DIGITALCOMMONS
—@WAYNESTATE—

Journal of Modern Applied Statistical Methods

Volume 14 | Issue 2 Article 3

11-1-2015

Vol. 14, No. 2 (Full Issue)

JMASM Editors

Recommended Citation

Editors, JMASM (2015) "Vol. 14, No. 2 (Full Issue)," *Journal of Modern Applied Statistical Methods*: Vol. 14 : Iss. 2 , Article 3.
DOI: 10.22237/jmasm/1446350520
Available at: http://digitalcommons.wayne.edu/jmasm/vol14/iss2/3

Journal of Modern Applied Statistical Methods
November 2015, Vol. 14, No. 2, ii-v

ISSN 1538 − 9472

Journal of
Modern Applied Statistical Methods

JMASM (ISSN 1538−9472, http://digitalcommons.wayne.edu/jmasm) is an independent, open access electronic journal, published biannually in May and November by JMASM Inc. (PO Box 48023, Oak Park, MI, 48237) in collaboration with the Wayne State University Library System. *JMASM* seeks to publish (1) new statistical tests or procedures, or the comparison of existing statistical tests or procedures, using computer-intensive Monte Carlo, bootstrap, jackknife, or resampling methods, (2) the study of nonparametric, robust, permutation, exact, and approximate randomization methods, and (3) applications of computer programming, preferably in Fortran (all other programming environments are welcome), related to statistical algorithms, pseudo- random number generators, simulation techniques, and self-contained executable code to carry out new or interesting statistical methods.

Journal correspondence (other than manuscript submissions) and requests for advertising may be forwarded to ea@jmasm.com. See back matter for instructions for authors.

בס"ד
Journal of Modern Applied Statistical Methods
November 2015, Vol. 14, No. 2, ii-vi

ISSN 1538 − 9472

Journal of Modern Applied Statistical Methods

Vol. 14, No. 2

⁓ November 2015 ⁓

Table of Contents

Invited Articles

Regular Articles

Copyright © 2015 JMASM, Inc.
ISSN 1538 − 9472

Statistical Software Applications and Review

Algorithms and Code

Journal of Modern Applied Statistical Methods
November 2015, Vol. 14, No. 2, 2-8.

ISSN 1538 − 9472

Invited Article
Inferences About the Skipped Correlation Coefficient: Dealing with Heteroscedasticity and Non-Normality

Rand Wilcox
University of Southern California
Los Angeles, CA

A common goal is testing the hypothesis that Pearson's correlation is zero and typically this is done based on Student's T test. There are, however, several well- known concerns. First, Student's T is sensitive to heteroscedasticity. That is, when it rejects, it is reasonable to conclude that there is dependence, but in terms of making a decision about the strength of the association, it is unsatisfactory. Second, Pearson's correlation is not robust: it can poorly reflect the strength of the association. Even a single outlier can have a tremendous impact on the usual estimate of Pearson's correlation, which can result in a poor indication of the strength of the association among the bulk of the points. Numerous robust correlation coefficients have been proposed that deal with outliers among the marginal distributions, but these methods do not take into account the overall structure of the data in terms of dealing with outliers. A skipped correlation addresses this concern and methods for testing the hypothesis that this correlation is zero have been studied. However, there are serious limitations associated with one of these methods and extant studies regarding an alternative percentile bootstrap method do not address practical concerns reviewed in the paper. A minor goal is to report situations where this percentile bootstrap method can be unsatisfactory. The main result is that an alternative percentile bootstrap method performs well in simulations.

Keywords: Robust measures of association, level robust methods, non-normality, heteroscedasticity

Introduction

A basic goal is testing the hypothesis that the strength of the association between two random variables is zero. Certainly the best-known strategy is to test the hypothesis that Pearson's correlation is zero, using Student's T test.

Dr. Wilcox is Professor of Psychology at the University of Southern California. Email him at rwilcox@usc.edu.

$$H_0 : \rho = 0 \tag{1}$$

There are, however, well known concerns with this approach. First, Student's T assumes homoscedasticity. In practical terms, it provides a reasonable test of the hypothesis that two variables are independent, but in terms of making inferences about ρ, it can be unsatisfactory. For example, even when the null hypothesis is true, the probability of rejecting can increase as the sample size increases when there is heteroscedasticity (e.g., Wilcox, 2012). Roughly, the reason is that Student's T uses the wrong standard error when there is heteroscedasticity, given the goal of testing (1).

Another concern is that r, the usual estimate of ρ, is not robust. Even a single outlier can result in a poor reflection of the strength of the association among the bulk of the points. Numerous robust estimators have been proposed for dealing with outliers among the marginal distributions (e.g., Wilcox, 2012, chapter 9). Certainly the two best-known approaches are Kendall's tau and Spearman's rho. But a known concern with these measures of association is that they do not deal with outliers in a manner that takes into account the overall structure of the data. That is, based on the random sample (X_1, Y_1), …, (X_n, Y_n), situations are encountered where no outliers are detected among $X_1, \ldots, X_n$, ignoring Y, and no outliers are detected among Y_1, …, Y_{n}, ignoring X, yet there are outliers that can have a substantial impact on Kendall's tau, Spearman's rho and other measures of association that do not deal with the overall structure of the data (e.g., Wilcox, 2012, chapter 9). A measure of the strength of an association that deals with this issue is the skipped correlation coefficient. The basic strategy is to use some outlier detection method that takes into account the overall structure of the data, remove any outliers that are found, and then compute Pearson's correlation using the remaining data.

There are many outlier detection methods that take into account the overall structure of the data. In the context of a skipped correlation, a projection type outlier detection method has been the focus of attention. No single outlier detection method dominates, but the projection-type method used here appears to perform relatively well in terms of avoiding masking and detecting truly unusual points (e.g., Wilcox, 2012). Masking refers to missing outliers due to their very presence. For example, in the univariate case, detecting outliers using the mean and standard deviation can result in masking. The basic problem is that outliers inflate the sample standard deviation, which in turn can result is missing even extreme outliers.

Based on the projection type method for detecting outliers, let ξ denote the population analog of the skipped correlation and consider the goal of testing

$$H_0 : \xi = 0 \tag{2}$$

A very simple approach is described in Wilcox (2012, Section 9.4.4). However, the method is limited to testing at the $\alpha = 0.05$ level and it assumes homoscedasticity. More recently, Pernet, Wilcox and Rousselet (2013) studied a bootstrap method when sampling from a bivariate normal distribution. But the impact of non-normality and heteroscedasticity was not addressed. A minor goal in this paper is to report results indicating situations where the Pernet et al. method can be unsatisfactory when dealing with non-normality and heteroscedasticity. The primary goal is to report simulation results on an alternative bootstrap method that provides good control over the Type I error probability for a broader range of situations.

Description of the methods to be compared

This section describes the projection outlier detection method followed by the two percentile bootstrap methods that were studied when testing (2). For brevity, just an outline of the method is provided. Complete computational details can be found in Wilcox (2012, section 6.4.9). Included is an R function called outpro for applying it, which is used here.

The projection method begins by estimating the center of the data cloud, say $\hat{\theta}$. Here this is done using the marginal medians. Then for fixed i, project all n points onto the line connecting $\hat{\theta}$ and (X_i, Y_i). Based on the projected points, let D_j ($j = 1, \ldots, n$) be the distance between the projection of (X_j, Y_j) and the center, $\hat{\theta}$. Next, check for outliers using the usual boxplot rule based on the D_j values. That is, if q_1 and q_2 are estimates of the lower and upper quartiles, respectively, based on $D_1, \ldots, D_n$, declare D_j an outlier if $D_j < 1.5(q_2 - q_1)$ or if $D_j > 1.5(q_2 - q_1)$, in which case (X_j, Y_j) is declared an outlier as well. This process is performed for each i ($i = 1, \ldots, n$) and (X_j, Y_j) is declared an outlier if its projected distance is flagged as an outlier for any i.

The percentile bootstrap method used by Pernet et al. (2013) is applied as follows:

1. Remove any points flagged as outliers using the projection method. Let m denote the sample size after outliers are removed.
2. Generate a bootstrap sample from the remaining data by resampling with replacement m points.
3. Compute Pearson's correlation based on this bootstrap sample yielding r^*.
4. Repeat steps 2-3 and B times yielding $r_1{}^*, \ldots, r_B{}^*$.
5. Put the values $r_1{}^*, \ldots, r_B{}^*$ in ascending order and label the results $r^*_{(1)} \leq \cdots \leq r^*_{(B)}$.
6. Let $l = \alpha B/2$, rounded to the nearest integer and $u = B - 1$. Then the $1 - \alpha$ confidence interval for ξ is taken to be $(r_{(l+1)}, r_{(u)})$. This will be called method B1 henceforth.

An unusual feature of method B1 is that the process of generating bootstrap samples does not exactly mimic the manner in which the data are generated and the skipped correlation is computed. A percentile bootstrap method that does mimic the way data are generated, labeled method B2 here, begins by generating a bootstrap sample from all n points, removing any points flagged as outliers and then computing $\hat{\xi}^*$, Pearson's correlation based on the remaining data. That is, in the description of method B1, replace steps 1-3 with

1. Generate a bootstrap sample by resampling with replacement n points from the entire sample of size n.
2. Remove any points from the bootstrap sample in step 1 that are flagged as outliers using the projection method.
3. Compute Pearson's correlation using the points not flagged as outliers in step 2.

As done in step 4 of method B1, this process is repeated B times only now the results are labeled $\hat{\xi}^*_1, \cdots, \hat{\xi}^*_B$. The $1 - \alpha$ confidence interval for ξ is taken to be $\left(\hat{\xi}^*_{(l+1)}, \hat{\xi}^*_u\right)$.

It is noted that a p-value is readily computed when testing (2), which is motivated by general results in Liu and Singh (1997). Let Q^* be the proportion of $\hat{\xi}^*$ values that are less than zero. Then a p-value is $p = \min(2Q^*, (1 - 2Q^*)$.

Simulation results

Four types of distributions are considered: normal, symmetric and heavy-tailed (roughly meaning that outliers tend to be common), asymmetric and relatively light-tailed, and asymmetric and relatively heavy-tailed. More specifically, g-and-h distributions (Hoaglin, 1985) are used, which arise as follows. Let *Z* be a random variable having a standard normal distribution and let

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

If $g = 0$

$$W = Z\exp\left(h\frac{Z^2}{2}\right)$$

Then *W* has a g-and-h distribution, where *g* and *h* are parameters that determine the first four moments. The four distributions used here are the standard normal ($g = h = 0$), a symmetric heavy-tailed distribution ($h = .2$, $g = 0$), an asymmetric distribution with relatively light tails ($h = 0$, $g = .2$), and an asymmetric distribution with heavy tails ($g = h = .2$). Table 1 summarizes the skewness (γ_1) and kurtosis (γ_2) of these distributions.

The number of bootstrap samples was taken to be $B = 1000$. Bradley (1978) suggests that as a general guide, when testing at the .05 level, the actual level should be between .025 and .075. Preliminary simulations based on $B = 500$ indicated that method B2 does not satisfy this criterion; increasing *B* to 1000 gave more satisfactory results.

