightarrow BX_f + CY_f$, for any non-singular square matrix C, it can be assumed, without loss of generality, that $B = 0$ and $\Sigma = I_m$ to derive the predictive distribution of Y_f. With this assumption, the joint density function of (Y, Y_f) becomes

$$
f(Y, Y_f) \quad \propto \quad g\{tr(YY' + Y_f Y'_f)\} \qquad (29)
$$

Because $YY' = S_Y + \hat{B}XX'\hat{B}'$ and, using the invariant differential in Fraser and Ng (1980), the joint density function of $\hat{B}_Y$, S_Y and Y_f is obtained from (29) as:

$$
f(\hat{B}_Y, S_Y, Y_f) \propto |S_Y|^{-\frac{n-p-k-1}{2}} \; g\{tr(S_Y + \hat{B}_Y XX'\hat{B}'_Y + Y_f Y'_f)\} \qquad (30)
$$

Making the transformation $R = S_Y^{-\frac{1}{2}}(Y_f - \hat{B}_Y X_f)$, followed by the Jacobian of the transformation is $|S_Y|^{\frac{n_f}{2}}$, the joint density of $\hat{B}_Y$, S_Y, R is:

$$
\begin{aligned}
&f(\hat{B}_Y, S_Y, R) \propto \\
&|S_Y|^{-\frac{n+n_f-p-k}{2}} \; g\{tr(S_Y + \hat{B}_Y XX'\hat{B}'_y \qquad (31)\\
&+(S_Y^{\frac{1}{2}}R + \hat{B}_Y X_f)(S_Y^{\frac{1}{2}}R + \hat{B}_Y X_f)')\}
\end{aligned}
$$

The matrix expression in (31) can be rewritten as:

$$
\begin{aligned}
&S_Y + \hat{B}_Y XX'\hat{B}_Y \\
&+(S_Y^{\frac{1}{2}}R + \hat{B}_Y X_f)(S_Y^{\frac{1}{2}}R + \hat{B}_Y X_f)' = \qquad (32)\\
&(I_m + RHR')S_Y \\
&+tr(\hat{B}_Y + S_Y^{\frac{1}{2}}RX'_f A^{-1})A(\hat{B}_Y + S_Y^{\frac{1}{2}}RX'_f A^{-1})'.
\end{aligned}
$$

Making the following transformation

$$
\begin{aligned}
Y &= QVQ' \\
Z &= U + V^{\frac{1}{2}} W X'_f A^{-1},
\end{aligned}
\qquad (33)
$$

and following procedures similar to the Bayesian Approach, the the joint density function of R, Y and Z is obtained as follows:

$$
p(R,Y,Z \mid D) \propto |I_m + RHR'|^{\frac{n+n_f-m}{2}} \, |Y|^{\frac{n+n_f-m-p-1}{2}} \, g\{tr(Y) + tr(ZAZ'\} \qquad (34)
$$

Integrating (34) with respect to Y and Z yields the density function of Y_f as:

$$
p(Y_f \mid D) \propto |I_m + S_Y^{-1}(Y_f - B_Y X_f)(I_{n_f} - X_{f'} A^{-1} X_f)(Y_f - B_Y X_f)'|^{\frac{n+n_f-m}{2}}, \qquad (35)
$$

which is a Matric-t density. The predictive distribution of the future responses for given data is an $m \times n_f$ dimensional matric-t distribution with $(n-p-m+1)$ degrees of freedom. Thus, the predictive distribution under the structural relation, Bayesian and classical approaches are the same.

Predictive Distribution of Future Regression Matrix

Based on the results in Kibria (2006), the joint density function of error statistics B_E, S_E, B_{E_f} and S_{E_f} are obtained as:

$$
p(B_E, S_E, B_{E_f}, S_{E_f} \mid E, X, X_f) \propto |S_E|^{\frac{n-m-p-1}{2}} \, |S_{E_f}|^{\frac{n_f-m-p-1}{2}} \times g\left\{tr\left(B_E XX'B'_E + S_E + B_{E_f} X_f X_{f'} B_{E_f}{}' + S_{E_f}\right)\right\}. \qquad (36)
$$

The structural relation of model (1) yields

$$
B_E = \Sigma^{-\frac{1}{2}}(B_Y - \beta) \quad and \quad S_E = \Sigma^{-1} S_Y, \qquad (37)
$$

and the Jacobian of the transformation $J\{[B_E, S_E] \to [\beta, \Sigma]\}$ is equal to $|S_Y|^{\frac{m+1}{2}} |\Sigma|^{-\left(\frac{p}{2}+m+1\right)}$. Thus, the joint density of β, Σ, B_{E_f}, and S_{E_f} is obtained as:

$$
\begin{aligned}
p(\beta, \Sigma, B_{E_f}, S_{E_f} \mid E, X, X_f) &\propto |S_{E_f}|^{\frac{n_f-m-p-1}{2}} \, |\Sigma|^{-\frac{n+m+1}{2}} \, g\left\{tr\Sigma^{-1}\left((B-\beta)XX'(B-\beta)'\right.\right. \\
&\quad \left.\left. + \; S + B_{E_f} X_f X_{f'} B_{E_f}{}' + S_{E_f}\right)\right\},
\end{aligned}
\qquad (38)
$$

where $B_Y = B$ and $S_Y = S$ for notational convenience. Similarly, the structural relation of the model (7) yields

$$
B_{E_f} = \Sigma^{-\frac{1}{2}}(B_{Y_f} - \beta)
$$

and

$$
S_{E_f} = \Sigma^{-1} S_{Y_f}, \qquad (39)
$$

where B_{Y_f} is the regression matrix for the future model, and S_{Y_f} is the Wishart matrix for the future responses. If the Jacobian of the transformation $J\{[B_{E_f}, S_{E_f}] \to [B_f, S_f]\}$ is equal to $|\Sigma|^{-\frac{p+m+1}{2}}$, then the joint density function of β, Σ, B_f, and S_f is obtained as

$$
\begin{aligned}
p(\beta, \Sigma, B_f, S_f \mid Y, X, X_f) &\propto |S_f|^{\frac{n_f-m-p-1}{2}} \, |\Sigma|^{-\frac{n+n_f+m+1}{2}} \\
&\quad g\left\{tr\left(\Sigma^{-1}\left[(B-\beta)XX'(B-\beta)'\right.\right.\right. \\
&\quad \left.\left.\left. + S + (B_f-\beta)X_f X_{f'}(B_f-\beta)' + S_f\right]\right)\right\},
\end{aligned}
$$

(40)

where $B_{Y_f} = B_f$ and $S_{Y_f} = S_f$.

The marginal density function of β, B_f and S_f is obtained from (40) as

$$\begin{aligned} p(\beta, B_f, S_f \mid Y, X, X_f) \propto{} & |S_f|^{\frac{n_f - m - p - 1}{2}} \int_\Sigma |\Sigma|^{-\frac{n+n_f+m+1}{2}} \\ & g\left\{tr\left(\Sigma^{-1}\left[(B-\beta)XX'(B-\beta)'\right.\right.\right. \\ & \left.\left.\left. + S + (B_f-\beta)X_f X'_f (B_f-\beta)' + S_f\right]\right)\right\} d\Sigma. \end{aligned} \tag{41}$$

To evaluate the integral in (41), let $\Sigma^{-1} = \Lambda$, then

$$d\Sigma = |\Lambda|^{-(m+1)}\, d\Lambda,$$

therefore,

$$\begin{aligned} p(\beta, B_f, S_f \mid Y, X, X_f) \propto{} & |S_f|^{\frac{n_f - m - p - 1}{2}} \int_\Lambda |\Lambda|^{\frac{n+n_f-m-1}{2}} \\ & g\left\{tr\left(\Lambda\left[(B-\beta)XX'(B-\beta)'\right.\right.\right. \\ & \left.\left.\left. + S + (B_f-\beta)X_f X'_f (B_f-\beta)' + S_f\right]\right)\right\} d\Lambda, \end{aligned} \tag{42}$$

Following Ng (2002), consider G to be a nonsingular matrix of order m such that

$$G^T G = \begin{bmatrix} (B-\beta)XX'(B-\beta)' + S \\ +(B_f-\beta)X_f X_{f'}(B_f-\beta)' + S_f \end{bmatrix}.$$

The transformation, $W = G\Lambda G^T$ has the Jacobian of the transformation as $|G^T G|^{-\frac{m+1}{2}}$, and integrating the above with respect to W yields the marginal density of β, B_f and S_f as,

$$\begin{aligned} p(\beta, B_f, S_f \mid Y, X, X_f) \propto{} & |S_f|^{\frac{n_f - m - p - 1}{2}} \\ & \begin{bmatrix} (B-\beta)XX'(B-\beta)' + S \\ +(B_f-\beta)X_f X_{f'}(B_f-\beta)' + S_f \end{bmatrix}^{-\frac{n+n_f}{2}} \\ & \int_\Sigma g\{tr(W)\} |W|^{\frac{n+n_f-m-1}{2}} dW \\ \propto{} & |S_f|^{\frac{n_f - m - p - 1}{2}} \\ \times{} & \begin{bmatrix} (B-\beta)XX'(B-\beta)' + S \\ +(B_f-\beta)X_f X'_f (B_f-\beta)' + S_f \end{bmatrix}^{-\frac{n+n_f}{2}}. \end{aligned} \tag{43}$$

The density function in (43) can further be expressed as

$$\begin{aligned} p(\beta, B_f, S_f \mid Y, X, X_f) \propto{} & |S_f|^{\frac{n_f - m - p - 1}{2}} \\ & \left[(\beta - FA^{-1})A(\beta - FA^{-1})' + S\right. \\ & \left. + (B_f - B)H^{-1}(B_f - B)' + S_f\right]^{-\frac{n+n_f}{2}}, \end{aligned} \tag{44}$$

where $F = BXX' + B_f X_f X'_f$, $A = XX' + X_f X'_f$ and $H = [XX']^{-1} + [X_f X'_f]^{-1}$.

The marginal density function of B_f and S_f are obtained by integrating β using matric-t argument (Press, 1982) from (44) as

$$
\begin{aligned}
p(B_f, S_f \mid Y, X, X_f) &\propto \int_\beta p(\beta, B_f, S_f \mid D)\, d\beta \\
&\propto |S_f|^{\frac{n_f - m - p - 1}{2}} \int_\beta \Big[(\beta - FA^{-1}) A (\beta - FA^{-1})' \\
&+ S + (B_f - B) H^{-1} (B_f - B)' + S_f\Big]^{-\frac{n+n_f}{2}} d\beta \\
&\propto |S_f|^{\frac{n_f - m - p - 1}{2}} \left[S + (B_f - B) H^{-1} (B_f - B)' + S_f\right]^{-\frac{n+n_f-p}{2}} .
\end{aligned}
\tag{45}
$$

Finally, the predictive distribution of the future regression matrix B_f is obtained as

$$
\begin{aligned}
p(B_f \mid Y, X, X_f) &\propto \\
\int_{S_f} |S_f|^{\frac{n_f - m - p - 1}{2}} &\left[S + (B_f - B) H^{-1} (B_f - B)' + S_f\right]^{-\frac{n+n_f-p}{2}} dS_f \\
&= \frac{\Gamma_m\left(\frac{n}{2}\right) |H|^{\frac{m}{2}}}{\pi^{\frac{mp}{2}} \Gamma_m\left(\frac{n-p}{2}\right)} |S|^{-\frac{p}{2}} \left|I_m + S^{-1}(B_f - B) H^{-1} (B_f - B)'\right|^{-\frac{n}{2}},
\end{aligned}
\tag{46}
$$

which is a Matric-t density. Thus the predictive distribution of the future regression matrix for given data is an $m \times p$ dimensional matric-t distribution with $(n - p - m + 1)$ degrees of freedom. That is

$$B_f \sim t_{m \times n_f}(B, H, S_Y, n - p - m + 1).$$

The predictive distribution of B_f is identical to that obtained under the assumption of matric normal error (Haq 1982). Thus, the predictive distribution of the future regression matrix is unaffected by departures from normality, or are dependent but uncorrelated assumptions to the elliptically contoured distribution. It may be concluded that the predictive distributions of a future regression matrix under structural, Bayesian and classical approaches are the same.

Conclusions

The predictive distribution of future responses for observed information under assumptions of multivariate elliptically contoured error distributions were considered, and the structural, Bayesian and classical approaches all resulted in the same predictive distributions. The predictive distributions under the elliptical errors assumption are identical to those obtained under independent normal errors or matric-t errors, thus showing robustness with respect to departure from the reference case of independent sampling from the matric normal or dependent, but uncorrelated sampling from matric-t distributions to elliptically contoured distributions. In the Classical approach, mild restictions were adopted, whereas the structural relation did not need those restrictions. The predictive distribution of the future regression matrix was also obtained as matric t. When $n_f = 1$, the predictive distribution of a single future response from a multivariate elliptically contoured distribution is obtained as a multivariate t distribution with $n - p - m + 1$ degrees of freedom. Findings in this article are more general, and include a linear model as a special case, as well as a variety of symmetric distributions. It is also noted that using the predictive distribution one can construct the β expectation tolerance regions for future response(s). In both application and theoretical aspects, these findings have potential applications in many areas of statistics.

There is great interest in the statistical literature toward robust statistical methods to represent strongly asymmetric data as adequately as possible and, at the same time, reduce the unrealistic ordinary normal or Student t assumptions. In scientific fields, such as gold concentration in soil samples (Galea-Rojas, et al., 2003), arsenate in water samples (Ripley & Thomson, 1987), cholesterol in blood samples (Lachos & Bolfarine, 2007) and many other situations, the data follow asymmetric distributions.

In such cases, normal or t distributions do not work well. Instead, certain types of skewed distributions are proposed in the literature to study the skewed data. These distributions allow for skewness and contain the normal or t distribution as a proper member or as a limiting case. Various kinds of skew distributions exist in the literature: skew-symmetric distributions (Gomez, et al., 2007), skew normal distribution (Azzalini, 1985, 1986), multivariate skew normal (Azzalini & Dalla Valle, 1996; Azzalini & Capitanio, 1999; Gupta, et al., 2004), skew *t* distribution (Jones & Faddy, 2003), generalized skew-*t* distribution (Theodossion, 1998), skew multivariate *t* (Azzalini & Capitanio, 2003; Gupta 2003), skew elliptical distribution (Branco & Dey, 2001; Dey & Liu, 2005; Fang 2003, 2005a, 2005b; Sahu & Chai, 2005), generalized skew elliptical distribution (Genton & Loperfido, 2005). The location and scale parameters of skewed elliptical distributions control the skewness and maintains the symmetry of the elliptical distributions.

They also provide an opportunity to study the robustness of normal theory procedures when both skewness and kurtosis are different from the normal. The skewed elliptical distributions are more useful to fit real data (Arnold & Beaver, 2000). Genton and Genton (2004) give an excellent review about skew-elliptical distributions and provide many new developments, including theoretical results and applications of skewed-elliptical distributions with real life data. Regression analysis with skewed elliptical distributions have been considered by Sahu, et al., (2003), for example. Unfortunately, predictive inferences with skewed elliptical models are limited or not available in the literature. It is necessary and to derive the predictive distribution when the error of the model follows the skewed elliptical distribution.

References

Anderson, T. W., & Fang, K. T. (1990). Inference in multivariate elliptically contoured distribution based on maximum likelihood. In *Statistical Inference in Elliptically Contoured and Related Distribution*, K. T. Fang & T. W. Andeson (*Eds.*), 201-216. NY: Allerton Press.

Arellano-Valle, R. B., Pino, G., & Iglesias, P. (2006). Bayesian inference in spherical linear models: robustness and conjugate analysis. *Journal of Multivariate Analysis*, *97*, 179-197.

Arnold, B. C., & Beaver, R. J. (2000). The skew-Cauchy distribution. *Statistical Probabilty. Letters*, *49*, 285-290.

Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scand.J.Statist.* 12, 171-178.

Azzalini, A. (1986). Further results on a class of distributions which includes the normal ones. *Statistica* 46, 199-208

Azzalini, A., & Capitanio, A. (1999). Statistical applications of the multivariate skew normal distribution. *Journal of the Royal Statistical Society*, *61*(B), 579-602.

Azzalini, A., & Capitanio, A. (2003). Distributions generate by perturbation of symmetry with emphasis on a multivariate skew *t* distribution. *Journal of the Royal Statistical Society*, *65*(B), 367-389.

Azzalini, A., & Dalla Valle, A. (1996). The multivariate skew-normal distribution. *Biometrika*, *83*, 715-726.

Branco, M. D., & Dey, D. K., (2001.) A general class of multivariate skew-elliptical distributions. *Journal of Multivariate Analysis*, *79*, 99-113.

Chib, S., Tiwari, R. C., & Jammalamadaka, S. R. (1988). Bayes prediction in regressions with elliptical errors. *Journal of Econometrics*, *38*, 349-360.

Dey, D. K. & Liu, J. (2005). A new construction for skew multivariate distributions. *Journal of Multivariate Analysis*, *95*, 323-344.

Fang, B. Q. (2003). The skew elliptical distributions and their quadratic forms. *Journal of Multivariate Analysis*, *87*, 298-314.

Fang, B. Q. (2005a). Noncentral quadratic forms of the skew elliptical variables. *Journal of Multivariate Analysis*, *95*, 410-430.

Fang, B. Q. (2005b). Invariant distribution of the multivariate tests in the skew elliptical model. *Statistical Methodology, 2*, 285-296.

Fang, K-T, Kotz, S., & Ng, V. M. (1990). *Symmetric Multivariate and related distributions*. London: Chapman and Hall.

Fang, K-T, & Li, R. (1999). Bayesian statistical inference on elliptical matrix distributions. *Journal of Multivariate Analysis*, *70*, 66-85.

Fraser, D. A. S., & Haq, M. S. (1969). Structural probability and prediction for the multivariate model. *Journal of the Royal Statistical Society*, *31*(B), 317-331.

Fraser, D. A. S., & K. W. Ng (1980). Multivariate regression analysis with spherical error. In *Multivariate analysis*, V. P. R. Krishnaiah (*Ed.*), 369-386. NY: North-Holland Publishing Co.

Galea-Rojas M., de Castilho M., Bolfariane, H., & De Castro M. (2003). Detection of analytical bias. *Analyst*, *128*, 1073-1081.

Geisser, S. (1965). Bayesian estimation in multivariate analysis. *Annals of Mathematical Statistics*, *36*, 150-159.

Genton, G. G. and and Genton, M. G. G. (2004). *Skew-Elliptical Distributions and Their Applications: A Journey Beyond Normality*. NY: Chapman and Hall.

Genton, M. G., & Loperfido, N. M. R. (2005). Generalized skew-elliptical distributions and their quadratic forms. *Annals of the Institute of Statistical Mathematics*, 57 (*2*), 389-401.

Gnanadesikan, R. (1977). *Methods for Statistical Data Analysis of Multivariate Observations*. NY: Wiley.

Gomez, H. W., Venegas, O., & Bolfarine, H. (2007). Skew-symmetric distributions generated by the distribution function of the normal distribution. *Environmetrics*, *18*, 395-407.

Gupta, A. K. (2003). Multivariate Skew t-Distribution. *Statistics*, *37*, 359-363.

Gupta, A. K., Gonzalez-Fariars, G., & Dominguez-Molina, J. A. (2004). A multivariate skew normal distribution. *Journal of Multivariate Analysis*, *89*, 181-190.

Guttman, I., & Hougaard, P. (1985). Studentization and prediction problems in multivariate multiple regression. *Communications in Statistics-Theory and Methods*, *14*, 1251-1258.

Haq, M. S. (1982). Structural relationships and prediction for the multivariate models. *Statistische Hifte*, *23*, 218-227.

Jones, M. C., & Faddy, M. J. (2003). A skew extension of the t-distribution, with application. *Journal of the Royal Statistical Society*, *65*(B), 159-174.

Kibria, B. M. G. (2003). Robust predictive inference for the multivariate linear models with elliptically contoured distribution. *Far East Journal of Theoretical Statistics*, *10*(1), 11-24.

Kibria, B. M. G. (2006). The matric t distribution and its applications in predictive inference. *Journal of Multivariate Analysis*, *97*, 785-795.

Kibria, B. M. G., & Haq, M. S. (1998). Marginal likelihood and prediction for the multivariate ARMA(1,1) linear models with multivariate t error distribution. *Journal of Statistical Research*, *32*(1), 71-80.

Kibria, B. M. G., & Haq, M. S. (1999a). The multivariate linear model with multivariate t and intra-class covariance structure, *Statistical Paper*, *40*, 263-276.

Kibria, B. M. G., & Haq, M. S. (1999b). Predictive inference for the elliptical linear model. *Journal of Multivariate Analysis*, *68*, 235-249.

Kibria, B. M. G., & Haq, M. S. (2000). The multivariate linear model with matric T error variables. *Journal of Applied Statistical Sciences*, *4*, 277-290.

Kibria, B. M. G., & Nadarajah, S. (2006). The predictive distribution for the heteroscedastic multivariate linear model with elliptically contoured distributions. *Journal of Statistical Research*, *40*(1), 35-43.

Kibria, B. M. G., Sun, L., Zidek, J. V., & Nhu, L. (2002). Bayesian spatial prediction of random space-time fields with application to mapping $PM_{2.5}$ exposure. *Journal of the American Statistical Association*, *97*, 112-124.

Kowalski, J., Mendoza-Blanco, J. R., Tu, X. M., & Gleser, E. J. (1999). On the difference in inference and prediction between the joint and independent t-error models for seemingly unrelated regression. *Communications in Statistics- Theory and Methods*, *29*, 2119-2140.

Kubokawa, T., & Srivastava, M. S. (2001). Robust improvement in estimation of a mean matrix in an elliptically contoured distribution. *Journal of Multivariate Analysis*, *76*, 138-152.

Lachos, V.H. and Bolfarine H. (2007). Skew-Probit Measurement Error Models. *Statistical Methodology*, 4, 1-12.

Ng, V. M. (2000). A note on predictive inference for multivariate elliptically contoured distributions. *Communications in Statistics-Theory and Methods*, *29*, 477-2000.

Ng, V. M. (2002). Robust Bayesian Inference for seemingly unrelated regression with elliptical errors. *Journal of Multivariate Analysis*, *80*, 122-135.

Press, J. (1982). *Applied multivariate analysis: Using bayesian and frequentist methods of inference*. FL: Robert E. Krieger Publishing Company.

Ripley, B., & Thompson, M. (1987). Regression techniques for detection of analytic bia. *Analyst*, *122*, 377-383.

Sahu, S. K., Dey, D. K., & Branco, M. D. (2003). A new class of multivariate skew distributions with applications to bayesian regression models. *The Canadian Journal of Statistics*, *31*(2), 129-150.

Sahu, S. K., & Chai, H. S. (2005) A new skew-elliptical distribution and its properties. S3RI Methodology Working Papers, M05/19). Southampton, UK, Southampton Statistical Sciences Research Institute, 33pp. http://eprints.soton.ac.uk/18180/.

Sutradhar, B. C., & M. M. Ali (1988). A generalization of the Wishart distribution for the elliptical model and its moments for the multivariate *t* model. *Journal of Multivariate Analysis*, *29*, 155-162.

Thabane, L. (2000). *Contributions to bayesian statistical inference*. Unpublished Ph. D. Thesis. Department of Statistical and Actuarial Sciences. London, Ontario, Canada.

Thabane, L. and Haq, M. S. (2003). The generalized multivariate modified bessel distributionsand its Bayesian applications. *Journal of Applied Statistical Science*, 11(3), 255-267.

Theodossion, P. (1998). Financial data and the skewed generalized *t* distribution. *Management Science*, *44*, 1650-1661.

Zellner, A. (1976). Bayesian and non-Bayesian analysis of the regression model with multivariate Student-*t* error terms. *Journal of the American Statistical Association*, *71*, 400-405.

Zellner, A., & Chetty, V. K. (1965). Prediction and decision problems in regression models from the Bayesian point of view. *Journal of the American Statistical Association*, *60*, 608-616.

Journal of Modern Applied Statistical Methods
November, 2008, Vol. 7, No. 2, 546-560

1538 – 9472/08/$95.00

Delete and Revise Procedures for Two-Stage Short-Run Control Charts

Matthew E. Elam
Texas A&M University-Commerce

This article investigates the effect different delete and revise procedures have on the performance of two-stage short-run control charting methodology in the second stage of its two stage procedure. Five variables control chart combinations, six delete and revise procedures, and various out-of-control situations in both stages are considered.

Key words: Delete and revise, two-stage, short-run, control chart, probability of detection, run length, false alarm, computer program, FORTRAN.

Introduction

Control charting in short-run situations has received much attention in the literature. In a short-run situation, little or no historical information is available about a process in order to estimate process parameters to begin control charting. The application of two-stage control charting, which is used to determine the initial state of the process and the control limits for testing the future performance of the process, to short-run situations has resulted in a Shewhart-based control chart methodology with control chart factors for finite numbers of subgroups (Hillier, 1969; Yang & Hillier, 1970).

The recent extension of this methodology to $(\overline{X}, s)$ (Elam & Case, 2005a) and (X, MR) (Elam & Case, 2008) control charts, as well as the computerization of the control chart factor calculations for two-stage short run $(\overline{X}, R)$ (Elam & Case, 2001), $(\overline{X}, v)$ and $(\overline{X}, \sqrt{v})$ (Elam & Case, 2003), $(\overline{X}, s)$ (Elam & Case, 2005b), and (X, MR) (Elam & Case, 2006) has allowed for its further examination. Of particular interest is the effect that different delete and revise (D&R) procedures have on the performance of the methodology in the second stage of the two-stage procedure. A D&R procedure removes out-of-control subgroups in stage one, allowing the data used to construct stage two control limits to be considered in-control. The removal of data in stage one becomes a more serious issue in a short-run situation because the less data available to construct stage two control limits, the less reliable they will be.

This article considers six different D&R procedures for establishing control of a process in the first stage of the two-stage procedure. The first D&R procedure (D&R 1) is from Hillier (1969), Ryan (1989), & Montgomery (1997). It executes as follows:

i. Deletes out-of-control (OOC) initial subgroups on either the control chart for centering or spread entirely (i.e., if a subgroup shows OOC on either control chart, it is deleted from both charts).
ii. Recalculates control limits for both charts using the subgroups remaining after step i.
iii. Determines OOC subgroups.
iv. Repeats steps i-iii until no initial subgroups show OOC on either chart.

The second D&R procedure (D&R 2) is from Pyzdek (1993). It executes as follows:

i. Deletes OOC initial subgroups on the control chart for spread.
ii. Recalculates control limits for the control chart for spread using the subgroups remaining after step i.
iii. Determines OOC subgroups.
iv. Repeats steps i-iii until no initial subgroups show OOC on the control chart for spread.

Matthew E. Elam is an Associate Professor of Industrial Engineering and is an ASQ Certified Quality Engineer. Email: Matthew_Elam@tamu-commerce.edu.

v. Determines control limits for the chart for centering using the parameter estimate for spread obtained after completing steps i-iv and the overall average obtained from all of the initial subgroups.
vi. Repeats steps i-ii for the control chart for centering until no initial subgroups show OOC.

The third D&R procedure (D&R 3) is from Case (1998). It deletes OOC initial subgroups on the control chart for spread just once. No D&R is performed on the control chart for centering.

The fourth D&R procedure (D&R 4) is from Doty (1997). It does not perform D&R. This means that all initial subgroups are used to determine second stage control limits for both the control charts for centering and spread.

The fifth D&R procedure (D&R 5) is a hybrid of D&R 1 in that it iterates only once. It deletes OOC initial subgroups on either the control chart for centering or spread entirely (i.e., if a subgroup shows OOC on either control chart, it is deleted from both charts). D&R is performed just once.

The sixth D&R procedure (D&R 6) is a hybrid of D&R 2 in that it iterates only once. It executes as follows:

i. Deletes OOC initial subgroups on the control chart for spread once.
ii. Determines the control limits for the chart for centering by using the parameter estimate for spread obtained after completing step i and the overall average obtained from all initial subgroups.
iii. Performs step i for the control chart for centering.

Any of the six D&R procedures may be used on two-stage short-run $(\overline{X}, R)$, $(\overline{X}, v)$, $(\overline{X}, \sqrt{v})$, and $(\overline{X}, s)$ control charts. However, only D&Rs 2, 3, 4, and 6 may be used on two-stage short-run (X, MR) control charts because the MR values are calculated from two consecutive X values, thus no single MR value can be associated with a single X value. Consequently, D&Rs 1 and 5, which delete OOC initial subgroups on either the control chart for centering or spread entirely (i.e., if a subgroup shows OOC on either control chart, it is deleted from both charts), cannot be used on two-stage short-run (X, MR) control charts.

Methodology

The methodology for investigating the effect these six D&R procedures have on the performance of two-stage short-run control charting in its second stage consists of three elements. The main element is the computer program that simulates two-stage short-run variables control charting. The second element, which is included in the operation of the program, is the measurements used to determine which D&R procedure establishes the most reliable second stage control limits. The third element, which is explained using sample runs from the program, is the interpretation of the results from the program.

Measurements

The computer program presented here uses two sets of measurements to provide information that may be used to determine the reliability of second stage control limits. The first set of measurements is: the probability of detection (POD), the average run length (ARL), and the standard deviation of the run length (SDRL). The second set of measurements is: the probability of a false alarm (P(false alarm)), the average probability of a false alarm (APFL), and the standard deviation of the probability of a false alarm (SDPFL).

The POD is the probability that a control chart will signal, within a given number of subgroups following a shift, that a process is out-of-control (OOC). Additionally, if a process is in-control (IC), the POD may be interpreted as the probability of a Type I error (i.e., the probability of a false alarm) within a given number of subgroups starting with the first subgroup drawn from the process.

Using the POD allows for the characterization of the run length (RL) distribution. This is particularly useful in a short-run situation because it is desirable to know, for small numbers of subgroups, the probability of detecting a special cause signal or a false alarm. Using the ARL, which is the average number of subgroups that must be plotted on a control chart before an OOC condition is indicated, in a short-run situation is not appropriate because a short-run may not last long enough to achieve the ARL. Additionally, as will be shown, the ARL can be misleading in choosing the appropriate D&R procedure.

The POD may be expressed mathematically as:

$$\mathrm{POD} = \mathrm{P}(\mathrm{RL} \leq t) \qquad (1)$$

where RL is the run length (in number of subgroups), t is the subgroup number, and P(RL ≤ t) is the probability that the RL is less than or equal to subgroup number t. As calculated by the computer program herein, for an OOC situation in the second stage of the two-stage procedure, the subgroup count starts at one at the first OOC subgroup. For an IC situation, the subgroup count starts at one with the first subgroup drawn from the process in the second stage.

Each time the program simulates two-stage short-run variables control charting an RL value is determined. As the simulation is repeated, RL and RL^2 values are summed, and counts for the number of RLs less than or equal to each integer value in the interval [1, 50000] are kept. Once the repeating of the simulation is complete, the two sums are used to calculate the ARL and the SDRL, which is the standard deviation of the number of subgroups that must be plotted on a control chart before an OOC condition is indicated. The counts are used to determine the POD values.

For an OOC situation in the second stage of the two-stage procedure, it is desirable to have the highest possible POD values and the lowest possible ARL. For an IC situation in the second stage, it is desirable to have the lowest possible POD values and the highest possible ARL.

The probability of a false alarm (P(false alarm)) is the probability of a control chart indicating an OOC condition when none exists. Hillier's (1969) methodology, upon which the two-stage short-run variables control charts are based, allowed for the specification of the desired P(false alarm), that is, the desired Type I error probability.

The computer program presented here calculates the P(false alarm) when an OOC situation occurs beyond the first subgroup drawn from the process in the second stage of the two-stage procedure. Each time the program simulates two-stage short-run variables control charting under these conditions, a value for P(false alarm) is determined. As the simulation is repeated, P(false alarm) and P(false alarm)^2 values are summed. Once the repeating of the simulation is complete, these two sums are used to calculate the APFL and the SDPFL. It is desirable for the P(false alarm) values, and consequently the APFL, to be as low as possible.

The Computer Program

The computer program that simulates two-stage short-run variables control charting is available starting at http://program.20m.com. It is coded in FORTRAN (1999). The program is intended to simulate two-stage short-run variables control charting of a process before initiating it so that a decision can be made regarding which D&R procedure to use when performing two-stage short-run variables control charting during the early run of the process. The D&R procedures provided by the program were described earlier; each segment of the program and its operation is now detailed.

The main program cc (control charting) includes the data entry, file setup, subroutine calls, summations of various values determined by the subroutines, final ARL, SDRL, P(false alarm), APFL, and SDPFL calculations, and the output of information to a file. It is the only segment of the program requiring user interaction.

The following inputs (in order of appearance in the program) are requested from the user in the main program cc:

- The process mean and standard deviation.
- The number of times to replicate the two-stage short-run control charting procedure.
- The control chart combination: $(\overline{\mathrm{X}}, \mathrm{R})$, $(\overline{\mathrm{X}}, \mathrm{v})$, $(\overline{\mathrm{X}}, \sqrt{\mathrm{v}})$, $(\overline{\mathrm{X}}, \mathrm{s})$, or (X, MR).
- The subgroup size (not applicable to (X, MR) control charts).
- The number of subgroups for Stage 1.
- The choice of simulating the process in Stage 1 as IC or OOC. If OOC is chosen, the user is requested to enter the choice of a sustained shift in the mean, the standard deviation, or both. Once a shift type is selected, the program prompts for the shift

size (in the same units as the parameter that has shifted) and the number of the first subgroup after the shift in Stage 1.

- The choice of simulating the process in Stage 2 as IC or OOC. If OOC is chosen, the user is requested to enter the choice of a sustained shift in the mean, the standard deviation, or both. Once the user chooses a shift type, the program prompts for the shift size (in the same units as the parameter that has shifted) and the number of the first subgroup after the shift in Stage 2.
- The choice of using a different starting value for seed for the Marse-Roberts Uniform (0, 1) random variate generator (Marse & Roberts, 1983) coded as subroutine random in module random_mod.
- The D&R procedure (entered as 1, 2, 3, 4, 5, or 6). The program describes the execution of each D&R procedure in detail for the user.
- The name (including the location) of the text file (extension .txt) containing the two-stage short-run control chart factors for the control chart combination entered earlier.
- The name (including the location) of the text file that will store the results from the program.

The second to last bullet point above requires further explanation. Appendix A shows the five input files that were used to generate the results in this study. The first input file contains the first and second stage short-run control chart factors for $(\overline{X}, R)$ charts from Table A4 in Elam & Case (2001) for n=3 and m: 1-5. The second input file contains the first and second stage short-run control chart factors for $(\overline{X}, v)$ charts from Table A.4 in Elam & Case (2003) for n=3 and m: 1-5. The third input file contains the first and second stage short-run control chart factors for $(\overline{X}, \sqrt{v})$ charts, also from Table A.4 in Elam & Case (2003) for n=3 and m: 1-5. The fourth input file contains the first and second stage short-run control chart factors for $(\overline{X}, s)$ charts from Table A.4 in Elam & Case (2005b) for n=3 and m: 1-5. The fifth input file contains the first and second stage short-run control chart factors for (X, MR) charts from Table 3 in Elam & Case (2006) for m: 2-15.

The only difference between the appearance of the input files and their corresponding tables in their respective references is that the first stage short-run control chart factors in the first row of each input file are set to zero. This is required in order for the program to correctly read the second stage short-run control chart factors from these input files when m=1 (in the case of $(\overline{X}, R)$, $(\overline{X}, v)$, $(\overline{X}, \sqrt{v})$, and $(\overline{X}, s)$ control charts) or m=2 (in the case of (X, MR) control charts).

When data entry is complete, the first replication of the two-stage short-run control charting procedure begins as program execution proceeds from main program cc to module Stage_1 and the subroutine for the control chart combination entered by the user. Each of the five subroutines for Stage 1 control charting performs the following tasks:

- Reads first stage short-run control chart factors from the input file.
- Generates first stage subgroups.
- Constructs first stage control limits.
- Determines OOC subgroups.

The tasks in the last two points use Hillier's (1969) approach. When Stage 1 control charting is complete, program execution returns to main program cc.

Once program execution returns to main program cc, it immediately proceeds to module D_and_R and the subroutine for the D&R procedure selected by the user. When the D&R procedure is complete, program execution returns to main program cc. At this point, the program assumes that control of the process has been established.

Once program execution returns to main program cc, required summations are calculated and required variable assignments are made. Program execution then proceeds to module Stage_2 and the subroutine for the control chart combination entered by the user. Each of the five subroutines for Stage 2 control charting performs the following tasks:

- Reads second stage short-run control chart factors from the input file.
- Constructs second stage control limits.
- Generates second stage subgroups.

- Determines the run length (RL) and, if applicable, if a false alarm occurs.

The calculations in the point above are based on the signaling capabilities of combined control charts for centering and spread; i.e., a signal occurs if a subgroup plots OOC on either the control chart for centering or the control chart for spread. The number of the first subgroup that signals is the RL value. The second stage control limits are not updated as subgroups are accumulated. When an RL value is determined, Stage 2 control charting is complete and program execution returns to main program cc.

In main program cc after Stage 2 control charting, required summations are calculated. When this is complete, execution returns to the location in main program cc immediately before the five subroutine calls for Stage 1 control charting to begin the second replication. The entire procedure for two-stage short-run control charting repeats for the number of times entered by the user.

After the last replication, program execution in main program cc proceeds to writing the following information to the output file:

- The process mean and standard deviation.
- The number of replications of the two-stage short-run control charting procedure that was carried out.
- The control chart combination ($(\overline{X}, R)$, $(\overline{X}, v)$, $(\overline{X}, \sqrt{v})$, $(\overline{X}, s)$, or (X, MR)).
- The subgroup size (not applicable to (X, MR) control charts).
- The number of subgroups for Stage 1.
- The D&R procedure.
- The state of the process in Stage 1: IC or OOC. If it is OOC, then the type of sustained shift, the shift size (in the same units as the parameter that has shifted), and the number of the first subgroup after the shift in Stage 1 are given.
- The state of the process in Stage 2: IC or OOC. If it is OOC, then the type of sustained shift, the shift size (in the same units as the parameter that has shifted), and the number of the first subgroup after the shift in Stage 2 are given.
- The ARL and SDRL.
- The APFL and SDPFL (if applicable).
- A table of POD values.

The information in the first eight bullet points was entered by the user. The values in the last three bullet points are calculated by the program.

In addition to these calculated values, the computer program determines counts of the number of occurrences of certain events (when applicable). These events are as follows:

- The number of times out of the total number of replications that D&R 1 iterated more than once.
- The number of times out of the total number of replications that D&R 2 iterated more than once for the control chart for spread as well as for the control chart for centering.
- The number of times out of the total number of replications the program skipped a replication because the number of subgroups dropped to zero (for two-stage short-run $(\overline{X}, R)$, $(\overline{X}, v)$, $(\overline{X}, \sqrt{v})$, $(\overline{X}, s)$, and (X, MR) control charts) or one (for two-stage short-run (X, MR) control charts) after OOC subgroups were deleted in a D&R procedure.
- The number of times out of the total number of replications a D&R procedure was stopped because the number of subgroups dropped to one (for two-stage short-run $(\overline{X}, R)$, $(\overline{X}, v)$, $(\overline{X}, \sqrt{v})$, and $(\overline{X}, s)$ control charts) or two (for two-stage short-run (X, MR) control charts) after OOC subgroups were deleted.

These counts, if applicable, are also written to the output file.

Once the above information, applicable calculations, and applicable counts have been written to the output file, execution of the computer program is complete.

Results

Fourteen pairs of tables (Tables 1a-14b) were constructed from output files generated from sample runs of the computer program. Tables 1a and 1b are shown here. Tables 2a-14b are available starting at http://program.20m.com. For example, Tables 12a and 12b were constructed using Sample Output File #1 (see Appendix B) and Sample Output Files #s 2-6 (available starting at http://program.20m.com). In addition to the notation already introduced in this article, Tables 1a-14b use the following notation:

- MN: a sustained shift in the mean
- SD: a sustained shift in the standard deviation
- MS: a sustained shift in both the mean and the standard deviation
- Replications (skipped): the number of replications carried out and, in parentheses, the number of replications skipped because the number of subgroups dropped to zero (for two-stage short-run $(\overline{X}, R)$, $(\overline{X}, v)$, $(\overline{X}, \sqrt{v})$, $(\overline{X}, s)$, and (X, MR) control charts) or one (for two-stage short-run (X, MR) control charts) after OOC subgroups were deleted in a D&R procedure.
- Stops: the number of times out of the total number of replications carried out that a D&R procedure was stopped because the number of subgroups dropped to one (for two-stage short-run $(\overline{X}, R)$, $(\overline{X}, v)$, $(\overline{X}, \sqrt{v})$, and $(\overline{X}, s)$ control charts) or two (for two-stage short-run (X, MR) control charts) after OOC subgroups were deleted.