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

g	***h***	$\boldsymbol{\kappa_2}$	$\boldsymbol{\kappa_1}$
0.0	0.0	0.00	3.00
0.0	0.2	0.00	21.46
0.2	0.0	0.61	3.68
0.2	0.2	2.81	155.98

Observations were generated according to the model $Y = \lambda(X)\varepsilon$, where both *X* and ε have one of the g-and-h distributions in Table 1 and $\lambda(X)$ is used to model heteroscedasticity. Three choices for $\lambda(X)$ were used: $\lambda(X) \equiv 1$ (homosecdasticity), $\lambda(X) = |X| + 1$ (so the conditional variance of Y, given X, is smallest when X is close to its mean), and $\lambda(X) = 1/(|X| + 1)$ (in which case the conditional variance of Y, given X, is largest when X is close to its mean. For convenience these three choices for λ will be called variance patterns (VP) 1, 2 and 3, respectively.

The simulation estimates of the actual Type I error probabilities were based on 2,000 replications. A common suggestion is that ideally, simulation estimates be based on 10,000 replications. However, when using method B2, a single replication takes a little over 14 seconds using the software R on a MacBook Pro with a 2.5 GHz processor. So 10,000 replications would require over 38 hours of execution time. To add perspective on the precision of the estimates, assuming Bradley's criterion is reasonable, consider the issue of whether the actual level is less than or equal .075. Using the method in Pratt (1968), it can be seen that based on a two-sided .95 confidence interval for the actual level, the confidence interval will not contain .075 if $\hat{\alpha} \leq .063$. In a similar manner, based on a two-sided .95 confidence interval, the confidence interval for the actual level does not contain .025 if $\hat{\alpha} \geq .0325$.

Table 2. Estimated Type I error probabilities, n = 40, α = .05

g	*h*	VP	B2	B1
0.0	0.0	1	0.022	0.066
		2	0.022	0.071
		3	0.028	0.055
0.0	0.2	1	0.022	0.070
		2	0.024	0.080
		3	0.024	0.046
0.2	0.0	1	0.027	0.066
		2	0.024	0.072
		3	0.030	0.056
0.2	0.2	1	0.021	0.072
		2	0.024	0.080
		3	0.022	0.045

Table 2 shows the estimated Type I error probabilities when $n = 40$ and $\alpha = .05$. As can be seen, method B2 tends to be conservative, meaning that the estimated Type I error probability is always less than the nominal .05 level. The estimates are consistently close to .025 over all of the situations considered. So there is some possibility that the actual level drops below .025, but there is no strong indication that this is the case. In contrast, the estimates using method B1 are always greater than or equal to .05 with the two largest estimates equal to .08. So all indication are that in terms of avoiding a Type I error probability greater than the nominal level, B2 performs better than B1.

Concluding remarks

Some positive features of method B1 are that it reduces execution time compared to method B2 and it performs reasonably well in simulations when there is homoscedasticity and sampling is from a bivariate normal distribution. For most situations, it was estimated that the actual level using method B1 is less than .075, but for variance pattern VP 2 this is not the case when dealing with distributions with heavy-tails. In contrast, method B2 avoids Type I error probabilities greater than .05 among all of the situations considered, the only concern being that the actual level was estimated to be as low as .022 with a sample size of $n = 40$. That is, there is some possibility that B2 does not satisfy Bradley's criterion that the actual level should be at least .025. The main result for the goal of avoiding an actual level well above .05, all indications are that B2 is preferable to B1.

References

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

Hoaglin, D. C. (1985) Summarizing shape numerically: The g-and-h distributions. In D. Hoaglin, F. Mosteller and J. Tukey (Eds.) *Exploring Data Tables, Trends, and Shapes.* (pp. 461-515). New York: Wiley.

Liu, R. G. & Singh, K. (1997). Notions of limiting P values based on data depth and bootstrap. *Journal of the American Statistical Association, 9*(2), 266–277.

Pernet, C. R., Wilcox, R. & Rousselet, G A. (2013). Robust correlation analyses: false positive and power validation using a new open source Matlab toolbox. *Frontiers in Quantitative Psychology and Measurement*. doi:10.3389/fpsyg.2012.00606

Pratt, J. W. (1968). A normal approximation for binomial, F, beta, and other common, related tail probabilities, I. *Journal of the American Statistical Association*, *63*, 1457–1483.

Wilcox, R. R. (2012). *Introduction to robust estimation and hypothesis testing*. (3rd Ed.). San Diego, CA: Academic Press.

Journal of Modern Applied Statistical Methods
November 2015, Vol. 14, No. 2, 9-26.

ISSN 1538 − 9472

Invited Article
Resolving the Issue of How Reliability Is Related to Statistical Power: Adhering to Mathematical Definitions

Donald W. Zimmerman
Carleton University
Ottawa, ON, CAN

Bruno D. Zumbo
University of British Columbia
Vancouver, BC, CAN

Reliability in classical test theory is a population-dependent concept, defined as a ratio of true-score variance and observed-score variance, where observed-score variance is a sum of true and error components. On the other hand, the power of a statistical significance test is a function of the total variance, irrespective of its decomposition into true and error components. For that reason, the reliability of a dependent variable is a function of the ratio of true-score variance and observed-score variance, whereas statistical power is a function of the sum of the same two variances. Controversies about how reliability is related to statistical power often can be explained by authors' use of the term "reliability" in a general way to mean "consistency," "precision," or "dependability," which does not always correspond to its mathematical definition as a variance ratio. The present note shows how adherence to the mathematical definition can help resolve the issue and presents some derivations and illustrative examples that have further implications for significance testing and practical research.

Keywords: Reliability, power, hypothesis test, error of measurement, true score, error score, observed score, difference score

The relation between the reliability of measurement, as the concept is defined in classical test theory, and the power of statistical hypothesis tests, has been investigated for many years and has engendered controversy that has not been

Professor Donald W. Zimmerman passed away in December 2013 at age 82. His scholarly work was known for a precision of thought and a desire for mathematical rigor, never shying away from controversial topics. This is one of the last collaborative papers with Professor Zumbo.

Dr. Zumbo is Paragon UBC Professor of Psychometrics & Measurement with affiliate appointments in the Department of Statistics and the Institute of Applied Mathematics at the University of British Columbia, and an Associate Editor of this journal. Email him at bruno.zumbo@ubc.ca.

completely resolved. Overall & Woodward (1975, 1976) observed that the paired-samples *t* test based on difference scores can under some conditions have maximum power when the reliability of differences is zero. That finding led to discussion as to how the power of the *t* test and other familiar hypothesis tests depends on the reliability of dependent variables in experiments (Cleary & Linn, 1959; Collins, 1996; Feldt & Brennan, 1989; Fleiss, 1976; Hopkins & Hopkins, 1979; Kopriva & Shaw, 1991; Levin, 1986; Mellenbergh, 1996, 1999; Subkoviak & Levin, 1977; Sutcliffe, 1958; Zimmerman & Williams, 1986; Zimmerman, Williams, & Zumbo, 1993), with presentation of various inconsistent points of view.

The methods introduced by Cohen (1988) have been applied widely to calculate the power of familiar hypothesis tests used in educational and psychological research. In the case of tests based on the normal distribution, such as the Student *t* and ANOVA *F* tests, those methods provide a good approximation to exact results obtained from noncentral *t* and *F* distributions. However, the concept of test reliability and validity defined in classical test theory has not been employed in power analysis with the same degree of precision (see Thomas & Zumbo, 2012).

Researchers and test users often associate the concept of reliability with terms such as dependability, precision, repeatability, and so on, assuming they are consistent with the mathematical definition in classical test theory. The classical definition is based on the decomposition of scores in a population of individuals into true scores and error scores and the relative variability of those components. In the traditional theory, each individual's test score is a sum of a true score and an error score, $X = T + E$, and the total variance (or observed-score variance) with respect to a population of individuals is a sum of the variances of the components, $\sigma_X^2 = \sigma_T^2 + \sigma_E^2$. Finally, *reliability* is defined as the ratio of the true-score variance and the total variance, $\rho = \sigma_T^2 / \sigma_X^2 = \sigma_T^2 / \left(\sigma_T^2 + \sigma_E^2\right)$, or equivalently as $\rho^2\left(X, T_X\right)$, the squared correlation between observed scores and true scores (Gulliksen, 1950; Novick, 1966; Lord & Novick, 1968). It is also worth noting that the numerical value of reliability can always be found solely from the ratio of σ_T and σ_E, although the combined values of the two standard deviations may differ in size. This can be seen by defining $\psi = \sigma_T / \sigma_E$ and dividing both the numerator and denominator of $\sigma_T^2 / \left(\sigma_T^2 + \sigma_E^2\right)$ by $\sigma_T\ \sigma_E$ to obtain $\rho = \psi / \left(\psi + \psi^{-1}\right)$.

The fact that reliability in classical test theory is a *population-dependent* concept has been emphasized by Mellenbergh (1996, 1999). The concept does not apply to an individual examinee, and this fact is important in considering statistical power. Because reliability is defined as a *ratio* of two components of variance with respect to a population, a given numerical value of reliability can be associated with many different combinations of values of true-score variance and error-score variance. That fact has been at the root of many problems in analyzing how reliability is related to statistical power.

Reliability and variance heterogeneity

A familiar formula in classical test theory enables one to find reliability in one population with a particular observed-score variance when knowing reliability in another population with a different observed-score variance. The formula is

$$\rho_2 = 1 - \frac{\sigma^2_{X_1}}{\sigma^2_{X_2}}\left(1 - \rho_1\right) \tag{1}$$

where the subscripts 1 and 2 denote the respective populations. This equation was derived under the assumption that the change in observed-score variance is accounted for by a change in true-score variance, while error-score variance remains constant (Gulliksen, 1950, p 111; Lord & Novick, 1968, p 130).

In contrast to the familiar approach, if a change in observed-score variance is accounted for by a change in error-score variance, while true-score variance remains constant, the results are described by the equation

$$\rho_2 = \frac{\sigma^2_{X_1}}{\sigma^2_{X_2}}\rho_1 \tag{2}$$

which can be derived easily, although equation (1) is prominent in test theory. Whether it is more reasonable to regard a difference in the observed scores of two groups as resulting from different true-score variances or different error-score variances is problematic. Curiously, test theorists have assumed constant error-score variance in deriving equation (1), but when considering how reliability influences statistical power, have adopted implicitly the assumption underlying the relatively unknown equation (2).

It is well understood in statistics that the power of an hypothesis test is inversely proportional to the variance of any dependent variable, assuming that other determinants, including significance level, sample size, and directionality of the hypothesis, remain constant. Expressed otherwise, the power of an hypothesis test is inversely proportional to the *observed-score variance* considered in test theory, irrespective of how that variance is partitioned into true score variance and error-score variance. For this reason, if observed-score variance does not change, the power of a significance test remains the same, even when the value of the reliability coefficient changes extensively over a wide range.