The sample runs of the program that generated the information in Tables 1a-14b assumed the following:

- The process mean and standard deviation are always 0.0 and 1.0, respectively.
- The planned number of replications is always 5,000.
- The subgroup size n is always 3 (not applicable to (X, MR) control charts).
- The number of Stage 1 subgroups (denoted by m) is always 5 for two-stage short-run $(\overline{X}, R)$, $(\overline{X}, v)$, $(\overline{X}, \sqrt{v})$, and $(\overline{X}, s)$ control charts and it is always 15 for two-stage short-run (X, MR) control charts. This is why the first four sample input files (see Appendix A have two-stage short-run control chart factors for $(\overline{X}, R)$, $(\overline{X}, v)$, $(\overline{X}, \sqrt{v})$, and $(\overline{X}, s)$ charts for m up to and including m=5 and the fifth sample input file (see Appendix A) has two-stage short-run control chart factors for (X, MR) charts for m up to and including m=15.
- A shift in the mean is always of size 1.5 (same units as the mean).
- A shift in the standard deviation is always of size 1.0 (same units as the standard deviation).
- A shift in Stage 1 always occurs between subgroups 2 and 3.
- A shift in Stage 2 always occurs between subgroups 10 and 11.
- The process is IC immediately before Stage 2 control charting begins.

Sample Runs for an IC Process in Stages 1 and 2

The first 28 sample runs of the program are for the IC process during both Stage 1 and Stage 2 control charting. Two-stage short-run control charting for $(\overline{X}, R)$, $(\overline{X}, v)$, $(\overline{X}, \sqrt{v})$, $(\overline{X}, s)$, and (X, MR) charts was simulated using all six D&R procedures for each control chart combination. The results of these simulations appear in Tables 1a-5b.

Because the process is being simulated as IC in Stage 2, it is desirable for the ARL values in Tables 1a-5a to be as high as possible. Also, it is desirable for the P(RL≤t) values in Tables 1b-5b to be as low as possible (because they correspond to probabilities of false alarms within t or less subgroups after starting Stage 2 control charting), especially for small numbers of subgroups (because a short-run situation is in effect).

Based on both of these criteria, the information in Tables 1a-5b indicates that D&R 4 is, for the most part, the D&R procedure of choice. The only exception is in Table 3a, where D&R 1 is the D&R procedure of choice based on the ARL. This implies that, under the assumptions of this simulation, it is preferable to use subgroups that signal false alarms in the construction of second stage control limits. The cost, in terms of the loss in reliability of second stage control limits, is higher by throwing out subgroups that signal false alarms than it is by including them in the construction of second stage control limits.

Table 1a: ARL, SDRL, Replications, and Stops for Two-Stage Short-Run $(\overline{X}, R)$ Control Charts with Stage 1: IC and Stage 2: IC

D&R Procedure	**ARL**	**SDRL**	**Replications (Skipped)**	**Stops**
1	552.89	701.12	5000 (0)	0
2	550.10	702.51	4999 (1)	1
3	552.87	701.72	5000 (0)	0
4	560.49	702.22	5000 (-----)	-----
5	552.08	700.49	5000 (0)	0
6	552.03	700.61	5000 (0)	0
# of Times D&R 1 Iterated More Than Once: 22 # of Times D&R 2 Iterated More Than Once for the R Control Chart: 8 # of Times D&R 2 Iterated More Than Once for the $\overline{X}$ Control Chart: 70				

Table 1b: P(RL≤t) for Two-Stage Short-Run $(\overline{X}, R)$ Control Charts with Stage 1: IC and Stage 2: IC

t	**Delete and Revise (D&R) Procedure**					
	1	**2**	**3**	**4**	**5**	**6**
1	0.00940	0.01000	0.00900	0.00740	0.00820	0.00860
2	0.01640	0.01760	0.01600	0.01260	0.01520	0.01560
3	0.02540	0.02741	0.02520	0.02040	0.02440	0.02500
4	0.03360	0.03561	0.03300	0.02700	0.03260	0.03300
5	0.03820	0.04061	0.03760	0.03140	0.03700	0.03760
6	0.04400	0.04721	0.04400	0.03580	0.04320	0.04420
8	0.05380	0.05761	0.05460	0.04520	0.05320	0.05480
10	0.06400	0.06721	0.06480	0.05420	0.06380	0.06500
15	0.08880	0.09182	0.08880	0.07820	0.08840	0.08920
20	0.11040	0.11462	0.11100	0.09960	0.11000	0.11180
30	0.14040	0.14423	0.14100	0.12980	0.13960	0.14180
40	0.16480	0.16863	0.16520	0.15360	0.16420	0.16620
50	0.19180	0.19584	0.19160	0.17980	0.19120	0.19320
100	0.27440	0.27806	0.27460	0.26480	0.27440	0.27520
200	0.40740	0.41148	0.40800	0.40060	0.40820	0.40820
300	0.50200	0.50630	0.50340	0.49600	0.50360	0.50380
400	0.57760	0.58192	0.57900	0.57320	0.57900	0.57940
500	0.63500	0.63773	0.63640	0.63120	0.63600	0.63680
750	0.74900	0.75075	0.74840	0.74560	0.74920	0.74860
1000	0.82100	0.82156	0.82060	0.81840	0.82120	0.82080
2000	0.95460	0.95479	0.95460	0.95280	0.95460	0.95480
3000	0.98480	0.98480	0.98480	0.98440	0.98500	0.98500
5000	0.99840	0.99840	0.99840	0.99860	0.99840	0.99840
7500	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000

Comparing results in Tables 1a-5a reveals that two-stage short-run $(\overline{\mathrm{X}}, \mathrm{s})$ control charts have the highest ARL for D&R 4. Comparing results in Tables 1b-5b reveals that two-stage short-run $(\overline{\mathrm{X}}, \sqrt{\mathrm{v}})$ control charts have, for most of the shown values of t, the lowest P(RL≤t) values for D&R 4. These results imply that, under the assumptions of this simulation, different control chart combinations are preferable depending on the measurement used.

The information in Tables 1b-4b also indicates that the P(RL≤t) values when t=1 are reasonably close to the theoretical probability of a false alarm. Assuming independence between the control charts for centering and spread, the theoretical P(false alarm) may be calculated by:

$$P(\mathit{false}\ \text{alarm}) = \alpha_{Cen} + (\alpha_{SpreadUCL} + \alpha_{SpreadLCL}) - \alpha_{Cen} \times (\alpha_{SpreadUCL} + \alpha_{SpreadLCL}) \tag{2}$$

where α_{Cen} is the P(false alarm) on the control chart for centering, $\alpha_{SpreadUCL}$ is the P(false alarm) on the control chart for spread above the upper control limit (UCL), and $\alpha_{\mathrm{SpreadLCL}}$ is the P(false alarm) on the control chart for spread below the lower control limit (LCL). For the sample runs of the program, $\alpha_{\mathrm{Cen}} = 0.0027$, $\alpha_{\mathrm{SpreadUCL}} = 0.005$, and $\alpha_{\mathrm{SpreadLCL}} = 0.001$. This means that P(false alarm), as calculated by equation (2), is equal to 0.0086838.

For example, the P(RL≤t) value from Table 1b for D&R 1 and t=1 is 0.00940. The fact that this value is reasonably close to the theoretical probability of a false alarm is not surprising. As mentioned previously, Hillier's (1969) methodology, upon which the two-stage short-run variables control charts are based, allowed for the specification of the desired probability of a false alarm.

In Table 5b, each of the P(RL≤ t) values for t=1 are much lower than 0.0086838. The closest one is 60.847% smaller than 0.0086838. However, these lower P(RL≤t) values for t=1 come at the price of having the lowest ARL for D&R 4 among Tables 1a-5a. This is an example of the tradeoff mentioned by Del Castillo (1995) between having a low probability of a false alarm and a high probability of detecting a special cause signal inherent with two-stage short-run control charts.

The information in Tables 1a-5a also indicates that D&R 1 and D&R 2 are iterating more than once. These multiple iterations seem to create conditions causing replications to be skipped and the chosen D&R procedure to be stopped. Also, if confidence intervals were constructed using the ARL and SDRL values in Tables 1a-5a, then, depending on the confidence level chosen, the ARL results in Tables 1a-5a may not be statistically significantly different.

Sample Runs for an OOC Process in Stage 1 and an IC Process in Stage 2

The next 18 sample runs of the program were for the process being OOC during Stage 1 control charting and IC during Stage 2 control charting. Two-stage short-run control charting for $(\overline{\mathrm{X}}, \mathrm{R})$ charts was simulated using all six D&R procedures for each OOC condition (MN, SD, MS). The results of these simulations are shown in Tables 6a-8b.

Because the process is being simulated as IC in Stage 2, it is desirable for the ARL values in Tables 6a-8a to be as high as possible. Also, it is desirable for the P(RL≤t) values in Tables 6b-8b to be as low as possible (because they correspond to probabilities of false alarms within t or less subgroups after starting Stage 2 control charting), especially for small numbers of subgroups (since a short-run situation is in effect).

Based on the ARL, Tables 6a-8a indicate that D&R 1 was the procedure of choice, regardless of the OOC condition in Stage 1. However, the SDRL values for D&R 1 are higher than those for the other D&R procedures. The ARL for D&R 1 in Table 7a is higher than the ARL values for D&R 1 in Tables 6a and 8a. The ARL for D&R 1 in Table 6a is the lowest of the three. These results imply that, under the assumptions of this simulation, the type of OOC condition in Stage 1 has an effect on the IC ARL in Stage 2. Additionally, the ARL values for each of the six D&R procedures in Table 1a are higher than the respective ARL values in Tables 6a-8a. This result implies that, under the assumptions of this simulation, an OOC condition in Stage 1 caused a reduction in the IC ARL in Stage 2, regardless of the D&R procedure used.

The choice of the appropriate D&R procedure based on the P(RL≤t) values in Tables 6b-8b varies depending on the OOC condition as well as the subgroup number t. In Table 6b, D&R 4 results in the lowest P(RL≤t) values for shown values of t ≤ 10. For shown values of t > 10, D&R 1 is the D&R procedure of choice. In Table 7b, D&R 4 again results in the lowest P(RL≤t) values, but for shown values of t ≤ 300. For most of the shown values of t ≥ 300, D&R 1 is the D&R procedure of choice. In Table 8b, D&R 1 results in the lowest P(R ≤t) values for each of the shown values of t except t: 30, 40, 50. Because D&R 1 is not the procedure of choice in Tables 6b and 7b for shown values of t ≤ 10 and t ≤ 200, respectively, this is an example of how the ARL can be misleading in choosing the appropriate D&R procedure to use in a short-run situation.

The results from Tables 6b and 7b imply that, under the assumptions of this simulation, it is preferable to use subgroups that signal shifts in either the mean or the standard deviation in the construction of second stage control limits. The cost, in terms of the loss in reliability of second stage control limits, is higher by throwing out subgroups that signal shifts in either the mean or the standard deviation than it is by including them in the construction of second stage control limits.

The P(RL≤t) values for shown values of t ≤ 300 for D&R 4 and for shown values of t ≥ 300 for D&R 1 in Table 7b are lower than the lowest P(RL≤t) values in Tables 6b and 8b. The lowest P(RL≤t) values in Table 6b are higher than those in Tables 7b and 8b. These results imply that, under the assumptions of this simulation, the type of OOC condition in Stage 1 has an effect on the P(RL≤t) values in Stage 2. Additionally, the lowest P(RL≤t) values in Table 1b are higher than those in Table 7b for shown values of t ≤ 200 and in Table 8b for shown values of t ≤ 100. These results imply that, under the assumptions of this simulation, having Stage 1 IC does not necessarily result in Stage 2 control limits with the lowest P(RL≤t) values.

An issue of concern is the P(RL≤t) values when t=1. In Table 6b, each of these values is much larger than 0.0086838, the theoretical probability of a false alarm. The closest one is 396.140% larger than 0.0086838. In Table 7b, each of these values is much smaller than 0.0086838. The closest one is 241.217% smaller than 0.0086838. In Table 8b, some of these values are reasonably close to 0.0086838, although others are not. These results are in contrast to the P(RL≤t) values when t=1 in Table 1b. Clearly, under the assumptions of this simulation, an OOC condition as well as the type of OOC condition in Stage 1 has a significant effect on the P(RL≤t) values when t=1 in Stage 2.

Again, the information in Tables 6a-8a indicates that D&R 1 and D&R 2 are iterating more than once. These multiple iterations seem to create conditions causing replications to be skipped and the chosen D&R procedure to be stopped. Also, if confidence intervals were constructed using the ARL and SDRL values in Tables 6a-8a, then, depending on the confidence level chosen, the ARL results in Tables 6a-8a may not be statistically significantly different.

Sample Runs for an IC Process in Stage 1 and an OOC Process in Stage 2

The next 18 sample runs of the program were for the process being IC during Stage 1 control charting and OOC during Stage 2 control charting. Two-stage short-run control charting for $(\overline{X}, R)$ charts was simulated using all six D&R procedures for each OOC condition (MN, SD, MS). The results of these simulations are shown in Tables 9a-11b.

Because the process is being simulated as OOC in Stage 2, it is desirable for the ARL and the APFL values in Tables 9a-11a to be as low as possible. Also, it is desirable for the P(RL≤t) values in Tables 9b-11b to be as high as possible (because they correspond to probabilities of detecting special causes within t or less subgroups after the shift in Stage 2), especially for small numbers of subgroups (because a short-run situation is in effect).

Based on the ARL, D&R 2 (in Tables 9a and 11a) and D&R 4 (in Table 10a) are the D&R procedures of choice. The ARL for D&R 2 in Table 11a is lower than the ARL values for D&Rs 2 and 4 in Tables 9a and 10a, respectively. The ARL for D&R 2 in Table 9a is the highest of the three (it is 1423.680% larger than the ARL for D&R 2 in Table 11a). These results imply that, under the assumptions of this simulation, the type of OOC condition in Stage 2 has an effect on the OOC ARL in Stage 2. As expected, the ARL values for each of the six D&R procedures in Tables 9a-11a are much lower than the respective ARL values in Table 1a.

Based on the APFL, Tables 9a-11a indicate that D&R 4 is the D&R procedure of choice regardless of the OOC condition in Stage 2. This reaffirms the statement that, in terms of the APFL, it is preferable to use subgroups that signal false alarms in the construction of second stage control limits. Also, the APFL values for D&R 4 are reasonably close to 0.0086838, the theoretical probability of a false alarm. However, the APFL values for the other D&R procedures are slightly inflated.

The choice of the appropriate D&R procedure based on the P(RL≤t) values varies depending on the OOC condition as well as the subgroup number t. In Table 9b, D&R 2 results in the highest P(RL≤t) values for shown values of t ≤ 200 (except t=4). In Table 10b, D&Rs 5 (for shown values of t ≤ 10 (except t=1)), 2 (for shown values of t ≥ 15 and t ≤ 100), and 4 (for shown values of t ≥ 200) result in the highest P(RL≤t) values. In Table 11b, D&Rs 2 (for shown values of t ≤ 200, except t=1) and 4 (for shown values of t ≥ 100) result in the highest P(RL≤t) values. Because the ARL value in Table 10a is not the lowest for D&R 2 or D&R 5, this is another example of how the ARL can be misleading in choosing the appropriate D&R procedure in a short-run situation.

The largest P(RL≤t) values in Table 11b are larger than the largest P(RL≤t) values in Tables 9b and 10b. The largest P(RL≤t) values in Table 9b are lower than those in Tables 10b and 11b. These results imply that, under the assumptions of this simulation, the type of OOC condition in Stage 2 has an effect on the P(RL≤t) values in Stage 2. As expected, the P(RL≤t) values for each of the six D&R procedures in Tables 9b-11b are much higher than the respective P(RL≤t) values in Table 1a.

The information in Tables 9a-11b presents another example of the tradeoff mentioned by Del Castillo (1995) between having a low probability of a false alarm and a high probability of detecting a special cause signal inherent with two-stage short-run control charts. Although D&R 4 results in the lowest APFL values regardless of the OOC condition in Stage 2, it also results in the lowest P(RL≤t) values for many of the shown values of t in Tables 9b and 10b.

Again, the information in Tables 9a-11a indicates that D&R 1 and D&R 2 are iterating more than once. These multiple iterations seem to create conditions causing replications to be skipped and the chosen D&R procedure to be stopped. Also, if confidence intervals were constructed using the ARL and SDRL values in Tables 9a-11a, then, depending on the confidence level chosen, the ARL results in Tables 9a-11a may not be statistically significantly different.

Sample Runs for an OOC Process in Stages 1 and 2

The final 18 sample runs of the program were for the process being OOC during both Stage 1 and Stage 2 control charting. Two-stage short-run control charting for $(\overline{X}, R)$ charts was simulated using all six D&R procedures for each OOC condition (MN, SD, MS) in Stage 1 and one OOC condition (MN) in Stage 2. The results of these simulations are shown in Tables 12a-14b.

Because the process was simulated as OOC in Stage 2, it is desirable for the ARL and the APFL values in Tables 12a-14a to be as low as possible. Also, it is desirable for the P(RL≤t) values in Tables 12b-14b to be as high as possible (because they correspond to probabilities of detecting special causes within t or less subgroups after the shift in Stage 2), especially for small numbers of subgroups (because a short-run situation is in effect).

Based on the ARL, D&R 2 (in Tables 12a and 14a) and D&R 3 (in Table 13a) are the

D&R procedures of choice. The ARL for D&R 3 in Table 13a is lower than the ARL values for D&R 2 in Tables 12a and 14a. The ARL for D&R 2 in Table 14a is the highest of the three. These results imply that, under the assumptions of this simulation, the type of OOC condition in Stage 1 has an effect on the OOC (MN) ARL in Stage 2. Additionally, the ARL values for each of the six D&R procedures in Table 9a are much lower than the respective ARL values in Tables 12a-14a. This implies that, under the assumptions of this simulation, an OOC condition in Stage 1 causes an increase in the OOC (MN) ARL in Stage 2, regardless of the D&R procedure used.

Based on the APFL, Tables 12a-14a indicate that D&R 4 is the procedure of choice regardless of the OOC condition in Stage 1. This implies that, under the assumptions of this simulation, it is preferable to use subgroups that signal shifts in the mean, the standard deviation, or both in the construction of second stage control limits. The cost, in terms of the loss in reliability of second stage control limits, is higher by throwing out subgroups that signal shifts in the mean, the standard deviation, or both than it is by including them in the construction of second stage control limits. Additionally, comparing the APFL results in Table 9a with those in Tables 12a-14a reveals that, under the assumptions of this simulation, an MN in Stage 1 has the effect of increasing the APFL (see Table 12a) and an SD in Stage 1 has the effect of decreasing the APFL (see Table 13a).

An issue of concern is the differences in the APFL values from 0.0086838, the theoretical probability of a false alarm. The APFL value for D&R 4 in Table 12a is 369.424% larger than 0.0086838. The APFL values for D&R 4 in Tables 13a and 14a are 65.683% and 33.209%, respectively, smaller than 0.0086838. These results are somewhat consistent with those regarding the P(RL≤t) values when t=1 in Tables 6b-8b. Clearly, under the assumptions of this simulation, the type of OOC condition in Stage 1 has a significant effect on the APFL values before the shift in Stage 2.

Based on the P(RL≤t) values, D&R 2 is the appropriate procedure for most of the shown values of t regardless of the OOC condition in Stage 1. Because Table 13a indicates that D&R 3 is the D&R procedure of choice, this is another example of how the ARL can be misleading in choosing the appropriate D&R procedure in a short-run situation. The fact that the largest P(RL≤t) values in Table 14b are lower than those in Tables 12b and 13b for most of the shown values of t implies that, under the assumptions of this simulation, the type of OOC condition in Stage 1 has an effect on the P(RL≤t) values in Stage 2.

Additionally, the largest P(RL≤t) values in Table 9b are larger than those in Tables 12b-14b. This result implies that, under the assumptions of this simulation, an OOC condition in Stage 1 decreases the P(RL≤t) values in Stage 2; this is not desirable due to the MN in Stage 2. However, it is desirable for Stage 2 IC as was the case in comparing results in Table 1b to those in Tables 6b-8b. Clearly, under the assumptions of this simulation, when one is interested in detecting MN in Stage 2, it is highly desirable to have the process IC when drawing first stage subgroups.

The information in Tables 12a-14b presents another example of the tradeoff mentioned by Del Castillo (1995) between having a low probability of a false alarm and a high probability of detecting a special cause signal inherent with two-stage short-run control charts. Although D&R 4 results in the lowest APFL values regardless of the OOC condition in Stage 1, it also results in the lowest P(RL≤t) values for many of the shown values of t in Tables 13b and 14b.

Again, as in the three previous sub-sections, the information in Tables 12a-14a indicates that D&R 1 and D&R 2 are iterating more than once. These multiple iterations seem to create conditions causing replications to be skipped and the chosen D&R procedure to be stopped. Also, if confidence intervals were constructed using the ARL and SDRL values in Tables 12a-14a, then, depending on the confidence level chosen, the ARL results in Tables 12a-14a may not be statistically significantly different.

Conclusion

The interpretation of the sample runs of the computer program establish the fact that no hard and fast rules can be developed regarding which D&R procedure is appropriate when performing two-stage short-run variables control charting. Under the assumptions of the simulations performed, the choice of the appropriate D&R procedure varies both among and within measurements, among control chart combinations, among IC and various OOC conditions in both stages, and among numbers of subgroups plotted in Stage 2. It may be possible that the choice of the appropriate D&R procedure varies among shift sizes and the timing of shifts, though this was not investigated.

If decisions cannot be made regarding values for these variables, then extensive sample runs similar to the ones in the previous section need to be performed. However, if certain values for these variables are desired, then the process of making sample runs and interpreting their results is much simpler.

References

Case, K. E. (1998). The delete and revise (D&R) procedure. In *Class Notes / Lecture Material*. Stillwater, Oklahoma: Oklahoma State University.

Del Castillo, E. (1995). Discussion. *Journal of Quality Technology, 27(4)*, 316-321.

Doty, L. A. (1997). *SPC for short run manufacturing*. Cincinnati, OH: Hanser Gardner Publications.

Elam, M. E. & Case, K. E. (2001). A computer program to calculate two-stage short-run control chart factors for $(\overline{X}, R)$ charts. *Quality Engineering*, *14(1)*, 77-102.

Elam, M. E. & Case, K. E. (2005b). A computer program to calculate two-stage short-run control chart factors for $(\overline{X}, s)$ charts. *Quality Engineering*, *17(2)*, 259-277.

Elam, M. E. & Case, K. E. (2006). A computer program to calculate two-stage short-run control chart factors for (X, MR) charts. *Journal of Statistical Software, 15(11)*, http://www.jstatsoft.org/.

Elam, M. E. & Case, K. E. (2003). A computer program to calculate two-stage short-run control chart factors for $(\overline{X}, v)$ and $(\overline{X}, \sqrt{v})$ charts. *Quality Engineering*, *15(4)*, 609-638.

Elam, M. E. & Case, K. E. (2005a). Two-stage short-run $(\overline{X}, s)$ control charts. *Quality Engineering*, *17(1)*, 95-107.

Elam, M. E. & Case, K. E. (2008). Two-stage short-run (X, MR) control charts. *Journal of Modern Applied Statistical Methods*, *7(1)*, 275-285.

FORTRAN (1999). *Essential Lahey FORTRAN 90 compiler release 4.00c*. Incline Village, NV: Lahey Computer Systems, Inc.

Hillier, F. S. (1969). $\overline{X}$- and R-chart control limits based on a small number of subgroups. *Journal of Quality Technology*, *1(1)*, 17-26.

Marse, K. & Roberts, S. D. (1983). Implementing a portable FORTRAN uniform (0, 1) generator. *Simulation*, *41(4)*, 135-139.

Montgomery, D. C. (1997). *Introduction to statistical quality control*, 3rd Ed. New York, NY: John Wiley and Sons, Inc.

Pyzdek, T. (1993). Process control for short and small runs. *Quality Progress*, *26(4)*, 51-60.

Ryan, T. P. (1989). *Statistical methods for quality improvement*. New York, NY: John Wiley and Sons, Inc.

Yang, C.-H. & Hillier, F. S. (1970). Mean and variance control chart limits based on a small number of subgroups. *Journal of Quality Technology*, *2(1)*, 9-16.

Appendix A

Sample Input File Containing First and Second Stage Short Run Control Chart Factors for $(\overline{X}, R)$ Charts for n=3 and m: 1-5

```
0.00000     0.00000     0.00000     8.35221     14.34466     0.03152
1.56033     1.86966     0.06112     2.70257     5.65885      0.03337
1.35226     2.21659     0.04924     1.91239     4.27295      0.03407
1.25601     2.35005     0.04491     1.62151     3.74247      0.03443
1.20246     2.41685     0.04267     1.47271     3.46631      0.03465
```

Sample Input File Containing First and Second Stage Short Run Control Chart Factors for $(\overline{X}, v)$ Charts for n=3 and m: 1-5

```
0.00000    0.00000    0.00000       17.69484    199.00000    0.00100100
2.87519    1.99000    0.00200000    4.97997     26.28427     0.00100075
2.40967    2.78787    0.00150038    3.40779     14.54411     0.00100067
2.20599    3.31601    0.00133378    2.84792     11.04241     0.00100063
2.09497    3.67043    0.00125047    2.56580     9.42700      0.00100060
```

Sample Input File Containing First and Second Stage Short Run Control Chart Factors for $(\overline{X}, \sqrt{v})$ Charts for n=3 and m: 1-5

```
0.00000     0.00000     0.00000     17.69484     15.91775     0.03570
2.87519     1.59177     0.05046     4.97997      5.45415      0.03365
2.40967     1.77629     0.04121     3.40779      3.97519      0.03297
2.20599     1.89811     0.03807     2.84792      3.42822      0.03263
2.09497     1.97649     0.03648     2.56580      3.14794      0.03243
```

Sample Input File Containing First and Second Stage Short Run Control Chart Factors for $(\overline{X}, s)$ Charts for n=3 and m: 1-5

```
0.00000     0.00000     0.00000     15.68165     14.10674     0.03164
2.95828     1.86761     0.06134     5.12390      5.60680      0.03348
2.57119     2.21123     0.04940     3.63621      4.24135      0.03417
2.39128     2.34285     0.04505     3.08713      3.71725      0.03453
2.29099     2.40840     0.04280     2.80588      3.44396      0.03476
```

Sample Input File Containing First and Second Stage Short Run Control Chart Factors for (X, MR) Charts for m: 2-15

```
0.00000     0.00000     0.00000     204.19466     127.32134     0.00157
22.24670    2.95360     0.00235     31.46159      26.11886      0.00157
10.72641    3.58790     0.00209     13.84773      13.20218      0.00157
7.34996     3.83736     0.00196     9.00182       9.27880       0.00157
5.87022     3.89898     0.00188     6.94574       7.52080       0.00157
5.06862     3.89368     0.00183     5.85274       6.55349       0.00157
4.57470     3.86822     0.00179     5.18723       5.95038       0.00157
4.24308     3.83885     0.00177     4.74391       5.54166       0.00157
4.00644     3.81088     0.00175     4.42928       5.24776       0.00157
3.82972     3.78583     0.00173     4.19525       5.02691       0.00157
3.69307     3.76385     0.00171     4.01479       4.85521       0.00157
3.58441     3.74470     0.00170     3.87161       4.71806       0.00157
3.49606     3.72800     0.00169     3.75537       4.60610       0.00157
3.42287     3.71338     0.00168     3.65920       4.51303       0.00157
```

Appendix B: Sample Output File #1

```
----------------------------------------
mean: ....................   0.00000
standard deviation: ......   1.00000
# of replications of
  two stage procedure: ... 4996
Control chart combination: (Xbar, R)
n: .......................    3
m (Stage 1): .............    5
D&R procedure: ...........    1
----------------------------------------
Stage 1: shift size of     1.50000 (same
         units as the mean) in the mean
         between subgroups   2 and   3.

Stage 2: shift size of     1.50000 (same
         units as the mean) in the mean
         between subgroups  10 and  11.
---------------------------------------------

---------------------------------------------------
Out-of-Control (OOC) Average Run Length (ARL) and
Standard Deviation of the Run Length (SDRL) results
---------------------------------------------------
ARL (in number of subgroups):     464.85809
SDRL (in number of subgroups):    693.88171
---------------------------------------------------

------------------------------------------------
The Average Probability of a False Alarm (APFL)
and the Standard Deviation of the Probability of
a False Alarm (SDPFL) in the first  10 subgroups
before the shift in Stage 2:
------------------------------------------------
APFL:  0.03813
SDPFL: 0.11174
------------------------------------------------
    Starting at subgroup  11 in Stage 2:
-------------------------------------------
  t       Number of RLs <= t     P(RL <= t)
-----     ------------------     ----------
    1             90               0.01801
    2            162               0.03243
    3            236               0.04724
    4            290               0.05805
    5            340               0.06805
    6            384               0.07686
    7            422               0.08447
    8            463               0.09267
    9            508               0.10168
   10            548               0.10969
```

Appendix B: Sample Output File #1 *(continued)*

```
   15             674              0.13491
   20             793              0.15873
   30            1002              0.20056
   40            1162              0.23259
   50            1277              0.25560
   75            1550              0.31025
  100            1781              0.35649
  200            2432              0.48679
  300            2893              0.57906
  400            3259              0.65232
  500            3504              0.70136
  750            3997              0.80004
 1000            4296              0.85989
 2000            4814              0.96357
 3000            4934              0.98759
 4000            4973              0.99540
 5000            4984              0.99760
 7500            4994              0.99960
10000            4995              0.99980
20000            4996              1.00000
30000            4996              1.00000
40000            4996              1.00000
50000            4996              1.00000
-------------------------------------------

The first D&R procedure iterated more than
  once a total of 111 time(s).

Replications skipped   4 time(s)
  because the number of subgroups dropped
  to zero after out-of-control (OOC)
  subgroups were deleted.

D&R procedure 1 stopped  12 time(s)
  because the number of subgroups dropped
  to one after out-of-control (OOC)
  subgroups were deleted.
```

Journal of Modern Applied Statistical Methods
November, 2008, Vol. 7, No. 2, 561-576

1538 – 9472/08/$95.00

Data Mining CEO Compensation

Susan M. Adams   Atul Gupta   Dominique M. Haughton   John D.Leeth
Bentley University

The need to pre-specify expected interactions between variables is an issue in multiple regression. Theoretical and practical considerations make it impossible to pre-specify all possible interactions. The functional form of the dependent variable on the predictors is unknown in many cases. Two ways are described in which the data mining technique Multivariate Adaptive Regression Splines (MARS) can be utilized: first, to obtain possible improvements in model specification, and second, to test for the robustness of findings from a regression analysis. An empirical illustration is provided to show how MARS can be used for both purposes.

Key words: data mining, interactions, modeling, multivariate adaptive regression splines (MARS), multiple regression

Introduction

The use of multiple regression analysis is widespread in empirical research. To use multiple regression analysis the full set of independent variables affecting the dependent variable must first be identified and all of the expected interactions among these explanatory variables specified. Since both theoretical and practical considerations make it impossible to pre-specify all possible interactions, the explanatory power of any given regression specification will be limited. In addition, while theory may provide guidance as to which predictors to use in a model, the functional form of the dependent variable on the predictors is unknown in many cases. This article describes two ways in which the data mining technique Multivariate Adaptive Regression Splines (MARS) can be utilized: first, to obtain possible improvements in model specification, and second, to test for the robustness of findings from a regression analysis. An empirical illustration of how MARS can be used for both purposes is then provided.

The intuition underlying MARS is straightforward; the algorithm examines the data for all possible interactions among the specified explanatory variables and for non-linear relations between the dependent and explanatory variables and, in general, yields substantial improvements in explanatory power. Findings from the MARS analysis can be used in two possible ways. First, MARS may yield insight into possible empirical relationships that exist in data, but which have not been identified by the researcher. Such relationships can be examined for theoretical content and used to improve the specification of the regression model.

A second useful application of MARS is in the context of testing for the robustness of findings from a particular regression. For example, consider a research study interested in examining the relationship between employee gender and compensation. Because compensation is expected to depend on a variety of characteristics, the typical regression model includes a set of explanatory variables and a dummy variable to capture the gender effect.

Susan M. Adams is a Professor in the Department of Management. Email: sadams@bentley.edu. Atul Gupta is a Professor in the Department of Finance. Email: agupta@bentley.edu. Dominique M. Haughton is a Professor in the Department of Mathematical Sciences. Email: dhaughton@bentley.edu. John D. Leeth is a Professor in the Department of Economics. Email: jleeth@bentley.edu.

The sign and statistical significance of the dummy variable and the explanatory power of the entire model depend on three factors: the choice of explanatory variables, the set of interactions included in the model, and the specified functional form of the dependent variable in terms of the predictors. While MARS can add no insight into the choice of explanatory variables, it can test for all possible interactions among the explanatory variables, the preponderance of which have not been included in a normal regression analysis.

Moreover, MARS uses splines (understood here to be piecewise-linear functions) to allow for possible non-linearities in the data. Given that MARS will generally yield a substantial improvement in explanatory power, a finding that the sign and statistical significance of a variable of interest (the dummy variable for employee gender in our example) remains unchanged serves as a useful test for the robustness of the findings from the original regression. Normally, researchers using regression analysis provide the results from several model specifications to demonstrate the empirical strength of their conclusions. MARS provides a more structured approach to this model specification procedure and, thereby, generates a more powerful test of robustness.

Methodology

The Data

Standard and Poor's (S&P) ExecuComp database was used to examine the compensation of male and female CEOs. This database tracks a variety of corporate data for the 1500 largest companies in the U.S. from 1992 to 2003 and personal and compensation data for their associated CEOs. From 1992 through 2003, 56 women served as CEOs of the top 1500 Standard & Poor's companies in the United States; in contrast, 4,242 men served as corporate CEOs over the same time period. The ExecuComp database yielded 214 individual executive/year observations for female CEOs and 18,179 observations for male CEOs. The CEOs are scattered across 369 4-digit SIC industries. To control for possible industry effects in salary determination, analysis focused on CEOs employed in the forty-one 4-digit SIC industries with at least one female CEO.

Table 1 gives a summary of the variables used in the analysis. The left-hand side of the table provides information on the OLS sample and the right-hand side provides information on the MARS sample. To be included in an OLS regression an observation must have a complete set of information on all explanatory variables. The MARS sample is larger because the MARS procedure explicitly controls for missing values, allowing all observations with information on total compensation to be included in the analysis, an important advantage of MARS over OLS.

The dependent variable used was the logarithm of the CEO's total compensation for the year, which includes salary, bonus, restricted stock, stock options (evaluated using the Black-Scholes procedure), long-term incentive payouts, and other types of compensation. The independent variables are fairly standard. Most studies of wages and salaries include information on human capital such as education, general labor market experience, and experience within a specific company (Topel, 1991; Willis, 1986). The ExecuComp data does not provide information on education and measures of experience are somewhat spotty. To capture human capital characteristics included in the analysis are age and the number of years the person has served as CEO. (For some CEOs, the data lists the date the person started working for the company. Unfortunately, the information was available for only 59.2 percent of the sample and so was not used in the analysis.)

Because economic theory indicates that investments in human capital should have positive but diminishing returns, also included were squares of age and years as CEO. While early studies of the pay-performance relationship found little evidence of such a link (see Jensen & Murphy, 1990), some recent work documents that CEO compensation is related to company size and company performance (see Bebchuk & Grinstein, 2005). Company size is measured using the dollar value of sales revenue and company performance using the return on assets.

Table 1: Means (Standard Deviations)

Variable	OLS Sample Men	OLS Sample Women	Difference in Means (absolute t/z statistic)	MARS Sample Men	MARS Sample Women	Difference in Means (absolute t/z statistic)
Total Compensation (thousands of 2003 $)	5,036 (17,332)	4,926 (9,402)	110 (0.15)	4,797 (16,589)	4,768 (9,257)	28 (0.04)
Log Total Compensation	7.69 (1.176)	7.73 (1.151)	-0.04 (0.43)	7.64 (1.170)	7.68 (1.157)	-0.04 (0.45)
Age	53.96 (7.396)	51.14 (7.396)	2.82 (5.30)**	54.13 (7.937)	51.03 (7.327)	3.09 (5.96)**
Years as CEO	8.11 (7.957)	7.97 (11.991)	0.14 (0.16)	8.06 (7.894)	7.93 (11.974)	0.13 (0.15)
Sales (billions of 2003 $)	2.93 (7.228)	2.70 (8.576)	0.23 (0.37)	2.85 (7.033)	2.63 (8.440)	0.22 (0.37)
Return on assets (Percent)	0.10 (29.129)	1.68 (15.773)	-1.58 (1.31)	0.44 (28.033)	1.82 (15.564)	-1.38 (1.19)
Manufacturing	0.406	0.405	0.001 (0.03)	0.413	0.423	-0.010 (0.77)
Transportation	0.149	0.049	0.100 (3.98)**	0.152	0.047	0.105 (4.23)**
Trade	0.072	0.195	-0.123 (6.35)**	0.070	0.188	-0.117 (6.29)**
Finance	0.049	0.078	-0.029 (1.88)	0.051	0.075	-0.024 (1.56)
Services	0.324	0.273	0.051 (1.53)	0.314	0.268	0.047 (1.43)
Number	3,689	205		4,058	213	

Finally, to control for differences in pay across industries and over time the OLS analysis includes binary variables measuring the company’s 1-digit SIC code and a linear time trend. The MARS analysis permits a more detailed investigation of industry and time effects. The MARS procedure includes a categorical variable representing 41 different 4-digit SIC industries and a categorical variable representing 12 different years. All dollar figures for total compensation and sales revenue have been adjusted to correct for the impact of inflation and are stated in 2003 dollars.

Table 1 uncovers only a few statistically significant differences in means or proportions between male and female CEOs. Within the four-digit SIC industries examined, female CEOs are a few years younger than their male counterparts and the companies they operate are more likely to be involved in trade and less likely to be involved in transportation. In terms of compensation, the data provide no evidence that male and female CEOs are paid differently.

The MARS methodology

The MARS algorithm, proposed by Friedman in 1991, relies on the following basic ideas:

For each continuous independent variable, MARS creates a piecewise linear function with too many change points (knots) to begin with, and then prunes unnecessary knots by a backward procedure. Consider the functions BF3 and BF4 (Basis Functions 3 and 4) identified by MARS (definitions of all Basis Functions are given in Appendix A). These two functions are preceded by BF1, as follows:

```
BF1 = (SALES > .);
BF3 = max(0, SALES – 1.747087) * BF1;
BF4 = max(0, 1.747087 - SALES ) * BF1;
```

BF1 is zero whenever the variable SALES is missing, and one otherwise. The functions BF3 and BF 4, taken together, define a piecewise linear function of SALES, with a break point (otherwise referred to as a knot or a change point) at about 1.75 billion dollars. Note that BF3 is zero when SALES is less than 1.747, and BF4 is zero when SALES is greater than 1.747. Basis functions are chosen by MARS to achieve the best fit in a regression of the dependent variable on the Basis Functions. Of course, without any restriction on over-fitting, better and better fits will be attained by using more and more Basis Functions breaking at more and more knots. MARS uses a backward stepwise method to eliminate Basis Functions and knots which contribute least to the fit of the model.

For each independent categorical variable, MARS groups categories and creates dummy variables which correspond to these groups in such a way as to yield the best fit possible. For instance, the Basis Function BF5, given by the expression is:

```
BF5 = (SICNEW = 1 OR SICNEW = 2 OR
SICNEW = 5 OR SICNEW = 13 OR SICNEW
= 15 OR SICNEW = 16 OR SICNEW = 21 OR
SICNEW = 22 OR SICNEW = 23 OR SICNEW
= 25 OR SICNEW = 26 OR SICNEW = 27 OR
SICNEW = 28 OR SICNEW = 29 OR SICNEW
= 31 OR SICNEW = 32) * BF1;
```

BF5 equals one if the SICNEW code for an observation is one of those listed in the expression (1, 2, 5, 13, …, etc.), zero otherwise. This means that, of all the ways MARS considered to create a dummy variable that would represent a group of industries, the grouping in BF5 is one of the groupings it found would yield the best fit with the dependent variable. Other industry groupings are identified and expressed in other Basis Functions.

MARS looks for interactions among independent variables, by introducing into the model the product of two variables, if such an interaction leads to a sufficient improvement in the model. For example, the Basis Functions BF23 and BF24 represent an interaction of age with the number of years as CEO since BF21 includes BF18 in its expression, which in turns includes age. An interesting aspect is that MARS can (and often will) create interactions, not between original variables, but between restrictions of these variables to a particular range as is done in BF23 and BF24. BF23 (respectively BF24) interacts age with number of years as CEO, but only beyond 12 years as CEO (respectively up to 12 years as CEO), and in any case only up to ages of 43 years. BF23 and BF24 have a different coefficient in the final model, so the strength of the interaction depends on the range of years as CEO involved in the interaction: it is stronger (.030) for BF24 than for BF23 (.019).

To summarize, MARS ends up with a collection of Basis Functions, which are transformations of independent variables taking into account non-linearities and interactions. MARS then estimates a least-squares model with a parsimonious set of Basis Functions as independent variables. Parsimony is achieved by removing Basis Functions, knots and interactions which do not contribute sufficiently to the model fit.

MARS, in essence, is an OLS procedure, but with judicious transformations of the independent variables. Risks of overfitting are controlled in various ways by the algorithm (Friedman, 1991, Section 3.6). To take into account the fact the data are used not only to estimate the coefficients of the Basis Functions but to create these Basis Functions in the first place, a penalized sum of squared residuals is minimized to select the final model (in least squares regression, a non-penalized sum of squares would be used). This is achieved by minimizing a quantity referred to as the Generalized Cross Validation (GCV) criterion equal to $(1/N)\ SSR/[1-C(M)/N]^2$ (see Friedman, 1991, p. 20), where N is the number of observations, *SSR* is the residual sum of squares, and $C(M)$ is a measure of the complexity of a model with M Basis Functions. The complexity $C(M)$, which would equal M in usual least squares modeling, is defined to be equal to $M + dM$, where d is a penalty for each additional Basis Function.

The parameter D can be determined in a number of ways: a value of 3 has been recommended on the basis of simulations in Friedman (1991), but a larger value may be appropriate for larger sample sizes. An alternative, used in this article, is to determine the parameter d via ten-fold cross validation (not to be confused with the GCV mentioned above, the GCV does not actually involve cross-validation). Ten-fold cross- validation involves randomly dividing the data into ten parts, building the model – with various values of the parameter d – with nine tenths of the data, and evaluating the performance of the model on the remaining tenth. This is done ten times, for each tenth in turns, and the performance averaged out over the ten runs. The value of d yielding the best performance is selected, and the GCV criterion is computed with this value of d. A clearly over-fitting model is first built, and Basis Functions are removed one after the other, yielding a sequence of models with a decreasing number of Basis Functions. A model is selected from that sequence which minimizes the GCV criterion.

A convenient place to get information with introductions to the MARS methodology, white papers, and useful references is the Salford Systems Web site (www.salford-systems.com). The article by De Veaux, et al. (1993) includes a good introduction to MARS, albeit in the context of chemical engineering, and contrasts the MARS methodology with that of neural networks. The article by Sephton (2001) gives an introduction to MARS and evaluates how well MARS performs at forecasting recessions; the author finds that for the time series considered for predicting recessions, MARS yields a better in-sample, but a worse out-of-sample performance than for instance probit regression (with a dependent variable of 1 if a time period was in recession, 0 if not); this may indicate that the MARS models used in this context were over-fitting the data to some extent. This is the reason why it is recommended in the literature (Deichman, et. al. 2002; Munoz & Felicisimo 2004) to evaluate MARS on validation samples, independent of the sample used to build the data, in order to select a MARS model that will not over-fit the data and will predict well on validation samples.

This approach is adopted in Deichman, et al. (2002) where MARS is used in the context of direct response modeling; the authors find that response models which use MARS Basis Functions perform better than alternatives on independent validation samples. Munoz & Felicisimo contrast a MARS methodology with several alternatives and reach two interesting conclusions: one is that MARS yields the best predictive power, and the other is that an independent validation sample is truly needed (cross-validation is not sufficient).

The issue of over-fitting is considered later in this article and will explain why in our case over-fitting does not risk calling results into question. Finally, an article where MARS is used in analyses of living standards in Vietnam (see for example Deichman, et. al. (2001)), where interesting interactions are revealed between regions of the country and other predictors when modeling the logarithm of household expenditure per capita, indicating that such models of household wealth are likely to differ across regions, with the importance of some predictors varying across these regions.

Results

Table 2 presents the OLS results. As is typical, several specifications to check for robustness are included. The first specification includes only human capital characteristics, while the second augments these characteristics with information on the company. The third specification controls for differences in pay by industry and over time and the fourth specification interacts each independent variable with the binary variable indicating the gender of the CEO. The last specification is a test to determine if any significant differences exist in how male and female CEOs are paid across the variables considered. In standard parlance, it is a test to determine if it is permissible to pool male and female CEOs in the same sample.

The results in Table 2 appear remarkably robust. In none of the first three specifications is the female binary variable statistically significant, indicating no difference in pay between male and female CEOs. Although in the fourth specification the F-statistic indicates male and female CEOs are paid differently, the only statistically significant difference in CEO pay is in the transportation industry, but the positive interaction term points to female CEOs earning more than their male counterparts. In short, in terms of pay the data provide no evidence of discrimination against women once they have made it to the highest rung of the corporate ladder. Almost all other studies of gender differences in compensation find women earning far less than men, controlling for other factors including occupation and title (Bertrand & Hallock, 2001).

The other variables in Table 2 are also robust across the four empirical specifications in terms of statistical significance and absolute size. In all four specifications general experience as measured by age raises log total compensation but at a decreasing rate (the coefficient on age is significantly positive and the coefficient on age squared is significantly negative). Company size as measured by sales and company performance as measured by return on assets significantly boost CEO compensation. The positive coefficient on time demonstrates a substantial yearly increase in real CEO compensation and the negative coefficients on transportation and trade shows CEOs in these industries earn less, all else equal, than CEOs in manufacturing (the excluded category). The other variables are insignificant across all four specifications.

Appendix A presents the full set of MARS results. The MARS model explains about 46 percent of the variability in (logged) total compensation, compared to about 17 percent for the OLS model. This improvement is due (in part) to the fact that MARS identifies groups of industries for which the compensation model differs, a matter very much at the heart of compensation modeling, and successfully includes interactions of these industry groupings and other independent variables.

Most important to our analysis, gender does not enter the model at all once the above mentioned interactions are included. Even following a very structured approach for determining model specification, an approach which investigates hundreds of possible interactions among the independent variables and allows for complex non-linear relationships to exist between the dependent and independent variables, the data still uncovers no difference in how male and female CEOs are compensated. A maximum of 80 basis functions were allowed to be used in this MARS model, and ten-fold cross-validation were used to evaluate models considered by MARS. The maximum number of basis functions allowed (80) is sufficient for MARS to build a large enough model from which to prune to get a satisfactory final model (such a maximum should be at least as large as about twice the number of basis functions in the final model; in this case the final model contains 33 basis functions, so an initial maximum of 80 basis functions is ample). To determine how much to prune (in other words how many basis functions to drop) to yield a final model, MARS uses as a measure of performance a modified R-square measure referred to as the Generalized Cross Validation (GCV) criterion; the GCV incorporates a cost per basis function into its formula; the higher the cost, the smaller the number of basis functions in the final model. The choice of that cost is quite crucial, and is performed here by ten-fold cross validation, which consists in splitting the data into ten parts, using 9/10 of the data to build the model and the remaining tenth to evaluate candidate models corresponding to different choices of cost in order to select the cost that yields the best performance on the held out tenth of the data. Typically, and here as well, each tenth of the data plays the role of a hold-out sample in turns and performance is judged on all ten such samples. The absence of a gender effect in CEO

Table 2: OLS Results on Log Total Compensation ($2003)

	(1)	(2)	(3)	(4)
Constant	3.118 (2.31)*	4.442 (3.33)**	3.806 (2.91)**	3.773 (2.79)**
Female	0.041 (0.24)	0.035 (0.22)	-0.060 (0.35)	2.193 (0.36)
Age	0.166 (3.29)**	0.117 (2.38)*	0.133 (2.71)**	0.134 (2.65)**
Age squared	-0.001 (3.10)**	-0.001 (2.36)*	-0.001 (2.63)**	-0.001 (2.55)*
Years CEO	-0.011 (0.96)	-0.004 (0.33)	-0.011 (1.03)	-0.015 (1.31)
Years CEO squared	0.000 (0.77)	0.000 (0.45)	0.000 (0.89)	-0.000 (1.05)
Sales (billions 2003$)		0.049 (5.94)**	0.049 (5.76)**	0.049 (5.35)**
Return on assets		0.003 (2.80)**	0.004 (3.30)**	0.004 (3.30)**
Time			0.048 (6.13)**	0.050 (6.24)**
Transportation			-0.669 (7.92)**	-0.694 (8.12)**
Trade			-0.289 (2.24)*	-0.330 (2.37)*
Finance			-0.030 (0.19)	0.007 (0.04)
Service			-0.070 (0.84)	-0.083 (0.98)
Age×Female				-0.86 (0.36)
Age squared×Female				0.001 (0.30)
Years CEO×Female				0.104 (1.95)
Years CEO squared×Female				-0.002 (1.90)
Sales×Female				-0.002 (0.13)
Return on assets×Female				-0.005 (0.96)
Time×Female				-0.034 (0.80)
Transportation×Female				0.967 (2.72)**
Trade×Female				0.504 (1.31)
Finance×Female				-0.586 (1.23)
Service×Female				0.264 (0.69)

Table 2: OLS Results on Log Total Compensation ($2003) (continued)

	(1)	(2)	(3)	(4)
R-squared	0.02	0.12	0.17	0.18
F-statistic: all coefficients = 0	3.04**	8.59**	14.44**	10.50**
F-statistic: female and female interaction terms = 0				2.93**

* significant at 5%; ** significant at 1% *Note*: The t statistics are calculated using standard errors that correct for heteroskedasticity and the correlation among observations for the same individual. Industry results are measured relative to the excluded category, manufacturing.

compensation is robust across empirical specifications.

Tables 3 and 4 summarize the MARS results. To simplify matters, the two tables present results only for observations in the data set where none of the independent variables are missing. When one or more independent variables is missing, the model adjusts for that in the equations (see for example BF1 in Appendix A, which captures the fact that the variable SALES is not missing), but the adjustments involve a fairly small number of observations (see Table 1).

An examination of the basis functions in Appendix A reveals that, for observations without missing values, MARS identifies fourteen groups of Standard Industry Codes (SIC) among which it determines that the models for (log of) total compensation differ. Table 3 categorizes each of the 41 4-digit SIC industries by MARS-created SIC group. The first column of the table lists the industry's 1-digit SIC code, the second column provides a description of the 4-digit SIC industry, and the final columns of the table identify which of the 14 broadly related MARS industries each 4-digit SIC industry belongs. The effects of the various industry variables on total compensation depend on these industry groups; as seen in Appendix A that a 4-digit industry can appear in multiple MARS groupings since different industry groupings can interact with different independent variables.

Generally, researchers investigating industry effects classify firms based on the firm's 1-digit or 2-digit SIC code. The OLS analysis in Table 2 allows CEO compensation to shift upward or downward depending on the firm's 1-digit SIC industry. The Swiss-cheese appearance of Table 3 indicates, at least in terms of CEO compensation, that industry effects are far more complex than a simple upward or downward shift in compensation. Multiple industry interactions exist among the independent variables and the interactions are not grouped according to 1-digit or 2-digit SIC industry.

Table 4 presents the impact of each of the independent variables by industry. The notation with a plus sign (+) as a superscript indicates the expression in brackets is evaluated only for observations where the expression is positive. The expression is set equal to zero for all other observations. Blanks in the table indicate that the coefficient of the expression in the 1st column is zero for that particular industry group. For example, Panel A demonstrates that, as estimated in the MARS model, in SIC1 a one percentage point increase in a company's return on assets (ROA) raises total CEO compensation by 1.4 percent (0.014 log points) when ROA is below 7.047 percent but by 3.6 percent (0.035 log points) when ROA is above 7.047 percent. (In a log-linear specification a one-unit change in an independent variable causes a $e^{\hat{\beta}}-1$ percentage change in the dependent variable, where $\hat{\beta}$ is the estimated parameter. For small values, β is approximately equal to the percentage change.) In the second SIC group a one percentage point increase in ROA has no impact on log total CEO compensation when ROA is below 1.206 percent but, surprisingly, reduces total CEO compensation by 8.0 percent (0.077 log points) when ROA is above 1.206 percent. MARS uncovers no significant impact on CEO compensation from higher ROAs in the other 12 industry groups. The OLS regressions presented in Table 2 model pay for performance as a general phenomenon across industries. The MARS methodology, in contrast, discovers ROA affecting CEO pay in only a few 4-digit SIC industries, meaning that pay for performance is far more limited than one might have originally thought.

The second panel in Table 4 reveals that in all industry groups except for SIC5 and to some extent SIC3, CEO compensation rose over time. The coefficient on year is generally zero from 1992 to 1997 but positive for years 1998 to 2003. The parameter of 0.206 on the years 1998 to 2003 indicates that, all else equal, CEOs earned about 23 percent more in these years than in the years from 1992 to 1997 in industry groups other than SIC3, SIC4, and SIC5. The largest jump in salaries over time occurs in SIC4 where the impact of year moves from a -0.439 log points for years 1992 to 1997 to a +0.206 log points for years 1998 to 2003. Other studies also find a rise in CEO salaries in the 1990s (Bebchuk & Grinstein, 2005). The MARS results indicate not a general upward trend in CEO compensation in the 1990s, as implied by the OLS results in Table 2, but a structural break in compensation occurring in 1998.

As can be seen in Panel C, the impact of an additional year of CEO experience (YRSCEO) depends on the age of the CEO, a rough proxy for general labor market experience, and the overall level of CEO experience. For CEOs younger than 43 an additional year of CEO experience lowers total compensation for individuals serving as CEO for less than 12 years but raises it for individuals serving as CEO for more than 12 years. For CEOs older than 43 an additional year of CEO experience has no impact on total compensation except in SIC2 where the impact of greater CEO experience is positive and SIC3 where the impact of greater CEO experience is negative for individuals serving as CEO for less than 1.63 years.

The MARS results on CEO experience are in contrast to the OLS results in Table 2. OLS finds no impact of CEO experience on total compensation, while MARS discovers additional CEO experience raising compensation in some cases but lowering it in others. The counterintuitive results of CEO experience reducing compensation apply to very few observations in the sample. Only 249 of the sample observations are for CEOs younger than 43 with less than 12 years of CEO experience (6.4 percent) and only 340 observations are for CEOs with less than 1.63 years CEO experience in industry group SIC3 (8.7 percent).

The positive impact of CEO experience on compensation pertains to many more observations: 926 observations in SIC2 have more than 0.583 years of CEO experience and are older than 43 (23.8 percent) and 30 observations are for CEOs younger than 43 with more than 12 years CEO experience (0.8%). For the remaining 2,163 observations (55.5 percent) MARS finds no impact on compensation from greater CEO experience. In other words, the MARS results indicate for the vast majority of CEOs greater CEO experience has either a positive or a neutral impact on compensation although for a few CEOs in some industries and at some levels of general and CEO-specific experience greater years heading the company reduces compensation.

Panel D shows the impact on CEO compensation from increases in company size as measured by sales revenue. As can be seen, the impact of company size depends on the company's current level of sales, the age of the CEO, and the industry. Ignoring the age effect, an increase in sales has a larger impact when a company is small, sales less than $1.7471 billion (70.9 percent of the sample), than when it is large, sales greater than $1.7471.

Age augments the impact of sales on CEO compensation for CEOs older than 43 in companies with less than $8.1352 billion in sales revenue and for CEOs younger than 43 in companies with less than $4.4857 billion in sales revenue. Evaluated at the mean age of 53.8, a $1 billion dollar increase in sales revenue raises CEO compensation in most industry groups by 75.5 percent for companies with sales of less than $1.7471 billion, by 6.3 percent for companies with sales between $1.7471 billion and $8.1352 billion, and by 1.82 percent for companies with sales greater than $8.1352 billion. Mathematically, company size appears to raise CEO compensation but at a decreasing rate.

Table 3. MARS identified industry groups

1-digit SIC Industry	4-digit SIC Industry	SIC1 BF5	SIC2 BF25	SIC3 BF6	SIC4 BF13	SIC5 BF61	SIC6 BF73	SIC7 BF45	SIC8 BF43	SIC9 BF11	SIC10 BF7	SIC11 BF57	SIC12 BF19	SIC13 BF51	SIC14 BF75
Mfg	Broadwoven Fabric Mills, Cotton	■					■			■	■	■			
Mfg	Apparel & Other Finished Prods of Fabrics & Similar Mat'l	■										■			■
Mfg	Men's & Boys' Furnishings, Work Clothing, & Allied Garments		■	■		■	■	■	■	■	■	■	■		
Mfg	Newspapers: Publishing or Publishing & Printing			■	■		■	■			■			■	■
Mfg	Commercial Printing	■					■					■			■
Mfg	Pharmaceutical Preparations			■	■			■	■					■	
Mfg	Biological Products, (No Diagnostic Substances)		■	■	■	■		■				■		■	■
Mfg	Perfumes, Cosmetics & Other Toilet Preparations		■	■	■			■				■			
Mfg	Pottery & Related Products		■	■			■	■	■				■		
Mfg	Special Industry Machinery, NEC			■				■	■		■				
Mfg	Computer & Office Equipment			■	■		■	■			■				
Mfg	Computer Peripheral Equipment, NEC			■		■	■	■	■		■				
Mfg	Electric Housewares & Fans	■	■		■		■		■		■				
Mfg	Telephone & Telegraph Apparatus			■					■				■	■	
Mfg	Motor Vehicle Parts & Accessories	■							■			■			
Mfg	Motor Homes	■							■		■				■
Mfg	Electromedical & Electrotherapeutic Apparatus			■	■		■		■			■			
Mfg	Dolls & Stuffed Toys			■			■	■	■						■
Mfg	Miscellaneous Manufacturing Industries			■		■	■	■	■			■		■	
Trans, Comm & Utilities	Communications Services, NEC			■		■	■		■		■		■		
Trans, Comm & Utilities	Electric Services	■			■				■		■			■	
Trans, Comm & Utilities	Natural Gas Distribution	■			■	■	■			■	■	■			■
Trade	Retail-Apparel & Accessory Stores	■				■		■				■		■	■
Trade	Retail-Women's Clothing Stores			■	■		■	■			■	■			
Trade	Retail-Furniture Stores	■				■				■	■				
Trade	Retail-Drug Stores & Proprietary Stores	■										■			
Trade	Retail-Jewelry Stores	■				■	■			■					
Trade	Retail-Catalog & Mail-Order Houses	■	■							■	■	■			
Finance, Ins, Real Estate	Savings Institution, Federally Chartered	■										■			■
Finance, Ins, Real Estate	Patent Owners & Lessors		■	■					■						■
Services	Services-Personal Services	■				■	■		■			■			■
Services	Services-Help Supply Services	■	■				■						■		■
Services	Services-Computer Programming, Data Processing, etc.			■	■				■				■		
Services	Services-Prepackaged Software		■	■		■			■	■			■		■
Services	Services-Computer Integrated Systems Design			■	■		■	■	■		■				
Services	Services-Telephone Interconnect Systems		■	■											
Services	Services-Business Services, NEC		■	■	■		■	■				■	■	■	
Services	Services-Medical Laboratories		■	■	■			■	■		■	■			■
Services	Services-Child Day Care Services			■				■			■	■			
Services	Services-Research, Accounting, Engineering, Management			■					■				■		■
Services	Services-Commercial Physical & Biological Research			■					■		■	■			■

Table 4: MARS Results on Log Total Compensation

Panel A: Return on Assets (ROA)

	SIC1	SIC2	SIC3	SIC4	SIC5	SIC6	SIC7	SIC8	SIC9	SIC10	SIC11	SIC12	SIC13	SIC14
	BF5	BF25	BF6	BF13	BF61	BF73	BF45	BF43	BF11	BF7	BF57	BF19	BF51	BF75
$(\text{ROA}-7.047)^{+}$	0.035													
$(7.047-\text{ROA})^{+}$	-0.014													
$(\text{ROA}-1.206)^{+}$		-0.077												

Panel B: Year

	SIC1	SIC2	SIC3	SIC4	SIC5	SIC6	SIC7	SIC8	SIC9	SIC10	SIC11	SIC12	SIC13	SIC14
	BF5	BF25	BF6	BF13	BF61	BF73	BF45	BF43	BF11	BF7	BF57	BF19	BF51	BF75
Yrs 98-03	0.206	0.206	0.206	0.206	0.206	0.051	0.206	0.206	0.206	0.206	0.206	0.206	0.206	0.206
Yrs 92-97				-0.439	0.225									
Yrs 92,93,95,98,03			-0.228											

Panel C: CEO Tenure (YRSCEO)

	SIC1	SIC2	SIC3	SIC4	SIC5	SIC6	SIC7	SIC8	SIC9	SIC10	SIC11	SIC12	SIC13	SIC14
	BF5	BF25	BF6	BF13	BF61	BF73	BF45	BF43	BF11	BF7	BF57	BF19	BF51	BF75
$(\text{YRSCEO}-.583)^{+}$		0.018												
$(1.63-\text{YRSCEO})^{+}$			0.350											
$(\text{YRSCEO}-12.0)^{+}$x $(43-\text{AGE})^{+}$	0.019	0.019	0.019	0.019	0.019	0.019	0.019	0.019	0.019	0.019	0.019	0.019	0.019	0.019
$(12-\text{YRSCEO})^{+}$x $(43-\text{AGE})^{+}$	0.030	0.030	0.030	0.030	0.030	0.030	0.030	0.030	0.030	0.030	0.030	0.030	0.030	0.030

Panel D: Sales Revenue (SALES), in billions 2003 $

	SIC1	SIC2	SIC3	SIC4	SIC5	SIC6	SIC7	SIC8	SIC9	SIC10	SIC11	SIC12	SIC13	SIC14
	BF5	BF25	BF6	BF13	BF61	BF73	BF45	BF43	BF11	BF7	BF57	BF19	BF51	BF75
$(\text{SALES}-1.7471)^{+}$	0.018	0.018	0.018	0.018	0.018	0.018	0.018	0.018	-0.019	0.018	0.018	0.018	0.018	0.077
$(1.7471-\text{SALES})^{+}$	-0.519	-0.519	-0.519	-0.519	-0.519	-0.519	-0.519	-0.821	-0.519	-0.519	-0.519	-0.519	-0.519	-0.519
$(0.2881-\text{SALES})^{+}$		-3.000												
$(8.1352-\text{SALES})^{+}$ x$(\text{AGE}-43)^{+}$	-0.004	-0.004	-0.004	-0.004	-0.004	-0.004	-0.004	-0.004	-0.004	-0.004	-0.004	-0.004	-0.004	-0.004
$(4.4857-\text{SALES})^{+}$ x$(43-\text{AGE})^{+}$	0.059	0.059	0.059	0.059	0.059	0.059	0.059	0.059	0.059	0.059	0.059	0.059	0.059	0.059

Panel E: Age of the CEO (AGE)

	SIC1	SIC2	SIC3	SIC4	SIC5	SIC6	SIC7	SIC8	SIC9	SIC10	SIC11	SIC12	SIC13	SIC14
	BF5	BF25	BF6	BF13	BF61	BF73	BF45	BF43	BF11	BF7	BF57	BF19	BF51	BF75
$(\text{AGE}-43)^{+}$	0.035	0.035	0.035	0.035	0.035	0.035	0.017	0.035	0.035	0.035	0.035	-0.016	0.035	0.035
$(43-\text{AGE})^{+}$													0.201	
$(\text{AGE}-54)^{+}$	-0.052	-0.052	-0.052	-0.052	-0.052	-0.052	-0.052	-0.052	-0.052	-0.052	-0.052	-0.052	-0.052	-0.052
$(\text{AGE}-43)^{+}$x $(8.1352-\text{SALES})^{+}$	-0.004	-0.004	-0.004	-0.004	-0.004	-0.004	-0.004	-0.004	-0.004	-0.004	-0.004	-0.004	-0.004	-0.004
$(43\text{-AGE})^{+}$x $(4.4857\text{-SALES})^{+}$	0.059	0.059	0.059	0.059	0.059	0.059	0.059	0.059	0.059	0.059	0.059	0.059	0.059	0.059
$(43-\text{AGE})^{+}$x $(\text{YRSCEO}-12)^{+}$	0.019	0.019	0.019	0.019	0.019	0.019	0.019	0.019	0.019	0.019	0.019	0.019	0.019	0.019
$(43-\text{AGE})^{+}$x $(12-\text{YRSCEO})^{+}$	0.030	0.030	0.030	0.030	0.030	0.030	0.030	0.030	0.030	0.030	0.030	0.030	0.030	0.030

Note: Table 3 lists the specific 4-digit SIC industries comprising each SIC industry grouping. A superscript on a bracketed term indicates the expression is evaluated only for observations where the expression is positive. The expression equals zero for all other observations. Blanks in the table indicate the associated industry effect is zero. The table presents results only for observations with information on all independent variables. Appendix A presents the full set of MARS results including the impact of missing values.

The OLS results in Table 2 examine the impact of sales revenue on CEO compensation but the impact of sales is assumed to be linear. The MARS results suggest that a more appropriate specification would include sales revenue and sales revenue squared to allow for the positive but diminishing returns from company size. (When both sales revenue and sales revenue squared are included in the OLS regression both coefficients are highly statistically significant (p values < 0.001) but the inclusion alters the size and significance of the other coefficients only slightly.) In Table 2 across all specifications, an additional $1 billion of sales revenue creates a 5.0 percent increase in CEO compensation. In Table 3, an additional $1 billion of sales revenue creates in most industries a 6.3 percent increase in CEO compensation when evaluated at the sample means of age and sales revenue ($2.914 billion).

The final panel of Table 4 reports the impact of age on CEO compensation. The last four rows of the Panel E simply duplicate the interactive results on age and sales and age and years as CEO discussed previously. Across all age groups higher sales revenue either expands the positive impact of age on total compensation or contracts the negative impact – the interaction between age and sales is positive. Surprisingly, for CEOs younger than 43 an additional year of general experience as measured by age reduces total compensation, all else equal. The reduction is smaller as years as CEO expands for CEOs serving for fewer than 12 years but is larger as years as CEO expands for CEOs serving more than 12 years. In all but SIC13 an additional year of general experience raises total compensation by 1.42 percent for CEOs from 43 to 54 but reduces total compensation by 7.56 percent for CEOs older than 54 when evaluated at the mean level of sales. The influence of age on total compensation is not impacted by years as CEO for CEOs older than 43.

The stereotypical age/earnings profile has a worker's earnings rising steeply early in his or her career, leveling off over time, and then declining. Researchers include age and age squared as independent variables in an OLS analysis of earnings to capture the positive but diminishing impact of general experience on earnings and to allow for the possibility of earnings hitting a peak at some point. Based on the OLS results, CEO compensation hits a peak somewhere between 54.9 and 57.2 years of age depending on the empirical specification. Although the MARS results do not reproduce the standard leveling off of earnings, they do indicate an earnings peak at age 54, a result largely consistent with the OLS analysis.

Conclusion

In most empirical investigations theory guides the selection of independent variables but rarely dictates the functional relationship between the dependent and the independent variables or specifies all possible interactions among the independent variables. Consequently, researchers generally present several sets of results generated using slightly different estimating relationships to demonstrate that the conclusions of the analysis are robust to model specification. Multivariate Adaptive Regression Splines (MARS) is a data mining technique that examines data for all possible interactions among specified explanatory variables and for non-linear relations between the dependent and explanatory variables. By using MARS researchers can check for the robustness of their empirical findings in a highly structured manner, thereby providing a more convincing case that the results are insensitive to model specification. Additionally, MARS may uncover relationships that can be examined for theoretical content and aid future research in the area.

As an example of how MARS can be used as a procedure to check for robustness and as an aid in future research, we examine data on CEO compensation to determine if pay differences exist between men and women. Most studies find men out earn women by a sizable margin even after controlling for differences in education, experience, and occupation (Altonji & Blank, 1999; Bertrand & Hallock, 2001; Stanley & Jarrell, 1998). Using standard OLS analysis we find no evidence male CEOs have an advantage over female CEOs in terms of compensation. Across the four empirical specifications we examine female CEOs earn the same or more than male CEOs, all else equal. In the MARS methodology the variable representing gender never enters the model indicating that no significant pay difference exists between male and female CEOs. The MARS model controls for observable characteristics and considers all possible interactions among the observable characteristics and total compensation in addition to potential nonlinearities in the relationships between the observable characteristics and total compensation. In short, the absence of a gender effect on CEO compensation is robust.

In terms of the other factors affecting CEO compensation, OLS generates a fairly standard picture of CEO compensation. All else equal, CEOs leading larger companies as measured by sales revenue, more profitable companies as measured by return on assets, and who have more general labor market experience as measured by age earn more than CEOs leading smaller companies, less profitable companies, and who have less general labor market experience. Over time CEO compensation has expanded by almost 5 percent per year in real terms and CEOs in transportation and trade earn less than CEOs in manufacturing. Inconsistent with the human capital model of earnings, OLS finds no reward for CEO experience.

The MARS results are generally consistent with the OLS results but with some important distinctions. Similar to OLS, MARS finds sizable differences in CEO compensation across industries. Unlike OLS, the MARS grouping of industries is unrelated to a broader industry classification such as a 1- or 2-digit SIC code. Further, the MARS industry effects do not simply increase or decrease compensation but instead interact with the other independent variables, suggesting the underlying model of compensation varies by industry groupings. However, note that these industry groupings are not the recognized industry groups based on 1- or 2-digit SIC codes. Similar to OLS, MARS shows CEO compensation rising over time, but unlike OLS the rise is not gradual. In most of the MARS industry groups a structural break in compensation occurs in 1998 causing CEO pay to jump by about 23 percent. In the OLS analysis, the impact of return on assets is modeled as a general phenomenon across industries. The MARS analysis finds return on assets raising CEO compensation but in only one broad industry grouping – meaning pay for performance is fairly limited. The OLS analysis uncovers a positive, linear relationship between sales revenue and CEO compensation. The MARS results suggest sales revenue has a positive but diminishing impact on CEO compensation. In the OLS analysis, the number of years a person has served as CEO appears to have no impact on compensation, while MARS finds CEO experience raising total compensation but only in a few industry groupings. Finally, OLS indicates a CEO's age, a proxy for general labor market experience, raises total compensation but at a decreasing rate, a result in line with the human capital model and the stereotypical age/earnings profile. MARS finds a far more complex relationship with compensation falling, rising, and then falling again as the CEO ages. Both the OLS and the MARS results imply CEO compensation peaks at around 54 years of age.

It is not suggested that MARS be used as a replacement to the standard procedures of model building and hypothesis testing. Instead, MARS may be viewed as a complement to the more traditional methods of analysis. There are implications for practicing managers to consider when evaluating the use of MARS and OLS. For the manager who wants to understand the dynamics of executive compensation, the MARS model provides more details about the specifics related his or her particular situation (e.g., the industry grouping formed by MARS and corresponding interactions). By examining data for unanticipated and possibly complex interactions among the independent variables and for potential nonlinear relationships between the dependent and independent variables, MARS allows researchers to conduct a structured test of robustness and determine important areas for future research. In particular, the MARS analysis of CEO compensation suggests additional work is required to determine the factors causing the compensation explosion in 1998, the reasons for the paucity of pay for performance, and the elements generating common compensation practices across industries.

References

Altonji, J. G., & Blank, R. M. (1999). Race and gender in the labor market. In O. Ashenfelter & D. Card (Eds.), *Handbook of Labor Economics*, *3c*, Amsterdam: Elsevier Science B. V., 3143-3259.

Bebchuk, L. A., & Grinstein, Y. (2005). The growth of executive pay. *Oxford Review of Economic Policy*, *21*, 283-303.

Bertrand, M., & Hallock, K.F. (2001). The gender gap in top corporate jobs. *Industrial and Labor Relations Review*, *55*, 3-21.

De Veaux, R. D., Psichogios, D. C. & Ungar, L. H. (1993). A comparison of two non-parametric schemes: MARS and neural networks. *Computers in Chemical Engineering, 17*, 819-837.

Deichman, J., Haughton, D., Phong, N. & Tung, P.D. (2001). A graphical and statistical analysis of the correlates of poverty in Vietnam in 1993 and 1998. In *Living standards during an economic boom in Vietnam 1993-1998*, UNDP and GSO, Hanoi. Available at http://www.undp.org.vn/undp/docs/2001/living/lse.pdf, accessed January 3rd 2006.

Deichman, J., Eshghi, A., Sayek, S. & Teebagy, N. (2002). Application of multiple adaptive regression splines (MARS) in direct response modeling. *Journal of Interactive Marketing, 16*, 15-27.

Friedman, J. H. (1991). Multivariate Adaptive Regression Splines. *Annals of Statistics, 19*, 1-141.

Jensen, M. C., & Murphy, K. J. (1990). Performance, pay, and top management incentives. *Journal of Political Economy 98*, 225-264.

Munoz, J., & Felicisimo, A. (2004). Comparison of statistical methods commonly used in predictive modeling, *Journal of Vegetation Science, 15*, 285-292.

Salford Systems (2006). http://www.salford-systems.com/mars.php, accessed January 2nd 2006.

Sephton, P. (2001). Forecasting recessions: can we do better on MARS? *Federal Reserve Bank of St. Louis Review, 83*, 39-49.

Stanley, T. D., & Jarrell, S. B. (1998). Gender wage discrimination bias? A meta-regression analysis, *Journal of Human Resources, 33*(4), 947-973.

Willis, R. J. (1986), Wage determinants: A survey and reinterpretation of human capital earnings functions. In O. C. Ashenfelter & R. Layard (Eds.), *Handbook of labor economics*, *1*, Amsterdam: Elsevier Science B.V., 525-602.

Appendix A: The MARS model; basis functions and estimated equation

Basis Functions

BF1 = (SALES > .);

BF3 = max(0, SALES – 1.747087) * BF1;

BF4 = max(0, 1.747087 - SALES) * BF1;

BF5 = (SICNEW = 1 OR SICNEW = 2 OR SICNEW = 5 OR SICNEW = 13 OR SICNEW = 15 OR SICNEW = 16 OR SICNEW = 21 OR SICNEW = 22 OR SICNEW = 23 OR SICNEW = 25 OR SICNEW = 26 OR SICNEW = 27 OR SICNEW = 28 OR SICNEW = 29 OR SICNEW = 31 OR SICNEW = 32) * BF1;

BF6 = (SICNEW = 3 OR SICNEW = 4 OR SICNEW = 6 OR SICNEW = 7 OR SICNEW = 8 OR SICNEW = 9 OR SICNEW = 10 OR SICNEW = 11 OR SICNEW = 12 OR SICNEW = 14 OR SICNEW = 17 OR SICNEW = 18 OR SICNEW = 19 OR SICNEW = 20 OR SICNEW = 24 OR SICNEW = 30 OR SICNEW = 33 OR SICNEW = 34 OR SICNEW = 35 OR SICNEW = 36 OR SICNEW = 37 OR SICNEW = 38 OR SICNEW = 39 OR SICNEW = 40 OR SICNEW = 41) * BF1;

BF7 = (SICNEW = 1 OR SICNEW = 3 OR SICNEW = 4 OR SICNEW = 10 OR SICNEW = 11 OR SICNEW = 12 OR SICNEW = 13 OR SICNEW = 16 OR SICNEW = 20 OR SICNEW = 21 OR SICNEW = 22 OR SICNEW = 24 OR SICNEW = 25 OR SICNEW = 28 OR SICNEW = 35 OR SICNEW = 38 OR SICNEW = 39 OR SICNEW = 41);

BF9 = (YEAR = 1998 OR YEAR = 1999 OR YEAR = 2000 OR YEAR = 2001 OR YEAR = 2002 OR YEAR = 2003) * BF1;

BF10 = (YEAR = 1992 OR YEAR = 1993 OR YEAR = 1994 OR YEAR = 1995 OR YEAR = 1996 OR YEAR = 1997) * BF1;

BF11 = (SICNEW = 1 OR SICNEW = 3 OR SICNEW = 22 OR SICNEW = 25 OR SICNEW = 27 OR SICNEW = 28 OR SICNEW = 34) * BF3;

BF13 = (SICNEW = 4 OR SICNEW = 6 OR SICNEW = 7 OR SICNEW = 8 OR SICNEW = 11 OR SICNEW = 13 OR SICNEW = 17 OR SICNEW = 21 OR SICNEW = 22 OR SICNEW = 24 OR SICNEW = 33 OR SICNEW = 35 OR SICNEW = 37 OR SICNEW = 38) * BF10;

BF15 = (AGE > .) * BF1;

BF16 = (AGE = .) * BF1;

BF17 = max(0, AGE - 43.000) * BF15;

BF18 = max(0, 43.000 - AGE) * BF15;

BF19 = (SICNEW = 3 OR SICNEW = 9 OR SICNEW = 14 OR SICNEW = 20 OR SICNEW = 32 OR SICNEW = 33 OR SICNEW = 34 OR SICNEW = 37 OR SICNEW = 40) * BF17;

BF21 = (YRSCEO > .) * BF18;

BF23 = max(0, YRSCEO - 11.997) * BF21;

BF24 = max(0, 11.997 - YRSCEO) * BF21;

BF25 = (SICNEW = 3 OR SICNEW = 7 OR SICNEW = 8 OR SICNEW = 9 OR SICNEW = 13 OR SICNEW = 28 OR SICNEW = 30 OR SICNEW = 32 OR SICNEW = 34 OR SICNEW = 36 OR SICNEW = 37 OR SICNEW = 38);

BF27 = (SALES > .) * BF25;

BF30 = max(0, 0.288101 - SALES) * BF27;

BF32 = max(0, 8.135196 - SALES) * BF17;

BF33 = (YEAR = 1992 OR YEAR = 1993 OR YEAR = 1995 OR YEAR = 1998 OR YEAR = 2003) * BF6;

BF35 = (ROA > .) * BF5;

BF37 = max(0, ROA - 7.047) * BF35;

BF38 = max(0, 7.047 - ROA) * BF35;

BF39 = (YRSCEO > .) * BF25;

BF40 = (YRSCEO = .) * BF25;

BF41 = max(0, YRSCEO - 0.583) * BF39;

BF43 = (SICNEW = 3 OR SICNEW = 6 OR SICNEW = 9 OR SICNEW = 10 OR SICNEW = 12 OR SICNEW = 13 OR SICNEW = 14 OR SICNEW = 15 OR SICNEW = 16 OR SICNEW = 17 OR SICNEW = 18 OR SICNEW = 19 OR SICNEW = 20 OR SICNEW = 21 OR SICNEW = 30 OR SICNEW = 31 OR SICNEW = 33 OR SICNEW = 34 OR SICNEW = 35 OR SICNEW = 37 OR SICNEW = 38 OR SICNEW = 40 OR SICNEW = 41) * BF4;

BF45 = (SICNEW = 3 OR SICNEW = 4 OR SICNEW = 6 OR SICNEW = 7 OR SICNEW = 8 OR SICNEW = 9 OR SICNEW = 10 OR SICNEW = 11 OR SICNEW = 12 OR SICNEW = 18 OR

SICNEW = 19 OR SICNEW = 23 OR SICNEW = 24 OR SICNEW = 35 OR SICNEW = 37 OR SICNEW = 38 OR SICNEW = 39) * BF17;

BF47 = (AGE > .) * BF39;

BF49 = max(0, AGE - 54.000) * BF47;

BF51 = (SICNEW = 4 OR SICNEW = 6 OR SICNEW = 7 OR SICNEW = 14 OR SICNEW = 19 OR SICNEW = 21 OR SICNEW = 23 OR SICNEW = 37) * BF21;

BF53 = (ROA > .) * BF40;

BF55 = max(0, ROA - 1.206) * BF53;

BF57 = (SICNEW = 1 OR SICNEW = 2 OR SICNEW = 3 OR SICNEW = 5 OR SICNEW = 7 OR SICNEW = 8 OR SICNEW = 15 OR SICNEW = 17 OR SICNEW = 19 OR SICNEW = 22 OR SICNEW = 23 OR SICNEW = 24 OR SICNEW = 26 OR SICNEW = 28 OR SICNEW = 29 OR SICNEW = 31 OR SICNEW = 37 OR SICNEW = 38 OR SICNEW = 39 OR SICNEW = 41) * BF1;

BF59 = (SICNEW = 12 OR SICNEW = 19 OR SICNEW = 24 OR SICNEW = 26 OR SICNEW = 30 OR SICNEW = 34) * BF16; BF61 = (SICNEW = 3 OR SICNEW = 7 OR SICNEW = 12 OR SICNEW = 19 OR SICNEW = 20 OR SICNEW = 22 OR SICNEW = 23 OR SICNEW = 25 OR SICNEW = 28 OR SICNEW = 32 OR SICNEW = 35) * BF10;

BF63 = (YRSCEO = .) * BF9;

BF64 = (YRSCEO > .) * BF9;

BF66 = max(0, 4.485668 - SALES) * BF21;

BF67 = (SICNEW = 2 OR SICNEW = 5 OR SICNEW = 6 OR SICNEW = 7 OR SICNEW = 14 OR SICNEW = 16 OR SICNEW = 21 OR SICNEW = 31 OR SICNEW = 34 OR SICNEW = 40) * BF63;

BF73 = (SICNEW = 1 OR SICNEW = 3 OR SICNEW = 4 OR SICNEW = 5 OR SICNEW = 9 OR SICNEW = 11 OR SICNEW = 12 OR SICNEW = 13 OR SICNEW = 17 OR SICNEW = 18 OR SICNEW = 19 OR SICNEW = 20 OR SICNEW = 22 OR SICNEW = 24 OR SICNEW = 27 OR SICNEW = 31 OR SICNEW = 32 OR SICNEW = 35 OR SICNEW = 37) * BF64;

BF75 = (SICNEW = 2 OR SICNEW = 4 OR SICNEW = 5 OR SICNEW = 7 OR SICNEW = 16 OR SICNEW = 18 OR SICNEW = 22 OR SICNEW = 23 OR SICNEW = 29 OR SICNEW = 30 OR SICNEW = 31 OR SICNEW = 32 OR SICNEW = 34 OR SICNEW = 38 OR SICNEW = 40 OR SICNEW = 41) * BF3;

BF77 = (YRSCEO > .) * BF6;

BF80 = max(0, 1.626 - YRSCEO) * BF77;

Estimated Equation

Y = 6.661 + 2.206 * BF1 + 0.0177346 * BF3 - 0.518625 * BF4 - 1.014 * BF5 - 0.399 * BF7 + 0.206 * BF9 - 0.203566 * BF11 - 0.439 * BF13 + 0.035 * BF17 - 0.051 * BF19 - 0.595 * BF21 + 0.019 * BF23 + 0.030 * BF24 + 0.408 * BF25 – 3.000 * BF30 - 0.00353013 * BF32 - 0.228 * BF33 + 0.035 * BF37 - 0.014 * BF38 + 0.018 * BF41 - 0.301961 * BF43 - 0.018 * BF45 - 0.052 * BF49 + 0.201 * BF51 - 0.077 * BF55 - 0.158 * BF57 - 0.869 * BF59 + 0.225 * BF61 + 0.0589534 * BF66 - 0.762 * BF67 - 0.155 * BF73 + 0.0590723 * BF75 + 0.350 * BF80;

Appendix B: Variables

AGE = age of the CEO.