Although equations (1) and (2) show how reliability changes as observed-score variance changes, for present purposes in considering statistical power, we need just the reverse, that is, equations showing how observed-score variance changes as reliability changes. Simply rearranging equations (1) and (2), we can write

$$\frac{\sigma_{X_2}^2}{\sigma_{X_1}^2} = \frac{1-\rho_1}{1-\rho_2}, \text{ and} \tag{3}$$

$$\frac{\sigma_{X_2}^2}{\sigma_{X_1}^2} = \frac{\rho_1}{\rho_2} \tag{4}$$

These forms show immediately that, if error-score variance is constant, observed-score variance is proportional to reliability, and, if true-score variance is constant, observed-score variance is inversely proportional to reliability. In turn, because of what is known about power functions, that means that, if error-score variance is constant, statistical power is inversely proportional to reliability, and, if true-score variance is constant, statistical power is directly proportional to reliability.

It is possible for a test to have high reliability and still have low power, or, conversely, to have low reliability and have high power (see, for example, the paradox originally discussed by Overall and Woodward (1975, 1976) in the context of difference scores). Furthermore, it is possible for the same reliability coefficient to be associated with different degrees of power and for different reliability coefficients to result in the same power.

A simple example illustrates some possibilities. Table 1 compares hypothetical tests, each having a large number of scores with distributions like those shown in the table. In section A, the test on the left apparently has high true scores and low error scores, so that its reliability might be expected to be high, but, because the *variance* of T_1 is much higher than that of E_1, reliability is only .096. In the test on the right, the reverse is true, and the reliability is .904, even though the true scores at first glance look small. Nevertheless, despite the difference in reliability, the two tests have the same statistical power, because the observed-score variances are the same. In section B, the two tests have the same reliability, .645, because the variances of *T* and *E*, although different, have the same ratio. However, the observed-score variances are different, and the statistical power of the test on the left is greater.

Table 1. A) Score components of two tests having substantially different reliability coefficients and the same statistical power; B) Score components of two tests having the same reliability coefficients and substantially different statistical power.

A						**B**					
Score Components			**Score Components**			**Score Components**			**Score Components**		
T_1	**E_1**	**X_1**	**T_2**	**E_2**	**X_2**	**T_1**	**E_1**	**X_1**	**T_2**	**E_2**	**X_2**
100	1	101	0	99	99	50	5	55	100	10	110
101	6	107	5	100	105	52	6	58	104	12	116
100	2	102	1	99	100	51	4	55	102	8	110
102	7	109	6	101	107	53	6	59	106	12	118
101	2	103	1	100	101	52	4	56	104	8	112
100	7	107	6	99	105	50	6	56	100	12	112
102	1	103	0	101	101	53	5	58	106	10	116
100	6	106	5	99	104	51	6	57	102	12	114
Variance of T_1 − 0.786			Variance of T_2 − 7.429			Variance of T_1 − 1.429			Variance of T_2 − 5.714		
Variance of E_1 − 7.429			Variance of E_2 − 0.786			Variance of E_1 − 0.786			Variance of E_2 − 3.143		
Variance of X_1 − 8.214			Variance of X_2 − 8.214			Variance of X_1 − 2.214			Variance of X_2 − 8.857		
Reliability − .096			Reliability − .904			Reliability − .645			Reliability − .645		

Power as a composite function of reliability

For investigating the relation of reliability and power, it is more convenient to examine changes in reliability with changes in true-score variance and error-score variance, as opposed to changes in observed-score variance as given by equations (1) and (2). It is then possible to express observed-score variance as a 1-1 function of reliability, provided either true-score variance or error-score variance is held constant. Then, because power is a 1-1 function of observed-score variance, it is possible in turn to express power as a composite function. Under those conditions, power is a monotonic decreasing function of observed-score variance and a monotonic increasing or decreasing function of reliability depending on which component is constant. Of course, the form of the functions depends on properties of the particular hypothesis test considered.

First, begin with the equations $\rho_1 = 1 - \sigma^2_{E_1} / \left(\sigma^2_T + \sigma^2_{E_1} \right)$ and $\rho_2 = 1 - \sigma^2_{E_2} / \left(\sigma^2_T + \sigma^2_{E_2} \right)$, solve both for σ^2_T, assumed to be constant, and set the two expressions equal. The result is

$$\frac{\rho_1 \sigma^2_{E_1}}{1 - \rho_1} = \frac{\rho_2 \sigma^2_{E_2}}{1 - \rho_2}$$

Then, solving for ρ_2 gives the result

$$\rho_2 = \frac{1}{1 - \dfrac{\sigma^2_{E_2}}{\sigma^2_{E_1}} \left(1 - \dfrac{1}{\rho_1} \right)} \tag{5}$$

This equation indicates how reliability changes as the variance of the error component changes, while the true-score variance remains fixed.

Alternatively, if σ^2_T changes while σ^2_E is constant, a similar derivation give $\rho_1 = \sigma^2_{T_1} / \left(\sigma^2_{T_1} + \sigma^2_E \right)$ and $\rho_2 = \sigma^2_{T_2} / \left(\sigma^2_{T_2} + \sigma^2_E \right)$, so that $\sigma^2_{T_1} \left(1 - \rho_1 \right) / \rho_1 = \sigma^2_{T_2} \left(1 - \rho_2 \right) / \rho_2$. Solving for ρ_2 gives the result

$$\rho_2 = \frac{1}{1 - \dfrac{\sigma^2_{T_1}}{\sigma^2_{T_2}} \left(1 - \dfrac{1}{\rho_1} \right)} \tag{6}$$

This equation indicates how reliability changes as true-score variance changes, while error-score variance is constant. Equations (5) and (6) clearly indicate that changes in reliability resulting from changes in either true-score variance or error-score variance depend only on the *ratios* $\sigma^2_{E_2} / \sigma^2_{E_1}$ or $\sigma^2_{T_1} / \sigma^2_{T_2}$ relating the old and new score components and not on the individual variances considered separately.

Changes in observed score variability and power with changes in reliability

Table 2 contains results found from equations (5) and (6). The first row at the top, labeled "Initial ρ" is the value of the reliability coefficient, denoted by ρ_1 in the equations, and the entries in the right-hand section of the table are the values of the new reliability coefficient, ρ_2, after a designated change in the error-score variance or true-score variance. The ratio of old-to-new error-score variance, $\sigma^2_{E_1} / \sigma^2_{E_2}$, is located in the first column, and the entry in the table gives the value of the new reliability after the change, assuming that true-score variance remains constant. The same entry in the table is also the value of the new reliability if a change shown by the adjacent entry in the second column is made in the ratio $\sigma^2_{T_1} / \sigma^2_{T_2}$, assuming that error-score variance remains constant. That is, the ratios in the second columns are inverses of those in the first column, and the same change in reliability corresponds to both ratios.

Table 2. Modification of reliability and observed-score variance by changes in error-score variance ($\sigma^2_{E_1} / \sigma^2_{E_2}$) and in true-score variance ($\sigma^2_{T_1} / \sigma^2_{T_2}$). Entries in the five right-hand columns are the modified reliability values (ρ_2) corresponding to variances and variance ratios in the first four columns.

				Initial Reliability (ρ_1)				
$\sigma^2_{E_1} / \sigma^2_{E_2}$	σ^2	$\sigma^2_{T_1} / \sigma^2_{T_2}$	σ^2	.10	.30	.50	.70	.90
0.250	5.000	4.000	1.250	.027	.097	.200	.368	.692
0.286	4.500	3.500	1.286	.031	.109	.222	.400	.720
0.333	4.000	3.000	1.333	.036	.125	.250	.438	.750
0.400	3.500	2.500	1.400	.043	.146	.286	.483	.783
0.500	3.000	2.000	1.500	.053	.176	.333	.538	.818
0.667	2.500	1.500	1.667	.069	.222	.400	.609	.857
1.000	2.000	1.000	2.000	.100	.300	.500	.700	.900
1.500	1.667	0.667	2.500	.143	.391	.600	.778	.931
2.000	1.500	0.500	3.000	.182	.462	.667	.824	.947
2.500	1.400	0.400	3.500	.217	.517	.714	.854	.957
3.000	1.333	0.333	4.000	.250	.562	.750	.875	.964
3.500	1.286	0.286	4.500	.280	.600	.778	.891	.969
4.000	1.250	0.250	5.000	.308	.632	.800	.903	.973

The values of ρ_2 in the right-hand section always *increase* as values of $\sigma^2_{E_1} / \sigma^2_{E_2}$ increase and also as those of $\sigma^2_{T_1} / \sigma^2_{T_2}$ *decrease*. At the same time, the values of σ^2 decrease (and therefore power increases), as those of ρ_1 increase, and vice versa. Also, the same values of ρ_2 are associated with different values of σ^2 (and therefore power).

The relationship can be seen in more detail by plotting graphs of some power functions obtained from simulations. Figure 1 plots power functions of the one-sample Student *t* test under conditions where reliability was either increased or reduced by changing one component of the observed-score variance while the other remained constant. These simulations were programmed using *Mathematica*, version 4.1 (Wolfram, 1999), together with *Mathematica* statistical add-on packages. The program performed *t* tests on sums of "true-score" and "error-score" random variables, selected from $N(0,1)$ and multiplied by constants in order to determine means, variances, and reliabilities. The means increased in increments of $.32\sigma$, and each data point in the figure was found from 20,000 iterations of the sampling procedure.

In both sections of the figure, the true-score and error-score variances were initially equal, so that reliability was .50. The middle curves with filled circles represent these initial reliabilities. In the upper section, reliability was increased to .80 in two ways. In the top curve in that section (triangular symbols), error-score variance was reduced, while true-score variance was constant. In the lower section (square symbols), true-score variance was increased while error-score variance was constant.

In the lower graph, reliability was decreased to .20 in two ways. In the top curve (square symbols), true-score variance was reduced while error-score variance was constant. In the lower curve (triangular symbols), error-score variance was increased while true-score variance was constant. All these curves, with shapes typical of power curves, show that the sum of the two variance components, that is, the observed-score variance, determined the power of the hypothesis test irrespective of how reliability changed as a result of a change in the ratio of the two components.