NEWSIC = 4-digit SIC industry. NEWSIC is a categorical variable ranging from 1 (Broadwoven Fabric Mills, Cotton) to 41 (Services-Commercial Physical & Biological Research). See Table 3 for a complete listing of the 4-digit SIC industries.

ROA = return on assets.

SALES = sales revenue in billions of 2003 $.

Y = log of total compensation.

YEAR = observation year.

YRSCEO = years serving as CEO.

Journal of Modern Applied Statistical Methods
November, 2008, Vol. 7, No. 2, 577-587

1538 – 9472/08/$95.00

Application of Dynamic Poisson Models to Japanese Cancer Mortality Data

Shuichi Midorikawa
Tokyo University of Science

Etsuo Miyaoka
Tokyo University of Science

Bruce Smith
Dalhousie University

A dynamic Poisson model is used with a Bayesian approach to modeling to predict cancer mortality. The complexity of the posterior distribution prohibits direct evaluation of the posterior, and so parameters are estimated by using a Markov Chain Monte Carlo method. The model is applied to analyze lung and stomach cancer data which have been collected in Japan.

Key words: Dynamic Poisson model, Markov Chain Monte Carlo, cancer mortality data, age-period-cohort model prediction.

Introduction

The number of cases of stomach cancer in the Japanese male population is tabulated in Table 1, by five year period, and by 5 year age group. Periods are identified in the Table by their central year. For example, the period labeled 1970 includes all data for the years 1968 through 1972, inclusive. These data were obtained from the Japanese Ministry of Health and Welfare. (http://wwwdbtk.mhlw.go.jp/toukei/index.html)

The goal of this article is the development trend models for these data, and in particular, the development of methods for short to medium term prediction, which will be important from the perspective of public health planning. The entries in Table 1 for the 2005 and 2010 periods are, in fact, predictions, calculated as described in section 5 below. In assessing trends in such data, care must be taken to accommodate for trends in the underlying population structure. In particular, the reduction in numbers of cancers at increased age is due primarily to the reduction in the associated number of individuals at risk. To accommodate for the number of individuals at risk, we focus on the incidence rate, equal to the number of events divided by the number at risk. Table 2 shows the incidence rate as numbers of stomach cancers per million males, calculated by dividing the raw incidence numbers from Table 1 by the population size (the total number of males in the associated age group), and multiplying by one million. The population cohort numbers were obtained from the Japanese Ministry of Internal Affairs and Communications.

There are some notable trends in the incidence rate data of Table 2. In particular, except for the oldest few age groups, the incidence rate is increasing with age, in each period. On the other hand, at least up until age 65 or 70, the incidence rate within age group appears to be more or less decreasing over time. Incidence rates for female stomach cancer, and for male and female lung cancer, were similarly calculated, and are illustrated in the appendix, together with predictions for the periods centered at 2005 and 2010. Patterns similar to the males for the rate of female stomach cancer are noted (Table 5), and with respect to rates of lung cancer in both males and females, the data appear to show increasing rates over age group, and over time (Tables 6 and 7).

Shuichi Midorikawa is in the Department of Mathematics. Email: shuu@m2.dion.ne.jp Etsuo Miyaoka is Professor of Statistics and Probability in the Department of Mathematics. Email:miyaoka@rs.kagu.tus.ac.jp. Bruce Smith is Professor in the Department of Mathematics and Statistics. Email: Bruce.Smith@Dal.Ca.

Table 1: Numbers of cases of stomach cancer - males

Age	5 year period centered at												
Group	1950	1955	1960	1965	1970	1975	1980	1985	1990	1995	2000	2005	2010
15-19	9	9	15	17	19	9	5	5	11	5	4	5	5
20-24	27	30	46	72	80	59	28	21	18	20	18	18	18
25-29	65	106	127	158	162	162	104	87	49	49	30	45	45
30-34	166	196	300	353	346	309	308	196	102	77	70	87	95
35-39	359	395	470	615	628	562	526	453	315	207	142	171	165
40-44	788	854	790	781	1004	1003	799	719	646	494	303	322	300
45-49	1406	1568	1517	1309	1387	1638	1583	1192	1101	1027	724	766	581
50-54	2206	2470	2488	2402	1991	1922	2465	2203	1772	1626	1608	1491	1231
55-59	3024	3398	3717	3666	3214	2871	2632	3253	2992	2592	2458	2209	2068
60-64	3602	4125	4569	4993	4638	4201	3603	3467	4263	4034	3408	3310	2929
65-69	3465	4195	4799	5483	5699	5334	5013	4082	4081	5210	5237	4213	3998
70-74	2505	3244	4147	4483	5228	5594	5472	4952	4258	4869	6009	4757	4257
75-79	1063	1743	2289	3010	3459	4132	4743	4702	4613	4571	4859	4234	4005
80-84	261	479	829	1034	1457	2000	2636	3273	3512	4073	3977	2976	2971
85-89	61	78	170	241	313	536	804	1283	1727	2361	2767	1747	1580
90-	3	15	16	28	35	77	134	269	460	806	1184	628	598

Table 2: Stomach cancer - males, rate per million

Age	5 year period centered at										
Group	1950	1955	1960	1965	1970	1975	1980	1985	1990	1995	2000
15-19	2	2	3	3	4	2	1	1	2	1	1
20-24	7	7	11	16	15	13	7	5	4	4	4
25-29	23	28	31	38	36	30	23	22	12	11	6
30-34	70	70	80	85	83	67	57	43	26	19	16
35-39	151	170	170	164	153	134	115	84	70	53	35
40-44	358	367	347	286	275	244	193	160	121	110	78
45-49	696	734	672	588	522	450	394	294	246	194	163
50-54	1283	1280	1219	1105	930	740	698	565	444	370	310
55-59	2193	2114	2062	1899	1584	1395	1055	959	791	667	575
60-64	3246	3362	3178	3072	2656	2183	1864	1476	1318	1121	911
65-69	4353	4564	4673	4498	4090	3411	2890	2305	1864	1744	1562
70-74	4636	5463	5979	5682	5455	4892	4170	3332	2735	2521	2253
75-79	3971	5096	6076	6661	6517	6021	5607	4718	3855	3644	2997
80-84	2730	3596	4901	5531	6048	6511	6326	5978	5176	4957	4355
85-89	2489	2304	3527	4008	4381	5321	5805	6696	6259	6540	5806
90-	706	2573	1937	2040	2003	3554	4041	4873	5647	6894	6715

The numbers of deaths from cancer represent count data, and as such, statistical models for counts, rates or proportions are appropriate. Cancer mortality rates have often been modeled using a classical age-period-cohort model, which is a type of Poisson regression model, and was used to make predictions for lung cancer mortality rates in England and Wales (Osmond, 1985), for example. In particular, for the data in Tables 1 and 2, there are 16 age groups, 11 periods (the 5 year time intervals), and 26 cohorts. Individual cohorts are represented as diagonal slices in the Table. For example, in Table 2, two cohorts are identified by boldface type. The oldest cohort includes those individuals who where were 90 years or older in the period labeled 1950, and this is the only period in which data was recorded for this cohort. The youngest cohort includes those individuals who were 15-19 in the period labeled 2000, and there is again only one year of incidence data for this cohort. There are 6 cohorts which include a maximum of 11 periods of incidence data.

Let $i \in (1,2,\ldots,16)$ index age group, where age group 1 includes 15-19 year olds, age group 2 includes 20-24 year olds, and so on; $j \in (1,2,\ldots,11)$ index 5 year period, with period 1 centered at 1950, period 2 centered at 1955, and so on; and $k \in (1,2,\ldots,26)$ index cohort, where, for example, cohort 26 includes individuals who were 15-19 in 2000, cohort 2 includes individuals 85-89 in 1950, and so on.

Let Y_{ijk} denote the number of cases in age group i, period j and cohort k. The classical age-period-cohort model assumes that Y_{ijk} is a Poisson random variable with mean λ_{ijk}, where

$$\log(\lambda_{ijk}) = \log(n_{ijk}) + \alpha_i + \beta_j + \gamma_k. \quad (1)$$

Here α_i, β_j and γ_k are the effects of age group i, period j and cohort k respectively. The size of the population at risk, assumed to be known without error from census data, is denoted as n_{ijk}, and was used to transform the raw incidence data in Table 1 to the rates in Table 2. Inclusion of the offsets n_{ijk} in the model for the Poisson mean implies that we are effectively modeling incidence rates λ_{ijk} / n_{ijk}, thereby correcting for the number at risk.

It is clear that the parameterization is not identifiable, as we are using three co-ordinates to index into a two dimensional Table of counts. In particular, $k = 16 - i + j$.

Detailed discussions of this model, including identifiability issues, are included, for example, in Osmond and Gardner (1982), Clayton and Schifflers (1987a, b), and Holford (1991), and various methods have been suggested to overcome the non-identifiability problem, for example, imposing constraints on the parameters (Osmond & Gardner, 1982; Holford, 1991), or restricting consideration to certain estimable functions of the parameters (Clayton & Schifflers, 1987a; Holford, 1991). Clayton & Schifflers (1987a, b) advised the use of a reduced age-period or age-cohort model whenever possible and the use of the full age-period-cohort model only when no other model provides a satisfactory fit. Tango (1985) showed that nonlinear effect parameters can be uniquely determined by imposing restrictions on each block of parameters, for example, $\sum \alpha_i = \sum \beta_j = \sum \gamma_l = 0$, with the nonlinear age effects being specified as:

$$\tilde{\alpha}_j = \alpha_j - \frac{\sum_{j=1}^{A} L(j,A)\alpha_j}{\sum_{j=1}^{A} L(j,A)^2}, \quad (2)$$

$$\text{where } L(j,A) = j - \left(\frac{A+1}{2}\right).$$

It is important to note that while individual age, period and cohort parameters are not identifiable, forward prediction is possible (Holford, 1985).

Different cohorts are typically unequally represented in age-period-cohort data. In the present case, there are single observations on cohorts 1 and 26, two observations on cohorts 2 and 25, eleven observations on each of cohorts 11 through 16, and so on. Therefore, the precision of estimated cohort effects will differ markedly, which has important consequences for prediction. For example, simple predictive models that carry forward estimated cohort effects may lead to predictions with a high degree of variability. Recently, Bayesian models have been used to smooth predictions by incorporating a priori beliefs about the smoothness of the model parameters. Berzuini and Clayton (1994) predicted lung cancer mortality rates using a Bayesian age-period-cohort model. Besag, et al. (1995) fit a Bayesian logistic regression to prostate cancer mortality rates in the USA, with age, period and cohort as explanatory variables, and Bray (2000, 2002) used Gaussian autoregressive priors for incidence rates of Hodgkin's disease.

This paper is organized as follows: A dynamic Poisson model and a dynamic age-period-cohort model are specified, Markov Chain Monte Carlo is reviewed and the estimation method is discussed in detail, a prediction method is described, and the result of the analysis of Japanese cancer data is provided. Finally, concluding remarks are given.

Model Specification

Throughout in this section, y_t denotes the t-th in a sequence of observations, $t = 1, \ldots, T$, θ_t is a p-dimensional parameter vector, F_t is a known p-dimensional vector of regressors, G_t is a known $p \times p$ matrix, w_t is a p-dimensional vector of errors with covariance matrix W, and $g(\cdot)$ is a link function.

Dynamic Poisson Model

A dynamic Poisson model is a state space time series model consisting of observation and system equations, as follows:

Observation equation:

$$P(y_t \mid \lambda_t) = \frac{\exp(-\lambda_t)\lambda_t^{y_t}}{y_t!}, g(\lambda_t) = F_t'\theta_t. \quad (3)$$

System (state) equation:

$$\theta_t = G_t\theta_{t-1} + \eta_{t,} \quad \eta_t \sim N_p[0, W]. \quad (4)$$

When there is no system equation, the dynamic Poisson model becomes the usual Poisson regression model. The dynamic Poisson model is a particular case of the general state space model, discussed, for example, in Kitagawa and Gersch (1996). There is currently much activity in the development of algorithms for general state space models, focusing primarily on so-called particle filters. For example, see Kitagawa (1998) or Doucet, et al. (2001).

Dynamic Age-Period-Cohort Model

To incorporate the age-period-cohort model within the dynamic Poisson model, let $i(t)$, $j(t)$ and $k(t)$ denote the age, period and cohort indices associated with observation Y_t, and denote the associated age, period and cohort effects as $\alpha_{i(t)}$, $\beta_{j(t)}$ and $\gamma_{i(t)}$. Assume $F_t' = (\log(n_t), 1, 1, 1)$, where n_t is the number at risk for observation t, and let $\theta_t' = (1, \alpha_{i(t)}, \beta_{j(t)}, \gamma_{k(t)})$.

Let $\eta_t' = (0, \eta_{i(t)}^{\alpha}, \eta_{j(t)}^{\beta}, \eta_{k(t)}^{\gamma})$, and $G_i = I$ be the 4×4 identity matrix. In this case the dynamic state space model is specified by the following observation and system equations.

Observation equation:

$$\begin{aligned} P(y_t \mid \lambda_t) &= \frac{\exp(-\lambda_t)\lambda_t^{y_t}}{y_t!}, \\ \log(\lambda_t) &= \log(n_t) + \alpha_{i(t)} + \beta_{j(t)} + \gamma_{k(t)}. \end{aligned} \quad (5)$$

System (state) equation:

$$\begin{aligned} \alpha_{i(t)} &= \alpha_{i(t)-1} + \eta_{i(t)}^{\alpha}, \quad \eta_{i(t)}^{\alpha} \sim N[0, W_\alpha], \\ \beta_{j(t)} &= \beta_{j(t)-1} + \eta_{j(t)}^{\beta}, \quad \eta_{j(t)}^{\beta} \sim N[0, W_\beta], \\ \gamma_{k(t)} &= \gamma_{k(t)-1} + \eta_{k(t)}^{\gamma}, \quad \eta_{k(t)}^{\gamma} \sim N[0, W_\gamma]. \end{aligned} \quad (6)$$

This assumes that the system equation corresponds to three independently evolving random walks for age, period and cohort effect - the same model as considered by Knorr-Held and Rainer (2001). The state variables $\{\alpha_{i(t)}, \beta_{j(t)}, \gamma_{k(t)}, t = 1, \ldots, T\}$ take the form of time varying parameters, while the variances W_α, W_β, and W_γ are assumed not to depend on time.

In addition to the observation and system equations, a Bayesian dynamic age-period-cohort model requires the specification of prior distributions for model parameters. However, because of the recursive nature of the state equation, the Bayesian model requires prior distributions only for W_α, W_β, W_γ, α_0, β_0, γ_0. Where $N[\mu,\sigma^2]$ denotes the normal distribution with mean μ and variance σ^2 and $IG[\nu,S]$ denotes the inverse gamma distribution with scale parameter S and shape parameter ν, assume the following prior distributions: $\alpha_0 \sim N[\mu_\alpha, R_\alpha]$, $\beta_0 \sim N[\mu_\beta, R_\beta]$, $\gamma_0 \sim N[\mu_\gamma, R_\gamma]$, $W_\alpha \sim IG[\nu_\alpha, S_\alpha]$, $W_\beta \sim IG[\nu_\beta, S_\beta]$ and $W_\gamma \sim IG[\nu_\gamma, S_\gamma]$. Non-informative priors are achieved by letting R_α^{-1}, R_β^{-1}, R_γ^{-1}, ν_α, ν_β, ν_γ, S_α, S_β, and $S_\gamma \to 0$. Other prior was applied to the dynamic age-period-cohort model, but the result was similar to non-informative priors.

Where there are A age groups, P periods and C cohorts, it follows that the joint posterior for $\alpha_0,\alpha_1,\ldots,\alpha_A$, $\beta_0,\beta_1,\ldots,\beta_P$, $\gamma_0,\gamma_1,\ldots,\gamma_C$, W_α, W_β and W_γ is given by:

$$\pi(\alpha_0,\ldots,\alpha_A,\beta_0,\ldots,\beta_P,\gamma_0,\ldots,\gamma_C,W_\alpha,W_\beta,W_\gamma \mid y) \propto$$
$$\prod_{t=1}^{T}\frac{\exp(-\lambda_t)\times\lambda_t^{y_t}}{y_t!}\times\prod_{j=1}^{A}(W_\alpha)^{\frac{1}{2}}\exp(-\frac{1}{2}W_\alpha^{-1}(\alpha_j-\alpha_{j-1}))$$
$$\times\prod_{k=1}^{P}(W_\beta)^{\frac{1}{2}}\exp(-\frac{1}{2}W_\beta^{-1}(\beta_k-\beta_{k-1}))$$
$$\times\prod_{l=1}^{C}(W_\gamma)^{\frac{1}{2}}\exp(-\frac{1}{2}W_\gamma^{-1}(\gamma_l-\gamma_{l-1}))$$
$$\times\exp\left[-\frac{1}{2}R_\alpha^{-1}(\alpha_0-\mu_\alpha)^2\right]\times\exp\left[-\frac{1}{2}R_\beta^{-1}(\beta_0-\mu_\beta)^2\right]$$
$$\times\exp\left[-\frac{1}{2}R_\gamma^{-1}(\gamma_0-\mu_\gamma)^2\right]$$
$$\times f_{IG}(W_\alpha;\nu_\alpha,S_\alpha)\times f_{IG}(W_\beta;\nu_\beta,S_\beta)\times f_{IG}(W_\gamma;\nu_\gamma,S_\gamma). \tag{7}$$

where $\lambda_t = n_t \exp(\alpha_{i(t)} + \beta_{j(t)} + \gamma_{k(t)})$, and $f_{IG}(\cdot;\nu,S)$ is the inverse gamma density function with parameters ν and S.

More generally, the independence structure of the priors could be removed by assuming that $W = (W_\alpha, W_\beta, W_\gamma)'$ follows a trivariate inverse Wishart distribution with kernel $|W|^{-\frac{1}{2}\nu W}\exp\left(-\frac{1}{2}tr\left(W^{-1}S_W\right)\right)$, and that $(\alpha_0,\beta_0,\gamma_0)'$ has a trivariate Gaussian distribution with mean vector μ and covariance matrix R.

Estimation Method

A Bayesian approach is taken to estimate parameters using posterior means. As analytical calculation of integrals with respect to the posterior distribution is typically intractable, a Markov chain Monte Carlo method has been used to approximate the posterior means. The Gibbs sampler was used to generate samples from the joint posterior distribution. General discussions of the Gibbs sampler are provided, for example, by Geman and Geman (1984) and Gammerman (1997). The WinBugs implementation was used to carry out computations (Spiegelhalter, et al., 2003), with non-informative hyper-priors referred to previously.

As described, Tango (1985) was followed in defining nonlinear age and period effects after applying zero sum constraints. Such mean constraints were also used by Berzuini and Clayton (1994) and Bray (2000, 2002).

In order to assess convergence of the sampler, two chains of 10,000 iterations were run from different initial values and time series plots of the MCMC samples were examined. As an example, Figure 1 shows a plot of the sampled values of γ_1, for the male stomach cancer data. And Figure 2 shows the autocorrelation function of γ_1. The plot suggests that convergence was achieved, and it was confirmed that all other parameters were convergent in the same manner.

Prediction

Osmond (1985) used a standard age-period-cohort model (1) to project lung cancer mortality rates for England and Wales. In this method, unknown period and cohort effects for future periods are estimated using linear regression, while estimated age effects need not be extrapolated.

Figure 1: Time series plots of MCMC iterations for γ_1.

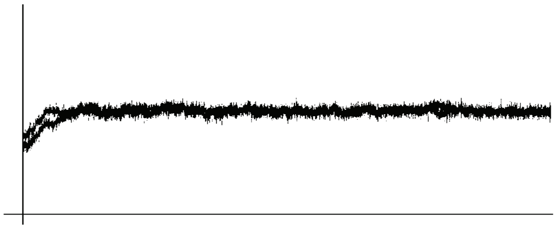
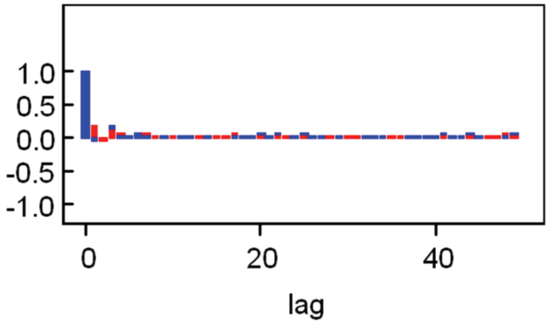


Figure 2: The autocorrelation function of γ_1.



A criticism of the regression, while estimated age effects need not be extrapolated. A criticism of the method is the arbitrariness introduced by the choice of past values to use in the regression, and the type of regression model (e.g., weighted or unweighted).

More recently, parametric bootstrap methods have been used to make projections, for example, by Berzuini and Clayton (1994) and Bray (2000, 2002). In particular, to obtain a prediction for $\tilde{\lambda}_{T+1}$ given data $Y_1, \ldots, Y_T$, sample

$$\tilde{\theta}_{T+1} \sim N\left[G_{T+1}\hat{\theta}_T, \hat{W}\right], \qquad (8)$$

where $\hat{W}$ and $\hat{\theta}_T$ are estimates based on $Y_1, \ldots, Y_T$. Then set

$$\tilde{\lambda}_{T+1} = F'_{T+1}\tilde{\theta}_{T+1}. \qquad (9)$$

This process is repeated J times leading to $\left\{\tilde{\lambda}^{(j)}_{T+1}, j = 1, \ldots, J\right\}$, which are then averaged to provide the overall prediction $\tilde{\lambda}_{T+1}$ of Y_{T+1}. The prediction at time $T+2$ is then based on the combined data $Y_1, \ldots, Y_T$ and prediction $\tilde{\lambda}_{T+1}$. In carrying out the calculations, $J = 100$ was used. The Table 3 shows predicted values and simulated 95% prediction intervals for male stomach cancer in 2005.

Predictions were also made using the traditional age-period-cohort model (1). To estimate age and period effects, a simple linear regression was used on one previous period or age group.

Table 3: Predicted value and simulated 95% prediction intervals for male stomach cancer in 2005

Age Group	Lower	Predicted Value	Upper
15-19	5.580305162	5.745869588	5.91469553
20-24	17.96144371	18.07019342	18.17972061
25-29	45.17424256	45.46220645	45.7517975
30-34	86.88730352	87.43865363	87.9901386
35-39	170.0798945	171.1925396	172.2686358
40-44	320.5989165	322.6740741	324.7104773
45-49	762.0522524	766.9412402	771.9716268
50-54	1481.543376	1491.174459	1500.812711
55-59	2195.522161	2209.505762	2223.767449
60-64	3288.881735	3310300914	3331.643898
65-69	4186.238369	4213.082323	4240.180761
70-74	4727.281932	4757.428491	4788.496251
75-79	4207.928407	4234.477476	4262.194609
80-84	2957.042947	2976.174262	2995.429859
85-89	1736.78588	1747.876571	1759.049125
90-	624.2118163	628.1242619	632.1725958

To assess the adequacy of models for fitting and prediction, the first nine periods for model fitting were used and projections for the tenth and eleventh periods were constructed. Estimates and predictions were compared with observed counts using the following estimates of residual and prediction error.

$$\text{Scaled residual error} = \sum_{i=1}^{9} \frac{(y_i - \hat{y}_i)^2}{\hat{y}_i}, \quad (10)$$

$$\text{Scaled prediction error} = \sum_{j=1}^{11} \frac{(y_j - \tilde{y}_j)^2}{\tilde{y}_j}, \quad (11)$$

where $\hat{y}_i$ is the fitted value for period i and $\tilde{y}_j$ is the predicted value for period j. Table 4 shows these estimates of residual and prediction errors for the age-period-cohort model and a dynamic Poisson models. The estimates of residual error are consistently a bit smaller for the age-period-cohort model, as compared to the dynamic Poisson model.

For the male stomach cancer data, the estimated prediction error is a bit smaller using the age-period-cohort model. However, in the other three cases, the prediction error is smaller using the dynamic Poisson model, and dramatically so in the case of male lung cancer. This suggests that the dynamic Poisson model is the preferred method for making future predictions.

The latter two columns of Tables 1, 5, 6 and 7 contain predictions of lung and stomach cancer rates to the periods centered at 2005 and 2010 using the dynamic Poisson model.

Modeling Variance Heterogeneity

Thus far, we have assumed constant variances for each of the system variables $\alpha_{i(t)}$, $\beta_{i(t)}$, and $\gamma_{i(t)}$ of the dynamic Poisson model. Under this assumption, we have observed that some estimated variances were very large, leading to imprecision of predictions. For example, children born during the years when war occurred, might be faced high risk, then the cohort effect become extremely large than children born at another time. In an attempt to reduce the variability in predictions, the model has been generalized to include non-constant variance, as follows.

$$\begin{aligned} \alpha_{i(t)} &= \alpha_{i(t)-1} + \eta^{\alpha}_{i(t)}, \quad \eta^{\alpha}_{i(t)} \sim N\left[0, W_{\alpha_{i(t)}}\right], \\ \beta_{j(t)} &= \beta_{j(t)-1} + \eta^{\beta}_{j(t)}, \quad \eta^{\beta}_{j(t)} \sim N\left[0, W_{\beta_{j(t)}}\right], \quad (12) \\ \gamma_{k(t)} &= \gamma_{k(t)-1} + \eta^{\gamma}_{k(t)}, \quad \eta^{\gamma}_{k(t)} \sim N\left[0, W_{\gamma_{k(t)}}\right], \end{aligned}$$

Again using non-informative priors, this led, for example, to estimates $\hat{W}_{\beta_1}, \hat{W}_{\beta_2}, \ldots, \hat{W}_{\beta_P}$ for the P period effect variances, which were averaged to produce an overall estimate $\tilde{W}_{\beta} = {}^{1}\!/_{P} \sum_{j=1}^{P} \hat{W}_{\beta_j}$. This latter quantity was then used to predict the $N+1$'st period effect, as

$$\tilde{\beta}_{j(N+1)} \sim N\left[\tilde{\beta}_{j(N)}, \tilde{W}_{\beta}\right], \quad (13)$$

For moderately large P, $\tilde{W}_{\beta}$ should typically be less than $\hat{W}_{\beta}$, thereby increasing the stability of

Table 4: Scaled residual and scaled prediction error

		Scaled Residual	Scaled Prediction Error
Stomach Man	Dynamic Poisson model	6.62314	79.83967
	Age-Period-Cohort model	6.575341	73.35086
Stomach Woman	Dynamic Poisson model	8.740782	80.4833
	Age-Period-Cohort model	8.169002	100.3244
Lung Man	Dynamic Poisson model	2.066267	48.64581
	Age-Period-Cohort model	2.26717	234.6294
Lung Woman	Dynamic Poisson model	1.150069	49.20646
	Age-Period-Cohort model	1.12215	61.80142

the forecast. Indeed, a small $\widetilde{W}_\beta$ could be obtained when the number of death at a specific period group was increased. Table 5 shows the estimated variances for the fixed and heterogeneous variance models.

Conclusion

In the data sets considered, it was observed that the classical age-period-cohort model provided a better fit to past data than did the dynamic age-period-cohort model. On the other hand, when the focus is on making projections, it was found that the classical age-period-cohort model, which makes strong parametric and regression assumptions, was out performed by the dynamic model. Under the assumption of homogeneous error variances in the system equations of the dynamic age-period-cohort model, large standard errors were observed in several cases. It is possible that at least some of this imprecision is the result of natural variation in the Monte Carlo algorithm. Further research will focus on incorporating heterogeneous variances into the model.

The focus has been on the dynamic Poisson model, but the dynamic model can be extended in a straightforward manner to incorporate generalized linear models.

Table 5: The value of $\widetilde{W}_\beta$ and $\hat{W}_\beta$ to each data in Japan

		Variance For Period	Variance For Cohort
Stomach Man	Homogeneous	0.001324854 $\hat{W}_\beta$	0.034891835 $\hat{W}_\gamma$
	Heterogeneous	0.003901684 $\widetilde{W}_\beta$	0.02093981 $\widetilde{W}_\gamma$
Stomach Woman	Homogeneous	0.001282216 $\hat{W}_\beta$	0.034843206 $\hat{W}_\gamma$
	Heterogeneous	0.003719764 $\widetilde{W}_\beta$	0.029620221 $\widetilde{W}_\gamma$
Lung Man	Homogeneous	0.035637919 $\hat{W}_\beta$	0.049188392 $\hat{W}_\gamma$
	Heterogeneous	0.031894849 $\widetilde{W}_\beta$	0.040851123 $\widetilde{W}_\gamma$
Lung Woman	Homogeneous	0.049800797 $\hat{W}_\beta$	0.039339103 $\hat{W}_\gamma$
	Heterogeneous	0.043756821 $\widetilde{W}_\beta$	0.028132685 $\widetilde{W}_\gamma$

References

Berger, J. (1985) *Statistical decision theory and Bayesian analysis* (2nd ed). NY:Springer Verlag.

Berzuini, C., & Clayton, D. (1994). Bayesian analysis of survival on multiple time scales. *Statistics in Medicine*, *13*, 823-838.

Besag, J., Green, P., Higdon, D., & Mengersen, K. (1995) Bayesian computation and stochastic systems. *Statistical Science*, *10*, 3-66.

Bray, I., Brennan, P., & Boffetta, P. (2000). Projections of alcohol- and tobacco-related cancer mortality in central Europe. *International Journal of Cancer*, *87*, 122-128.

Bray, I. (2002). Application of Markov chain Monte Carlo methods to projecting cancer incidence and mortality. *Journal of the Royal Statistical Society*, *51*, Part 2, 151-164.

Clayton, D., & Schifflers, E. (1987a). Models for temporal variation in cancer rates II: age-period-cohort models. *Statistics in Medicine*, *6*, 469-481.

Clayton, D., & Schifflers, E. (1987b). Models for temporal variation in cancer rates I: age-period-cohort models. *Statistics in Medicine*, *6*, 449-467.

Doucet, A., Freitas, N., & Gordon, N., eds. (2001). *Sequential Monte Carlo methods in practice*. NY: Springer-Verlag.

Frome, E. L. (1983). The analysis of rates using Poisson regression models. *Biometries*, *39*, 665-674.

Gamerman, D. (1998).Markov chain Monte Carlo for dynamic generalized linear models. *Biometrika*, *85*, 215-227.

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transcation on Pattern Analysis and Machine Intelligence*, *6*, 721-741.

Gamerman, D. (1997). *Markov chain Monte Carlo*. Boca Raton, FL: Chapman & Hall/CRC.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, *57*, 97-109.

Holford, T. R. (1985). An alternative approach to statistical age-period-cohort analysis. *Journal of Chronic Diseases*, *38*, 831-836.

Holford, T. R. (1991). Understanding the effects of age, period and cohort on incidence and mortality rates. *Annual Review of Public Health*, *12*, 425-457.

Kitagawa, G. (1998). Self-organising state space model. *Journal of the American Statistical Association*, *93*, 1203-1215.

Kitagawa, G., & Gersch, W. (1996). *Smoothness Priors Analysis of Time Series*, NY: Springer-Verlag.

Knorr-Held, L., & Rainer, E. (2001) Projection of lung cancer mortality in West Germany: a case study in Bayesian prediction. *Biostatistics*, *2*, 109-129.

Nelder, & Wedderburn, (1972). Generalized linear models. *Journal of the Royal Statistical Society, Series A*, *135*, 370-384.

Osmond, C., & Gardner, M. J. (1982). Age, period and cohort models applied to cancer mortality rates. *Statistics in Medicine*, *1*, 245-259.

Spiegelhalter, D., Thomas, A., & Best, N. (2003). *WinBUGS Version 1.3 User Manual* Medical Research Council Biostatistics Unit, Cambridge, UK.

Tango, T. (1985) Estimation of age, period and cohort effects (in Japanese). *Applied statistics in Japanese*, *14*, 45-49.

West, M., Harrison, P. J., & Migon, H. S. (1985). Dynamic generalized linear models and Bayesian forecasting (with discussion). *Journal of the American Statistical Association*, *80*, 73-83.

West, M., & Harrison, P. J. (1997). *Bayesian Forecasting and Dynamic Models.* (2^{nd} ed.). NY: Springer Verlag.

Appendix

Age Group	Table 6: Stomach cancer - females, count 5 year period centered at												
	1950	1955	1960	1965	1970	1975	1980	1985	1990	1995	2000	2005	2010
15-19	9	9	10	11	23	8	5	5	1	5	4	3	3
20-24	39	43	55	65	86	63	43	21	26	10	9	15	15
25-29	115	119	153	198	232	246	179	123	75	64	48	43	54
30-34	242	289	389	432	380	465	431	293	180	126	98	90	93
35-39	415	495	597	672	651	629	619	561	387	237	162	204	152
40-44	679	784	857	857	937	874	745	693	677	448	306	330	318
45-49	902	982	1152	1139	1194	1152	1003	870	714	831	596	641	489
50-54	1166	1331	1398	1588	1430	1421	1296	1116	938	824	890	997	881
55-59	1490	1675	1779	1836	1956	1658	1634	1423	1180	1072	987	1251	1272
60-64	1891	1951	2124	2217	2339	2274	2031	1903	1583	1432	1291	1533	1657
65-69	2139	2337	2366	2592	2766	2726	2579	2117	1991	1910	1638	1895	1972
70-74	1743	2111	2462	2628	2976	3009	2957	2706	2211	2340	2131	2091	2221
75-79	959	1517	1858	2036	2362	2653	2974	2864	2727	2545	2610	2247	2199
80-84	306	622	964	1076	1354	1596	2076	2388	2606	2976	2820	2123	2033
85-89	81	132	268	343	420	569	845	1301	1693	2194	2685	1557	1502
90-	9	22	45	63	74	123	202	373	585	1054	1592	1503	775

Age Group	Table 7: Lung cancer - males, count 5 year period centered at												
	1950	1955	1960	1965	1970	1975	1980	1985	1990	1995	2000	2005	2010
15-19	9	5	10	6	5	1	5	1	1	5	4	5	5
20-24	4	9	9	18	11	10	4	5	5	5	5	7	7
25-29	6	12	17	21	18	22	23	12	17	14	10	13	14
30-34	3	14	27	34	34	42	49	41	39	49	57	28	29
35-39	10	19	31	68	66	80	115	157	149	102	97	101	67
40-44	20	47	62	99	168	202	207	275	363	287	288	282	247
45-49	43	105	174	190	226	415	450	450	577	757	635	710	627
50-54	85	195	323	407	431	451	933	983	918	1138	1463	1444	1449
55-59	107	250	550	618	826	904	1218	1818	2020	1831	2210	2453	2578
60-64	153	362	669	1078	1249	1534	1856	2321	3655	3760	3352	3787	4310
65-69	175	398	734	1135	1742	2180	2727	3172	4165	6044	5804	4898	5772
70-74	111	306	603	935	1501	2408	3316	4228	4675	6105	8193	6549	6157
75-79	58	125	309	570	854	1636	2784	4018	5022	5703	7326	6881	6576
80-84	9	36	106	198	306	589	1340	2355	3495	4730	5445	4958	5375
85-89	0	6	17	27	63	142	342	849	1466	2206	3146	2580	2720
90-	1	0	5	3	8	15	74	153	313	660	1017	790	874

Age Group	Table 8: Lung cancer - females, count 5 year period centered at												
	1950	1955	1960	1965	1970	1975	1980	1985	1990	1995	2000	2005	2010
15-19	5	5	5	6	5	0	0	1	1	0	0	4	4
20-24	1	5	9	10	6	0	4	5	0	5	0	6	6
25-29	4	8	5	13	10	17	14	8	8	5	10	9	9
30-34	6	20	31	37	34	33	43	18	27	36	30	25	26
35-39	11	31	43	68	49	72	78	96	89	65	75	61	54
40-44	19	50	50	88	107	98	121	151	164	169	146	147	111
45-49	22	67	126	106	166	200	207	220	258	379	261	299	276
50-54	46	85	139	232	211	297	350	374	425	551	667	557	542
55-59	50	106	194	293	337	412	492	587	618	733	806	908	822
60-64	49	120	250	356	482	545	688	843	886	1020	1069	1284	1368
65-69	49	139	251	374	506	671	932	1044	1239	1416	1601	1751	1892
70-74	44	113	203	344	483	728	1083	1353	1562	1767	2018	2239	2384
75-79	19	51	147	270	344	557	1001	1392	1811	2205	2459	2903	2894
80-84	6	16	63	109	194	277	610	1061	1382	2104	2453	3672	3113
85-89	0	5	17	27	43	117	190	480	864	1320	1989	3538	3327
90-	2	0	6	3	10	21	54	125	285	580	1096	2253	2493

Journal of Modern Applied Statistical Methods
November, 2008, Vol. 7, No. 2, 588-596

1538 – 9472/08/$95.00

A Methodology to Improve PCI Use in Industry

Milind A. Phadnis
The University of Alabama at Birmingham

Matthew E. Elam
Texas A&M University-Commerce

This article presents the development of a methodology using decision trees to resolve issues in industry with using process capability indices (PCIs). The methodology forms the structure of a prototype decision support system (PDSS) for PCI selection, calculation, and interpretation. Download instructions for the PDSS are available at http://program.20m.com.

Key words: Process capability index; decision tree; control chart; normality check; decision support system.

Introduction

Process capability may be defined as the ability of a process to achieve a certain objective. Process capability indices (PCIs) have been used for some time to provide a quantitative measure of this ability. Many PCIs have been developed in the literature for different situations encountered by industry. However, industry has not been able to achieve the full benefit from using PCIs for the following reasons:

- Abuse of PCIs by violating their underlying statistical assumptions;
- Lack of practical usage of multivariate PCIs and their interpretations;
- Unavailability of PCIs for data limited (short-run) situations;
- Shortcomings in software packages capable of calculating PCIs; and
- Lack of appropriate usage of PCIs in data with asymmetric specifications.

This article details a methodology for resolving the above mentioned issues. It makes use of a top-down decision making approach to select the appropriate PCI(s) regarding particular kinds of data. It also makes use of the latest theory available in the statistical literature pertaining to the definitions and properties of various PCIs. The methodology was developed by considering the situations in which industry needs PCI results, determining the PCIs available for these situations, and determining the decision-making process for handling these situations simultaneously.

The methodology forms the structure of a prototype decision support system (PDSS) built in order to facilitate easy usage in industry (Phadnis, Elam, Fonseca, Batson, & Adams, 2005). The PDSS analyzes the process data, verifies the statistical assumptions necessary for handling different types of process data, selects the most appropriate PCI(s) depending on the process parameters, calculates the PCI(s), provides a practical interpretation of the PCI(s), and guides the user towards the source of corrective action needed, if any. Visual Basic 6.0 and Microsoft Excel 2002 were used to design the PDSS so that it has a user-friendly graphical interface, portability, and ease of use for industry. The PDSS requires the user to enter only elementary characteristics of the collected process data, the process data itself, and the process's engineering specifications. Instructions for downloading the PDSS are available at http://program.20m.com.

Milind A. Phadnis is a Research Assistant and is pursuing a PhD. in Biostatistics. Email him at phadnismilind@gmail.com. Matthew E. Elam is an Associate Professor of Industrial Engineering at Texas A&M University-Commerce and is an ASQ Certified Quality Engineer. Email him at Matthew_Elam@tamu-commerce.edu.

Methodology

After considering the situations in which industry needs PCI results and studying the properties of the various PCIs available in the literature, the decision tree shown in Figure 1 was constructed as the backbone of the complete structure of the methodology. This decision tree presents a basic overview of the formulations used in constructing the methodology and can be further expanded into various branches and sub-branches. Thus, whenever branching is possible, a series of asterisks "*" is placed in the corresponding block to denote the same, and this particular block has been further expanded in subsequent figures in the Appendix.

As shown in Figure 1, the constructed methodology is equipped to handle the following types of data collected by the user:

- Type 1: univariate sufficient data (total number of observations ≥ 50), which also involves Appendix Figure 2;
- Type 2: univariate short-run data (total number of observations < 50), which also involves Appendix Figure 3; and
- Type 3: multivariate sufficient data (total number of observations ≥ 100), which also involves Appendix Figure 4.

The methodology adopted for selecting and evaluating PCIs is different for each of the above mentioned data types.

Type 1: Univariate Sufficient Data ($\geq$ 50 Observations)

The classifications of sufficient data as that with at least 50 observations, and a short-run situation as that with less than 50 observations, are based on the fact that the statistical properties of the commonly used PCIs do not permit calculation of an index when less than 50 observations are available as noted by Deleryd & Vannman (1998). Univariate data may further be classified into data collected in subgroups and data collected as individual observations. Each of these cases is discussed below. (See Figure 1 and Appendix Figure 2.)

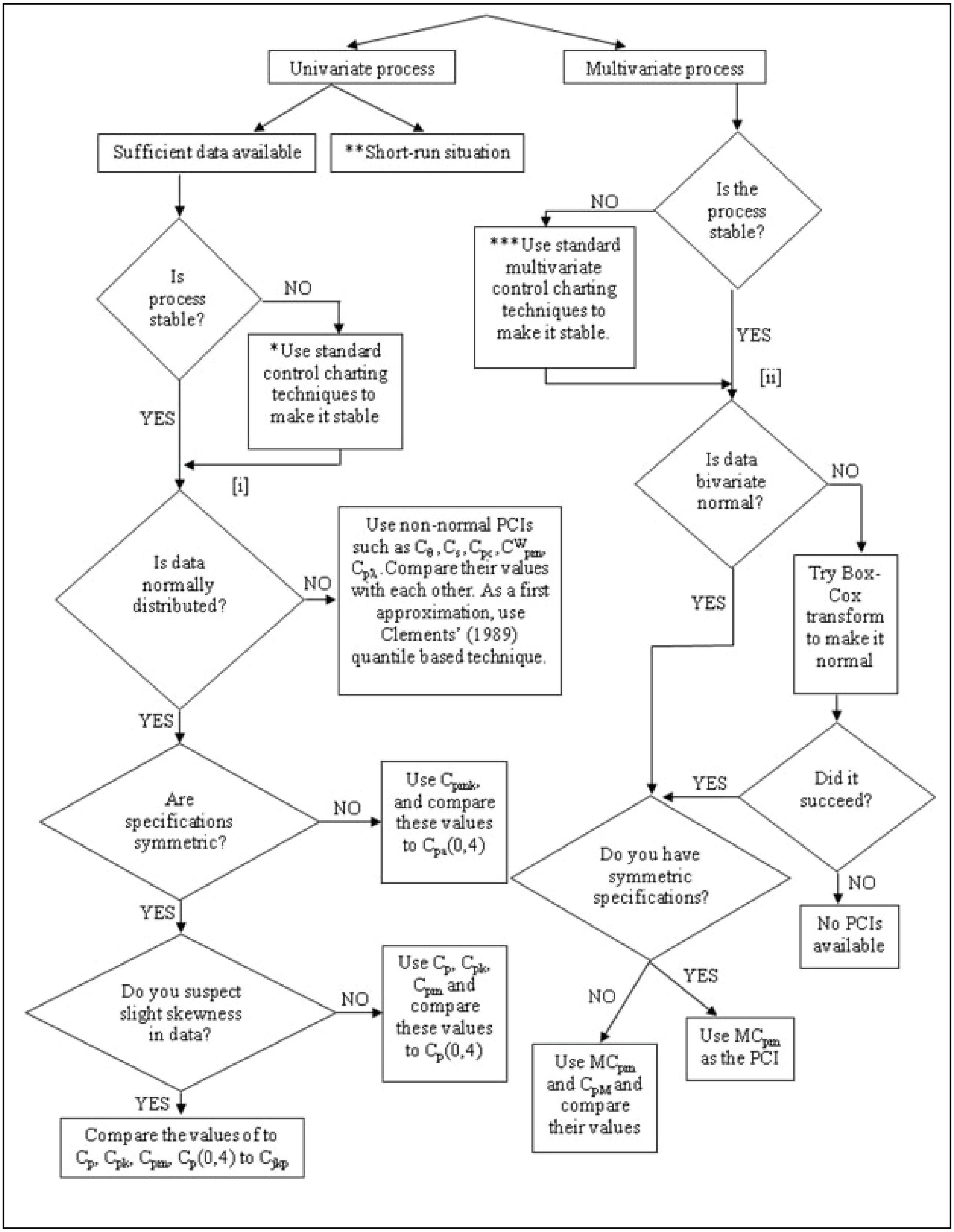
m Subgroups of Equal Size n

The data used to calculate any PCI must come from a stable process (i.e., a process governed by a single probability distribution). Statistical control charting with a delete and revise (D&R) procedure is one way to ensure this. In a D&R procedure, the data used to construct the control charts is also plotted on the charts to retrospectively test if the process was in control while the initial data was being obtained. Any points that plot outside the control limits are deleted and the remaining data is used to construct revised control charts. One of the several variations of the D&R procedure repeats this process until no points plot beyond the control limits, at which time the remaining data would be considered stable or in control.

For $2 \leq n \leq 10$, the usual $\overline{X}$ and R control charts (Montgomery, 2001) are used to perform control charting in order to establish control of the data. For $n > 10$, the usual $\overline{X}$ and S charts are used to perform control charting as the range method for estimating σ loses statistical efficiency for moderate to large subgroup sizes, as mentioned in Montgomery (2001).

Once the above procedure is completed, the remaining data is subjected to a normality check via the Kolmogorov-Smirnov (K-S) test, the procedure for which can be found in any standard statistical text, such as Ebeling (2000). If the normality assumption is satisfied, the decision tree approach makes use of the PCIs as shown in Figure 1 for this situation in order to evaluate process capability. PCIs like C_p, C_{pk}, C_{pm} (Kotz & Lovelace, 1998), and $C_p(0,4)$ (Vannman, 1993) are used when the target value is equal to the midpoint of the specifications (target = midpoint). These values are compared to C_{jkp} (Kotz & Lovelace, 1998) if doubt of slight skewness exists in the data. If not, C_p, C_{pk}, and C_{pm} are compared to C_p (0,4). If the target value is not equal to the midpoint of the specifications, PCIs such as C_{pmk} (Kotz & Lovelace, 1998), C'_{pm} (Perakis & Xekalaki, 2003), and C_{pa} (0,4) (Vannman, 1997) are used to evaluate process capability.

Figure 1: Main Decision Tree


Univariate process
Multivariate process
Sufficient data available
**Short-run situation
Is the process stable?
NO
***Use standard multivariate control charting techniques to make it stable.
YES
[ii]
Is process stable?
NO
*Use standard control charting techniques to make it stable
YES
[i]
Is data bivariate normal?
NO
Is data normally distributed?
NO
Use non-normal PCIs such as C_θ , C_s , C_{pc} , C^W_{pm}, $C_{p\lambda}$. Compare their values with each other. As a first approximation, use Clements' (1989) quantile based technique.
YES
Try Box-Cox transform to make it normal
YES
Are specifications symmetric?
NO
Use C_{pmk}, and compare these values to $C_{pa}(0,4)$
YES
Did it succeed?
Do you have symmetric specifications?
NO
No PCIs available
YES
Do you suspect slight skewness in data?
NO
Use C_p, C_{pk}, C_{pm} and compare these values to $C_p(0,4)$
YES
NO
Use MC_{pm} as the PCI
Use MC_{pm} and C_{pM} and compare their values
YES
Compare the values of to C_p, C_{pk}, C_{pm}, $C_p(0,4)$ to C_{jkp}

If the normality assumption is not satisfied, non-normal PCIs such as C_θ, C_s, C_{pc}, C^W_{pm}, and $C_{p\lambda}$ (Kotz & Lovelace, 1998) are used to evaluate process capability. Because there is no evidence in the statistical literature as to which of these indices is better for a particular situation, the values of these indices are compared with each other as per the methodology.

m Subgroups of Variable Sizes with Maximum Subgroup Size n

In this case, the usual $\overline{X}$ and S control charts for variable subgroup sizes (Montgomery, 2001) are used to perform control charting in order to establish control of the data. Once the process data is stable, the methodology proceeds with normality, symmetric specification, and skewness checks as described previously. The appropriate PCI(s) are then selected.

m Individual Observations

In this case, the usual Individuals (X) and Moving Range (MR) control charts (Montgomery, 2001) are used in order to establish control of the data. The moving range used here is defined by the equation:

$$MR_i = \left|x_i - x_{i-1}\right| \tag{1}$$

where x_i and x_{i-1} are two successive observations collected as individual process data.

The PCI selection procedures for the data remaining after the D&R procedure are performed in the same manner discussed above. However, it is necessary to ascertain that individual observations obtained are normally distributed even before control limits for these charts are calculated, because even for moderate departures from normality the use of the X and MR charts is not appropriate. Hence, if the data collected is not normally distributed, it should be transformed to another variable that is approximately normally distributed (this was not an issue in previous descriptions because the Central Limit Theorem could be invoked subgrouped data).

If the normality assumption is satisfied, the methodology suggests the continuation of the PCI selection procedure as mentioned earlier. However, if the normality assumption is not satisfied, the data should undergo a Box-Cox transformation of the type in equation:

$$Y = \left(X^{\lambda} - 1\right)/\lambda \tag{2}$$

where the optimal value of λ is determined by an iterative procedure using the following steps as mentioned by Johnson & Wichern (2003):

1. Construct a normal probability plot of the individual observations and determine the correlation coefficient, r.
2. For different values of λ ranging from -2 to 2, determine the value of r. Determine r_{max}, the maximum value of r among all the values calculated.
3. The value of λ which gives r_{max} is used for the transformation in accordance with the following values of λ: 2 (square transformation), 1 (use the original data), 1/2 (square root transformation), 0 (logarithm transformation), -1/2 (reciprocal square root transformation), and -1 (reciprocal transformation).

The transformed data is again checked for normality. If the transformed data is found to be normally distributed, the PCI selection procedure is conducted using the methods explained previously. However, if that is not the case, the data is considered to be strongly non-normal. As a result, control charting cannot be done and PCIs cannot be selected.

Type 2: Univariate Short Run Data (< 50 Observations)

In this case, the data may have been collected either in m subgroups each of size n or as individual observations. The following procedure is adopted for evaluating PCIs in this situation. (See Figure 1 and Appendix Figure 3.)

m Subgroups of Equal Size n

The control charting procedure adopted in this case for establishing control of the data is the short run $\overline{X}$ and S control charts from Elam & Case (2005a, 2005b). Once this procedure is completed, the remaining data is checked for normality via the Kolmogorov-Smirnov (K-S) test and the correlation coefficient test (Johnson

& Wichern, 2003) for normality at the specified level of significance. The underlying reason for using both tests is that, for a small number of observations, the correlation coefficient test is considered to be a very powerful test for normality. If the remaining data are found to be normally distributed, short-run PCIs such as C_{sp}, C_{spk}, and C_{spm} are used to evaluate process capability as mentioned by Balamurali (2003). According to this procedure, the remaining data are bootstrapped into 1,000 resamples, each of which are equal to the total number of observations in the remaining data. These are then used to calculate the short-run PCIs, and the standard bootstrap method is used to construct a 95% confidence interval for each index.

If the remaining data is found to be non-normal at the specified level of significance, the Box-Cox transformation is used to transform the original non-normal data to normal data. If the transformation is successful (the transformed data is subjected to the K-S test and the correlation coefficient test for normality), short-run PCIs as discussed above are evaluated. If the transformation is unsuccessful, the short-run PCIs are still evaluated. It should be noted, however, that the results obtained from PCI calculations may be inaccurate, as for a non-normal process, the coverage percentage points for 95% confidence limits might indicate a high proportion of values that are significantly different from the expected value of the index at the specified level of significance.

m Individual Observations

The control charting procedure adopted in this case for establishing control of the data is the short run X and MR control charts from Elam & Case (2008, 2006). Once this procedure is completed, the remaining data is subjected to the same procedures as related earlier in the m Individual Observations, starting with the normality check. The short-run PCIs discussed previously are used to evaluate process capability.

Type 3. Multivariate Sufficient Data (Observations ≥ 50) (See Figure 1 and Appendix Figure 4.)

m Subgroups of Size n

In this case, the usual Hotelling T^2 control chart (Montgomery, 2001) is used along with the usual bivariate control chart for dispersion (Johnson & Wichern, 2003) to conduct control charting for establishing control of the data. The remaining data are subjected to a bivariate normality check because the PCIs to be calculated are strictly based on the assumption of bivariate normality. This bivariate normality check is performed by:

$$(X-\mu)' S^{-1}(X-\mu) \leq \chi_2^2(0.5) \qquad (3)$$

The average μ and variance-covariance matrix S are for the remaining data grouped together. If approximately 50% of the remaining data grouped together satisfies equation (3) the data is considered to be bivariate normal as per Johnson & Wichern (2003).

If the bivariate normality assumption is satisfied, the bivariate PCIs C_{pM} and MC_{pm} (for bivariate process data with asymmetric specifications) and MC_{pm} (for bivariate process data with symmetric specifications) are evaluated as shown in Wang, Hubele, Lawrence, Miskulin & Shahriari (2000). If the bivariate normality assumption is not satisfied, the Box-Cox transformation of the data is performed. The optimal value of λ is the one that maximizes the following equation:

$$l(\lambda) = (-n/2)\ln\left((1/n)\sum_{j=1}^{n}\left[x^{(\lambda)}{}_{j} - \overline{x}_{j}^{(\lambda]}\right]^{2}\right) +$$

$$(\lambda - 1)n\sum_{j=1}^{n}\ln\left[x_{j}\right] \qquad (4)$$

where n is the total number of filtered observations, $x^{(\lambda)} = (x^{\lambda} - 1)/\lambda$ if $\lambda \neq 0$, and $x^{(\lambda)} = \ln(x)$ if $\lambda = 0$. If, after the above procedure, bivariate normality is not satisfied, then it is not possible to calculate a bivariate PCI.

m Individual Observations

In the case of individual observations of bivariate data, the usual T^2 control chart for individual observations (Johnson & Wichern, 2003) is used to establish control of the data.

Once this has been accomplished, the bivariate data is subject to a bivariate normality check in accordance with the procedure discussed herein. The PCI selection procedure continues similarly to the case for bivariate data collected in subgroups.

Results and Conclusion

The methodology used in formulating a decision tree approach in order to aid industry practitioners regarding the selection of a PCI has been discussed; the main advantage of this methodology that it offers a structured approach for programming the same into a decision support system for easy usage in industry. By incorporating such a methodology into a computer program with the capability to select, calculate, and interpret the appropriate PCI(s) for the situation under consideration, the problems industry experiences with PCIs, as noted in the Introduction, are alleviated. As all statistical assumptions have been taken into consideration while developing this methodology, a robust structure to the application of PCI usage in industry has been accomplished.

References

Balamurali, S. (2003). Bootstrap confidence limits for short-run capability indices. *Quality Engineering*, *15(4)*, 643-648.

Clements, J. (1989). Process capability calculations for non-normal distributions. *Quality Progress*, September, 95-100.

Deleryd, M. & Vannman, K. (1998). Process capability studies for short production runs. *International Journal of Reliability, Quality, and Safety Engineering*, *5(4)*, 383-401.

Ebeling, C. E. (2000). *Reliability and maintainability engineering*. New Delhi: Tata McGraw-Hill.

Elam, M. E. & Case, K. E. (2005a). Two-stage short-run ($\bar{X},s$) control charts. *Quality Engineering*, *17(1)*, 95-107.

Elam, M. E. & Case, K. E. (2005b). A computer program to calculate two-stage short-run control chart factors for ($\bar{X},s$) charts. *Quality Engineering*, *17(2)*, 259-277.

Elam, M. E. & Case, K. E. (2008). Two-stage short-run (X,MR) control charts. *Journal of Modern Applied Statistical Methods, 7(1)*, 275-285.

Elam, M. E. & Case, K. E. (2006). A computer program to calculate two-stage short-run control chart factors for (X,MR) charts. *Journal of Statistical Software*, *15(11)*, http://www.jstatsoft.org/.

Johnson, R. A. & Wichern, D. W. (2003). *Applied multivariate statistical analysis*. New York: Pearson Education.

Kotz, S. & Lovelace, C. (1998). *Process capability indices in theory and practice*. New York: Oxford University Press Inc.

Montgomery, D. C. (2001). *Introduction to statistical quality control*. 4th ed. New York: John Wiley and Sons.

Perakis, M. & Xekalaki, E. (2003). On a process capability index for asymmetric specifications. *Communication in Statistics – Theory and Methods*, *32(7)*, 1459-1492.

Phadnis, M. A., Elam, M. E., Fonseca, D. J., Batson, R. G., & Adams, B. M. (2005). A prototype DSS for PCI selection, calculation, and interpretation. *Proceedings of International Conference, Institute of Industrial Engineers*, Atlanta, USA.

Vannman, K. (1993). A unified approach to capability indices. *Statistica Sinica*, *22(2)*, 537-560.

Vannman, K. (1997). A general class of capability indices in the case of asymmetric tolerances. *Communication in Statistics – Theory and Methods*, *26(10)*, 2381-2396.

Wang, F. K., Hubele, N. F., Lawrence, F. P., Miskulin, J. D., & Shahriari, H. (2000). Comparison of three multivariate process capability indices. *Journal of Quality Technology*, *32(3)*, 263-275.

Figure 2: Decision Tree for Univariate Data

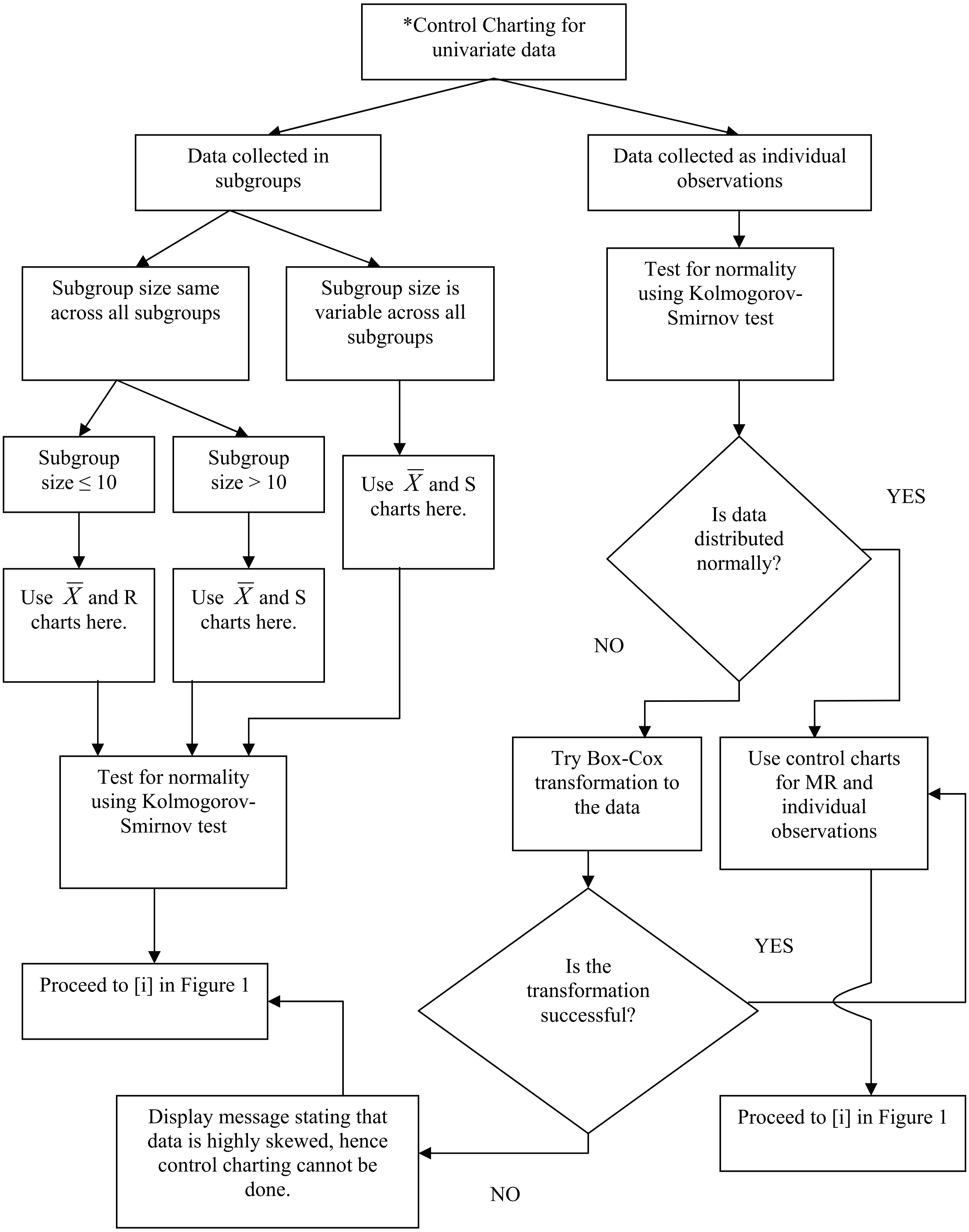

*Control Charting for univariate data
Data collected in subgroups
Data collected as individual observations
Subgroup size same across all subgroups
Subgroup size is variable across all subgroups
Test for normality using Kolmogorov-Smirnov test
Subgroup size ≤ 10
Subgroup size > 10
Use $\overline{X}$ and S charts here.
Is data distributed normally?
YES
Use $\overline{X}$ and R charts here.
Use $\overline{X}$ and S charts here.
NO
Try Box-Cox transformation to the data
Use control charts for MR and individual observations
Test for normality using Kolmogorov-Smirnov test
YES
Proceed to [i] in Figure 1
Is the transformation successful?
Display message stating that data is highly skewed, hence control charting cannot be done.
Proceed to [i] in Figure 1
NO

Figure 3: Decision Tree for the Short-Run Situation

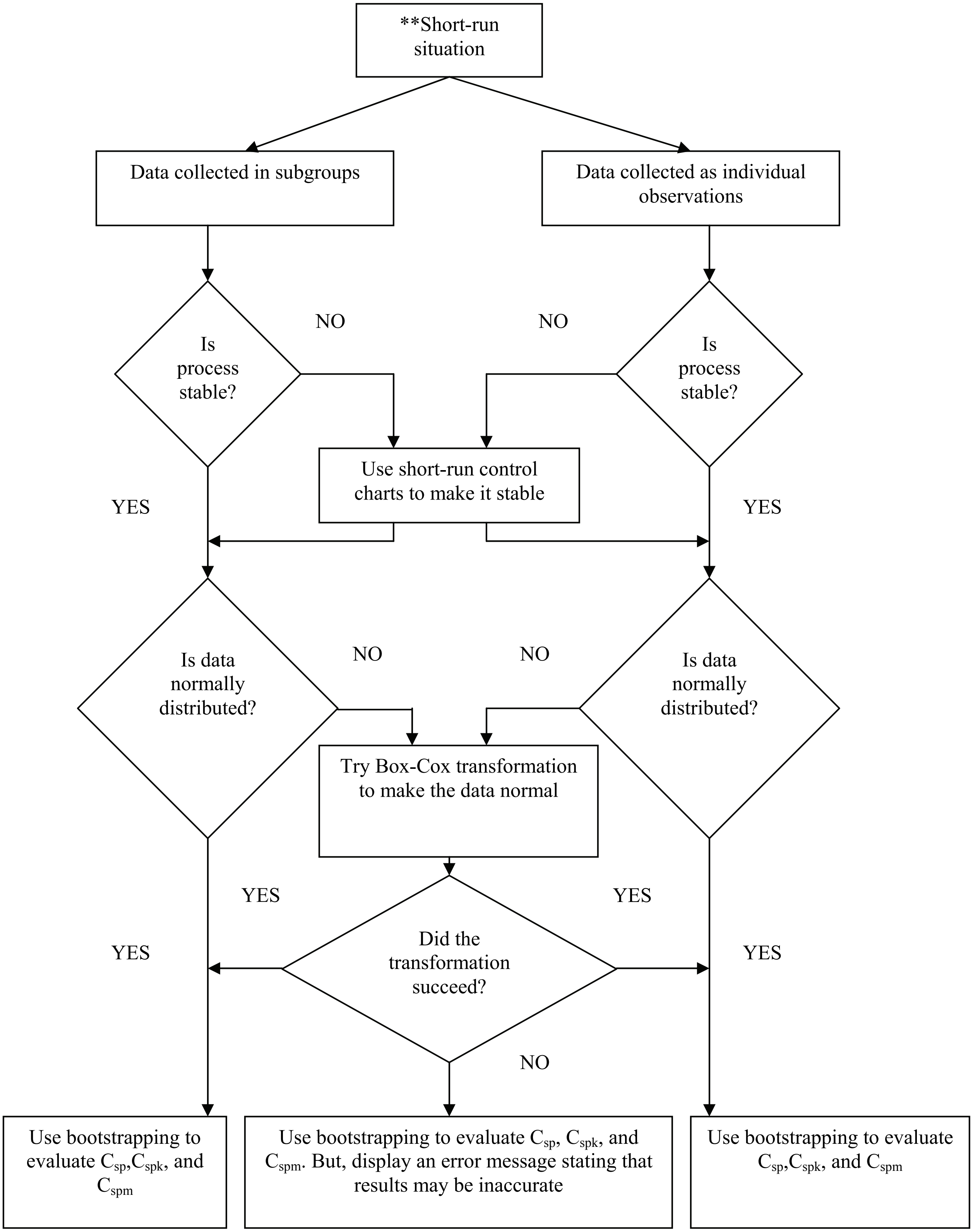
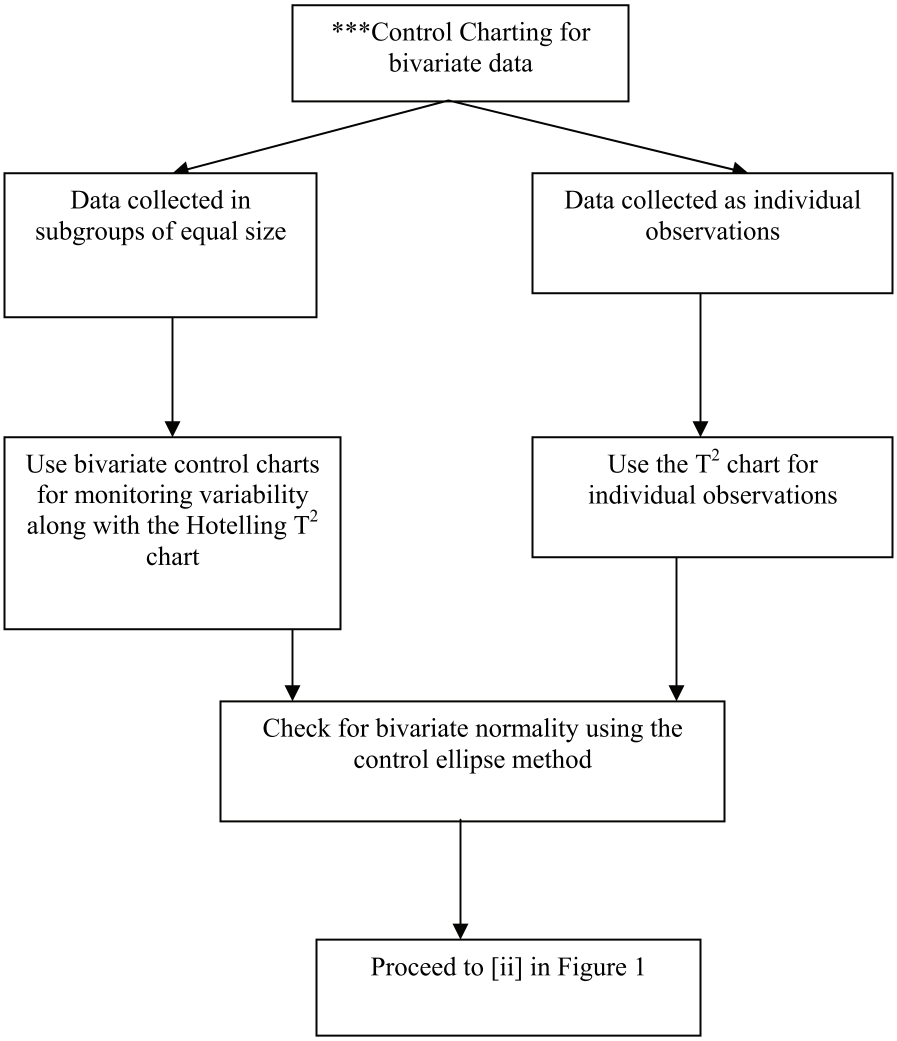

**Short-run situation
Data collected in subgroups
Data collected as individual observations
Is process stable?
NO
NO
Is process stable?
Use short-run control charts to make it stable
YES
YES
Is data normally distributed?
NO
NO
Is data normally distributed?
Try Box-Cox transformation to make the data normal
YES
YES
YES
Did the transformation succeed?
YES
NO
Use bootstrapping to evaluate C_{sp},C_{spk}, and C_{spm}
Use bootstrapping to evaluate C_{sp}, C_{spk}, and C_{spm}. But, display an error message stating that results may be inaccurate
Use bootstrapping to evaluate C_{sp},C_{spk}, and C_{spm}

Figure 4: Decision Tree for Bivariate Data


***Control Charting for bivariate data
Data collected in subgroups of equal size
Data collected as individual observations
Use bivariate control charts for monitoring variability along with the Hotelling T^2 chart
Use the T^2 chart for individual observations
Check for bivariate normality using the control ellipse method
Proceed to [ii] in Figure 1

Journal of Modern Applied Statistical Methods
November, 2008, Vol. 7, No. 2, 597-606

1538 – 9472/08/$95.00

The Multinomial Regression Modeling of the Cause-of-Death Mortality of the Oldest Old in the U.S.

Dudley L. Poston, Jr.
Texas A&M University

Hosik Min
University of Hawaii

The statistical modeling of the causes of death of the oldest old (persons aged 80 and over) in the U.S. in 2001 was conducted in this article. Data were analyzed using a multinomial logistic regression model (MNLM) because multiple causes of death are coded on death certificates and the codes are nominal. The percentage distribution of the 10 major causes of death among the oldest old was first examined; we next estimated a multinomial logistic regression equation to predict the likelihood of elders dying of one of the causes of death compared to dying of an "other cause." The independent variables used in the equation were age, sex, race, Hispanic origin, marital status, education, and metropolitan/non-metropolitan residence. Our analysis provides insights into the cause of death structure and dynamics of the oldest old in the U.S., demonstrates that MNLM is an appropriate statistical model when the dependent variable has nominal outcomes, and shows the statistical interpretation for complex results provided by MNLM.

Key words: multinomial regression, nominal outcome, logit, log odds, cause of death, mortality, oldest old, elderly, demography.

Introduction

Demographers use multinomial logistic regression models when a dependent variable has more than two nominal categories. The choice is between a logistic model and a probit model, because the nominal categories of a variable are assumed to be unordered and more than two. If the outcome is dichotomous, logistic models are preferred. If the outcome is ordered, ordered or probit models are most appropriate (Long & Freese, 2003).

Background information about the causes of death of the U.S. is helpful in understanding the logic of the data analysis. The National Center for Health Statistics specifies the causes of death based on ICD-10 system (the10th version of International Classification of Disease System). The causes are numerous and nominal. Although there is a ranking system for the causes of the death, the rank does not mean a certain cause is superior over another; they are ranked based on incidence alone. One cannot say that a death due to a certain disease is more meaningful than the others. This article examines the top 10 causes of death for persons aged 80 and older in the U.S., as well as the likelihood of dying of a particular cause versus other causes.

The best-fitting statistical model for handling a nominal outcome is the multinomial logistic regression model (MNLM). It is not always easy to use MNLM, because MNLM has many parameters and the dependent variables have more than two categories. In addition, these parameters sometimes lead to complex results, which are often difficult to interpret. Poston & Min (2004) employed multinomial logit models for South Korean and American decedents and found that various sociodemographic factors influenced dying of specific causes of death compared to others.

This article focuses mainly on methodological issues, namely, the appropriateness of multinomial logit models for

Dudley L. Poston, Jr. is a Professor in the Department of Sociology. Email: d-poston@tamu.edu. Hosik Min is a Faculty Assistant Specialist in the Center on The Family at the Mānoa campus. Email: hosik@hawaii.edu.

studying causes of death, and the interpretation of the results of such investigations. Thus, the goals were to examine the likelihood of dying of a certain cause versus other causes for the oldest old (age 80 years and over) in the United States and to offer an easily understandable interpretation of MNLM. This is a particularly important concern, given the expected increases in the numbers of persons aged 80 and over in the U.S. in the next few decades. In the year 2000, the U.S. had a population of over 13 million oldest old people, 1.5% of the total U.S. population (Hetzel & Smith, 2001). Projections are for 24 million oldest old population in 2050, over 6% of the total U.S. population (Census Bureau, 2000a; 2000b).

Given such tremendous increases predicted for the population of the oldest old in the next few decades, an analysis of cause-of-death mortality in the current American oldest old population is particularly relevant. A study of the dynamics of current causes of death should suggest patterns of mortality that may be anticipated in the U.S. as the numbers of oldest old increase by 200% over the next five decades.

Methodology

The data used in this article were obtained from death certificates filed in the U.S. The data were taken from 963,768 death certificates filed in 2001 for decedents age 80 and over (National Center for Health Statistics, 2003). The top 10 major causes of death for the oldest old Americans were heart disease; malignant neoplasms; cerebrovascular disease; chronic respiratory disease; Alzheimer's; influenza and pneumonia; diabetes; nephritis, nephrotic syndrome, and nephrosis; accidents; and septicemia.

Estimation from a multinomial logistic regression, which predicts the likelihood of dying among oldest old American decedents of one of the major causes of death, compared to dying of an other cause, provides the main focus of this research. The independent variables used in the multinomial logistic equations were age, sex, race, Hispanic origin, marital status, education, and metropolitan/non-metropolitan residence.

With respect to the statistical method used herein, consider as an example only three major causes of death, and a residual category of all other causes. Thus, think of the multinomial logistic regression equation as providing an estimate for each of the independent variables and a set of four logit coefficients corresponding to each of the four categories of the dependent variable as follows (Stata Corporation, 2003, Vol. 2, p. 506-507):

$$\Pr(Y=1) = \frac{e^{Xb(1)}}{e^{Xb(1)} + e^{Xb(2)} + e^{Xb(3)} + e^{Xb(4)}} \quad (1)$$

$$\Pr(Y=2) = \frac{e^{Xb(2)}}{e^{Xb(1)} + e^{Xb(2)} + e^{Xb(3)} + e^{Xb(4)}} \quad (2)$$

$$\Pr(Y=3) = \frac{e^{Xb(3)}}{e^{Xb(1)} + e^{Xb(2)} + e^{Xb(3)} + e^{Xb(4)}} \quad (3)$$

$$\Pr(Y=4) = \frac{e^{Xb(4)}}{e^{Xb(1)} + e^{Xb(2)} + e^{Xb(3)} + e^{Xb(4)}} \quad (4)$$

The multinomial model cannot be identified unless one of the logits in each set is set to zero. Strictly speaking, it does not matter which one is set to zero. If we set $\mathbf{b}^{(1)}$ to zero, then the remaining logit coefficients, $\mathbf{b}^{(2)}$, $\mathbf{b}^{(3)}$ and $\mathbf{b}^{(4)}$, will represent the change relative to the **y=1** category. In the example of cause-of-death mortality, $\mathbf{b}^{(1)}$ will be the logit referring to deaths due to all other causes, and $\mathbf{b}^{(2)}$, $\mathbf{b}^{(3)}$ and $\mathbf{b}^{(4)}$ will refer to deaths due to the three main causes being analyzed. Regarding the logit set to zero, its value becomes 1 because $\mathbf{e^0 = 1}$.

If $\mathbf{b}^{(1)}$ is set to zero, the equations for the four probabilities become:

$$\Pr(y=1) = \frac{1}{1 + e^{Xb(2)} + e^{Xb(3)} + e^{Xb(4)}} \quad (5)$$

$$\Pr(Y=2) = \frac{e^{Xb(2)}}{1 + e^{Xb(2)} + e^{Xb(3)} + e^{Xb(4)}} \quad (6)$$

$$\Pr(Y=3) = \frac{e^{Xb(3)}}{1+e^{Xb(2)}+e^{Xb(3)}+e^{Xb(4)}} \quad (7)$$

$$\Pr(Y=4) = \frac{e^{Xb(4)}}{1+e^{Xb(2)}+e^{Xb(3)}+e^{Xb(4)}} \quad (8)$$

In the actual multinomial logistic regression model, the top 10 causes of death and an 11th residual cause, i.e., dying of other causes, were used. Thus for each of the independent variables, 10 logits were formed from the contrasts of 10 non-redundant category pairs of the dependent variable modeling the logarithmic odds of dying of one of the 10 major causes of death versus dying of other causes. The estimated parameters are logit coefficients indicating the independent log odds of each independent variable being in the dependent variable category of interest, versus being in the base (or contrast) category of the dependent variable. The multinomial model was estimated using maximum likelihood procedures.

Separate logit coefficients were estimated for each independent variable for each of the dependent variable categories, excluding the outcome reference category. Thus the total number of parameters to be estimated was K × (J -1), where K is the number of independent variables and J is the number of categories in the dependent variable. As shown below, there were 14 independent variables and the dependent variable consisted of 10 specific causes of death and a residual category of other causes. Thus the multinomial logistic equation estimated 14 × (10-1) logits, for a total of 126 coefficients. The "biggest challenge in using the multinomial logistic regression model was that the model includes a lot of parameters, and it was easy to be overwhelmed by the complexity of the results" (Long & Freese, 2003, p. 189).

Results

In 2001, there were 963,768 death certificates filed for Americans age 80 and older in the U.S. As Table 1 shows, around 82% died from 10 main causes of death, as follows: heart disease, 36.7%; malignant neoplasms, 14.9%; cerebrovascular disease, 9.5%; chronic respiratory disease, 5.0%; Alzheimer's, 4.2%; influenza and pneumonia, 4.0%; diabetes, 2.5%; nephritis, nephrotic syndrome, and nephrosis, 1.9%; accidents, 1.8%, and septicemia, 1.4%. Around 18% of American oldest old decedents died of other causes.

The U.S. has low mortality levels in the general population as well as among the oldest old. The percentage of elderly decedents in 2001 was 80% of total deaths because America has completed the epidemiological transition (Omran, 1971; 1981). Omran's epidemiological transition describes and explains variations in countries' experiences of mortality changes through time. For example, at the first stage, mortality is high and fluctuating, precluding sustained population growth. At the second stage, mortality declines progressively, as epidemics decrease in frequency and magnitude, and life expectancy increases. As the gap between birth and death rates widens, rapid population growth ensues. In the third stage, mortality continues to decline and eventually approaches stability. Thus, mortality is low, life expectancy is high (over 70 years for both males and females), and deaths mainly occur from degenerative and man-made diseases (Olshansky & Ault, 1986).

Multinomial Logistic Regression Results

We have shown that in 2001 there were 10 principal causes of death responsible for more than 82% of the deaths of U.S. oldest old. The remaining 18% of the oldest old decedents died for some other reason, treated here as a residual category of other causes.

Seven major classes of variables were used to predict cause-of-death mortality. They are age, sex, race, Hispanic origin, marital status, education and metro/non-metropolitan residence. From these seven classes of independent variables, we have developed 14 dummy variables, which were scored 1 if yes, as follows: 1) Age 90-99, and 2) Age 100+ (with Age 80-89 used as the reference variable); 3) Female; 4) Whites; and 5) Blacks (with Other races used as the reference group); 6) Hispanic origin; 7) Married, 8) Divorced, and 9) Widowed (with Never Married used as the reference variable); 10) Elementary School, 11) Junior High School, 12) High School, and 13)

Table 1: Top 10 Causes of Death among the Oldest Old (80+): U.S., 2001

Cause of Death	Number of Decedents	Percent
Heart Disease	353,315	36.66
Malignant Neoplasms	143,915	14.93
Cerebrovascular Disease	91,848	9.53
Chronic Respiratory Disease	48,419	5.02
Alzheimer's	40,381	4.19
Influenza & Pneumonia	38,254	3.97
Diabetes	23,679	2.46
Nephritis, Nephrotic Syndrome & Nephrosis	18,200	1.89
Accidents	17,559	1.82
Septicemia	13,054	1.35
Other Causes	175,144	18.17
TOTAL	963,768	100.00

Note: National Center for Health Statistics, *2001 Multiple Cause-of-Death File, NCHS CD-ROM, Series 20, No. 10H*. Hyattsville, Maryland: National Center for Health Statistics, 2003.