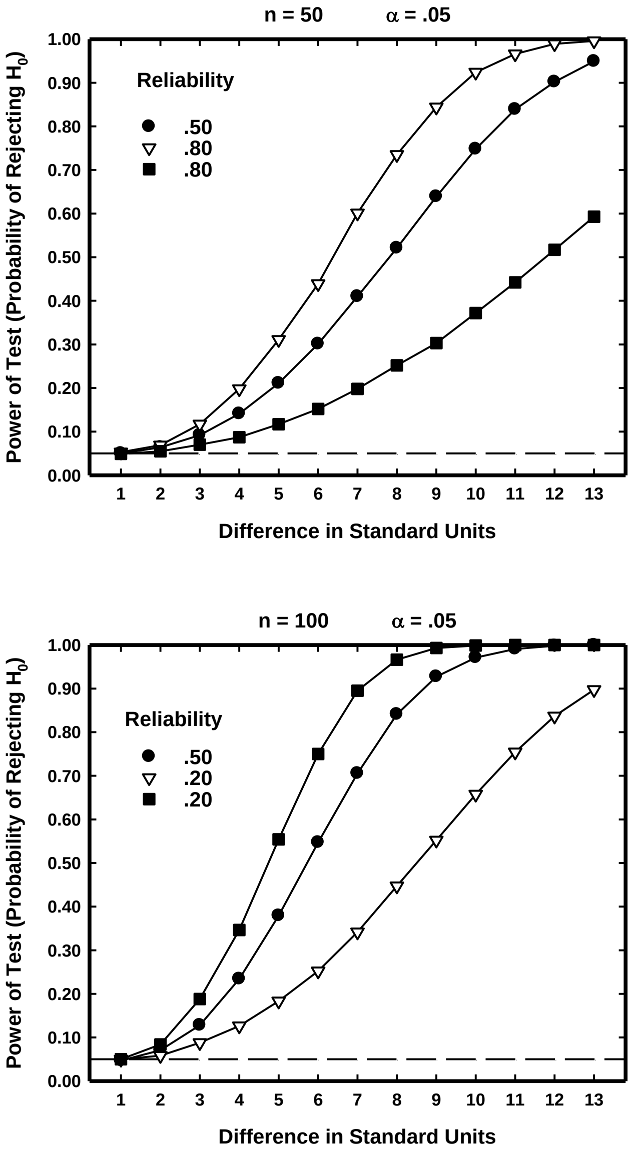


Figure 1. Power functions of the one-sample t test when reliability was increased or decreased by changing component variances. Upper graph: reliability was increased from .50 to .80. The middle curve is for ρ = .50. In the upper curve, error-score variance was reduced while true-score variance remained constant. In the lower curve, true-score variance was increased while error-score variance remained constant. Lower graph: Reliability was reduced from .50 to .20. The middle curve is for ρ = .50. In the upper curve, true-score variance was reduced while error-score variance remained constant. In the lower curve, error-score variance was increased while true-score variance remained constant.

Relations, functions, and composite functions

It is well known that statistical power is a function of several variables, some of which are under the direct control of an experimenter. These include sample size, N, the significance level, α, and the directionality of the hypothesis tested. Of course, different hypothesis tests, parametric and nonparametric, have different power characteristics under various conditions. The relations between N and power and between α and power are functional when the other variables are held constant; that is, each value in the domain of the relation is associated with a single value in its range. Some authors have considered it reasonable to add reliability to the list of determinants. However, as we have seen, reliability influences power only to the extent that it influences observed-score variance.

The association between reliability and power, therefore, is a mathematical relation, but it is not a *function* or a *functional relation*. However, it becomes functional if the variance of one of the two variables determining reliability is held constant. In that case, if the variance of one score component is held constant, power is a composite of two functions, the one between a score component and observed-score variance, and the one between observed-score variance and power. The range of the first function is the domain of the second.

As said before, still another way to express the same relationship is that, all other things equal, statistical power is a function of the sum of the variances of T and E, whereas reliability is a function of the ratio of those two variances. As noted earlier, reliability can be defined as $\psi/(\psi+\psi^{-1})$, where $\psi = \sigma_T/\sigma_E$. That definition makes it clear that reliability can be either large or small at the same time the sum, which determines power, is either large or small, independently of the ratio. The fact that power is determined by the observed-score variance, which is comprised of the sum in the denominator of the expression $\rho = \sigma_T^2 / \left(\sigma_T^2 + \sigma_E^2\right)$ shows that, for a fixed value of σ_E^2, power has its maximum value when $\rho = 0$. But for a fixed value of σ_T^2 power has a maximum when $\rho = 1$.

Reliability of difference scores and statistical power

In order to gain insight into paradoxes concerning difference scores, we shall pursue an approach similar to the above. Rather than directly seeking a relationship between the reliability of differences and the power of an hypothesis test employing differences, we first consider how both are related to observed-score variance and also the reliability coefficients of the two variables determining the differences.

Once again, beginning with what is known, the power of tests on difference scores, $X - Y$, is certainly a decreasing function of the variance of the difference scores. However, reliability depends on partitioning that variance into true and error components and finding ratios, which in turn depend on the similar ratios of both *X* and *Y*. In all cases, both reliability and the power of an hypothesis test can be considered joint functions of the true-score variance and error-score variance of the difference scores. However, power is determined uniquely by their sum and reliability by their ratio, just as in the case of a single variable *X*.

A familiar equation is

$$\rho_D = \frac{\sigma^2_{T_D}}{\sigma^2_D} = \frac{\sigma^2_{T_X} + \sigma^2_{T_Y} - 2\rho_{T_X T_Y}\sigma_{T_X}\sigma_{T_Y}}{\sigma^2_X + \sigma^2_Y - 2\rho_{XY}\sigma_X\sigma_Y} \tag{7}$$

where $D = X - Y$, T_X and T_Y are the true score components of *X* and *Y*, and ρ_D is the reliability of *D*. If $\sigma^2_{T_X} = \sigma^2_{T_Y}$ and $\sigma^2_{E_X} = \sigma^2_{E_Y}$, this equation can be solved for σ^2_D and substitutions made using $\rho_X = \sigma^2_{T_X} / \left(\sigma^2_{T_X} + \sigma^2_{E_X}\right)$. The result is

$$\sigma^2_D = \frac{2\sigma^2_{T_X}}{\rho_X}\left(1 - \rho_{T_X T_Y}\rho_X\right) \tag{8}$$

and an equivalent result is

$$\sigma^2_D = 2\left[\sigma^2_T\left(1 - \rho_{T_X T_Y}\right) + \sigma^2_E\right] \tag{9}$$

Although the assumption that variances of *X* and *Y* are equal is often unrealistic in practice, it suffices to indicate the form of the relation between reliability and statistical power. Next, the reliability of differences can be written in the form

$$\rho_D = \frac{\rho_X\left(1 - \rho_{T_X T_Y}\right)}{1 - \rho_{T_X T_Y}\rho_X}, \text{ or} \tag{10}$$

$$\rho_D = \frac{\sigma_T^2\left(1-\rho_{T_X T_Y}\right)}{\sigma_T^2\left(1-\rho_{T_X T_Y}\right)+\sigma_E^2} \qquad (11)$$

Equation (10) indicates that, if $\rho_{T_X T_Y} = 0$, the reliability of differences is the same as the common reliability of the components.

Equations (8), (9), (10), and (11) have the desirable feature that all combinations of values of the variables on the right-hand side of the equation yield meaningful values of ρ_D and σ_D^2. That is not true in the case of several well-known formulas that involve both ρ_{XY} and ρ_X, because the Cauchy-Schwarz inequality places limits on the values the two can have together (Zumbo, 1999). For example, the relation $\rho_D = \left(\rho_X - \rho_{XY}\right)/\left(1-\rho_{XY}\right)$ is not meaningful for all values of ρ_{XY} and ρ_X.

The above equations provide a convenient way to exhibit the relation between the reliability of differences and statistical power. Table 3 shows results of calculations using equations (9) and (11), comparing the reliability of component scores (ρ_X), the reliability of difference scores (ρ_D), and the observed variance of difference scores (σ_D^2), as a function of σ_T^2 while σ_E^2 is constant (upper section) and of σ_E^2 while σ_T^2 is constant (lower section).

If σ_E^2 is fixed, an increase in ρ_X comes from an increase in σ_T^2, and if σ_T^2 is fixed, it comes from a reduction in σ_E^2. Those outcomes are apparent in the table: As σ_T^2 increased from 0 to 1.8, the reliability coefficients ρ_X and ρ_D both increased, and also the variance of observed scores increased, so that statistical power *decreased*. The same was true for all three values of the correlation between true scores, $\rho(T_X,T_Y)$. On the other hand, as σ_E^2 increased from 0 to 1.8, ρ_X and ρ_D both decreased, but the variance of observed scores still increased, so that power again *decreased*. As σ_T^2 varied, power was greatest when the reliability of differences was 0. However, as σ_E^2 varied, power was greatest when the reliability of differences was 1.

Table 3. Changes in observed variance and reliability of difference scores associated with changes in reliability of component scores.

		$\rho(T_X,T_Y) = -.60$			$\rho(T_X,T_Y) = 0$			$\rho(T_X,T_Y) = .60$		
	σ_T^2	ρ_X	ρ_D	σ_D^2	ρ_X	ρ_D	σ_D^2	ρ_X	ρ_D	σ_D^2
$\sigma_E^2 = 1$	0.0	.000	.000	2.000	.000	.000	2.000	.000	.000	2.000
	0.2	.167	.242	2.640	.167	.167	2.400	.167	.074	2.160
	0.4	.286	.390	3.280	.286	.286	2.800	.286	.138	2.320
	0.6	.375	.490	3.920	.375	.375	3.200	.375	.194	2.480
	0.8	.444	.561	4.560	.444	.444	3.600	.444	.242	2.640
	1.0	.500	.615	5.200	.500	.500	4.000	.500	.286	2.800
	1.2	.545	.658	5.840	.545	.545	4.400	.545	.324	2.960
	1.4	.583	.691	6.480	.583	.583	4.800	.583	.359	3.120
	1.6	.615	.719	7.120	.615	.615	5.200	.615	.390	3.280
	1.8	.643	.742	7.760	.643	.643	5.600	.643	.419	3.440

		$\rho(T_X,T_Y) = -.60$			$\rho(T_X,T_Y) = 0$			$\rho(T_X,T_Y) = .60$		
	σ_E^2	ρ_X	ρ_D	σ_D^2	ρ_X	ρ_D	σ_D^2	ρ_X	ρ_D	σ_D^2
$\sigma_T^2 = 1$	0.0	1.000	1.000	3.200	1.000	1.000	2.000	1.000	1.000	0.800
	0.2	.833	.889	3.600	.833	.833	2.400	.833	.667	1.200
	0.4	.714	.800	4.000	.714	.714	2.800	.714	.500	1.600
	0.6	.625	.727	4.400	.625	.625	3.200	.625	.400	2.000
	0.8	.556	.667	4.800	.556	.556	3.600	.556	.333	2.400
	1.0	.500	.615	5.200	.500	.500	4.000	.500	.286	2.800
	1.2	.455	.571	5.600	.455	.455	4.400	.455	.250	3.200
	1.4	.417	.533	6.000	.417	.417	4.800	.417	.222	3.600
	1.6	.385	.500	6.400	.385	.385	5.200	.385	.200	4.000
	1.8	.357	.471	6.800	.357	.357	5.600	.357	.182	4.400

Consider now the relation between increases in reliability and power, reading from top to bottom in the columns in the upper section of the table and from bottom to top in the lower section. When the reliability coefficients of the component tests increased, the reliability of differences also increased, as long as just one column is considered. However, note that the same reliability of the components in many cases is associated with decidedly unlike reliabilities of the differences, depending on whether the change is attributable to a change in true-score variance or error-score variance. Often the values were far apart. Furthermore, the reliability of differences is either greater or less than that of the components, depending on whether the correlation between true scores, $\rho(T_X,T_Y)$, is positive or negative. As the absolute value of that correlation increases, the discrepancy is greater.

The observed scores of the differences, and hence the statistical power, *increases* as reliability increases if the change is attributable to a change in error-score variance and *decreases* if it is attributable to a change in true-score variance. That means that simply selecting a value of reliability, either of differences or the component tests, does not in itself provide information about the statistical power of the differences as a dependent variable. Just as in the case of a single test, the relation between reliability and power is not a *functional* relation unless the variance of one of the components of the scores is held constant.

These conclusions about the relation between power and the reliability of differences are consistent with results obtained by May & Hittner (2003), Overall & Woodward (1975, 1976), and Nicewander & Price (1978, 1983) using different methods. The so-called paradox of low reliability being associated with high power becomes more understandable from inspection of Table 3. That problem also is closely related to another issue that has been extensively treated in the literature, that of the reliability of differences often being considerably less than the reliability of the components. As the table shows, that is not always true, and again, looking at the reliability of the components alone, without further information, is one source of the trouble. The approach in Table 3, in which reliability coefficients are first related to the variances of true scores and error scores, makes it possible to focus on values that realistically would be likely to occur. At any rate, it is clear that an hypothesis test of differences can be powerful even if the reliability of a dependent variable is quite low.

How to increase statistical power: some practical implications

As mentioned before, a possible reason for the controversies surrounding the relation of reliability and statistical power is ambiguity about the precise meaning of the term “reliability” in practical research. The term often is used in a way that conforms to popular usage, and even to widespread usage in various scientific fields, but does not match the mathematical definition given in classical test theory. The root of the difficulty is the fact that reliability, as defined in test theory, is a property of populations of individuals, that is a ratio of statistics applicable to populations, but not to a single individual or experimental object. The “reliability” of a scientific instrument, especially in physical sciences, often refers to its consistency in measuring a single physical object of a certain kind, but that is not the way the term is used in classical test theory.

When one asks the question "How does reliability influence power?" investigators in psychology and education often assume the question is similar to "How does reliability influence validity?" or "How does test length influence reliability?" What is typically desired is a function relating changes in the first variable to changes in the second variable, and many such functions are known in test theory. On the other hand, a researcher in another field, or a statistician, may assume the question is similar to "How does sample size influence power?" or "How does the significance level influence power?" having in mind well-known functions relating those variables.

As emphasized in the present note, there is not a unique way of making the increments in reliability needed to exhibit power as a function of reliability. We can conclude that increasing an instrument's reliability will contribute to greater power in hypothesis testing only if the change occurs through a reduction of error-score variance that exceeds any increase in true-score variance occurring at the same time.

Suppose a researcher has a choice between two instruments, one with a known reliability coefficient of .90 and the other .80. Before assuming automatically that the first instrument is the better choice, it is prudent to look at the variance of scores that can be expected. If the instrument with lower reliability typically produces scores with considerably less variability, it could still be the better choice. That is especially true if the experiment is designed to detect possible differences among large groups of subjects with respect to an independent variable and is not concerned with short-term fluctuations in measures of individuals.

Another way to look at the problem is to recall that an hypothesis test is essentially a determination, based on probability, of whether or not a difference found between samples can be attributed to chance variability. However, an hypothesis test is blind to the partitioning of variability into contributions from separate components, such as "true scores" and "error scores." A test statistic such as t typically is computed as a ratio of an obtained value to an estimate of variability based on a sampling distribution.

Recommending that the reliability coefficient be increased whenever possible is not always good advice in hypothesis testing, although the conventional emphasis on practical measures to reduce error variance still applies. All other things being equal, the more error of measurement can be avoided in an experiment, the better, and that task certainly should be considered along with other well-known methods of increasing power (see, for example, Wilcox, 2003) that are useful in research. But reducing error is productive, we have seen, only if the same practical steps also reduce observed-score variance. If a more heterogeneous group is tested at the same time error of measurement is less, power does not necessarily increase. For practical usefulness, eliminating error and thereby increasing reliability for a particular population of examinees can be effective, provided the change is made without altering the population.

References

Cleary, T. A., & Linn, R. L. (1969). Error of measurement and the power of a statistical test. *British Journal of Mathematical and Statistical Psychology*, *22*(1), 49-55. doi:10.1111/j.2044-8317.1969.tb00419.x

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (3rd ed.). Englewood Cliffs, NJ: Lawrence Erlbaum Associates.

Collins, L. M. (1996). Is reliability obsolete? A commentary on “Are simple gain scores obsolete?” *Applied Psychological Measurement, 20*(3), 289-292. doi:10.1177/014662169602000308

Feldt, L. S. & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 105-146). New York: Macmillan.

Fleiss, J. J. (1976). Comment on Overall & Woodward’s asserted paradox concerning the measurement of change. *Psychological Bulletin*, *83*(5), 774-775. doi:10.1037/0033-2909.83.5.774

Gulliksen, H. (1950). *Theory of mental tests*. New York: Wiley.

Hopkins, K. D., & Hopkins, D. R. (1979). The effect of the reliability of the dependent variable on power. *Journal of Special Education, 13*(4), 463-466. doi:10.1177/002246697901300413

Kopriva, R. J., & Shaw, D. G. (1991). Power estimates: The effect of dependent variable reliability on the power of one-factor ANOVAs. *Educational and Psychological Measurement, 51*(3), 585-595. doi:10.1177/0013164491513006

Levin, J. R. (1986). Note on the relation between the power of a significance test and the reliability of the measuring instrument. *Multivariate Behavioral Research*, *21*(2), 255-261. doi:10.1207/s15327906mbr2102_6

Lord, F. M., & Novick, M. R. (1968). Statistical theories of mental test scores. Reading, MA: Addison-Wesley.

May, K., & Hittner, J. B. (2003). On the relation between power and reliability of difference scores. *Perceptual and Motor Skills, 97*(3.1), 905-908. doi:10.2466/PMS.97.7.905-908

Mellenbergh, G. J. (1996). Measurement precision in test score and item response models. *Psychological Methods, 1*(3), 293-299. doi:10.1037/1082-989X.1.3.293

Mellenbergh, G. J. (1999). A note on simple gain score precision. *Applied Psychological Measurement, 23*(1), 87-89. doi:10.1177/01466216990231007

Nicewander, W. A., & Price, J. M. (1978). Dependent variable reliability and the power of statistical tests. *Psychological Bulletin, 85*(2), 405-409. doi:10.1037/0033-2909.85.2.405

Nicewander, W. A., & Price, J. M. (1983). Reliability of measurement and the power of statistical tests. *Psychological Bulletin, 94*(3), 524-513. doi:10.1037/0033-2909.94.3.524

Novick, M. R. (1966). The axioms and principal results of classical test theory. *Journal of Mathematical Psychology, 3*(3), 1-18. doi:10.1016/0022-2496(66)90002-2

Overall, J. E., & Woodward, J. A. (1975). Unreliability of difference scores: A paradox for measurement of change. *Psychological Bulletin, 82*(1), 85-86. doi:10.1037/h0076158

Overall, J. E., & Woodward, J. A. (1976). Reassertion of the paradoxical power of tests of significance based on unreliable difference scores. *Psychological Bulletin, 83*(5), 776-777. doi:10.1037/0033-2909.83.5.776

Subkoviak, M. J., & Levin, J. R. (1977). Fallibility of measurement and the power of a significance test. *Journal of Educational Measurement, 14*(1), 47-52. doi:10.1111/j.1745-3984.1977.tb00028.x

Sutcliffe, J. P. (1958). Error of measurement and the sensitivity of a test of significance. *Psychometrika, 23*(1), 9-17. doi:10.1007/BF02288974

Thomas, D. R., & Zumbo, B. D. (2012). Difference scores from the point of view of reliability and repeated measures ANOVA: In defense of difference scores for data analysis. *Educational and Psychological Measurement, 72*(1), 37-43. doi: 10.1177/0013164411409929

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

Wolfram, S. (1999). *The Mathematica Book* (4th ed.). NY: Wolfram Media/Cambridge University Press.

Zimmerman, D. W., & Williams, R. H. (1986). Note on the reliability of experimental measures and the power of significance tests. *Psychological Bulletin, 100*(1), 123-124. doi:10.1037/0033-2909.100.1.123

Zimmerman, D. W., Williams, R. H., & Zumbo, B. D. (1993). Reliability of measurement and power of significance tests based on differences. *Applied Psychological Measurement*, *17*(1), 1-10. doi:10.1177/014662169301700101

Zumbo, B. D. (1999). The simple difference score as an inherently poor measure of change: Some reality, much mythology. In B. Thompson (Ed.). *Advances in Social Science Methodology*, *5*, (pp. 269-304). Greenwich, CT: JAI Press.

Journal of Modern Applied Statistical Methods
November 2015, Vol. 14, No. 2, 27-34.

ISSN 1538 − 9472

Invited Article
In (Partial) Defense of .05

Thomas R. Knapp
University of Rochester
Rochester, NY

Researchers are frequently chided for choosing the .05 alpha level as the determiner of statistical significance (or non-significance). A partial justification is provided.

Keywords: .05 level, statistical significance, R. A. Fisher

Introduction

For the last 50 or 60 years it has been fashionable to deride the insistence on using an alpha level of .05 for testing the statistical significance of a sample finding. It is commonplace to read critical comments such as “The current obsession with .05” (Skipper, Guenther, & Nass, 1967, p. 16; see also Labovitz, 1968) and “God loves the .06 nearly as much as the .05” (Rosnow & Rosenthal, 1989, p. 1277). In the spirit of Robinson, Funk, Halbur, and O'Ryan (2003) I would like to provide an explanation for ‘why .05?’ and an argument in favor of its prevailing use. Near the end of the paper I will give a similar argument for 95% confidence (.05's interval estimation counterpart), and I will conclude with a few cautionary statements regarding total devotion to .05 and/or 95%.

A bit of history

Although there is some evidence for earlier recommendations of .05 as a defensible level of statistical significance, most people claim that it was first suggested by Fisher (1926):

> [T]he evidence would have reached a point which may be called the verge of significance; for it is convenient to draw the line at about the

Dr. Knapp is Professor Emeritus of Education and Nursing at the University of Rochester and The Ohio State University. Email him at tknapp5@juno.com.

> level at which we can say 'Either there is something in the treatment or a coincidence has occurred such as does not occur more than once in twenty trials.' This level, which we may call the 5 per cent level point, would be indicated, though very roughly, by the greatest chance deviation observed in twenty successive trials... If one in twenty does not seem high enough odds, we may, if we prefer it, draw the line at one in fifty (the 2 per cent point) or one in a hundred (the 1 per cent point). Personally, the writer prefers to set the low standard of significance at the 5 per cent point, and ignore entirely all results which fail to reach this level. (p. 504)

There are several things to note about what Fisher said:

1. He used the interesting phrase “the verge of significance”. As far as I have been able to determine, none of his critics have commented about that choice of words.
2. He did not insist on .05, as the second part of the quote indicated. Many of his critics unfairly charged him with being unwavering regarding .05.
3. Surprisingly, he confused probability with odds (and high with low). The alpha level of .05 has to do with a probability of one in twenty; the corresponding odds are one to nineteen (in favor) or nineteen to one (against).

Fisher didn’t write about .05 being the probability of making a Type I error. That concept (along with the probability of making a Type II error) was yet to come in the Neyman-Pearson approach to hypothesis testing. Also yet to come were several acrimonious arguments between Fisher and W. S. Gosset (who had previously developed the t-test), between Fisher and Karl Pearson, and between Fisher and both Jerzy Neyman and Egon Sharpe Pearson (Karl’s son), as documented by Fienberg and Tanur (1966), Cowles and Davis (1982), Inman (1994), Wainer and Robinson (2003), and others.