College or More (with Illiterate used as the reference variable); and 14) Metropolitan Residence. These 14 dummy variables are the explanatory (X) variables used in the multinomial logistic regression.

In Table 2 frequency distributions for these 14 independent variables for the 963,768 American oldest old who died in 2001 are presented. The majority of these decedents were aged 80-89 (almost 69%). Almost thirty percent were aged 90-99, and over 1% were aged 100 and over.

Almost two thirds were females (62%). Whites were the majority among the American oldest old (92%). Almost 7% were African Americans. Only 1.4% of oldest old Americans were other races. The majority of the oldest old Americans were of non-Hispanic origin (97%). According to Rogers et al. (2000), non-Hispanic whites have lower morality risks than other groups, except Asian Americans. Asian American mortality is generally lower than that of non-Hispanic whites. Young Hispanic adults also have higher odds of mortality compared to non-Hispanic whites. African Americans suffer from the highest mortality risks compared to the other groups. Widowed is the leading category in these decedents' marital status (63%), and married is next (27%). Almost 5% of the oldest old American decedents were divorced, and another 5% were never married. Regarding education, less than 1% was illiterate and over two thirds had high school or more education.

Most of these American oldest old lived in metropolitan areas at the time of death (76%). In the multinomial logistic regression model, therefore, 10 logits (one for each of the independent variables) are estimated for each of the 10 causes of death, modeling the log odds of dying of a major cause versus dying of other causes. Each logit coefficient will represent the independent log odds of the independent variable of being in the dependent variable category of interest, versus being in the base (or contrast) category of the dependent variable. In the multinomial logistic equation we will estimate 14 × (10-1) logits, for a total of 126 coefficients.

Table 3 presents the results of the multinomial logistic regression analysis for America's oldest old who died in 2001. Ten logit coefficients were estimated for each of the

10 principal causes of death. Each logit coefficient represents the independent log odds of the independent variable of being in the particular cause-of-death category, versus being in the contrast category of the dependent variable of other causes. If no relationship exists, the coefficient would be 0. Negative coefficients indicate a negative association, that is, negative chances or log odds of being in the dependent variable category of interest, and positive coefficients indicate positive chances.

Table 2: Frequency Distributions for Explanatory Variables: The Oldest Old (80+) Decedents, U.S., 2001

Variable		Frequency	Percent
Age	80-89	661,738	68.66
	90-99	285,185	29.59
	100+	16,845	1.75
	TOTAL	963,768	100.00
Sex	Male	362,292	37.59
	Female	601,476	62.41
	TOTAL	963,768	100.00
Race	White	883,639	91.69
	Black	66,393	6.89
	Others	13,736	1.43
	TOTAL	963,768	100.00
Hispanic	Hispanic	27,969	2.90
	Non-Hispanic	935,799	97.10
	TOTAL	963,768	100.00
Marital Status	Never Married	50,273	5.22
	Married	259,311	26.91
	Divorced	44,857	4.65
	Widowed	609,327	63.22
	TOTAL	963,768	100.00
Education	Illiterate	6,384	0.66
	Elementary	74,942	7.78
	Junior High	198,781	20.63
	High	455,744	47.29
	College+	227,917	23.65
	TOTAL	963,768	100.00
Residence	Non-Metro	230,239	23.89
	Metro	733,529	76.11
	TOTAL	963,768	100.00

The second column of Table 3 presents the results of the log odds of dying of malignant neoplasms versus dying of other causes (the residual category). Malignant neoplasms were the second major cause of death among American oldest old who died in 2001, with 143,915 deaths from this cause. Table 1 also shows that the residual category of other causes was the cause of death of 175,144 American oldest old in 2001.

Ten independent variables were used to estimate the log odds of dying of malignant neoplasms versus dying of other causes. The first logit coefficient shown in the second column of Table 3 is -0.73 for age 90-99. This means that for decedents who were age 90-99 compared to those who were 80-89, there is a decrease of 0.73 in the log odds of dying of malignant neoplasms compared to dying of other causes. The second logit coefficient is -1.65; this means that for American decedents age 100 or more, compared to those who were 80-89, there is a decrease of 1.65 in their log odds of dying of malignant neoplasms compared to dying of other causes. Hence, the older the decedent, the less the log odds that the person died of malignant neoplasms compared to other causes. Each logit coefficient reflects the effect of the particular independent variable on the dependent variable, controlling for the effects on the dependent variable of the other independent variables in the multinomial logistic regression equation. Thus, the effects of age on malignant neoplasms mortality are independent of the effects on malignant neoplasms of sex, marital status, race, Hispanic origin, educational attainment, and residence.

The estimated parameter effects are interpreted straightforwardly when converted into odds ratios, which is done by exponentiating the coefficients. Odds ratios in the multinomial logistic regression equation are typically referred to as relative risk ratios. This is the relative risk, or the odds, of being in the dependent variable category of interest and not being in the contrast category of the dependent variable for the dummy independent variable

Table 3: Logit Coefficients from Multinomial Logistic Regression of Dying of 1 of 11 Causes vs. Dying of other causes, on Selected Social and Demographic Factors: Oldest Old Decedents, U.S., 2001

Independent Variables	Cause of Death				
	1	2	3	4	5
	Heart Disease	Malignant Neoplasms	Cerebrovascular Disease	Chronic Respiratory Disease	Alzheimer's
Age 80-89	Reference	Reference	Reference	Reference	Reference
Age 90-99	0.03***	-0.73***	-0.06***	-0.63***	0.26***
Age 100+	-0.00	-1.65***	-0.43***	-1.25***	0.43***
Female	-0.08***	-0.33***	0.22***	-0.38***	-0.23***
Other races	Reference	Reference	Reference	Reference	Reference
White	-0.01	-0.14***	-0.37***	0.10*	-0.50***
Black	0.06*	0.10**	-0.26***	-0.41***	-0.55***
Hispanic	0.10***	0.06*	0.01	-0.24***	0.27***
Never Married	Reference	Reference	Reference	Reference	Reference
Married	0.01	0.31***	0.21***	0.08**	-0.22***
Divorced	0.01	0.17***	0.12***	0.45***	-0.05
Widowed	0.04**	0.14***	0.15***	0.22***	-0.10***
Illiterate	Reference	Reference	Reference	Reference	Reference
Elementary	0.08*	0.17***	-0.00	0.03	-0.06
Junior High	0.07*	0.19***	0.02	-0.04	-0.07
High School	0.05	0.25***	-0.02	-0.00	-0.10
College or More	-0.05	0.24***	0.03	-0.19**	-0.12
Metro Residence	0.03***	0.03**	-0.08***	0.01	-0.03*
Intercept	0.65***	-0.12***	-0.51***	-1.08***	-0.75***
N	353,315	143,925	91,848	48,419	40,381

*p < .05; ** p < .01; ***p < .001; Reference group is dying of other causes (N = 175,144)

versus the reference category (Stata Corporation, 2003, p. 510-511).

The odds ratio for Age 90-99 is $e^{-0.73}$ = 0.48, which means that the odds of persons aged 90-99, compared to those age 80-89, dying of malignant neoplasms versus dying of other causes may be multiplied by 0.48, which means they decrease. The percentage amount of change may be determined in the odds by subtracting 1 from the odds ratio and multiplying the difference by 100: (0.48 -1) * 100 = -0.52. This indicates that the odds of dying of malignant neoplasms versus dying of other causes are 52% less for persons aged 90-99 compared to those aged 80-89. In contrast, the odds of dying of malignant neoplasms versus dying of other causes are 81% less for persons aged 100 and over compared to those aged 80-89, that is,

$$(e^{-1.65} -1) * 100 = -0.81.$$

Logits on Each of the 10 Causes of Death

The pattern of the effects shown for malignant neoplasms is one in which the log odds of 90-99 year-old decedents dying of malignant neoplasms versus dying of other causes are less than those of 80-89 year-old decedents, and the log odds of 100+ year-old decedents compared to those of 80-89 year-old decedents are more negative.

This pattern of increasingly negative log odds for decedents 100 years or older compared

Table 3 Continued. Logit Coefficients from Multinomial Logistic Regression of Dying of 1 of 11 Causes vs. Dying of other causes, on Selected Social and Demographic Factors: Oldest Old Decedents, U.S., 2001

Independent Variables	Cause of Death				
	6	7	8	9	10
	Influenza & Pneumonia	Diabetes	Nephritis, Nephrotic Syndrome & Nephrosis	Accidents	Septicemia
Age 80-89	Reference	Reference	Reference	Reference	Reference
Age 90-99	-0.57***	0.14***	-0.11***	-0.10***	-0.21***
Age 100+	-1.33***	-0.20***	-0.40***	-0.27***	-0.67***
Female	0.02	0.47***	-0.36***	-0.28***	-0.08***
Other races	Reference	Reference	Reference	Reference	Reference
White	-0.49***	0.53***	-0.19**	-0.28***	0.20*
Black	0.14**	0.29***	0.38***	-0.53***	0.86***
Hispanic	0.74***	-0.22***	0.11*	-0.06	0.07
Never Married	Reference	Reference	Reference	Reference	Reference
Married	0.23***	0.37***	0.04	-0.05	-0.19***
Divorced	0.18***	0.16***	0.01	-0.07	-0.13*
Widowed	0.25***	0.17***	0.06	-0.07*	-0.14***
Illiterate	Reference	Reference	Reference	Reference	Reference
Elementary	0.00	0.14	0.01	0.21	-0.01
Junior High	-0.03	0.17*	0.02	0.26*	-0.03
High School	-0.18*	0.20*	-0.08	0.27*	-0.06
College or More	-0.39***	0.28**	-0.28**	0.33**	-0.25*
Metro Residence	-0.14***	-0.00	-0.13***	-0.25***	0.10***
Intercept	-1.40***	-2.75***	-1.72***	-1.83***	-2.57***
N	38,254	23,679	18,200	17,559	13,054
Model chi-square (df) = 39905.58*** (140); Pseudo R^2 = .01					
*p < .05; **p < .01; ***p < .001; Reference group is Dying of other causes (N = 175,144)					

to 80-89 year-old decedents over the log odds of 90-99 year-old decedents compared to 80-89 year-old decedents is found in most of the other major causes of death except heart disease, Alzheimer's, and diabetes. For instance, the logit coefficients for 90-99 and 100+ year-old decedents for cerebrovascular disease are -0.06 and -0.43; chronic respiratory disease, -0.63 and 1.25; influenza and pneumonia, -0.57 and -1.33; nephritis, nephrotic syndrome and nephrosis -0.11 and -0.40; accidents -0.10 and -0.27; and for septicemia -0.21 and -0.67, respectively. However, this association does not hold for heart disease, Alzheimer's, and diabetes. Alzheimer's has a positive association with increasing age.

Sex was a dummy variable labeled female (female = 1, male = 0). The logit coefficient for female for malignant neoplasms is -0.33. Exponentiating the logit coefficient transforms it into an odds ratio; that is, $e^{-0.33}$ = 0.72. This means that the odds of females are 28% lower than the odds of males of dying of malignant neoplasms compared to dying of other causes; that is, $(e^{-0.33}-1) \times 100 = -28$. These negative odds of females, compared to males, are also found for heart disease; chronic respiratory disease; Alzheimer's; nephritis, nephrotic syndrome, and nephrosis; accidents; and septicemia. For the remaining causes of death, the odds of a female compared to a male dying of the specified cause versus dying of other causes are more. There is no statistically significant relationship for the cause of influenza and pneumonia.

Race was comprised of two dummy variables (White and Black), with the other races category used as the reference category. The logit coefficient for whites for malignant neoplasms is -0.14. Exponentiating the logit coefficient transforms it into an odds ratio; that is, $e^{-0.14} = 0.87$. This means that the odds of whites are 13% lower than the odds of other races of dying of malignant neoplasms compared to dying of other causes, that is, $(e^{-0.14}-1) \times 100 = -13$. These negative odds of whites, compared to other races are also found for cerebrovascular disease; Alzheimer's; influenza and pneumonia; nephritis, nephritic syndrome, and nephrosis; and accidents. For the remaining causes of death, except for heart disease, the odds of whites compared to other races of dying of the specified cause versus dying of other causes are more. There is no statistically significant relationship for this one cause of heart disease. The logit coefficient for blacks for malignant neoplasms is 0.10. Exponentiating the logit coefficient transforms it into an odds ratio, that is, $e^{0.10} = 1.11$. This means that the odds of blacks are 11% higher than the odds of other races of dying of malignant neoplasms compared to dying of other causes, that is, $(e^{0.10}-1) \times 100 = .11$. We find these positive odds of blacks, compared to other races for heart disease; influenza and pneumonia; diabetes, nephritis, nephrotic syndrome, and nephrosis; and septicemia. For the remaining causes of death, the odds of blacks compared to other races of dying of the specified cause versus dying of other causes are less.

The Hispanic origin dummy variable was labeled Hispanic (Hispanic = 1, non-Hispanic = 0). The logit coefficient for Hispanic origin for malignant neoplasms is 0.06. Exponentiating the logit coefficient transforms it into an odds ratio, that is, $e^{0.06} = 1.06$. This means that the odds for Hispanic origin are 6% higher than the odds for non-Hispanics dying of malignant neoplasms compared to dying of other causes, that is, $(e^{.06}-1) \times 100 = 6$. These positive odds for Hispanic origin, compared to non-Hispanic origin, are also found for heart disease; Alzheimer's; influenza and pneumonia; and nephritis, nephrotic syndrome, and nephrosis, accidents. For the remaining causes of death, except for cerebrovascular disease, accidents, and septicemia, the odds of Hispanics compared to non-Hispanics dying of the specified cause versus dying of other causes are less. There is no statistically significant relationship for cerebrovascular disease and accidents.

Regarding marital status, 6 of 10 causes of death with the marital status variable had significant relationships. The logit coefficient for married for malignant neoplasms is 0.31. Exponentiating the logit coefficient to an odds ratio, equals $e^{0.31} = 1.36$. This means that the odds of those who were married of dying of malignant neoplasms compared to dying of other causes, are 36% higher than the odds of those who never married; that is, $(e^{0.31}-1) \times 100 = 0.36$.

These positive associations are also found for cerebrovascular disease, chronic respiratory disease, influenza and pneumonia, and diabetes. Only Alzheimer's and septicemia show negative associations. The other remaining causes of death have no statistical relationships. This positive relationship between having been married and the odds of dying of malignant neoplasms versus dying of other causes are also similar to those who were divorced and widowed. The odds of those who were divorced compared to those who never married of dying of the specified cause versus dying of other causes are positive and significant for malignant neoplasms, cerebrovascular disease, chronic respiratory disease, influenza and pneumonia, and diabetes.

Only septicemia has a negative relationship. The other remaining causes of death have no significant relationships. The odds of those widowed compared to those never married of dying of the specified cause versus dying of other causes are positive and significant for heart disease, malignant neoplasms, cerebrovascular disease, chronic respiratory disease, influenza and pneumonia, and diabetes.

Alzheimer's, accidents, and septicemia have negative associations. The other remaining causes of death have no statistical relationships.

Education was comprised of four dummy variables (elementary school, junior high school, high school, college or more), with the illiterate category used as the reference category. Education is the least important variable. Only malignant neoplasms has positive and significant relationships for all education categories. However, the highest education category does not show higher log odds than high school education. All the other causes of death, other than malignant neoplasms, are either not statistically significant, or only one or two are significant.

The final explanatory variable pertains to metropolitan/non-metropolitan residence (scored 1 if the person was a metropolitan resident at the time of death, and 0 if a non-metropolitan resident). The logit coefficient for this variable and the likelihood of dying of malignant neoplasms was 0.03. This means that if the oldest old decedent was residing in a metropolitan area at the time of death, he/she had odds of dying of malignant neoplasms versus dying of other causes that are 3% more than those of a decedent who was living in a non-metropolitan area at the time of death, that is $(e^{0.03}-1) * 100 = 3$. This kind of positive association is also found for heart disease and septicemia. Five causes of death also have negative and statistically significant logits (cerebrovascular disease; Alzheimer's; influenza and pneumonia; nephritis, nephrotic syndrome, and nephrosis; and accidents). The other remaining causes of death have no statistically significant relationships.

At the base of Table 3 are two statistics that gauge the degree of fit of the overall model examined. The model chi-square statistic has a value of 39,905.58, with 140 degrees of freedom (one for each of the logits being estimated). These chi-square values are sufficiently large to reject the null hypothesis that the 126 logit coefficients are all zero. This finding is also shown by the fact that the majority of the logit coefficients are statistically significant.

A value of 0.01 for Pseudo R^2 statistic for the U.S. is also shown. Although this statistic does not have anywhere near as straightforward an interpretation as the explained variance interpretation that R^2 has in ordinary least squares regression, it is nevertheless a rough gauge of the degree of fit of the model used. With a low value of 0.01, these indicate that there are surely other independent variables, in addition to those used in Table 3, that are important in predicting the likelihood of oldest old Americans dying of a major cause of death instead of dying of all other causes.

Conclusion

In this article, the cause of death structure for the oldest old decedents in the United States in 2001 was examined. The three main causes-of-death for the American decedents, which comprised close to two thirds of all deaths, were heart disease, malignant neoplasms, and cerebrovascular disease. The top 10 causes of death accounted for over 82% of all deaths, and the residual category of other causes accounted for 18% of the deaths.

A multinomial logistic regression equation was estimated to predict the patterns of cause-of-death mortality for the 963,768 oldest old Americans who died in 2001. The primary goal was to ascertain which independent variables best predicted the log odds of dying of one of the major causes of death compared to dying of other causes. The best predictors were age, sex, race, Hispanic origin, and metropolitan residence. Marital status and education did not perform as well. In particular, education was found to be the least important variable in the multinomial equation.

Also, the older the American decedent, the less likely he/she would die of a major cause of death compared to other causes. This relationship was found for most of the 10 main causes of death. Regarding the independent variable of sex, it was found that females in the U.S. were less likely than males to die of one of seven main causes (heart disease; malignant disease; chronic respiratory disease; Alzheimer's; nephritis, nephrotic syndrome, and nephrosis; accidents; and septicemia).

With respect to race, whites had negative and statistically significant logits for 6 of the 10 causes of death (malignant neoplasms; cerebrovascular disease; Alzheimer's; influenza and pneumonia; nephritis, nephritic syndrome, and nephrosis; and accidents); three causes of death had positive and statistically significant logits (chronic respiratory disease, diabetes, and septicemia), while heart disease had no statistical relationships. It was also found that blacks had positive and statistically significant logits for 6 of the 10 causes of death (heart disease; malignant neoplasms; influenza and pneumonia; diabetes; nephritis, nephrotic syndrome, and nephrosis; and septicemia); the remaining four causes of death had negative and statistically significant logits (cerebrovascular disease, chronic respiratory disease, Alzheimer's, and accidents).

Hispanic origin had positive and statistically significant logits for 5 of the 10 causes of death (heart disease; malignant neoplasms; Alzheimer's; influenza and pneumonia; nephritis, nephrotic syndrome, and nephrosis; and accidents); two causes of death had negative and statistically significant logits (chronic respiratory disease and diabetes), and the remaining other causes of death had no statistical relationships.

Results indicated that metropolitan residence had negative and statistically significant logits for 5 of the 10 causes of death (cerebrovascular disease; Alzheimer's; influenza and pneumonia; nephritis, nephrotic syndrome, and nephrosis; and accidents); three causes of death had positive and statistically significant logits (heart disease, malignant neoplasms, and septicemia), and the remaining causes of death had no statistical relationships.

In the next 50 years, the number of oldest old persons in the U.S. is projected to increase almost two times, from 13 million in the year 2000 to almost 25 million in the year 2050. The analyses and results reported in this paper of the cause-of-death structure of the U.S. oldest old decedents in 2001 could well reflect the cause-of-death structure of the increasingly large numbers of oldest old decedents in future decades. The analyses of the dynamics of the current causes of death could suggest the patterns of mortality that may be anticipated for the growing population of oldest old Americans in the next several decades.

This study also demonstrates the appropriate usage of multinomial logistic regression models when the dependent variable has more than two nominal categories. It has been found that multinomial logistic modeling exhibits suitable statistical interpretations for complex results, weakening the criticism of using such a model for these types of data.

References

Hetzel, R., & Smith, A. (2001). *The 65 years and over population: 2000, Census 2000 brief.* U.S. Department of Commerce. Economics and Statistics Administration. U.S. Census Bureau. Washington D.C.

Long, J. S., & Freese, J. (2003). *Regression models for categorical dependent variables using Stata*. (Revised ed.). College Station, Texas: Stata Press.

National Center for Health Statistics. (2003). *1998 Multiple cause-of-death file, NCHS CD-ROM, Series 20, No. 10H.* Hyattsville, Maryland: National Center for Health Statistics.

Olshansky, S. J., & Ault, A. B. (1986). The fourth stage of the epidemiologic transition: The age of delayed degenerative disease." *Milbank Memorial Fund Quarterly,* 64, 355-391.

Omran, A. R. (1971). The epidemiologic transition: A theory of the epidemiology of population change. *Milbank Memorial Fund Quarterly,* 49, 509-538.

Omran, A. R. (1981). The epidemiologic transition. In Ross, J. A. (ed.), *International encyclopedia of population,* pp. 172-175. New York: The Free Press.

Poston, D. L., Jr., & Min, H. S. (2004). Cause of death mortality of the oldest old in the Republic of Korea in 2001, with comparison to the United States in 1998. In *International Conference on Current Issues of the Elderly in the 21st Century*, 1-19, Institute of Gerontology, Yeungnam University, October 31, 2003.

Rogers, R. G., Hummer, R. A., & Nam, C. B. (2000). *Living and dying in the USA*. Academic Press. San Diego, CA.

Stata Corporation. (2003). *Stata base reference manual, release 8* (*2*). College Station, Texas: Stata Corporation.

Journal of Modern Applied Statistical Methods
November, 2008, Vol. 7, No. 2, 607-615

1538 – 9472/08/$95.00

Construction of Insurance Scoring System using Regression Models

Noriszura Ismail Abdul Aziz Jemain
Universiti Kebangsaan Malaysia, Malaysia

This study suggests the regression models of Lognormal, Normal and Gamma for constructing insurance scoring system. The main advantage of a scoring system is that it can be used by insurers to differentiate between high and low risks insureds, thus allowing the profitability of insureds to be predicted.

Key words: Scoring system, insurance risks, regression model.

Introduction

One of the most recent developments in the U.S. and the European insurance industry is the rapidly growing use of a scoring system in pricing, underwriting and marketing of high volume and low premium insurance policies. In the Asian market, scoring system is still considered as relatively new, although several markets in the region have started utilizing the system especially in its rating of motor insurance premium. In Singapore for example, in 1992, the biggest private car insurer, NTUC Income, announced that it was changing from a tariff system to a scoring system, whereby the owners of newer cars and more expensive models would probably pay lower premiums (Lawrence 1996).

There are several advantages of utilizing scoring system in pricing, underwriting and marketing of insurance. The main advantage is that the scores may be used by insurers to differentiate between good and bad insureds, thus allowing the profitability of insureds to be predicted by using a specified list of rating factors such as driver's experience, vehicle's characteristics and scope of coverage. In addition to distinguishing the risks of insureds, insurers may also employ the scores to determine the amount of premium to be charged on each customer.

Several studies on scoring system have been carried out in the actuarial and insurance literatures. For example, Coutts (1984) proposed the Orthogonal Weighted Least Squares (OWLS) to convert premiums into scores; he examined the impact of changing several input assumptions such as inflation rates, base periods of TPBI claims, expenses and weights on the structure of scores. Brockman & Wright (1992) suggested Gamma regression model to convert premiums into scores, rationalizing that the variance of Gamma depends on the weights or exposures, and not on the magnitude of premiums.

In recent years, Miller & Smith (2003) analyzed the relationship between credit-based insurance scores and propensity of loss for private passenger automobile insurance, and found that insurance scores were correlated with propensity of loss due to the correlation between insurance scores and claim frequency rather than average claim severities. Anderson et al. (2004) suggested Generalized Linear Modeling (GLM) for deriving scores, and proposed the fitting of frequency and severity separately for each claim type as starting point. The expected claim costs resulting from frequency and severity fitting were then divided by the premiums to yield the expected loss ratios, and the profitability scores were derived by rescaling the loss ratios. Wu & Lucker (2004) reviewed the basic structure of several insurance credit scoring models in the

Noriszura Ismail is an Associate Professor in Actuarial Science Department. Her research areas are actuarial and statistical modeling in non-life insurance. Email: ni@ukm.my. Abdul Aziz Jemain is a Professor in Statistics Department. His research areas are climate modeling, actuarial modeling, medical statistics and social statistics. Email azizj@ukm.my.

U.S. by dividing scoring algorithms into two main categories; the rule-based approach which assigns scores directly to each rating factor, and the formula approach which determines scores using mathematical formulas. The minimum bias and GLM were suggested for the rule-based approach, whereas the Neural Networks (NN) and Multivariate Adaptive Regression Splines (MARS) were suggested for the formula approach. Wu & Guszcza (2004) studied the relationship between credit scores and insurance losses using data mining methodology along with several predictive modeling techniques such as NN, GLM, Classification and Regression Trees (CART) and MARS. Vojtek & Kocenda (2006) reviewed several methods of credit scoring employed by banks such as linear discriminant analysis (LDA), logit analysis, *k*-nearest neighbor classifier (*k*-NN) and NN to evaluate the applications of loans in Czech and Slovak Republics. Their results showed that the logit analysis and LDA methods were mainly used, the CART and NN methods were used only as supporting tools, and the *k*-NN method was rarely used in the process of selecting variables and evaluating the quality of credit scoring models.

The objective of this article is to suggest the Lognormal, Normal and Gamma regression models for the construction of insurance scoring system. Even though several actuarial studies have been carried out on the construction of scoring system, the detailed procedures of these methods have not been provided, with the exception of Coutts (1984) who proposed the use of Orthogonal Weighted Least Squares (OWLS) to convert premiums into scores. Although the Lognormal model proposed in this study is similar to the OWLS method proposed by Coutts (1984), the fitting procedure slightly differs. The OWLS method assumed that the weights were possible to be factorized and the fitted values were calculated using the estimated weights, whereas in this study, the fitting procedure does not require the weights to be factorized and the weights were not replaced by the estimated weights. This study also compares the Lognormal, Normal and Gamma regression models whereby the comparisons were centered upon three main elements; fitting procedures, parameter estimates and structure of scores.

Methodology

The response variable, independent variables and weight for the regression models are the premiums, rating factors and exposures respectively. The datasets are (g_i, e_i), where g_i and e_i respectively denote the premiums and the exposure in the i-th rating class, $i = 1,2,...,n$.

Appendix A shows a sample of rating factors, premiums and exposures for the data set. The premiums were written in Ringgit Malaysia (RM) currency based on motor insurance claims experience provided by an insurance company in Malaysia. The exposures were written in number of vehicle years, and the rating factors considered were scope of coverage (comprehensive, non-comprehensive), vehicle make (local, foreign), use-gender (private-male, private-female, business), vehicle year (0-1, 2-3, 4-5, 6+) and location (Central, North, East, South, East Malaysia).

Lognormal Model

Let the relationship between premiums, g_i and scores, s_i, be written as,

$$g_i = b^{s_i}, \qquad (1)$$

or,

$$\log_b g_i = s_i. \qquad (2)$$

In this study, the value of $b = 1.1$ was chosen for Equation (1) to accommodate the conversion of premiums ranging from RM30 to RM3,000 into scores ranging from 0 to 100. For example, the score corresponding to the premium amount of RM3,000 is equal to 84.

If the premium, G_i, is distributed as Lognormal with parameters s_i and $e_i^{-1}\sigma^2$, then $\log_{1.1} G_i$ is distributed as normal with mean s_i and variance $e_i^{-1}\sigma^2$, where the density is,

$$f(\log g_i; s_i) = \frac{1}{\sqrt{2\pi\sigma^2 e_i^{-1}}} \exp\left(-\frac{e_i(\log g_i - s_i)^2}{2\sigma^2} \right).$$

The relationship between scores, s_i, and rating factors, x_{ij}, may be written in a linear function as,

$$s_i = \mathbf{x_i^T \beta} = \sum_{j=1}^{p} \beta_j x_{ij} \ , \tag{3}$$

where $\mathbf{x_i}$ denotes the vector of explanatory variables or rating factors, and $\boldsymbol{\beta}$ the vector of regression parameters. In other words, $\beta_j, j = 1,2,...,p$, represents the individual score of each rating factor, and s_i represents the total scores of all rating factors.

The first derivatives of Equation (3) may be simplified into,

$$\frac{\partial s_i}{\partial \beta_j} = x_{ij} \, . \tag{4}$$

Therefore, the solution for $\boldsymbol{\beta}$ may be obtained from the maximum likelihood equation,

$$\frac{\partial \ell}{\partial \beta_j} = \sum_i e_i (\log g_i - s_i) x_{ij} = 0, \quad j = 1,2,...,p. \tag{5}$$

Since the maximum likelihood equation is also equivalent to the normal equation in standard weighted linear regression, $\boldsymbol{\beta}$ may be solved by using normal equation.

Normal Model

Assume that the premium, G_i, is distributed as normal with mean δ_i and variance $e_i^{-1}\sigma^2$, where the density function is,

$$f(g_i;\delta_i) = \frac{1}{\sqrt{2\pi\sigma^2 e_i^{-1}}} \exp\left(-\frac{e_i (g_i - \delta_i)^2}{2\sigma^2} \right).$$

The conversion of premiums into scores may be implemented by letting the relationship between scores (s_i) and fitted premium (δ_i) to be written in a log-linear function or multiplicative form. If the base value is equal to 1.1, the fitted premium is,

$$\delta_i = (1.1)^{s_i} , \tag{6}$$

where

$$s_i = \mathbf{x_i^T \beta} = \sum_{j=1}^{p} \beta_j x_{ij} \ .$$

The first derivative of Equation (6) is,

$$\frac{\partial \delta_i}{\partial \beta_j} = \log(1.1) \delta_i x_{ij} \ , \tag{7}$$

and the solution for $\boldsymbol{\beta}$ may be obtained from the maximum likelihood equation.

$$\frac{\partial \ell}{\partial \beta_j} = \sum_i e_i (g_i - \delta_i) \delta_i x_{ij} = 0, \quad j = 1,2,...,p. \tag{8}$$

The maximum likelihood equation shown by Equation (8) is not as straightforward to be solved compared to the normal equation shown in Equation (5). However, since Equation (8) is equivalent to the weighted least squares, the fitting procedure may be carried out by using an iterative method of weighted least squares (see McCullagh & Nelder, 1989; Mildenhall, 1999; Dobson, 2002; Ismail & Jemain, 2005; Ismail & Jemain, 2007). In this study, the iterative weighted least squares procedure was performed using SPLUS programming.

Gamma Model

The construction of a scoring system based on the Gamma Model is also similar to the Normal Model. Assume that the premium, G_i, is distributed as Gamma with mean δ_i and variance $v^{-1}\delta_i^2$, where the density function is,

$$f(g_i;\delta_i) = \frac{1}{g_i \Gamma(v)} \left(\frac{v g_i}{\delta_i} \right)^v \exp\left(-\frac{v g_i}{\delta_i} \right),$$

and v denotes the index parameter.

The conversion of premiums into scores may also be implemented by letting the relationship between scores (s_i) and fitted premiums (δ_i) to be written in a log-linear function or multiplicative form. Therefore, the first derivative is the same as Equation (7).

Assume that the index parameter, v, varies within classes, and can be written as $v_i = e_i \sigma^{-2}$. Therefore, the variance of the response variable is equal to $\sigma^2 \delta_i^2 e_i^{-1}$ and the solution for $\boldsymbol{\beta}$ may be obtained through the maximum likelihood equation,

$$\frac{\partial \ell}{\partial \beta_j} = \sum_i \frac{e_i (g_i - \delta_i) x_{ij}}{\delta_i}, \qquad j = 1,2,\ldots,p\,. \quad (9)$$

The maximum likelihood equation shown by Equation (9) is not as straightforward to be solved compared to the normal equation shown by Equation (5), and the fitting procedure may be carried out using an iterative method of weighted least squares.

Results

Scoring System based on Lognormal Model

The best model for lognormal regression may be determined by using standard analysis of variance. Based on the ANOVA results, all rating factors are significant, and 89.3% of the model's variations ($R^2 = 0.893$) can be explained by using the same rating factors.

The parameter estimates for the best regression model are shown in Table 1. The class for 2-3 year old vehicles is combined with 0-1 year old vehicles (intercept), and the classes for East and South locations are combined with Central location (intercept) to provide significant effects on all individual regression parameters.

The negative estimates are converted into positive values using the following procedure. First, the smallest negative estimate of each rating factor is transformed into zero by adding an appropriate positive value. Then, the same positive value is added to other estimates categorized under the same rating factor. Finally, the intercept is deducted by the total positive values which are added to all estimates. The final scores are then rounded into whole numbers to provide easier premium calculation and risk interpretation. The original estimates, modified estimates and final scores are shown in Table 2.

The final scores shown in Table 2 specify and summarize the degree of relative risks associated with each rating factor. For instance, the risks for foreign vehicles are relatively higher by four points compared to local vehicles, and the risks for male and female drivers who used their cars for private purposes are relatively higher by nine and five points compared to drivers who used their cars for business purposes. The goodness-of-fit of the scores in Table 2 may be tested by using two methods; (1) comparing the ratio of fitted over actual premium income, and (2) comparing the difference between fitted and actual premium income.

Table 3 shows the total difference of premium income and the overall ratio of premium income. The total income of fitted premiums is understated by RM560,380 or 0.2% of the total income of actual premiums.

Table 1: Parameter estimates for Lognormal Model

Parameters		Estimates	Std.dev.	*p*-values
β_1	Intercept	78.81	0.26	0.00
β_2	Non-comprehensive	-14.52	0.43	0.00
β_3	Foreign	4.23	0.26	0.00
β_4	Female	-4.30	0.28	0.00
β_5	Business	-9.25	0.53	0.00
β_6	4-5 years	-1.17	0.33	0.02
β_7	6+ years	-1.56	0.30	0.01
β_8	North	0.84	0.29	0.04
β_9	East Malaysia	-4.18	0.45	0.00

Table 2: Original estimates, modified estimates and final scores

Parameters	Original Estimates	Modified Estimates	Final Scores
Intercept (Minimum score)	78.81	49.30	49
Coverage:			
Comprehensive	0.00	14.52	15
Non-comprehensive	-14.52	0.00	0
Vehicle make:			
Local	0.00	0.00	0
Foreign	4.23	4.23	4
Use-gender:			
Private-male	0.00	9.25	9
Private-female	-4.30	4.95	5
Business	-9.25	0.00	0
Vehicle year:			
0-1 year & 2-3 years	0.00	1.56	2
4-5 years	-1.17	0.39	0
6+ years	-1.56	0.00	0
Vehicle location:			
Central, East & South	0.00	4.18	4
North	0.84	5.02	5
East Malaysia	-4.18	0.00	0

Table 3: Total premium income difference and overall premium income ratio

		Value
Total number of businesses/policies/exposures	$\sum_{i=1}^{240} e_i$	170,749
Total income from fitted premiums	$\sum_{i=1}^{240} e_i \hat{g}_i$	RM 275,269,816
Total income from actual premiums	$\sum_{i=1}^{240} e_i g_i$	RM 275,830,196
Total premium income difference	$\sum_{i=1}^{240} e_i (\hat{g}_i - g_i)$	- RM 560,380
Overall premium income ratio	$\frac{\sum_{i=1}^{240} e_i \hat{g}_i}{\sum_{i=1}^{240} e_i g_i}$	0.998

Therefore, the fitted premiums for all classes are suggested to be multiplied by a correction factor of 1.002 to match their values with the actual premiums.

Apart from differentiating between high and low risk insureds, a scoring system may also be used by insurers to calculate the amount of premium to be charged on each client. The procedure for converting scores into premium amounts involved two basic steps.

First, the scores for each rating factor are recorded and aggregated; then, the aggregate scores are converted into premium amount by using a scoring conversion table (a table listing the aggregate scores with associated monetary values). Table 4 shows a scoring conversion table, which is constructed using Equation (1).

Comparison of Scoring System based on Lognormal, Normal and Gamma Models

Comparison of parameter estimates resulted from Lognormal, Normal and Gamma regression models are shown in Table 5. The parameter estimates for Lognormal, Normal and Gamma models provided similar values, except for β_2 and β_5 which produced larger values in Normal and Gamma models compared to Lognormal model.

Table 4: Scoring conversion table

Aggregate Scores	Premium Amounts (RM)	Aggregate Scores	Premium Amounts (RM)
49	107	67	595
50	118	68	654
51	129	69	719
52	142	70	791
53	157	71	870
54	172	72	958
55	189	73	1053
56	208	74	1159
57	229	75	1274
58	252	76	1402
59	277	77	1542
60	305	78	1696
61	336	79	1866
62	369	80	2052
63	406	81	2258
64	447	82	2484
65	491	83	2732
66	540	84	3005

Table 5: Estimates for Lognormal, Normal and Gamma regression models

Parameters		Lognormal			Normal			Gamma		
		Est.	Std. Error	*p*-value	Est.	Std. Error	*p*-value	Est.	Std. Error	*p*-value
β_1	Intercept	78.81	0.26	0.00	79.02	0.01	0.00	78.89	0.02	0.00
β_2	Non-comp	-14.52	0.43	0.00	-12.79	0.05	0.00	-13.71	0.03	0.00
β_3	Foreign	4.23	0.26	0.00	4.02	0.01	0.00	4.19	0.02	0.00
β_4	Female	-4.30	0.28	0.00	-4.03	0.01	0.00	-4.25	0.02	0.00
β_5	Business	-9.25	0.53	0.00	-7.40	0.03	0.00	-8.55	0.04	0.00
β_6	4-5 years	-1.17	0.33	0.02	-1.17	0.01	0.00	-1.17	0.02	0.00
β_7	6+ years	-1.56	0.30	0.01	-2.10	0.01	0.00	-1.73	0.02	0.00
β_8	North	0.84	0.29	0.04	0.49	0.01	0.00	0.81	0.02	0.00
β_9	East M'sia	-4.18	0.45	0.00	-4.01	0.03	0.00	-4.21	0.03	0.00

Comparison of scoring system resulted from Lognormal, Normal and Gamma regression models are shown in Table 6. The scores for Lognormal range from 49 to 84, the scores for Normal range from 53 to 84, and the scores for Gamma range from 51 to 85. In terms of risk relativities, both Lognormal and Gamma models resulted in a relatively higher score for male driver, female driver and comprehensive coverage. Therefore, if an insurer is interested in charging higher premiums for male driver, female driver and comprehensive coverage, both Lognormal and Gamma models may be suitable for fulfilling this strategy. However, the difference between Lognormal and Gamma model is that the scores for low risk classes provided by Gamma are slightly higher compared to Lognormal.

Conclusion

This article shows the procedure for constructing insurance scoring systems using three different regression models; Lognormal, Normal and Gamma. The main advantage of a scoring system is that it may be used by insurers to differentiate between "good" and "bad" insureds, thus allowing the profitability of insureds to be predicted. In addition, the scoring system has an operational advantage of reducing premium calculations and can be treated as a more sophisticated device for customers to assess their individual risks.

Table 6: Scoring system for Lognormal, Normal and Gamma regression models

Rating factors	Scores		
	Lognormal	Normal	Gamma
Minimum scores	49	53	51
Coverage:			
Comprehensive	15	13	14
Non-comprehensive	0	0	0
Vehicle make:			
Local	0	0	0
Foreign	4	4	4
Use-gender:			
Private-male	9	7	9
Private-female	5	3	4
Business	0	0	0
Vehicle year:			
0-1 year	2	2	2
2-3 years	2	2	2
4-5 years	0	1	1
6+ years	0	0	0
Location:			
Central	4	4	4
North	5	5	5
East	4	4	4
South	4	4	4
East Malaysia	0	0	0

The relationship between aggregate scores and rating factors in Lognormal model was suggested as linear function or additive form, whereas the relationship between aggregate scores and rating factors in Normal and Gamma models were proposed as log-linear function or multiplicative form.

The best regression model for Lognormal model was selected by implementing the standard analysis of variance. The goodness-of-fit of scores estimates were tested by comparing the ratio of fitted over actual premium income and by comparing the difference between fitted and actual premium income.

Acknowledgements

The authors gratefully acknowledge the financial support received in the form of a research grant (IRPA RMK8: 09-02-02-0112-EA274) from the Ministry of Science, Technology and Innovation (MOSTI), Malaysia. The authors are also pleasured to thank Insurance Services Malaysia (ISM) for supplying the data.