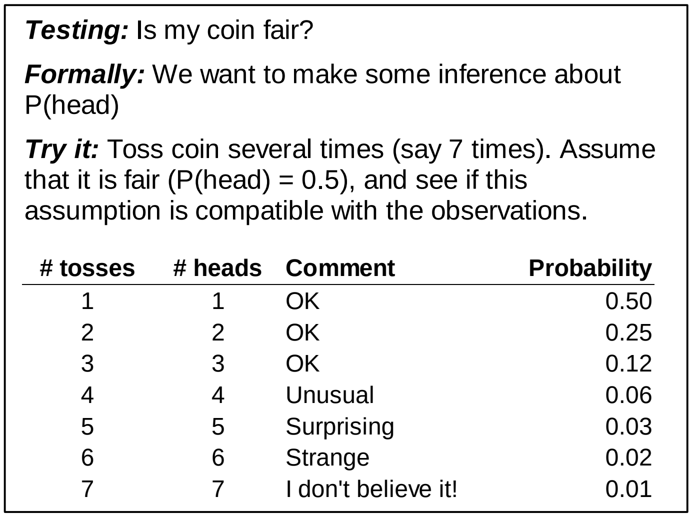
In the intervening years between 1926 and the present there were several criticisms of .05, e.g., Cohen (1994), along with some defenders, e.g., Robinson, et al. (2003). Cohen (1994) was particularly puzzling (see the collection of comments regarding it in the December, 1995 issue of *American Psychologist*). The title is difficult to understand. Was he trying to be clever in considering “The earth is round” as a null hypothesis that should be rejected at the .05 level, because it is actually slightly elliptical rather than perfectly round? He also made an error where he claimed many people believe a *p*-value is the probability that the null hypothesis is false. No; some people mistakenly believe that a *p*-value is the probability that the null hypothesis is true; no one believes p is the probability of a false null.

After discussing some of the historical origins of the use of an alpha level of .05, Robinson, et al. (2003) provided the results of empirical studies in which students were asked how many heads in each of the first n flips of a coin would lead them to claim that the coin was not "fair". The modal response in most of those studies was five. The probability of heads on the first five tosses of a fair coin is .03125, which is close to the traditional .05 (see Figure 1 below).

A rationale for .05

Although Fisher didn't use the following argument, some of the students in the Robinson, et al. (2003) studies apparently did, implicitly if not explicitly. (Comparable arguments have been made by Tintle, et al., 2014 and at the EMBstats website, http://www.embstats.com. See Figure 1 below for the latter.) Suppose you were asked your opinion about the fairness of a coin. You want to make a decision if its probability of landing as heads is equal to .5. How many heads would have to be obtained in the first five tosses for you to call a halt and conclude it's not a fair coin? The probability of one head in one toss of a fair coin is .5. (You wouldn't call a halt.) The probability of two heads in two tosses is .5 × .5 = .25, and the probability of three heads in three tosses is .5 × .5 × .5 = .125. (Still no clear decision to halt.) The probability of four heads in four tosses is .5 × .5 × .5 × .5 = .0625. (Perhaps the decision to halt is near, and note .0625 is close to .05.) If you want to wait for the result of one more toss, the probability of five heads in five tosses is .5 × .5 × .5 × .5 × .5 = .03125. At this point you are likely to claim that the coin is not fair. (The difference between .0625 and the .03125 is .046875, which is very close to .05.) However, you know you might be wrong.

Figure 1 details the argument presented at the EMBstats website. Note the interpretations of "Unusual" (for 4 heads in 4 tosses), "Surprising" (for 5 heads in 5 tosses), "Strange" (for 6 heads in 6 tosses), and "I don't believe it!" (for 7 heads in 7 tosses). Fisher's .05 would come between "Unusual" and "Surprising". He avoided the matter of proof and exhibited a commendable tolerance for uncertainty. Similarly, statisticians are so comfortable with uncertainty that they occasionally advocate the use of the randomized response technique for estimating a proportion where only some of the respondents to a survey actually answer the question of interest (Campbell & Joiner, 1973).



Testing: Is my coin fair?

Formally: We want to make some inference about P(head)

Try it: Toss coin several times (say 7 times). Assume that it is fair (P(head) = 0.5), and see if this assumption is compatible with the observations.

# tosses	# heads	Comment	Probability
1	1	OK	0.50
2	2	OK	0.25
3	3	OK	0.12
4	4	Unusual	0.06
5	5	Surprising	0.03
6	6	Strange	0.02
7	7	I don't believe it!	0.01

Figure 1. EMBstats dialogue on tossing a coin (http://www.embstats.com).

95% confidence intervals

In the last 25 or 30 years there has been a pronounced shift from an emphasis on significance testing to a preference for confidence intervals. Some methodologists suggest reporting both; some journal editors require it. (Reporting both is not a good idea. See the third statistics commandment in Knapp & Brown, 2014). But the continuing choice of 95% for confidence (the interval estimation counterpart to .05 for hypothesis testing) has not been subject to the same sort of scrutiny that has been directed at .05. Why is that?

Perhaps consumers are more convinced by a 95% confidence argument than by the .05 significance argument. Consider the coin-tossing problem above, but change it to a desire for estimating the degree of bias associated with the coin rather than testing its fairness. If the coin-tosser got five heads in five tosses and was interested in estimating the population proportion of heads for that coin, he could get a confidence interval by using Pezzullo's online computing routine (http://www.statpages.org) based on Clopper and Pearson's (1934) formulas, tables, and graphs.

For example, at http://www.statpages.org for Exact Binomial Confidence Intervals input 5 heads (the numerator) in 5 tosses (the denominator, chose 95% confidence (the default). The results returned are 1.0000 as the statistic and .4782 to 1.0000 as the confidence interval. A choice of 99% confidence (corresponding to .01 significance) or 99.9% confidence (corresponding to .001 significance) serves only to reduce the lower limit (.3466 for 99% and .2187 for 99.9%) and therefore provides more confidence. Could it be that some people regard 99% confidence intervals and 99.9% confidence intervals to be too wide and are willing to stick with 95% for its greater precision despite its lesser confidence?

Asterisks

Consider the still-common practice of labeling with a single asterisk a finding for which p < .05, two asterisks for p < .01, and three asterisks for p < .001 (or what Leahey, 2005 refers to as the three-star system", Abstract). That is not sound practice (see Slakter, Wu, & Suzuki-Slakter, 1991), because if an alpha of .05 has been used in a power analysis to select an appropriate sample size, then all that is necessary to determine is whether p is less than or greater than .05. (Similarly, for alphas of .01 and .001.) Some journal editors require the reporting of the actual p, and that is the preferred practice according to the American Psychological Association manual (APA, 2010), which is not perfect, but is more sound than using asterisks.

To be consistent, why aren't asterisks or similar symbols used in the tables where authors report 95%, 99%, or 99.9% confidence intervals? If this statistic is significant at the .05 level and that statistic is significant at the .01 level, doesn't it make sense to put a 95% confidence interval around the first statistic and a 99% confidence interval around the second statistic?

All of the references so far have been to journal articles. There are three books on this topic that are recommended: Fisher (1925), Salsburg (2001), and Moye (2006). These three authors addressed the choice of .05 for statistical significance. Fisher (1925) contained some of the same views later expressed in Fisher (1926). Salzburg related Fisher's classic experiment regarding a lady's ability to determine whether milk has been added to tea or tea added to milk. Moye provided a thorough discussion of the advantages and disadvantages of *p*-values (mostly disadvantages). Both Salzburg and Moye gave fascinating accounts of Fisher's battles with Neyman and Pearson (and with Gosset). Moye noted that Fisher was not wedded to .05, as stated above.

Some cautions

Continuing to emphasize .05 as the cut-off between statistical significance and non-significance is not all that bad. The same holds for continuing to emphasize 95% for confidence intervals. But there are exceptions.

1. If there might be very serious consequences should a Type I error be made, a more stringent alpha is necessary. For example, suppose a randomized clinical trial (RCT) were to be carried out comparing the effectiveness of a new and very expensive drug with an existing much less expensive drug. Suppose further that a decision might be made to reject the null hypothesis of no effect because of a statistically significant effect in favor of the new drug, but in reality it is no better. That could lead to the adoption of a drug that is not only no better than the existing drug but could result in an unnecessary cost of thousands or millions of dollars. In that case an argument could be made to use .01 or .001 or an even smaller significance level.
2. If the committing of a Type II error would have much greater consequences than a Type I error, the argument is reversed; i.e., change alpha to a more liberal level, such as .20. An example of this would be a medical diagnosis of no disease if a patient is in fact ill. Generally, it would be worse to not treat a patient who has a disease than to treat a patient when the disease is not present.
3. If the estimate of a population parameter must be both precise and defendable, a confidence coefficient of 99.9% might be chosen, as well as a huge sample size. For example, if an estimate of the proportion of people who are below the poverty line is to be made, we might want to do that in order to have both politically defensible and morally desirable evidence for so doing.

References

American Psychological Association. (2010). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: Author.

Campbell, C., & Joiner, B. L. (1973). How to get the answer without being sure you've asked the question. *The American Statistician, 27*(5), 229-231. doi:10.1080/00031305.1973.10479043

Clopper, C. J., & Pearson, E. S. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika, 26*, 404-413. doi:10.2307/2331986

Cohen, J. (1994). The earth is round (p<.05). *The American Psychologist, 49*(12), 997–1003. doi:10.1037/0003-066X.49.12.997

Cowles, M., & Davis, C. (1982). On the origins of the .05 level of statistical significance. *The American Psychologist, 37*(5), 553-558. doi:10.1037/0003-066X.37.5.553

Fienberg, S. E., & Tanur, J. M. (1996). Reconsidering the fundamental contributions of Fisher and Neyman on experimentation and sampling. *International Statistical Review / Revue Internationale de Statistique, 64*(3), 237-253. doi:10.2307/1403784

Fisher, R. A. (1925). *Statistical methods for research workers*. Edinburgh: Oliver and Boyd.

Fisher, R. A. (1926). The arrangement of field experiments. *Journal of the Ministry of Agriculture, 33*, 503-513.

Fowler, R. D. (Ed.) (1995). Bridging science and practice [Full issue]. *American Psychologist, 50*(12).

Inman, H. F. (1994). Karl Pearson and R. A. Fisher on statistical tests: A 1935 exchange from Nature. *The American Statistician, 48*(1), 2-11. doi:10.1080/00031305.1994.10476010

Knapp, T. R., & Brown, J. K. (2014). Ten statistics commandments that almost never should be broken. *Research in Nursing & Health, 37*(4), 347-351. doi:10.1002/nur.21605

Labovitz, S. (1968). Criteria for selecting a significance level: A note on the sacredness of .05. *The American Sociologist, 3*(3), 220–222. Retrieved from http://www.jstor.org/stable/27701367

Leahey, E. (2005). Alphas and asterisks: The development of statistical significance standards in sociology. *Social Forces, 84*(1), 1-24. doi:10.1353/sof.2005.0108

Moye, L. A. (2006). *Statistical reasoning in medicine: The intuitive p-value primer* (2nd. ed.). New York: Springer.

Robinson, D. H., Funk, D. C., Halbur, D., & O'Ryan, L. (2003). The .05 alpha level in educational research: Traditional, arbitrary, sacred, magical, or simply psychological? *Research in the Schools*, *10*(2), 79-86.

Rosnow, R. L., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *The American Psychologist, 44*(10), 1276-1284. doi:10.1037/0003-066X.44.10.1276

Salsburg, D. (2001). *The lady tasting tea*. New York: Freeman.

Skipper, J. K., Jr., Guenther, A. L., & Nass, G. (1967). The sacredness of .05: A note concerning the uses of statistical levels of significance in social science. *The American Sociologist, 2*(1), 16-18. Retrieved from http://www.jstor.org/stable/27701229

Slakter, M. J., Wu, Y.-W. B., & Suzuki-Slakter, N. S. (1991). *, **, and ***: Statistical nonsense at the .00000 level. *Nursing Research*, *40*(4), 248–249.

Tintle, N. L., Chance, B., Cobb, G., Rossman, A. Roy, S., Swanson, T., & VanderStoep, J. (2014). *Introduction to statistical inference*. (Preliminary edition). Hoboken, NJ: Wiley.

Wainer, H., & Robinson, D. H. (2003). Shaping up the practice of null hypothesis significance testing. *Educational Researcher, 32*(7), 22-30. doi:10.3102/0013189X032007022

Journal of Modern Applied Statistical Methods
November 2015, Vol. 14, No. 2, 35-52.

ISSN 1538 − 9472

An Empirical Study on Different Ranking Methods for Effective Data Classification

Ilangovan Sangaiah
K.L.N. College of
Engineering
Madurai, India

A. V. Antony Kumar
PSNA College of
Engineering and Technology
Dindigul, India

Appavu Balamurugan
K.L.N. College of
Information Technology
Tamil Nadu, India

Ranking is the attribute selection technique used in the pre-processing phase to emphasize the most relevant attributes which allow models of classification simpler and easy to understand. It is a very important and a central task for information retrieval, such as web search engines, recommendation systems, and advertisement systems. A comparison between eight ranking methods was conducted. Ten different learning algorithms (NaiveBayes, J48, SMO, JRIP, Decision table, RandomForest, Multilayerperceptron, Kstar) were used to test the accuracy. The ranking methods with different supervised learning algorithms give different results for balanced accuracy. It was shown the selection of ranking methods could be important for classification accuracy.

Keywords: Feature selection, Ranking Methods, Classification algorithms, Classification accuracy

Introduction

Ranking is a crucial part of information retrieval. It is able to compute sorted score when given document as objects. Ranking is a central issue in information retrieval, in which, given a set of objects (e.g., Documents), a score for each of them is computed and the objects are sorted according to the scores. Depending on the applications the scores may represent the degrees of relevance, preference, or importance. Ranking is a very important topic in feature selection. Although algorithms for learning ranking models have been intensively studied, this is not the case for feature selection, despite of its importance. The reality is that many

Ilangovan Sangaiah is a PhD candidate Department of Information and Communication Engineering at Anna University and faculty at KLN College of Engineering. Dr. A. Vincent Antony Kumar is Professor and head of Information technology. Dr. Appavu Alais Balamurugan S. is head of Information Technology. Email him at: *app_s@yahoo.com* .

feature selection methods used in classification are directly applied to ranking. Because of the striking differences between ranking and classification, it is better to develop different feature selection methods for ranking.

Feature selection has emerged as a successful mechanism in many machine learning applications. Feature selection is also desirable for learning to rank. First, as the numbers of useful features for ranking are continuously growing, the time of extracting such high-dimensional features has become a bottleneck in ranking.

High-dimensional features may be redundant or noisy, which results in poor generalization performance. Also, a ranking model with only a small set of features has less computational cost in prediction. Recently, considerable efforts have been made on feature selection for ranking. The main aim of this paper was to experimentally verify the impact of different ranking methods on classification accuracy.

The only way to be sure that the highest accuracy is obtained in practical problems is testing a given classifier on a number of feature subsets, obtained from different ranking indices. Diverse feature ranking and feature selection techniques have been proposed in the machine learning literature. The purpose of these techniques is to discard irrelevant or redundant features from a given feature vector. The usefulness of the following commonly used ranking methods in different datasets are considered:

1. Relief.
2. Gain Ratio (GR).
3. Information Gain (IG).
4. One-R.
5. Symmetrical Uncertainty (SU).
6. Chi-Squared.
7. Support Vector Machine (SVM).
8. Filter.

The results were validated using different algorithms for classification. A wide range of classification algorithms is available, each with its strengths and weaknesses. There is no single learning algorithm that works best on all supervised learning problems.

Review of the literature

A ranking is a task that applies machine learning techniques to learn good ranking predictors. It is a relationship between a set of items and a unit that refer to different values. Many learning-to-rank algorithms have been proposed. The two prime functions of ranking are to deliver highly relevant search results and to be fast in ranking results. Many feature selection and feature ranking methods have been proposed. Fuhr and Norbert (1989) introduced a Ranking OPRF method which uses the idea of Polynomial regression. Cooper, Gey and Dabney (1992) proposed a point wise SLR (Staged logistic regression ranking) method. A RELIEF ranking algorithm was proposed by Kira and Rendell (1992).

The strengths of relief is that, it is not dependent on heuristics, it requires only linear time in the number of given features and training instances, and it is noise-tolerant and robust to feature interactions, as well as being applicable for binary or continuous data. However, it does not discriminate between redundant features, and low numbers of training instances fool the algorithm. Robnik-Sikonja and Kononenko (2003), proposed some updates to the algorithm (RELIEF-F) in order to improve the reliability of the probability approximation, make it robust to incomplete data, and generalizing it to multi-class problems. Then the original Support Vector Machine algorithm (SVM) was invented by Vladimir N. Vapnik in 1992 (Cortes & Vapnik, 1995). This SVM is supervised learning models with associated learning algorithms that analyze data and recognize patterns, used for classification and regression analysis. SVMs are based on the concept of decision planes that define decision boundaries. A decision plane is one that separates between a set of objects having different class memberships. SVMs deliver state-of-the-art performance in real-world applications such as text categorization, hand-written character recognition, image classification, bio sequences analysis, etc., and are now established as one of the standard tools for machine learning and data mining.

Information Gain Another ranking method called as Information Gain (*IG*) evaluates the worth of an attribute by measuring the information gain with respect to the class. An attribute selection measure, based on pioneering work by Claude Shannon on information theory, which studied the value of the information content of messages. It is given by

$$IG = H(Y) - H\left(\frac{Y}{X}\right) = H(X) - H\left(\frac{X}{Y}\right)$$

IG is a symmetrical measure. The information gained about *Y* after observing *X* is equal to the information gained about *X* after observing *Y*. A weakness of the *IG* criterion is that, it is biased in favour of features with more values even when they are not more informative.

The attribute has the best score for the measure is chosen as the splitting attribute for the given tuple. Depending on the measure, either the highest or lowest score is chosen as the best attribute. The *IG* measure is biased toward tests with many outcomes. That is, it prefers to select attributes having large number of values.

Gain Ratio But Gain Ratio is the extension of *IG* which attempts to overcome this bias. It evaluates the worth of an attribute by measuring the gain ratio with respect to the class. The Gain Ratio is the non-symmetrical measure that is introduced to compensate for the bias of the *IG* (Hall & Smith, 1998). Gain Ratio is given by

$$G(R) = IG/H(X)$$

When the variable *Y* has to be predicted, we normalize the *IG* by dividing by the entropy of *X*, and vice versa. Due to this normalization, the *GR* values always fall in the range [0, 1]. A value of *GR* = 1 indicates that the knowledge of *X* completely predicts *Y*, and *GR* = 0 means that there is no relation between *Y* and *X*. In opposition to the *IG*, the *GR* favours variables with fewer values.

Symmetrical Uncertainty The Symmetrical Uncertainty criterion compensates for the inherent bias of *IG* by dividing it by the sum of the entropies of *X* and *Y* (Hall & Smith, 1998). It is given by

$$SU = 2\left(\frac{IG}{H(Y)+H(X)}\right)$$

SU takes values, which are normalized to the range [0, 1] because of the Correction factor 2. A value of *SU* = 1 means that the knowledge of one feature completely predicts, and the other *SU* = 0 indicates, that *X* and *Y* are uncorrelated. Similar to *GR*, the *SU* is biased toward features with fewer values.

Chi-squared Feature Selection via chi square test is another very commonly used method (Liu & Setiono, 1995). Chi-squared attribute evaluation evaluates the worth of a feature by computing the value of the chi-squared statistic with respect to the class. The initial hypothesis H_0 is the assumption that the two features are unrelated, and it is tested by chi squared

Formula:

$$x^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}}$$

where O_{ij} is the observed frequency and E_{ij} is the expected (theoretical) frequency, asserted by the null hypothesis. The greater the value of χ^2, the greater the evidence against the hypothesis H_0 is.

One-R OneR is a simple algorithm proposed by Holte (1993). It builds one rule for each attribute in the training data and then selects the rule with the smallest error. It treats all numerically valued features as continuous and uses a straightforward method to divide the range of values into several disjoint intervals. It handles missing values by treating "missing" as a legitimate value. This is one of the most primitive schemes. It produces simple rules based on feature only. Although it is a minimal form of classifier, it can be useful for determining a baseline performance as a benchmark for other learning schemes.

A pairwise RankSVM (Herbrich, Graepel & Obermayer, 2000) method was devised that out performs more naive approaches to ordinal regression such as Support Vector Classification and Support Vector Regression in the case of more than two ranks. In the year 2003, 2005 and 2006 a pairwise RankBoost, RankNet (Burges et al., 2005) and IR-SVM, Lambda Rank methods were developed. Subsequently, in 2007, the ranking methods Frank, GB Rank, ListNet, McRank, QBRank, RankCosine, RankGP, and RankRLS were innovated. In the year 2007 a listwise ranking methods ListNet, RankCosine, RankGPand, SVMmap (Yue, Finley, Radlinski, & Joachims, 2007) were introduced. Ranking Refinement method (2008) is a semi-supervised approach to learning to rank that uses Boosting. Then a list wise ranking methods LambdaMART (Wu, Burges, Svore, & Gao 2008), ListMLE, PermuRank, SoftRank and a pairwise ranking methods Ranking Refinement (Rigutini, Papini, Maggini, & Scarselli, 2008) SSRankBoost (Amini, Troung, & Goutte, 2008), SortNet (Rigutini et al., 2008) were developed in 2008. In 2009 MPBoost, BoltzRank and BayesRank (Kuo, Cheng, & Wang,

2009) later in 2010 NDCG Boost (Valizadegan, Jin, Zhang, & Mao, 2010), Gblend, IntervalRank (Moon, Smola, Chang, & Zhen, 2010) and CRR (Sculley, 2010) were discovered.