References

Anderson, D., Feldblum, S., Modlin, C., Schirmacher, D., Schirmacher, E., & Thandi, N. (2004). A practitioner's guide to generalized linear models. *Casualty Actuarial Society Discussion Paper Program*, 1-115.

Brockman, M. H., & Wright, T. S. (1992). Statistical motor rating: Making effective use of your data. *Journal of the Institute of Actuaries*. *119*(3), 457-543.

Coutts, S. M. (1984). Motor insurance rating, an actuarial approach. *Journal of the Institute of Actuaries*, *111*, 87-148. Dobson, A. J. (2002). *An introduction to Generalized Linear Models (second edition)*. NY: Chapman & Hall.

Ismail, N., & Jemain, A. A. (2005). Bridging minimum bias and maximum likelihood methods through weighted equation. *Casualty Actuarial Society Forum*, *Spring*, 367-394.

Ismail, N., & Jemain, A. A. (2007). Handling overdispersion with Negative Binomial and Generalized Poisson regression models. *Casualty Actuarial Society Forum*, *Winter*: 103-158.

Lawrence, B. (1996). Motor insurance in Singapore. In Low Chan Kee Ed., *Actuarial and insurance practices in Singapore*. 191-216. Addison-Wesley: Singapore.

McCullagh, P., & Nelder, J. A. (1989). *Generalized Linear Model.* (2nd ed.) London, UK: Chapman & Hall.

Mildenhall, S. J. (1999). A systematic relationship between minimum bias and generalized linear models. *Proceedings of the Casualty Actuarial Society*, *86*(164), 93-487.

Miller, M. J., & Smith, R. A. (2003). The relationship of credit-based insurance scores to private passenger automobile insurance loss propensity. *Presentation to NAIC*. July, 2003.

Vojtek, M., & Kocenda, E. (2006). Credit scoring models. *Czech Journal of Economics and Finance*, *56*(3-4), 152-167.

Wu, C. P., & Lucker, J. R. (2004, Winter). A view inside the "Black Box": A review and analysis of personal lines insurance credit scoring models filed in the state of Virginia. *Casualty Actuarial Society Forum*, 251-290.

Wu, C. P., & Guszcza, J. C. (2004, Winter). Does credit score really explain in losses? Multivariate analysis from a data mining point of view. *Casualty Actuarial Society Forum*, 113-138.

Appendix A: Rating factors, exposures and premium amounts for Malaysian data

Rating factors					Exposure (vehicle-year)	Premium amount (RM)
Coverage	Vehicle make	Use-gender	Vehicle year	Location		
Comprehensive	Local	Private-male	0-1 year	Central	4243	1811
				North	2567	2012
				East	598	1927
				South	1281	1869
				East Malaysia	219	983
			2-3 years	Central	6926	1704
				North	4896	1919
				East	1123	1854
				South	2865	1794
				East Malaysia	679	1301
			4-5 years	Central	6286	1613
				North	4125	1840
				East	1152	1770
				South	2675	1687
				East Malaysia	700	1162
			6+ years	Central	6905	1524
				North	5784	1790
				East	2156	1734
				South	3310	1633
				East Malaysia	1406	1144
		Private-female	0-1 year	Central	2025	1256
				North	1635	1343
				East	301	1396
				South	608	1289
				East Malaysia	126	787
			2-3 years	Central	3661	1210
				North	2619	1298
				East	527	1255
				South	1192	1212
				East Malaysia	359	942
			4-5 years	Central	2939	1139
				North	1927	1243
				East	439	1125
				South	959	1176
				East Malaysia	376	652
			6+ years	Central	2215	1072
				North	1989	1215
				East	581	1219
				South	937	1112
				East Malaysia	589	623
		Business	0-1 year	Central	290	722
				North	66	547
				East	24	107
				South	52	685
				East Malaysia	6	107
			2-3 years	Central	572	731
				North	148	630
				East	40	107
				South	91	657
				East Malaysia	17	107
			4-5 years	Central	487	654
				North	100	549
				East	40	540
				South	59	571
				East Malaysia	22	493
			6+ years	Central	468	567
				North	93	518
				East	33	562
				South	77	515
				East Malaysia	25	402

Journal of Modern Applied Statistical Methods
November, 2008, Vol. 7, No. 2, 616-631

1538 – 9472/08/$95.00

EARLY SCHOLARS

On Some Properties of Quasi-Negative-Binomial Distribution and Its Applications

Anwar Hassan
University of Kashmir
India

Sheikh Bilal
Degree College
India

The quasi-negative-binomial distribution was applied to queuing theory for determining the distribution of total number of customers served before the queue vanishes under certain assumptions. Some structural properties (probability generating function, convolution, mode and recurrence relation) for the moments of quasi-negative-binomial distribution are discussed. The distribution's characterization and its relation with other distributions were investigated. A computer program was developed using R to obtain ML estimates and the distribution was fitted to some observed sets of data to test its goodness of fit.

Key words: Simultaneous quasi-negative-binomial distribution, Borel-Tanner distribution, probability generating function, convolution property, characterization, chi-square fitting.

Introduction

The classical negative binomial distribution has become increasingly popular as a more flexible alternative to the Poisson distribution, especially in cases when it is doubtful whether the strict independence requirements for a Poisson distribution will be satisfied. In a classical negative binomial distribution the probability of success from trial to trial is assumed to be constant, but this assumption holds true only in the case of chance mechanism and is not realistic for many practical situations. Most living beings use past experiences (successes or failures) and wisdom to help determine future strategies to achieve goals, thus, the probability of success or failure does not remain constant. It is generally felt that the probability of a success depends on the number of previous failures, and the quasi- negative-binomial distribution, as well as other distributions, takes this fact into consideration.

Much work has been done on the quasi-binomial distribution, but little has been done on quasi-negative-binomial distribution. The quasi-negative-binomial distribution has been obtained in different forms by Janardan (1975), Nandi & Das (1994), and Sen & Jain (1996), but has not to date been studied in detail. This article examines various aspects of this distribution. The distribution of the number of customers served in the queuing theory under certain assumptions, which gives rise to a quasi-negative-binomial distribution, was derived. It is also shown that the quasi-negative-binomial distribution belongs to a family of Abel series distributions. Some structural properties of the distribution are discussed, along with its relation with some other important distributions, and a characterization of the distribution is provided. A computer program written in R was developed to obtain ML estimates and the distribution was fitted to a number of data sets to show its superiority over other distributions.

Anwar Hassan is a Post Graduate in the Department of Statistics. Research interests: Probability and Lagrangian Probability distributions and Statistical Inference. E-mail: anwar.hassan2007@gmail.com. Sheikh Bilal Ahmad is in the Department of Statistics. Research interests: Probability distributions and Statistical Inference. E-mail: sbilal_sbilal@yahoo.com.

Methodology

Quasi-negative-binomial distribution (QNBD)

In the theory of queuing, suppose there exists a single queue beginning with r customers. Haight & Brever (1960) showed that, if it is first assumed that the random arrival time of a customer is at a constant rate (λ), and a constant amount of time is devoted to serving each customer (β), then the probability distribution of the total number of customers served before the queue vanishes is:

$$P(X:x)=\frac{r}{(x-r)!}\,x^{x-r-1}\,(\lambda\beta)^{x-r}\,e^{-\lambda\beta x} \qquad (2.1)$$

$$x=r,r+1,.......$$

This is known as the Borel-Tanner distribution and it gives the probability of a customer arriving during the period $(t,t+\Delta t)$ as $\lambda(\Delta t)+0(\Delta t)$, by assuming λ is constant, where $0(\Delta t)$ is the probability of two or more customers arriving in this period. This assumption, however, is not realistic. The random arrival time of customers is not at a constant rate, it varies from interval to interval of equal length. In order to make the formula more flexible it is allowed to vary in different intervals of equal length with a constant amount of time (β) spent serving each customer. Thus gives the probability distribution of total number of customers served before the queue vanishes as:

$$P(X:x)=E\left[\frac{r}{(x-r)!}\,x^{x-r-1}\,(\lambda\beta)^{x-r}\,e^{-\lambda\beta x}\right] \qquad (2.2)$$

where expectation is to be taken over λ. Suppose that the distribution of λ is a gamma variate with parameters (a, b), then the above equation becomes:

$$P(X=x)=\frac{r}{(x-r)!}\,x^{x-r-1}\,\beta^{x-r}\,\frac{b^a}{\Gamma(a)}\int_0^\infty \lambda^{a+x-r-1}\,e^{-\lambda(b+\beta x)}\,d\lambda \qquad (2.3)$$

$$=\frac{\beta r}{(x-r)!}\,(\beta x)^{x-r-1}\,\frac{b^a}{\Gamma(a)}\,\frac{\Gamma(a+x-r)}{(b+\beta x)^{a+x-r}}$$

Taking $x=x+r$, that is, starting with an idle queue the probability distribution becomes:

$$P(X:x)=\frac{\Gamma(a+x)}{\Gamma(a)\;x!}\;\frac{(\beta r)b^a(\beta r+\beta x)^{x-1}}{(b+\beta r+\beta x)^{a+x}}$$

$$x=0,1,2,..........$$

$$=\binom{a+x-1}{x}\frac{\theta_1\,(\theta_1+\theta_2 x)^{x-1}}{(1+\theta_1+\theta_2 x)^{a+x}}$$

$$x=0,1,2,..... \qquad (2.4)$$

where $\beta rb^{-1}=\theta_1, \beta b^{-1}=\theta_2$. The distribution represented by (2.4) is a quasi-negative-binomial distribution (QNBD). Hence, the distribution of the total number of customers served before the queue vanishes, assuming a start with an idle queue wherein the random arrival time of customers follows a gamma distribution and the time occupied in serving each customer is constant, is a QNBD.

Equation 2.3 clearly suggests that the quasi-negative-binomial distribution is a mixture of the Borel-Tanner distribution (2.1) with gamma $\gamma(a,b)$ as the mixing distribution. Another way of obtaining the QNBD (2.4) is to compound the restricted generalized Poisson model $(\theta,\alpha\theta)$ with the gamma distribution $\gamma(a,b)$, where $\theta_1=b^{-1}$ and $\theta_2=\alpha b^{-1}$. This is the method employed to obtain the probability generating function of the proposed model (2.4).

The Abel series distribution and QNBD.

Charalambides (1990) explored the use of the Abel series and introduced the family of Abel series distributions with applications to fluctuations of sample functions of stochastic processes. Nandi & Das (1994) defined a family of Abel series distributions for real valued parameters r and b by its probability function:

$$P(x) = \frac{r(r+bx)^{x-1}h(x,b)}{f(r)}$$
$$x = 0,1,2,........$$
$$(3.1)$$

where $h(x,b) \geq 0$; $r \geq 0$ if $b \geq 0$ and $r + xb \geq 0$ if $b \leq 0$; $f(r)$ is finite and positive function given by:

$$f(r) = \sum_{x=0}^{\infty} r(r+bx)^{x-1}h(x,b)$$

and

$$h(x,b) = \frac{1}{x!}\frac{d^x f(r)}{dr^x}\Big|_{r=-xb} \,.$$

Taking

$$f(r) = (c-r)^{-a}$$

results in

$$h(x,b) = \frac{(a+x-1)!}{(a-1)!x!}(c+bx)^{-a-x}$$

and, using (3.1), gives

$$P(X:x) = \binom{a+x-1}{x}\frac{r(r+bx)^{x-1}(c+bx)^{-a-x}}{(c-r)^{-a}}$$
$$x = 0,1,....... \,.$$
$$(3.2)$$

Finally, taking

$$\frac{r}{(c-r)} = \theta_1 \quad , \frac{b}{(c-r)} = \theta_2$$

and

$$\frac{c}{(c-r)} = 1 + \frac{r}{(c-r)} = 1 + \theta_1$$

the quasi-negative-binomial distribution (2.4) is obtained. Hence, the QNBD is a member of the Abel series of distributions.

Structural properties.

Some of the structural properties that describe the nature of the quasi-negative-binomial distribution were studied. These properties are described as follows:

Convolution property.

Using (3.2) it is possible to show that quasi-negative-binomial variates possess the important – and very desirable – convolution property given by Theorem 4.1: The sum of two independent quasi-negative-binomial variates X_1 and X_2 with parameters $(a_1, \theta_1, \theta_2)$ and $(a_2, \theta_1, \theta_2)$, respectively, is a quasi-negative-binomial variate with parameters $(a_1 + a_2, \theta_1, \theta_2)$.

Proof:

The sum of the probabilities of the QNBD equals unity, therefore from (3.2) the following results:

$$(c-r)^{-a} = \sum_{x=0}^{\infty}\binom{a+x-1}{x} r(r+bx)^{x-1}(c+bx)^{-a-x}$$
$$(4.1)$$

Considering the expansion of $(c-r)^{-(a_1+a_2)} = (c-r)^{-a_1}(c-r)^{-a_2}$ as a single series of Abel polynomials on the left-hand side and the product of two series of Abel polynomials on the right-hand side, using (4.1) and simplifying, the following identity is obtained:

$$\binom{a_1+a_2+x-1}{x} r(r+bx)^{x-1}(c+bx)^{-(a_1+a_2+x)}$$

$$= \sum_{t=0}^{x}\binom{a_1+t-1}{t}\binom{a_2+(x-t)-1}{x-t} r(r+bt)^{t-1} r(r+b(x-t))^{x-t-1}$$
$$(c+bt)^{-(a_1+t)}(c+b(x-t))^{-(a_2+x-t)}$$
$$(4.2)$$

This identity reduces to a Vandermonde-type identity on $b = 0$, Lagrangian Probability Distribution (Consul & Famoye, 2006).

Assuming the sum $X_1 + X_2 = x$, then by definition:

$$P\left((X_1 + X_2):x\right) = \sum_{t=0}^{x} P_t(a_1,r,b,c)\,P_{x-t}(a_2,r,b,c)$$

$$= \frac{1}{(c-r)^{-(a_1+a_2)}} \sum_{t=0}^{x} \binom{a_1+t-1}{t} \binom{a_2+(x-t)-1}{x-t} r(r+bx)^{t-1}$$

$$r(r+b(x-t))^{x-t-1}(c+bt)^{-(a_1+t)}(c+b(x-t))^{-(a_2+x-t)}$$

Using the result (4.2) in the above gives:

$$P\big((X_1+X_2):x\big)$$

$$= \binom{a_1+a_2+x-1}{x} \frac{(r+bx)^{x-1}(c+bx)^{-(a_1+a_2+x)}}{(c-r)^{-(a_1+a_2)}} .$$

Next, taking

$$\frac{r}{(c-r)} = \theta_1 \quad , \frac{b}{(c-r)} = \theta_2 \quad \text{and}$$

$$\frac{c}{(c-r)} = 1 + \frac{r}{(c-r)} = 1 + \theta_1$$

results in the convolution property:

$$P(X_1 + X_2 = x)$$

$$= \binom{a_1+a_2+x-1}{x} \theta_1 (\theta_1 + \theta_2 x)^{x-1}$$

$$(1+\theta_1+\theta_2 x)^{-(a_1+a_2+x)}$$

$$= P_X(a_1+a_2, \theta_1, \theta_2)$$

More generally the sum of n independent quasi-negative-binomial variates with parameters $(a_i, \theta_1, \theta_2)$, $i = 1,2,........n$ is also a quasi-negative-binomial variate with parameters $\left(\sum a_i, \theta_1, \theta_2\right)$.

Unimodality

The QNBD is unimodal according to the Lemma: if the mixing distribution is non-negative, continuous, and unimodal then the resulting distribution is unimodal. (Holgate, 1970) Thus, the proposed model is unimodal since the mixing distribution is the gamma distribution, which is unimodal.

Theorem 4.2: The QNB model (2.4) is unimodal for all values of (a, θ_1, θ_2) and the mode is $x = 0$ if $a\theta_1 < 1$ and, for $a\theta_1 > 1$, the mode is at some point $x = M$ such that

$$\frac{a\theta_1 - 1}{1-\theta_1} < M < \frac{\theta_1(a-1)}{1-\theta_1} .$$

Proof:

The QNBD model (2.4) gives the ratio

$$\frac{P(1)}{P(0)} = \frac{a\theta_1(1+\theta_1)^a}{(1+\theta_1+\theta_2)^{a+1}} < 1 \;\text{ If } a\theta_1 < 1,$$

$$\text{since } \frac{(1+\theta_1)^a}{(1+\theta_1+\theta_2)^{a+1}} < 1 \qquad \forall (a, \theta_1, \theta_2)$$

In general, the ratio of any two successive probabilities of QNBD (2.4) is:

$$\frac{P(x+1)}{P(x)}$$

$$= \frac{a+x}{x+1} \frac{(\theta_1+\theta_2+x\theta_2)^x}{(\theta_1+x\theta_2)^{x-1}} \frac{(1+\theta_1+x\theta_2)^{a+x}}{(1+\theta_1+\theta_2+x\theta_2)^{a+x+1}} \tag{4.3}$$

Since $\dfrac{(1+\theta_1+x\theta_2)^{a+x}}{(1+\theta_1+\theta_2+x\theta_2)^{a+x+1}} < 1$

$\forall (a, \theta_1, \theta_2)$, the ratio $\dfrac{P(x+1)}{P(x)} < 1$, if

$\dfrac{\theta_1(a+x)}{x+1} < \dfrac{(1+x\theta_2\theta_1{}^{-1})^{x-1}}{(1+\theta_2\theta_1{}^{-1}+x\theta_2\theta_1{}^{-1})^x}$, which is true only if $a\theta_1 < 1$ as

$$\frac{(1+x\theta_2\theta_1{}^{-1})^{x-1}}{(1+\theta_2\theta_1{}^{-1}+x\theta_2\theta_1{}^{-1})^x} < 1 \qquad \forall (\theta_1, \theta_2) .$$

Hence, for $a\theta_1 < 1$, the ratio $\dfrac{P(x+1)}{P(x)}$ is a non-increasing function, therefore the mode of the distribution is $x = 0$. Suppose $a\theta_1 > 1$ and the mode is at $x = M$, the ratio defined by (4.3) gives two inequalities:

$$\frac{P(M+1)}{P(M)} = \frac{a+M}{M+1}\,\frac{(\theta_1+\theta_2+M\theta_2)^{M}}{(\theta_1+M\theta_2)^{M-1}}\,\frac{(1+\theta_1+M\theta_2)^{a+M}}{(1+\theta_1+\theta_2+M\theta_2)^{a+M+1}} < 1 \qquad (4.4)$$

and

$$\frac{P(M)}{P(M-1)} = \frac{a+M-1}{M}\,\frac{(\theta_1+M\theta_2)^{M-1}}{(\theta_1-\theta_2+M\theta_2)^{M-2}}\,\frac{(1+\theta_1-\theta_2+M\theta_2)^{a+M-1}}{(1+\theta_1+M\theta_2)^{a+M}} > 1 \qquad (4.5)$$

By inequality (4.4):

$$\frac{\theta_1(a+M)}{M+1} < \frac{(1+M\theta_2\theta_1^{-1})^{M-1}}{(1+\theta_2\theta_1^{-1}+M\theta_2\theta_1^{-1})^{M}} < 1 \qquad (4.6)$$

since

$$\frac{(1+M\theta_2\theta_1^{-1})^{M-1}}{(1+\theta_2\theta_1^{-1}+M\theta_2\theta_1^{-1})^{M}} < 1 \quad \forall(\theta_1,\theta_2)$$

The inequality (4.6) gives the lower bond to M as:

$$M > \frac{a\theta_1-1}{1-\theta_1} \;. \qquad (4.7)$$

And by inequality (4.5):

$$\frac{a+M-1}{M}\,\frac{(\theta_1+M\theta_2)^{M-1}}{(\theta_1-\theta_2+M\theta_2)^{M-2}} > \frac{(1+\theta_1+M\theta_2)^{a+M}}{(1+\theta_1-\theta_2+M\theta_2)^{a+M-1}} \;.$$

Again

$$\frac{(1+\theta_1+M\theta_2)^{a+M}}{(1+\theta_1-\theta_2+M\theta_2)^{a+M-1}} > 1 \;\forall(a,\theta_1,\theta_2)$$

gives

$$\frac{\theta_1(a+M-1)}{M} > \frac{(1-\theta_2\theta_1^{-1}+M\theta_2\theta_1^{-1})^{M-2}}{(1+M\theta_2\theta_1^{-1})^{M-1}} \qquad (4.8)$$

because

$$a\theta_1 > 1,\; \frac{\theta_1(a+M-1)}{M} > 1 \;\forall(\theta_1,\theta_2)$$

and

$$\frac{(1-\theta_2\theta_1^{-1}+M\theta_2\theta_1^{-1})^{M-2}}{(1+M\theta_2\theta_1^{-1})^{M-1}} < 1 \quad \forall(\theta_1,\theta_2)$$

Thus, (4.8) can be written as:

$$\frac{\theta_1(a+M-1)}{M} > 1 > \frac{(1-\theta_2\theta_1^{-1}+M\theta_2\theta_1^{-1})^{M-2}}{(1+M\theta_2\theta_1^{-1})^{M-1}}$$

which gives the upper bond to M as:

$$M < \frac{\theta_1(a-1)}{1-\theta_1} \qquad (4.9)$$

By combining (4.7) and (4.9):

$$\frac{a\theta_1-1}{1-\theta_1} < M < \frac{\theta_1(a-1)}{1-\theta_1}$$

the proof is completed.

Probability generating function

Consul & Shenton (1972, 1974) showed that the derivation of the probability generating function (PGF) of a generalized Poisson variate is not straightforward and is based on the power series expansion of a function in terms of another variable (see GPD by Consul-1989). As they show, the PGF of a generalized Poisson variate $(\theta,\alpha\theta)$ is:

$$G_x(u) = E(u^X) = e^{\theta(t-1)}$$

where $t = ue^{\alpha\theta(t-1)}$, and u is a dummy variable.

Similar to the generalized Poisson variate, the PGF of QNBD also does not seem to be straightforward. Therefore, by compounding the restricted generalized Poisson model $(\theta, \alpha\theta)$ with the gamma distribution $\gamma(a,b)$, where $\theta_1 = b^{-1}$ and $\theta_2 = \alpha b^{-1}$, and using a theorem by Feller (1943), the PGF of a QNBD is:

$$G_x(u) = \frac{b^a}{\Gamma(a)} \int_0^\infty e^{-b\theta} \theta^{a-1} e^{\theta(t-1)} d\theta$$

$$= (1 + \theta_1 b - \theta_1 t)^{-a},\ \theta_1 = b^{-1} \text{ and } \theta_2 = \alpha b^{-1}$$

where $t = ue^{\alpha\theta(t-1)}$.

The function $t(u)$ can be written explicitly using Lagrange's Theorem (see Whittaker and Watson, 1927) as:

$$t = \sum_{n=1}^\infty e^{-n\alpha\theta} \frac{(n\alpha\theta)^{n-1}}{n!} u^n$$

Since θ is varying as gamma distribution $\theta \sim \gamma(a,b)$, the equation above gives:

$$t = \sum_{n=1}^\infty u^n \frac{(n\alpha)^{n-1}}{n!} \frac{b^a}{\Gamma(a)} \int_0^\infty e^{-\theta(b+n\alpha)} \theta^{a+n-2} d\theta$$

And, after simplification results in:

$$t = \sum_{n=1}^\infty u^n \frac{\Gamma(a+n-1)}{\Gamma(a)\ n!} \frac{b^a (n\alpha)^{n-1}}{(b+n\alpha)^{a+n-1}}$$

Taking $\theta_1 = b^{-1}$ and $\theta_2 = \alpha b^{-1}$:

$$t = \sum_{n=1}^\infty u^n \binom{a+n-2}{n-1} \frac{n^{n-2}\, \theta_2{}^{n-1}}{(1+n\theta_2)^{a+n-1}}$$

Hence the PGF of a QNBD (2.4) is:

$$G_x(u) = (1 + \theta_1 b - \theta_1 t)^{-a}$$

where

$$t = \sum_{n=1}^\infty u^n \binom{a+n-2}{n-1} \frac{n^{n-2}\, \theta_2{}^{n-1}}{(1+n\theta_2)^{a+n-1}}.$$

Recurrence relation between the moments

Suppose $\mu'_k(a,\theta_1)$ denotes the r^th^ moment about the origin of a QNBD (2.4), then

$$\mu'_k(a,\theta_1) = \theta_1 \sum_{x=0}^\infty x^k \frac{(a+x-1)!}{(a-1)!x!} \frac{\theta_1(\theta_1+\theta_2 x)^{x-1}}{(1+\theta_1+\theta_2 x)^{a+x}}$$

$$= \theta_1 \sum_{x=1}^\infty x^{k-1} \frac{(a+x-1)!}{(a-1)!(x-1)!} \frac{\theta_1(\theta_1+\theta_2 x)^{x-1}}{(1+\theta_1+\theta_2 x)^{a+x}}$$

taking $x = x+1$ and expanding $(x+1)^{k-1}$ results in:

$$\mu'_k(a,\theta_1) =$$

$$a\theta_1 \sum_{j=0}^{k-1} \binom{k-1}{j} \sum_{x=0}^\infty x^j (\overline{\theta_1+\theta_2} + \theta_2 x)$$

$$\frac{(\overline{a+1}+x-1)!}{(\overline{a+1}-1)!x!} \frac{\theta_1(\overline{\theta_1+\theta_2}+\theta_2 x)^{x-1}}{(1+\overline{\theta_1+\theta_2}+\theta_2 x)^{\overline{a+1}+x}}$$

Converting the above series into $\mu'_k(a,\theta_1)$ functions the recurrence relation

$$\mu'_k(a,\theta_1) = a\theta_1 \sum_{j=0}^{k-1} \binom{k-1}{j} \left[\begin{array}{l} \mu'_j(a+1, \theta_1+\theta_2) + \\ \dfrac{\theta_2}{(\theta_1+\theta_2)} \mu'_{j+1}(a+1, \theta_1+\theta_2) \end{array} \right] \quad (4.10)$$

is obtained.

Where $\mu'_j(a+1, \theta_1+\theta_2)$ is the j^th^ moment about the origin of a QNBD with parameters $(a+1, \theta_1+\theta_2, \theta_2)$. The relation (4.10) is used to determine the moments about the origin of a QNBD. Thus the mean of the distribution is:

$$\mu'_1 = a\theta_1 \left[1 + \frac{\theta_2}{(\theta_1+\theta_2)} \mu'_1(a+1, \theta_1+\theta_2) \right] \quad (4.11)$$

Using (4.10) recursively on the function μ_1', the mean is $\mu_1' = a\theta_1\ {}_2F_0[1,a+1,\_;\theta_2]$, where ${}_2F_0[1,a+1,\_;\theta_2]$ is a hypergeometric function defined by:

$$ {}_2F_0[1,a+1,\_;\theta_2] = \sum_{j=0}^{\infty} 1^{[j]}(a+1)^{[j]} \frac{\theta_2{}^j}{j!}. $$

The second moment about the origin is determined from (4.10) as:

$$ \mu_2' = a\theta_1 \begin{bmatrix} 1+\dfrac{\theta_1+2\theta_2}{\theta_1+\theta_2}\mu_1'(a+1,\theta_1+\theta_2)+ \\ \dfrac{\theta_2}{\theta_1+\theta_2}\mu_2'(a+1,\theta_1+\theta_2) \end{bmatrix} \tag{4.12} $$

Repeated use of (4.12) on the function μ_2' gives:

$$ \mu_2' = a\theta_1\{{}_2F_0[1,a+1,\_;\theta_2]+A_1\} \tag{4.13} $$

where

$$ \begin{aligned} A_1 = {} & \frac{\theta_1+2\theta_2}{\theta_1+\theta_2}\mu_1'(a+1,\theta_1+\theta_2)+(a+1) \\ & \theta_2\frac{\theta+3\theta_2}{\theta_1+2\theta_2}\mu_1'(a+2,\theta_1+2\theta_2) \\ & +(a+1)(a+2)\theta_2{}^2\frac{\theta_1+4\theta_2}{\theta_1+3\theta_2} \\ & \mu_1'(a+3,\theta_1+3\theta_2)+\dots \end{aligned} $$

Repeated use of (4.11) on the function μ_1' gives:

$$ \begin{aligned} A_1 = {} & (\theta_1+2\theta_2)(a+1)\ {}_2F_0[2,a+2,\_;\theta_2]+\theta_2{}^2 \\ & (a+1)(a+2)\ {}_2F_0[3,a+3,\_;\theta_2] \end{aligned} $$

On substituting the value of A_1 in (4.12) the second moment is obtained by:

$$ \begin{aligned} \mu_2' = {} & a\theta_1\ {}_2F_0[1,a+1,\_;\theta_2] \\ & +\theta_1(\theta_1+2\theta_2)a(a+1)\ {}_2F_0[2,a+2,\_;\theta_2] \\ & +\theta_1\theta_2{}^2a(a+1)(a+2)\ {}_2F_0[3,a+3,\_;\theta_2] \end{aligned} $$

Placing $k=3$ in (4.10) the third moment is obtained by:

$$ \mu_3' = a\theta_1 \left[\begin{aligned} & 1+\frac{2\theta_1+3\theta_2}{\theta_1+\theta_2}\mu_1'(a+1,\theta_1+\theta_2) \\ & +\frac{\theta_1+3\theta_2}{\theta_1+\theta_2}\mu_2'(a+1,\theta_1+\theta_2) \\ & +\frac{\theta_2}{\theta_1+\theta_2}\mu_3'(a+1,\theta_1+\theta_2) \end{aligned} \right] \tag{4.14} $$

Repeated use of (4.14) on the function μ_3' gives:

$$ \mu_3' = a\theta_1\{{}_2F_0[1,a+1,\_;\theta_2]+A_2+A_3\} \tag{4.15} $$

where

$$ \begin{aligned} A_2 = {} & \frac{2\theta_1+3\theta_2}{\theta_1+\theta_2}\mu_1'(a+1,\theta_1+\theta_2)+(a+1)\theta_2 \\ & \frac{2\theta_1+5\theta_2}{\theta_1+2\theta_2}\mu_1'(a+2,\theta_1+2\theta_2) \\ & +(a+1)(a+2)\theta_2{}^2 \\ & \frac{2\theta_1+7\theta_2}{\theta_1+3\theta_2}\mu_1'(a+3,\theta_1+3\theta_2)+\dots \end{aligned} \tag{4.16} $$

and

$$ \begin{aligned} A_3 = {} & \frac{\theta_1+3\theta_2}{\theta_1+\theta_2}\mu_2'(a+1,\theta_1+\theta_2)+(a+1)\theta_2 \\ & \frac{\theta_1+4\theta_2}{\theta_1+2\theta_2}\mu_2'(a+2,\theta_1+2\theta_2) \\ & +(a+1)(a+2)\theta_2{}^2 \\ & \frac{\theta_1+5\theta_2}{\theta_1+3\theta_2}\mu_2'(a+3,\theta_1+3\theta_2)+\dots \end{aligned} \tag{4.17} $$

Repeated use of (4.11) in (4.16) gives:

$$ \begin{aligned} A_2 = {} & (2\theta_1+3\theta_2)(a+1)\ {}_2F_0[2,a+2,\_;\theta_2]+2\theta_2{}^2 \\ & (a+1)(a+2)\ {}_2F_0[3,a+3,\_;\theta_2] \end{aligned} $$

Converting μ_2' functions on the right hand side of (4.17) into μ_1' functions by the repeated use of (4.12) and using (4.11) on the function μ_1' gives:

$$\begin{aligned}A_3 = {} & (\theta_1 + 3\theta_2)(a+1)\,{}_2F_0 \\ & [2, a+2, \_;\theta_2] + \theta_2^{\,2} a(a+1)(a+2)\,{}_2F_0[3, a+3, \_;\theta_2] \\ & +(\theta_1^{\,2} + 6\theta_1\theta_2 + 9\theta_2^{\,2})(a+1)(a+2)\,{}_2F_0[3, a+3, \_;\theta_2] \\ & +(3\theta_1\theta_2 + 10\theta_2^{\,2})(a+1)(a+2)(a+3)\,{}_2F_0[4, a+4, \_;\theta_2] \\ & +3\theta_2^{\,4}(a+1)(a+2)(a+3)(a+4)\,{}_2F_0[5, a+5, \_;\theta_2]\end{aligned}$$

Substituting the values of A_2 and A_3 into (4.15) results in:

$$\begin{aligned}\mu_3' = {} & a\theta_1\,{}_2F_0[1, a+1, \_;\theta_2] + \theta_1 \\ & 3(\theta_1 + 2\theta_2)a(a+1)\,{}_2F_0[2, a+2, \_;\theta_2] \\ & +\theta_1(\theta_1^{\,2} + 6\theta_1\theta_2 + 12\theta_2^{\,2})a(a+1) \\ & (a+2)\,{}_2F_0[3, a+3, \_;\theta_2] \\ & +\theta_1\theta_2(3\theta_1\theta_2 + 10\theta_2^{\,2})a(a+1) \\ & (a+2)(a+3)\,{}_2F_0[4, a+4, \_;\theta_2] \\ & +3\theta_1\theta_2^{\,4}a(a+1)(a+2)(a+3) \\ & (a+4)\,{}_2F_0[5, a+5, \_;\theta_2]\end{aligned}$$

Similarly the fourth moment can be determined from (4.10) as:

$$\begin{aligned}\mu_4' = a\theta_1 \Bigg[& 1 + \frac{3\theta_1 + 4\theta_2}{\theta_1 + \theta_2}\mu_1'(a+1, \theta_1 + \theta_2) \\ & + \frac{3(\theta_1 + 2\theta_2)}{\theta_1 + \theta_2}\mu_2'(a+1, \theta_1 + \theta_2) \\ & + \frac{\theta_1 + 4\theta_2}{\theta_1 + \theta_2}\mu_3'(a+1, \theta_1 + \theta_2) \\ & + \frac{\theta_2}{\theta_1 + \theta_2}\mu_4'(a+1, \theta_1 + \theta_2) \Bigg]\end{aligned} \tag{4.18}$$

Repeated use of (4.18) on the function μ_4' gives:

$$\begin{aligned}\mu_4' = a\theta_1 \Bigg[& {}_2F_0[1, a+1;\_, \theta_2] + \\ & \Bigg\{ \frac{3\theta_1 + 4\theta_2}{\theta_1 + \theta_2}\mu_1'(a+1, \theta_1 + \theta_2) \\ & + \frac{3\theta_1 + 7\theta_2}{\theta_1 + 2\theta_2}\theta_2(a+1)\mu_1'(a+2, \theta_1 + 2\theta_2) \\ & + \frac{3\theta_1 + 10\theta_2}{\theta_1 + 3\theta_2}\theta_2^{\,2}(a+1)(a+2)\mu_1'(a+3, \theta_1 + 3\theta_2) + \ldots \Bigg\} \\ & +3\Bigg\{ \frac{\theta_1 + 2\theta_2}{\theta_1 + \theta_2}\mu_2'(a+1, \theta_1 + \theta_2) \\ & + \frac{\theta_1 + 3\theta_2}{\theta_1 + 2\theta_2}\theta_2(a+1)\mu_2'(a+2, \theta_1 + 2\theta_2) \\ & + \frac{\theta_1 + 4\theta_2}{\theta_1 + 3\theta_2}\theta_2^{\,2}(a+1)(a+2)\mu_2'(a+3, \theta_1 + 3\theta_2) + \ldots \Bigg\} \\ & +\Bigg\{ \frac{\theta_1 + 4\theta_2}{\theta_1 + \theta_2}\mu_3'(a+1, \theta_1 + \theta_2) \\ & + \frac{\theta_1 + 5\theta_2}{\theta_1 + 2\theta_2}\theta_2(a+1)\mu_3'(a+2, \theta_1 + 2\theta_2) \\ & + \frac{\theta_1 + 6\theta_2}{\theta_1 + 3\theta_2}\theta_2^{\,2}(a+1)(a+2)\mu_3'(a+3, \theta_1 + 3\theta_2) + \ldots \Bigg\} \Bigg]\end{aligned}$$

Repeated use of (4.11), (4.12), and (4.14) recursively on the functions μ_1', μ_2' and μ_3' respectively with simplifications results in:

$$\begin{aligned}\mu_4' = {} & a\theta_1\,{}_2F_0[1, a+1, \_;\theta_2] + \theta_1 \\ & (7\theta_1 + 14\theta_2)a(a+1)\,{}_2F_0[2, a+2, \_;\theta_2] \\ & +\theta_1(6\theta_1^{\,2} + 36\theta_1\theta_2 + 6\theta_2 + 55\theta_2^{\,2})a(a+1) \\ & (a+2)\,{}_2F_0[3, a+3, \_;\theta_2] \\ & +\theta_1(\theta_1^{\,3} + 3\theta_1\theta_2 + 12\theta_1^{\,2}\theta_2 + 63\theta_1\theta_2^{\,2} \\ & +13\theta_2^{\,2} + 114\theta_2^{\,3})a(a+1)(a+2)(a+3) \\ & {}_2F_0[4, a+4, \_;\theta_2] + \theta_1\theta_2(6\theta_1^{\,2}\theta_2 \\ & +52\theta_1\theta_2^{\,2} + 131\theta_2^{\,3})a(a+1)(a+2) \\ & (a+3)(a+4)\,{}_2F_0[5, a+5, \_;\theta_2] + \theta_1\theta_2^{\,2} \\ & (15\theta_1\theta_2^{\,2} + 70\theta_2^{\,3})a(a+1)(a+2)(a+3) \\ & (a+4)(a+5)\,{}_2F_0[6, a+6, \_;\theta_2] + 15\theta_1\theta_2^{\,6} \\ & a(a+1)(a+2)(a+3)(a+4) \\ & (a+5)(a+6)\,{}_2F_0[7, a+7, \_;\theta_2]\end{aligned}$$

The moments about the origin can be easily verified for the negative-binomial distribution when $\theta_2 = 0$. Further, central moments can be obtained from the moments about origin, thus resulting in the variance:

$$\begin{aligned}\mu_2 = {} & a\theta_1\, {}_2F_0[1,a+1,\_;\theta_2] + \theta_1(\theta_1+2\theta_2) \\ & a(a+1)\, {}_2F_0[2,a+2,\_;\theta_2] \\ & +\theta_1\theta_2{}^2 a(a+1)(a+2)\, {}_2F_0 \\ & [3,a+3,\_;\theta_2] - \left[a\theta_1\, {}_2F_0[1,a+1,\_;\theta_2]\,\right]^2\end{aligned}$$

The third and fourth central moments are coming in messy forms and are not shown here.

Relation with other distributions.

Theorem 5.1: Let X = a quasi-negative-binomial variate with parameters (a, θ_1, θ_2). If $a \to \infty$ such that $a\theta_1 = \alpha$ and $a\theta_2 = \lambda$ show that X tends to generalized Poisson distribution with parameters (α, λ).

Proof:
The QNBD can be expressed as:

$$P(X:x) = \frac{a(a+1)...(a+x-1)}{x!} \frac{\theta_1(\theta_1+\theta_2 x)^{x-1}}{(1+\theta_1+\theta_2 x)^{a+x}} \qquad (5.1)$$

$$= \frac{(1+a^{-1}).......(1+(x-1)a^{-1}\ )}{x!} \frac{(a\theta_1)(a\theta_1+a\theta_2 x)^{x-1}}{1+(a+x)(\theta_1+\theta_2 x)+\dfrac{(a+x)(a+x-1)}{2!}(\theta_1+\theta_2 x)^2+...(\theta_1+\theta_2 x)^{a+x}}$$

Taking limit $a \to \infty$, such that $a\theta_1 = \alpha$ and $a\theta_2 = \lambda$ results in a generalized Poisson distribution with parameters (α, λ) as defined by Consul & Jain (1973).

Theorem 5.2: Let X = a quasi-negative-binomial variate with parameters (a, θ_1, θ_2). If $a \to \infty$ such that $a\lambda^{-1} = \alpha$, show that X tends to the Borel-Tanner distribution.

Proof:
Stating (5.1) as:

$$P(X:x) = \frac{a(a+1)......(a+x-1)}{x!} \frac{r(r+x)^{x-1}\lambda^a}{(r+\lambda+x)^{a+x}}$$

where

$$r = \frac{\theta_1}{\theta_2} \text{ and } \lambda = \frac{1}{\theta_2}.$$

Shifting the support of x from 0 to r, that is, $x = x - r$, results in:

$$P(X:x) = \frac{a(a+1)......(a+x-r-1)}{(x-r)!} \frac{rx^{x-r-1}\lambda^a}{(\lambda+x)^{a+x-r}} \\ x = r, r+1, r+2,...$$

$$= \frac{rx^{x-r-1}}{(x-r)!} \frac{a(a+1)......(a+x-r-1)}{\lambda^{x-r}(1+x\lambda^{-1})^{a+x-r}}$$

Taking the limit $a \to \infty$ in such a way so $a\lambda^{-1} = \alpha$ results in the Borel-Tanner distribution

$$P(X:x) = \frac{r}{(x-r)!} x^{x-r-1}\ e^{-\alpha x}\alpha^{x-r} \quad , \\ x = r, r+1, r+2,... \qquad .$$

Theorem 5.3: Let X = a quasi-negative-binomial variate with parameters (a, θ_1, θ_2). Show that zero-truncated quasi-negative-binomial distribution tends to quasi-logarithmic series distribution as $a \to 0$.

Proof:
The zero-truncated quasi-negative-binomial distribution is

$$P_1(x) = \frac{\Gamma(a+x)}{\Gamma(a)\Gamma(x+1)} \frac{\theta_1(\theta_1+\theta_2 x)^{x-1}}{[1-(1+\theta_1)^{-a}](1+\theta_1+\theta_2 x)^{a+x}}, \quad x=1,2,..... \tag{5.2}$$

Writing:

$$\Gamma(a)\left[1-(1+\theta_1)^{-a}\right] = \Gamma(a)\begin{bmatrix} 1-(1-a\theta_1+\frac{a(a+1)}{2!} \\ \theta_1^2 - \frac{a(a+1)(a+2)}{3!}\theta_1^3+... \end{bmatrix}$$

$$=\Gamma(a+1)\begin{bmatrix} -(\theta_1 - \frac{(a+1)}{2!}\theta_1^2 + \\ \frac{(a+1)(a+2)}{3!}\theta_1^3......) \end{bmatrix} \tag{5.3}$$

Substituting the value from (5.3) into (5.2) and taking limit $a \to 0$ the quasi-logarithmic series distribution is obtained:

$$P_1(x) = \frac{\theta_1(\theta_1+\theta_2 x)^{x-1}}{x[-\log(1-\theta_1)](1+\theta_1+\theta_2 x)^{a+x}}, \quad x=1,2,...$$

Theorem 5.4: If X_1 and X_2 are two independent quasi-negative-binomial variates with parameters $(n_1, \theta_1, \theta_2)$ and $(n_2, \theta_1, \theta_2)$, respectively, then the conditional probability of X_1, given $X_1+X_2=n$, gives a hypergeometric-QNBD.

Proof:

Because X_1 and X_2 are two independent quasi-negative-binomial variates, the conditional probability

$$P\left[X_1 : x \middle/ (X_1+X_2):n\right] = \frac{P(X_1=x, X_2=n-x)}{\sum_{x=0}^{n} P(X_1=x, X_2=n-x)}$$

can be written as

$$P\left[X_1 : x \middle/ (X_1+X_2):n\right]$$

$$\binom{n_1+x-1}{x}\frac{\theta_1(\theta_1+x\theta_2)^{x-1}}{(1+\theta_1+x\theta_2)^{n_1+x}}$$

$$= \frac{\binom{n_2+n-x-1}{n-x}\frac{\theta_1(\theta_1+(n-x)\theta_2)^{n-x-1}}{(1+\theta_1+(n-x)\theta_2)^{n_2+n-x}}}{\binom{n_1+n_2+n-1}{n}\frac{\theta_1(\theta_1+n\theta_2)^{n-1}}{(1+\theta_1+n\theta_2)^{n_1+n_2+n}}}$$

$$= \frac{\binom{n_1+x-1}{x}\binom{n_2+n-x-1}{n-x}}{\binom{n_1+n_2+n-1}{n}}$$

$$\frac{\theta_1(\theta_1+x\theta_2)^{x-1} (\theta_1+(n-x)\theta_2)^{n-x-1}(1+\theta_1+n\theta_2)^{n_1+n_2+n}}{(1+\theta_1+x\theta_2)^{n_1+x} (1+\theta_1+(n-x)\theta_2)^{n_2+n-x}(\theta_1+n\theta_2)^{n-1}}$$

Thus resulting in a new distribution, here called the hypergeometric QNBD. This probability distribution reduces to the classical hypergeometric distribution on $\theta_2 = 0$.

Some characterization.

A number of complex distributions can be reduced to the simpler form QNBD as shown in the following theorems.

Theorem: 6.1. If X is a quasi-inverse Polya variate with parameters (n, a, b, t), and if $b \to \infty$ such that $ab^{-1} = \lambda_1$ and $tb^{-1} = \lambda_2$ show that X approaches to quasi-negative-binomial variate.

Proof:

If X is a quasi-inverse Polya variate with parameters (n, a, b, t), then its probability mass function is:

$$P(X:x)=\frac{n}{n+x}\binom{n+x}{x}$$
$$\frac{a}{a+xt}\frac{(a+xt)^{[x]}(b+xt)^{[n]}}{(a+b+\overline{n+x}\,t)^{[n+x]}}$$
$$x=0,1,2,.....\qquad ,$$

which can be rewritten as:

$$P(X:x)=\frac{n(n+1)....(n+x-1)}{x!}$$
$$\frac{a(a+xt+1)...(a+xt+x-1)\,(b+xt)...(b+xt+n-1)}{(a+b+\overline{n+x}\,t)...(a+b+\overline{n+x}\,t+n+x-1)}\ .$$

Taking limit $b\to\infty$ such that $ab^{-1}=\lambda_1$ and $tb^{-1}=\lambda_2$ results in:

$$P(X:x)=\binom{n+x-1}{x}\frac{\lambda_1(\lambda_1+x\lambda_2)^{x-1}(1+\lambda_1+(n+x)\lambda_2)^{-(n+x)}}{(1+n\lambda_2)^{-n}.}$$

Incorporating $\theta_1=\lambda_1(1+n\lambda_2)^{-1}$ and $\theta_2=\lambda_2(1+n\lambda_2)^{-1}$, the QNBD (2.4) is obtained.

Theorem 6.2: If X is a generalized negative Polya-Eggenberger variate with parameters (n,β,α,γ), and if $\beta\to\infty$ such that $n\beta^{-1}=\lambda_1$ and $\gamma\beta^{-1}=\lambda_2$ show that X approaches to quasi-negative-binomial variate.

Proof:

The generalized negative Polya-Eggenberger distribution with parameters (n,β,α,γ) is:

$$P(X=x)=\frac{n}{(n+\beta x)}\binom{n+\beta x}{x}\frac{a^{[x]}\;b^{[n+\beta x-x]}}{(a+b)^{[n+\beta x]}}\quad ,$$
$$x=0,1,2,.......$$

This can be rewritten as:

$$P(X:x)=$$
$$\frac{n(n+\beta x-1)....(n+\beta x-x+1)}{x!}\qquad(6.1)$$
$$\frac{\alpha^{[x]}\gamma^{[n+\beta x-x]}}{(\alpha+\gamma)^{[n+\beta x]}}$$

Writing

$$\frac{\gamma^{[n+\beta x-x]}}{(\alpha+\gamma)^{[n+\beta x]}}=\frac{\gamma^{[\alpha]}\gamma^{[n+\beta x-x]}}{\gamma^{[n+\beta x-x+\alpha+x]}}$$
$$=\frac{\gamma^{[\alpha]}}{(\gamma+n+\beta x-x)^{[\alpha+x]}}$$
$$=\frac{\gamma(\gamma+1)...(\gamma+\alpha-1)}{(\gamma+n+\beta x-x)...(\gamma+n+\beta x-x+\alpha+x-1)}$$

On substituting this value into (6.1) and taking the limit $\beta\to\infty$, such that $n\beta^{-1}=\lambda_1$ and $\gamma\beta^{-1}=\lambda_2$ results in:

$$P(X:x)=\binom{\alpha+x-1}{x}\frac{\lambda_1(\lambda_1+x)^{x-1}\;\lambda_2^{\alpha}}{(\lambda_1+\lambda_2+x)^{\alpha+x}}$$
$$x=0,1,...$$

Taking $\theta_1=\lambda_1\lambda_2^{-1}$ and $\theta_2=\lambda_2^{-1}$ QNBD (2.4) is obtained.

Theorem 6.3: If X is a quasi-inverse hypergeometric variate with parameters (n, a, b, t), and if $b\to\infty$ such that $ab^{-1}=\lambda_1$ and $tb^{-1}=\lambda_2$, show that X approaches to quasi-negative-binomial variate.

Proof:

If X is a quasi-inverse hypergeometric variate with parameters (n, a, b, t) then its probability mass function is:

$$P(X:x) = \frac{n}{n+x}\,\frac{a}{a+xt}\,\frac{\binom{a+xt}{x}\binom{b+nt}{n}}{\binom{a+b+\overline{n+x}\,t}{n+x}}$$

$$x = 0,1,2,.....$$

Restating this as

$$P(X:x) = \frac{(n+x-1)!}{(n-1)!x!}\,\frac{a}{a+xt}$$

$$\frac{(a+xt-x+1)^{[x]}(b+nt-n+1)^{[n]}}{(a+b+(n+x)t-(n+x)+1)^{[n+x]}},$$

expanding

$$\frac{a}{a+xt}\,\frac{(a+xt-x+1)^{[x]}(b+nt-n+1)^{[n]}}{(a+b+(n+x)t-(n+x)+1)^{[n+x]}}$$

and taking limit $b \to \infty$, such that $ab^{-1} = \lambda_1$ and $tb^{-1} = \lambda_2$ the equation reduces to:

$$P(X:x) = \binom{n+x-1}{x}$$

$$\frac{\lambda_1(\lambda_1 + x\lambda_2)^{x-1}(1+\lambda_1+(n+x)\lambda_2)^{-(n+x)}}{(1+n\lambda_2)^{-n}.}.$$

Taking $\theta_1 = \lambda_1(1+n\lambda_2)^{-1}$ and $\theta_2 = \lambda_2(1+n\lambda_2)^{-1}$, the QNBD (2.4) is obtained.

Theorem 6.4: If X_1 and X_2 are two independent non-negative integer valued random variables such that

$$P\left[X_1 : 0 \big/ (X_1+X_2):x\right] = \frac{n_2^{[n_1]}}{(n_2+x)^{[n_1]}}\,\frac{(1+\theta_1+x\theta_2)^{n_1}}{(1+\theta_1)^{n_1}} \qquad \text{(i)}$$

and

$$P\left[X_1 : 1 \big/ (X_1+X_2):x\right] =$$

$$\frac{xn_2^{[n_1]}}{(n_2+x-1)(n_2+x)^{[n_1]}}$$

$$\frac{n_1\theta_1}{(1+\theta_1+\theta_2)^{n_1+1}}\,\frac{(\theta_1+(x-1)\theta_2)^{x-2}}{(\theta_1+x\theta_2)^{x-1}}$$

$$\frac{(1+\theta_1+x\theta_2)^{n_1+n_2+x}}{(1+\theta_1+(x-1)\theta_2)^{n_2+x-1}} \qquad \text{(ii)}$$

where $n_2^{[n_1]} = n_2(n_2+1)......(n_2+n_1-1)$ and $\theta_1 > 0,\ \theta_2 > 0, n_1 > 0,\ n_2 > 0$,show that X_1 and X_2 are two independent quasi-negative-binomial variates with parameters $(n_1, \theta_1, \theta_2)$ and $(n_2, \theta_1, \theta_2)$ respectively.

Proof:

Let $P(X_1 = x_1) = f(x_1)$ and $P(X_2 = x_2) = g(x_2)$. By condition (i)

$$\frac{f(0)g(x)}{\sum_{t=0}^{x} f(t)g(x-t)} = \frac{n_2^{[n_1]}}{(n_2+x)^{[n_1]}}\,\frac{(1+\theta_1+x\theta_2)^{n_1}}{(1+\theta_1)^{n_1}} \qquad (6.2)$$

and by condition (ii)

$$\frac{f(1)g(x-1)}{\sum_{t=0}^{x} f(t)g(x-t)} = \frac{xn_2^{[n_1]}}{(n_2+x-1)(n_2+x)^{[n_1]}}$$

$$\frac{n_1\theta_1}{(1+\theta_1+\theta_2)^{n_1+1}}$$

$$\frac{(\theta_1+(x-1)\theta_2)^{x-2}}{(\theta_1+x\theta_2)^{x-1}}\,\frac{(1+\theta_1+x\theta_2)^{n_1+n_2+x}}{(1+\theta_1+(x-1)\theta_2)^{n_2+x-1}} \qquad (6.3)$$

Dividing (6.2) by (6.3) results in:

$$\frac{f(0)g(x)}{f(1)g(x-1)} = \frac{(n_2+x-1)}{x}\,\frac{(1+\theta_1+\theta_2)^{n_1+1}}{n_1\theta_1(1+\theta_1)^{n_1}}\,\frac{(\theta_1+x\theta_2)^{x-1}}{(\theta_1+(x-1)\theta_2)^{x-2}}\,\frac{(1+\theta_1+(x-1)\theta_2)^{n_2+x-1}}{(1+\theta_1+x\theta_2)^{n_2+x}}$$

which gives a recurrence relation

$$g(x) = \frac{f(1)}{f(0)}\,\frac{(1+\theta_1+\theta_2)^{n_1+1}}{n_1\theta_1(1+\theta_1)^{n_1}}\,\frac{(n_2+x-1)}{x}\,\frac{(\theta_1+x\theta_2)^{x-1}}{(\theta_1+(x-1)\theta_2)^{x-2}}\,\frac{(1+\theta_1+(x-1)\theta_2)^{n_2+x-1}}{(1+\theta_1+x\theta_2)^{n_2+x}}\,g(x-1)$$

Repeated use of the equation above gives:

$$g(x) = \left[\frac{f(1)}{f(0)}\,\frac{(1+\theta_1+\theta_2)^{n_1+1}}{n_1\theta_1(1+\theta_1)^{n_1}}\right]^x \frac{(n_2+x-1)....n_2}{x(x-1)......1}\,\frac{(\theta_1+x\theta_2)^{x-1}}{\theta_1^{-1}}\,\frac{(1+\theta_1)^{n_2}}{(1+\theta_1+x\theta_2)^{n_2+x}}\,g(0)$$

Substituting $\dfrac{f(1)}{f(0)} = \dfrac{n_1\theta_1(1+\theta_1)^{n_1}}{(1+\theta_1+\theta_2)^{n_1+1}}$ results in:

$$g(x) = \binom{n_2+x-1}{x}\frac{(n_2+x-1)....n_2}{x(x-1)......1}\,\frac{\theta_1(\theta_1+x\theta_2)^{x-1}}{(1+\theta_1+x\theta_2)^{n_2+x}}\,(1+\theta_1)^{n_2}\,g(0)$$

f the above relation represents a probability mass function, then $\sum_x g(x) = 1 \Rightarrow g(0) = (1+\theta_1)^{-n_2}$ and this reduces the equation above to

$$g(x) = \binom{n_2+x-1}{x}\frac{(n_2+x-1)...n_2}{x(x-1)...1}\,\frac{\theta_1(\theta_1+x\theta_2)^{x-1}}{(1+\theta_1+x\theta_2)^{n_2+x}}$$
$$x = 0,1,2,...$$

This is a quasi-negative-binomial distribution with parameters $(n_2, \theta_1, \theta_2)$. Similarly it can be shown that $f(x)$ also represents a quasi-negative-binomial distribution with parameters $(n_1, \theta_1, \theta_2)$.

Goodness of Fit

Due to its complicated likelihood function, the maximum likelihood estimate of the parameters of the proposed distribution are not straightforward and require some iterative procedure such as Fisher's scoring method or the Newton-Rampson method for their solution. R-software provides one such solution. In R-software there exists the function nlm, which minimizes the negative log-likelihood function or equivalently maximizes the log likelihood function for estimating the parameters of the distribution by adopting the Newton-Rampson iterative procedure. A random start procedure is employed, that is, for a set of random starting points, the function nlm searches recursively until global maxima is reached. To verify that the global maximum has been found the gradient should be equal to zero. The closer the value of the random starting points to the ML estimate, the lesser number of iterations will be required to obtain the global maximum.

Two data sets examine the fitting of the proposed model and compare it with the negative binomial distribution and generalized negative binomial distribution defined by Jain & Consul (1971). A computer program was developed using R-software to estimate the parameters of the distribution by using the nlm function. The ML estimates of the parameters so obtained are shown at the bottom of the tables. It is evident from tables 4.1 and 4.2 that, in all cases, the Chi-square values of the proposed model give the best fit as compared to other distributions.

Table 4.1: Absenteeism among shift-workers in steel industry; data from Arbous & Sichel, 1954

Count	Observed Frequency	Expected Frequencies		
		NBD	GNBD Jain & Consul's (1971)	QNBD Proposed Model
0	7	12.02	10.51	10.47
1	16	16.16	17.45	16.05
2	23	17.77	20.38	18.55
3	20	18.08	20.80	19.19
4	23	17.65	19.88	18.72
5	24	16.80	18.34	17.63
6	12	15.72	16.56	16.24
7	13	14.52	14.78	14.74
8	09	13.28	13.08	13.23
9	09	12.06	11.53	11.80
10	08	10.89	10.13	10.46
11	10	09.78	08.89	9.25
12	08	08.75	07.79	8.15
13	07	07.80	16.83	7.18
14	02	06.93	05.99	6.31
15	12	06.14	05.26	5.55
16	03	05.43	04.61	4.88
17	05	04.79	04.05	4.30
18	04	04.22	03.56	3.79
19	02	03.17	03.14	3.34
20	02	03.23	02.76	2.94
21	05	02.86	02.43	2.60
22	05	02.50	02.15	2.30
23	02	02.91	01.90	2.04
24	01	01.91	01.68	1.81
25-48	16	12.77	13.50	16.48
TOTAL	248	248	248	248

ML Estimate		p=0.854
		n=1.576
			p=0.00010775
			β =5978.5288
			n=29337.08391
				$a = 2.0034559$
				$\theta_1 = 3.8528528$
				$\theta_2 = 0.0609776$

χ^2		14.92	27.79	11.18
d.f.		17	16	16

Table 4.2: Counts of numbers of European red mites on apple leaves; data from Garman, 1951

Count	Observed Frequency	Expected Frequencies		
		NBD	GNBD Jain & Consul's (1971)	QNBD Proposed Model
0	70	69.49	69.49	70.91
1	38	37.60	37.60	33.93
2	17	20.10	20.10	20.07
3	10	10.70	10.70	12.01
4	09	05.69	05.69	6.89
5	03	03.02	03.02	3.63
6	02	01.60	01.60	1.69
7	01	00.85	00.85	0.65
8	00	00.95	00.95	0.22
TOTAL	150	150	150	150

ML Estimates		p=0.5281
		n=1.0246
			p=0.52810
			β =1.000
			n=1.0246
				$a = 0.6268217$
				$\theta_1 = 2.3046227$
				$\theta_2 =$ -0.1785658

χ^2 d.f.		2.484 3	2.484 2	1.957 2

References

Charalambidies, C. A. (1990). Abel series distributions with applications to fluctuations of sample functions of stochastic processes, *Communications in Statistics- Theory and Methods*, *19*, 317-335.

Consul, P. C. (1974). *A simple urn model dependent on predetermined strategy*, *Sankhya, B, 36*, 391-399

Consul, P. C. (1989). *Generalized Poisson distributions properties and applications*, Marcel Dekker, Inc.: New York, NY.

Consul, P. C., & Famoye, F. (2006). *Lagrangian Probability Distributions*, Birkhauser: Boston, MA.

Consul, P. C., & Jain, G.C. (1973). A generalization of the Poisson distribution, *Techno-metrics*, *15*(4), 791-799.

Consul, P. C., & Shenton, L.R. (1972). Use of Lagrangian expansion for generating discrete generalized probability distributions, *SIAM J. Appl. Math.*, *23*(2), 239-248.

Consul, P. C., & Shenton, L.R. (1974). *On the probabilistic structure and properties of discrete Lagrangian distributions, statistical distribution in scientific work*, Vol. 1, G.P. Patil, S. Kotz, & J. K. Ord, Eds. D. Reidel Company: Boston, MA.

Feller, W. (1943): On a general class of contagious distributions, *Ann. Math. Stat. 14*, 389-400.

Gupta, R. C., & Ong, S. H. (2004). A new generalization of the negative binomial distribution, *Computational Statistics and Data Analysis*, *45*, 287-300.

Haight, F. A., & Brever, M. A. (1960). The Borel-Tanner distribution, *Biometrika*, *47*, 143-150.

Holgate, P. (1970). The modality of compound Poisson distribution, *Biometrika*, *57*, 665-667.

Janardan, K. G. (1975). Markov-Polya urn-model with pre-determined strategies, *Gujarat Statist. Rev.*, *2* (1), 17-32.

Johnson, N. L., & Kotz, S. (1992). *Univariate Discrete distributions*, second edition, John Willy & Sons, Inc.

Nand, S. B., & Das, K. K. (1994). *A family of Abel series distribution*, *Sankhya, B, 56*(2), 147-164.

Sen, K. & Jain, R. (1996). Generalized Markov-Polya urn-model with pre-determined strategies. *J. Statist. Plann. Infer.*, *54*, 119-133.

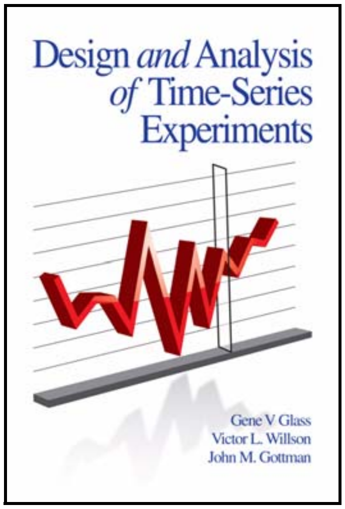
Whittaker, E. T., & Watson, G. N. (1927). *A course of modern analysis*, Cambridge University Press, Cambridge, MA.

It's Back!

Design and Analysis of Time-Series Experiments

(with a new Introduction by the first author)

Gene V Glass, *Arizona State University*
Victor L. Willson, *Texas A&M University*
John M. Gottman, *The Gottman Institute, Seattle, Washington*



Hailed as a landmark in the development of experimental methods when it appeared in 1975, *Design and Analysis of Time-Series Experiments* is available again after several years of being out of print.

Gene V Glass, Victor L. Willson and John M. Gottman have carried forward the design and analysis of perhaps the most powerful and useful quasi-experimental design identified by their mentors in the classic Campbell & Stanley text *Experimental and Quasi-experimental Design for Research* (1966). In an era when governments seek to resolve questions of experimental validity by fiat and the label "Scientifically Based Research" is appropriated for only certain privileged experimental designs, nothing could be more appropriate than to bring back the classic text that challenges doctrinaire opinions of proper causal analysis.

Glass, Willson & Gottman introduce and illustrate an armamentarium of interrupted time-series experimental designs that offer some of the most powerful tools for discovering and validating causal relationships in social and education policy analysis. Drawing on the ground-breaking statistical analytic tools of Box & Jenkins, the authors extend the comprehensive autoregressive-integrated-moving-averages (ARIMA) model to accommodate significance testing and estimation of the effects of interventions into real world time-series. Designs and full statistical analyses are richly illustrated with actual examples from education, behavioral psychology, and sociology.

Publication Date:
Winter 2009

ISBN's:
Paperback: 978-1-59311-980-5
Hardcover: N/A

Price:
Paperback: $39.99
Hardcover:

Trim Size: 6 X 9

Subject:
Education, Statistics

"…this book will come to be viewed as a true landmark. … [It] should stand the test of time exceedingly well." ~ James A. Walsh (*Educational & Psychological Measurement*, 1975)

"Ordinary least squares estimation is usually inapplicable because of autoregressive error…. Glass, Willson, and Gottman have assembled the best approach." ~Donald T. Campbell

Special Price: $25.99 paperbacks plus s/h

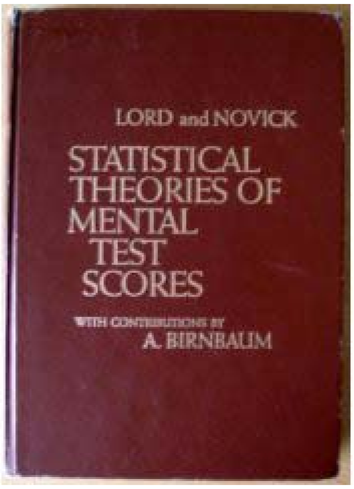
Book URL: http://www.infoagepub.com/products/content/p489c9049a428d.php

***IAP - Information Age Publishing,** PO Box 79049, Charlotte, NC 28271*
tel: 704-752-9125 fax: 704-752-9113 URL: www.infoagepub.com

It’s Back in Paperback!

Statistical Theories of Mental Test Scores

by **Frederic M. Lord** and **Melvin R. Novick**



A classic returns.

> …pioneering work…
> …comprehensive…
> …classic…
> …definitive…
> …unquestioned status and authority…

Tatsuoka was right:

> "This comprehensive and authoritative work is a major contribution to the literature of test theory. Without doubt it is destined to become a classic in the field." ~ Maurice Tatsuoka (1971)

One of the most important books in the history of psychometrics has been virtually unavailable to scholars and students for decades. A gap in the archives of modern test theory is now being filled by the release in paperback for the first time of the classic text, *Statistical Theories of Mental Test Scores,* by the late and honored statisticians and psychometricians, Frederic M. Lord and Melvin R. Novick. No single book since 1968 when Lord & Novick first appeared has had a comparable impact on the practice of testing and assessment.

Publication Date:
Spring 2008

ISBN’s:
Paperback: **978-1-59311-934-8**

Price:
Paperback: $59.99

Trim Size: 6 X 9

Subject:
Education, Statistics

Information Age Publishing is proud to make this classic text available to a new generation of scholars and researchers.

http://www.infoagepub.com/products/content/p4810c9a0891af.php

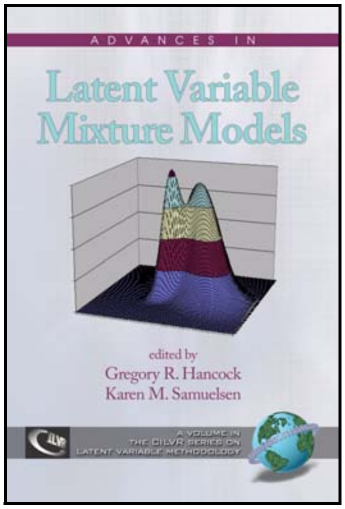
***IAP - Information Age Publishing,** PO Box 79049, Charlotte, NC 28271*
tel: 704-752-9125 fax: 704-752-9113 URL: www.infoagepub.com

INFORMATION AGE PUBLISHING, INC.
IAP

New Book Information

Advances in Latent Variable Mixture Models

Edited by **Gregory R. Hancock,** *University of Maryland, College Park,*
and **Karen M. Samuelsen,** *University of Georgia*



The current volume, *Advances in Latent Variable Mixture Models*, contains chapters by all of the speakers who participated in the 2006 CILVR conference, providing not just a snapshot of the event, but more importantly chronicling the state of the art in latent variable mixture model research. The volume starts with an overview chapter by the CILVR conference keynote speaker, Bengt Muthén, offering a "lay of the land" for latent variable mixture models before the volume moves to more specific constellations of topics. Part I, *Multilevel and Longitudinal Systems*, deals with mixtures for data that are hierarchical in nature either due to the data's sampling structure or to the repetition of measures (of varied types) over time. Part II, *Models for Assessment and Diagnosis*, addresses scenarios for making judgments about individuals' state of knowledge or development, and about the instruments used for making such judgments. Finally, Part III, *Challenges in Model Evaluation*, focuses on some of the methodological issues associated with the selection of models most accurately representing the processes and populations under investigation. It should be stated that this volume is not intended to be a first exposure to latent variable methods. Readers lacking such foundational knowledge are encouraged to consult primary and/or secondary didactic resources in order to get the most from the chapters in this volume. Once armed with that basic understanding of latent variable methods, we believe readers will find this volume incredibly exciting.

CONTENTS: Editors' Introduction, *Gregory R. Hancock and Karen M. Samuelsen.* Acknowledgments. Latent Variable Hybrids: Overview of Old and New Models, *Bengt Muthén.* **PART I: Multilevel and Longitudinal Systems.** Multilevel Mixture Models, *Tihomir Asparouhov and Bengt Muthén.* Longitudinal Modeling of Population Heterogeneity: Methodological Challenges to the Analysis of Empirically Derived Criminal Trajectory Profiles, *Frauke Kreuter and Bengt Muthén.* Examining Contingent Discrete Change Over Time with Associative Latent Transition Analysis, *Brian P. Flaherty.* Modeling Measurement Error in Event Occurrence for Single, Non-Recurring Events in Discrete-Time Survival Analysis, *Katherine E. Masyn.* **PART II: Models for Assessment and Diagnosis.** Evidentiary Foundations of Mixture Item Response Theory Models, *Robert J. Mislevy, Roy Levy, Marc Kroopnick, and Daisy Rutstein.* Examining Differential Item Functioning from a Latent Mixture Perspective, *Karen M. Samuelsen.* Mixture Models in a Developmental Context, *Karen Draney, Mark Wilson, Judith Glück, and Christiane Spiel.* Applications of Stochastic Analyses for Collaborative Learning and Cognitive Assessment, *Amy Soller and Ron Stevens.* The Mixture General Diagnostic Model, *Matthias von Davier.* **PART III: Challenges in Model Evaluation.** Categories or Continua? The Correspondence Between Mixture Models and Factor Models, *Eric Loken and Peter Molenaar.* Applications and Extensions of the Two-Point Mixture Index of Model Fit, *C. Mitchell Dayton.* Identifying the Correct Number of Classes in Growth Mixture Models, *Davood Tofighi and Craig K. Enders.* Choosing a "Correct" Factor Mixture Model: Power, Limitations, and Graphical Data Exploration, *Gitta H. Lubke and Jeffrey R. Spies.* About the Contributors.

Publication Date:
Fall 2007

ISBN's:
Paperback: **978-1-59311-847-1**
Hardcover: **978-1-59311-848-8**

Price:
Paperback: $39.99
Hardcover: $73.99

Trim Size: 6 X 9

Subject:
Education

Books of Related Interest:

Structural Equation Modeling: A Second Course
http://www.infoagepub.com/products/content/1-59311-015-4.php
2006 Paperback ISBN: 1-59311-014-6 $39.99 Hardcover ISBN: 1-59311-015-4 $73.95

Real Data Analysis
http://infoagepub.com/products/content/978-1-59311-565-4.php
2007 Paperback ISBN: 978-1-59311-564-7 $39.99 Hardcover ISBN: 978-1-59311-565-4 $73.95

IAP
IAP - Information Age Publishing, PO Box 79049, Charlotte, NC 28271
tel: 704-752-9125 fax: 704-752-9113 URL: www.infoagepub.com

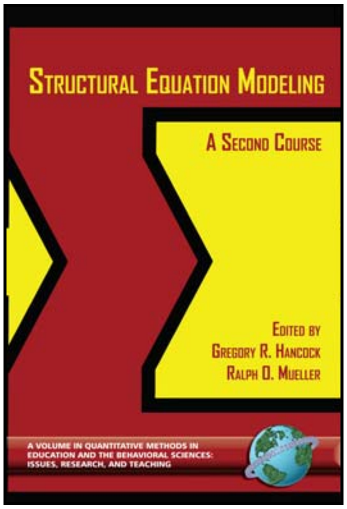
New Book Information

Structural Equation Modeling: A Second Course

Edited by **Gregory R. Hancock**, *University of Maryland*
and **Ralph O. Mueller,** *The George Washington University*

A volume in **Quantitative Methods in Education and the Behavioral Sciences: Issues, Research, and Teaching**
Series Editor **Ron Serlin,** *University of Wisconsin*

(sponsored by the Educational Statisticians, SIG)



"I believe that this volume represents a vital contribution to the field of SEM beyond the introductory level."

From the Preface by
Richard G. Lomax, *The University of Alabama*

This volume is intended to serve as a didactically-oriented resource covering a broad range of advanced topics often not discussed in introductory courses on structural equation modeling (SEM). Such topics are important in furthering the understanding of foundations and assumptions underlying SEM as well as in exploring SEM as a potential tool to address new types of research questions that might not have arisen during a first course. Chapters focus on the clear explanation and application of topics, rather than on analytical derivations, and contain syntax and partial output files from popular SEM software.

CONTENTS: Introduction to Series, *Ronald C. Serlin*. Preface, *Richard G. Lomax*. Dedication. Acknowledgements. Introduction, *Gregory R. Hancock & Ralph O. Mueller.* **Part I: Foundations.** The Problem of Equivalent Structural Models, *Scott L. Hershberger*. Formative Measurement and Feedback Loops, *Rex B. Kline*. Power Analysis in Covariance Structure Modeling, *Gregory R. Hancock*. **Part II: Extensions.** Evaluating Between-Group Differences in Latent Variable Means, *Marilyn S. Thompson & Samuel B. Green*. Using Latent Growth Models to Evaluate Longitudinal Change, *Gregory R. Hancock & Frank R. Lawrence*. Mean and Covariance Structure Mixture Models, *Phill Gagné*. Structural Equation Models of Latent Interaction and Quadratic Effects, *Herbert W. Marsh, Zhonglin Wen, & Kit-Tai Hau*. **Part III: Assumptions.** Nonnormal and Categorical Data in Structural Equation Modeling, *Sara J. Finney & Christine DiStefano*. Analyzing Structural Equation Models with Missing Data, *Craig K. Enders*. Using Multilevel Structural Equation Modeling Techniques with Complex Sample Data, *Laura M. Stapleton*. The Use of Monte Carlo Studies in Structural Equation Modeling Research, *Deborah L. Bandalos*. About the Authors.

Publication Date:
2005

ISBN's:
Paperback: **1-59311-014-6**
Hardcover: **1-59311-015-4**

Price:
Paperback: $39.99
Hardcover: $73.99

Subject:
Education, Statistics

Also Available:

Multilevel Modeling of Educational Data

2008 Paperback ISBN: 978-1-59311-684-2 $39.99 Hardcover ISBN: 978-1-59311-685-9 $73.99

Real Data Analysis

2007 Paperback ISBN: 978-1-59311-564-7 $39.99 Hardcover ISBN: 978-1-59311-565-4 $73.99

Series URL: http://www.infoagepub.com/products/series/serlin.html

IAP - Information Age Publishing, PO Box 79049, Charlotte, NC 28271
tel: 704-752-9125 fax: 704-752-9113 URL: www.infoagepub.com

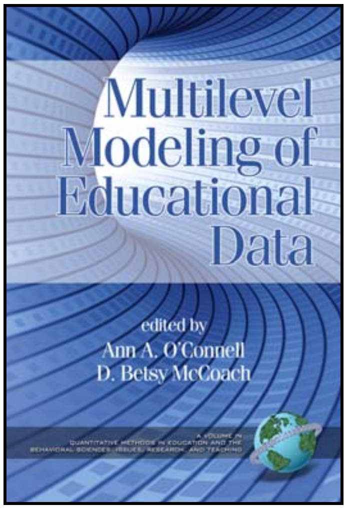
New Book Information

Multilevel Modeling of Educational Data

Edited by **Ann A. C'Connell,** *Ohio State University*
and **D. Betsy McCoach,** *University of Connecticut*

A volume in **Quantitative Methods in Education and the Behavioral Sciences: Issues, Research, and Teaching**
Series Editor **Ron Serlin,** *University of Wisconsin*

(sponsored by the Educational Statisticians, SIG)



Multilevel Modeling of Educational Data, co-edited by Ann A. O'Connell, Ed.D., and D. Betsy McCoach, Ph.D., is the next volume in the series: *Quantitative Methods in Education and the Behavioral Sciences: Issues, Research and Teaching* (Information Age Publishing), sponsored by the Educational Statisticians' Special Interest Group (Ed-Stat SIG) of the American Educational Research Association. The use of multilevel analyses to examine effects of groups or contexts on individual outcomes has burgeoned over the past few decades. Multilevel modeling techniques allow educational researchers to more appropriately model data that occur within multiple hierarchies (i.e.- the classroom, the school, and/or the district). Examples of multilevel research problems involving schools include establishing trajectories of academic achievement for children within diverse classrooms or schools or studying school-level characteristics on the incidence of bullying. Multilevel models provide an improvement over traditional single-level approaches to working with clustered or hierarchical data; however, multilevel data present complex and interesting methodological challenges for the applied education research community.

In keeping with the pedagogical focus for this book series, the papers this volume emphasize applications of multilevel models using educational data, with chapter topics ranging from basic to advanced. This book represents a comprehensive and instructional resource text on multilevel modeling for quantitative researchers who plan to use multilevel techniques in their work, as well as for professors and students of quantitative methods courses focusing on multilevel analysis. Through the contributions of experienced researchers and teachers of multilevel modeling, this volume provides an accessible and practical treatment of methods appropriate for use in a first and/or second course in multilevel analysis. A supporting website links chapter examples to actual data, creating an opportunity for readers to reinforce their knowledge through hands-on data analysis. This book serves as a guide for designing multilevel studies and applying multilevel modeling techniques in educational and behavioral research, thus contributing to a better understanding of and solution for the challenges posed by multilevel systems and data.

Publication Date:
Spring 2008

ISBN's:
Paperback: **978-1-59311-684-2**
Hardcover: **978-1-59311-685-9**

Price:
Paperback: $39.99
Hardcover: $73.99

Subject:
Education, Statistics

Also Available:

Real Data Analysis

2007 Paperback ISBN: 978-1-59311-564-7 $39.99 Hardcover ISBN: 978-1-59311-565-4 $73.95

Structural Equation Modeling: A Second Course

2005 Paperback ISBN: 1-59311-014-6 $39.99 Hardcover ISBN: 1-59311-015-4 $73.95

Series URL: http://www.infoagepub.com/products/series/serlin.html

Book URL: http://www.infoagepub.com/products/content/p478cb9504908a.php

IAP - Information Age Publishing, PO Box 79049, Charlotte, NC 28271
tel: 704-752-9125 fax: 704-752-9113 URL: www.infoagepub.com

New Book Information

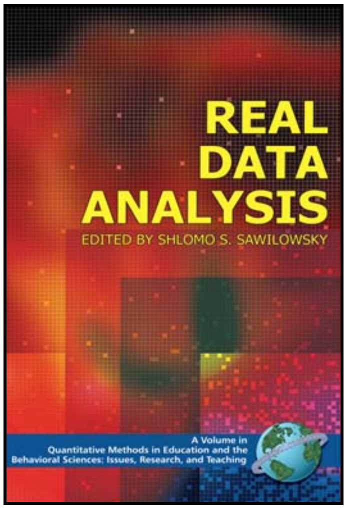
Real Data Analysis

Edited by **Shlomo S. Sawilowsky,** *Wayne State University*

A volume in **Quantitative Methods in Education and the Behavioral Sciences: Issues, Research, and Teaching**

Series Editor **Ron Serlin,** *University of Wisconsin*

(sponsored by the Educational Statisticians, SIG)


REAL DATA ANALYSIS
EDITED BY SHLOMO S. SAWILOWSKY
A Volume in Quantitative Methods in Education and the Behavioral Sciences: Issues, Research, and Teaching

The invited authors of this edited volume have been prolific in the arena of Real Data Analysis (RDA) as it applies to the social and behavioral sciences, especially in the disciplines of education and psychology. Combined, this brain trust represents 3,247 articles in refereed journals, 127 books published, US $45.3 Million in extramural research funding, 34 teaching and 92 research awards, serve(d) as Editor/Assistant Editor/Editorial Board Member for 95 peer reviewed journals, and provide(d) ad hoc reviews for 362 journals. Their enormous footprint on real data analysis is showcased for professors, researchers, educators, administrators, and graduate students in the second text in the AERA/SIG ES Quantitative Methods series.

CONTENTS:

Publication Date:
2007

ISBN's:
Paperback: **978-1-59311-564-7**
Hardcover: **978-1-59311-565-4**

Price:
Paperback: $39.99
Hardcover: $73.99

Subject:
Education, Statistics

Also Available:

Multilevel Modeling of Educational Data

2008 Paperback ISBN: 978-1-59311-684-2 $39.99 Hardcover ISBN: 978-1-59311-685-9 $73.99

Structural Equation Modeling: A Second Course

2005 Paperback ISBN: 1-59311-014-6 $39.99 Hardcover ISBN: 1-59311-015-4 $73.99

Series URL: http://www.infoagepub.com/products/series/serlin.html

IAP - Information Age Publishing, PO Box 79049, Charlotte, NC 28271
tel: 704-752-9125 fax: 704-752-9113 URL: www.infoagepub.com

Journal of Modern Applied Statistical Methods
November, 2008, Vol. 7, No. 2, 368-631

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@edstat.coe.wayne.edu. 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 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 Illustrator or 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.
8. 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 sections. Number all formulas, tables, figures, and graphs, but do not use italics, bold, or underline. Do not number references. Do not use footnotes or endnotes.
9. 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.
10. *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" instead of "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://tbf.coe.wayne.edu/jmasm. Mail subscription requests with remittances to JMASM, P. O. Box 48023, Oak Park, MI, 48237. Email journal correspondence, other than manuscript submissions, to jmasm@edstat.coe.wayne.edu.

Notice To Advertisers

Send requests for advertising information to jmasm@edstat.coe.wayne.edu.