Point wise approach It is assumed that each query-document pair in the training data has a numerical or ordinal score. Then learning-to-rank problem can be approximated by a regression problem-given a single query-document pair, predict its score.

Pairwise approach The learning-to-rank problem is approximated by a classification problem- learning a binary classifier that can tell which document is better in a given pair of documents. The goal is to minimize the average number of inversions in ranking.

List wise approach These algorithms try to directly optimize the value of one of the above evaluation measures, averaged over all queries in the training data. This is difficult because most evaluation measures are not continuous functions with respect to ranking model's parameters, and so continuous approximations or bounds on evaluation measures have to be used.

Proposed work and experimental results

Weka tool Data mining or “Knowledge Discovery in Databases” is the process of discovering patterns in large data sets with artificial intelligence, machine learning, statistics, and database systems. The overall goal of a data mining process is to extract information from a data set and transform it into an understandable structure for further use. In its simplest form, data mining automates the detection of relevant patterns in a database, using defined approaches and algorithms to look into current and historical data that can then be analyzed to predict future trends. A data mining tools predict future trends and behaviours by reading through databases for hidden patterns; they allow organizations to make proactive, knowledge-driven decisions and answer questions that were previously too time-consuming to resolve.

With Weka, Open Source software, patterns can be discovered in large data sets and extract all the information. It is a comprehensive tool for machine learning and data mining for predictive analytics. Weka is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a data set or called from your own JAVA code. It is also well suited for developing new machine learning schemes. It also brings great portability, since it was fully implemented in the JAVA programming language, plus supporting several standard data mining tasks. It contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. Different ranking methods can also be implemented using the data pre-processing tool which is available in Weka. It is also well-suited for developing new machine learning schemes.

Methodology

Datasets used in experiments

Five datasets are used: diabetes, segment-challenge, soybean, vote and ionosphere from the UCI data repository (Lichman, 2013). The first dataset is the diabetes data which has 768 instances and 9 attributes. The second data set segment-challenge has 1500 instances and 20 attributes. Similarly soybean, vote and ionosphere datasets have 683,435,351 instances and 36, 17, 35 attributes respectively. In Weka a wide range of classification algorithms is available for data analysis. From this wide range of learning algorithms, eight different algorithms are chosen and applied on all the five datasets for our study.

Table 1. Datasets used in the Experiment.

Sl.No	Name of the Dataset	No. of attributes	No. of Instances
1	Diabetes	9	768
2	segment-challenge	20	1500
3	soybean	36	683
4	vote	17	435
5	ionosphere	35	351

Table 2. Classification accuracy of different Classification algorithm without Ranking.

S. No.	Dataset	NB	J48	SMO	JRIP	DT	Rd.Frt	Mul.pr	Kstar
1	Diabetes	76.3	73.82	77.34	76.04	71.22	73.82	75.39	69.14
2	Segment-challenge	81.06	95.73	91.93	93.73	87.4	96.93	96.73	96.6
3	soybean	92.97	91.5	93.85	91.94	84.33	92.09	93.41	87.99
4	vote	90.11	96.32	96.09	95.4	94.94	95.63	94.71	93.33
5	ionosphere	82.62	91.45	88.6	89.74	89.45	92.87	91.16	84.61
Classification Average		84.61	89.76	89.56	89.37	85.47	90.27	90.28	86.33

Table 3. Processing Time of different Classification algorithm without Ranking.

S. No.	Dataset	NB	J48	SMO	JRIP	DT	Rd.Frt	Mul.pr	Kstar
1	Diabetes	0.02	0.04	0.26	0.06	0.09	0.13	1.96	0.0
2	Segment-challenge	0.02	0.09	1.85	0.55	0.49	0.26	17.06	0.0
3	soybean	0.0	0.03	4.77	0.11	0.81	0.33	97.25	0.0
4	vote	0.0	0.0	0.04	0.01	0.09	0.07	2.41	0.0
5	ionosphere	0.01	0.04	0.08	0.07	0.15	0.01	6.59	0.0
Average Processing Time		0.01	0.04	1.4	0.16	0.32	0.16	25.05	0

Table 4. Classification accuracy on selected features for Diabetes dataset.

Ranking Method	NB	J48	SMO	JRIP	DT	Rd. Frt	Mul. pr	Kstar	F.S. Avg.
Relief	75.4	74.3	76.4	74.1	73.0	73.4	74.7	69.0	73.8
GainRatio	75.5	74.9	76.2	75.9	72.4	72.0	76.3	71.4	74.3
InfoGain	75.4	74.3	76.0	75.1	72.1	72.0	77.2	71.6	74.2
OneR	75.5	74.9	76.2	76.2	72.4	72.6	76.0	71.4	74.4
SU	75.4	74.3	76.0	75.1	72.1	72.0	77.2	71.6	74.2
Chi-squared	75.4	74.3	76.0	74.9	71.6	71.2	76.7	71.6	74.0
SVM	77.2	74.9	76.8	74.2	72.7	72.4	75.1	71.9	74.4
Filter	75.4	74.3	76.0	75.1	72.1	72.0	77.2	71.6	74.2
Classification Avg.	75.7	74.5	76.2	75.1	72.3	72.2	76.3	71.3	

Table 5. Classification accuracy on selected features for segment-challenge dataset.

Ranking Method	NB	J48	SMO	JRIP	DT	Rd. Frt	Mul. pr	Kstar	F.S. Avg
Relief	73.3	94.6	83.1	93.8	87.0	96.2	95.6	96.9	90.1
GainRatio	66.4	89.2	77.4	86.6	82.8	90.6	86.3	92.1	84.3
InfoGain	76.9	94.8	89.6	93.9	87.0	96.2	85.3	97.1	91.4
OneR	75.0	94.9	87.6	93.6	87.0	96.4	95.5	97.0	90.9
SU	76.9	94.9	89.6	93.2	87.0	96.8	95.5	97.1	91.3
Chi-squared	66.4	89.2	77.6	88.0	95.6	82.8	88.9	95.1	85.5
SVM	82.0	94.6	90.7	93.4	88.2	96.7	96.0	95.1	92.2
Filter	76.9	94.8	89.6	93.9	87.0	96.2	95.3	95.7	91.4
Classification Avg.	74.2	93.4	85.7	92.1	87.7	94.4	93.9	96.0	

Table 6. Classification accuracy on selected features for soybean dataset.

Ranking Method	NB	J48	SMO	JRIP	DT	Rd. Frt	Mul. Pr	Kstar	F.S. Avg.
Relief	89.5	88.6	92.8	87.8	80.1	89.0	92.1	88.3	88.5
GainRatio	85.8	85.2	86.2	84.9	82.7	87.4	87.4	86.1	85.7
InfoGain	89.9	88.3	93.0	88.7	80.1	86.8	93.3	88.9	88.6
OneR	83.6	85.4	87.1	84.8	83.9	86.5	87.3	86.4	85.6
SU	89.8	90.3	93.4	89.8	82.4	88.3	93.6	90.5	89.8
Chi-squared	89.2	89.8	93.9	89.6	81.3	91.4	93.7	90.0	89.8
SVM	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Filter	89.9	88.3	93.0	89.9	80.1	86.8	93.3	88.9	88.8
Classification Avg.	88.2	88.0	91.3	87.9	81.5	88.0	91.5	88.4	

Table 7. Classification accuracy on selected features for vote dataset.

Ranking Method	NB	J48	SMO	JRIP	DT	Rd. Frt	Mul Pr	Kstar	F.S. Avg.
Relief	90.3	96.3	95.6	95.9	95.9	95.9	93.8	94.9	94.8
GainRatio	91.3	95.2	95.6	95.6	95.6	94.5	95.2	92.9	94.5
InfoGain	91.3	95.2	95.6	95.6	95.6	94.5	95.2	92.9	94.5
OneR	90.6	94.7	95.6	95.4	95.4	95.2	94.0	92.9	94.2
SU	91.3	95.2	95.6	95.6	95.6	94.1	95.2	92.9	94.4
Chi-squared	91.3	95.2	95.6	95.6	95.6	93.6	94.0	92.9	94.2
SVM	91.5	96.3	95.9	96.3	94.7	95.9	94.9	94.0	94.9
Filter	91.3	95.2	95.6	95.6	95.6	94.3	95.2	92.9	94.5
Classification Avg.	91.1	95.4	95.7	95.7	95.5	94.7	94.7	93.3	

Table 8. Classification accuracy on selected features for ionosphere dataset.

Ranking Method	NB	J48	SMO	JRIP	DT	Rd. Frt	Mul Pr	Kstar	F.S. Avg.
Relief	86.3	92.9	87.7	90.9	89.5	93.2	90.9	84.6	89.5
GainRatio	87.5	90.3	87.7	91.7	89.5	93.4	92.6	85.2	89.7
InfoGain	88.0	92.0	87.7	90.9	89.5	93.4	94.0	86.6	90.3
OneR	88.0	92.0	87.7	90.9	89.5	93.4	91.5	84.6	89.7
SU	88.0	92.0	87.7	90.9	89.5	93.4	92.0	86.3	90.0
Chi-squared	88.0	92.0	87.7	90.9	89.5	93.4	94.6	86.6	90.3
SVM	88.0	92.0	87.7	90.9	89.5	93.4	91.1	87.2	90.0
Filter	88.0	92.0	87.7	90.9	89.5	93.4	94.0	86.6	90.3
Classification Avg.	87.7	91.9	87.7	91.0	89.5	93.4	92.6	86.0	

Table 9. Average Classification accuracy on Full set with ranking.

Ranking Method	NB	J48	SMO	JRIP	DT	Rd. Frt	Mul. Pr	Kstar	F.S. Avg
Relief	84.51	89.65	89.56	89.062	85.65	89.89	72.53	86.82	85.96
GainRatio	84.61	89.65	89.53	88.77	85.084	91.00	90.52	87.082	88.28
InfoGain	84.61	89.70	89.37	89.46	85.22	90.62	90.33	86.82	88.27
OneR	84.61	89.76	89.55	88.91	85.38	90.91	90.25	85.16	88.07
SU	84.61	89.71	89.53	88.92	85.25	90.48	90.41	86.82	88.22
Chi-squared	84.61	89.71	89.56	89.19	85.35	90.55	90.47	86.38	88.23
SVM	82.52	89.33	88.49	88.60	85.81	90.45	89.39	85.81	87.55
Filter	84.61	89.71	89.56	89.46	85.22	90.62	90.34	86.56	88.26
Classification average	84.34	89.65	89.39	89.05	85.37	90.57	88.03	86.43	

Table 10. Average Classification accuracy on selected features with Ranking.

Ranking Method	NB	J48	SMO	JRIP	DT	Rd. Frt	Mul. Pr	Kstar	F.S Avg.
Relief	84.77	89.25	86.33	88.68	84.82	89.44	89.34	86.83	87.43
GainRatio	81.34	86.95	84.82	86.57	84.33	87.29	88.06	85.52	85.61
InfoGain	83.84	88.81	88.24	88.46	84.76	88.63	91.00	87.42	87.65
OneR	82.68	88.70	86.31	87.61	85.52	88.90	88.52	86.66	86.86
SU	84.10	89.30	88.55	88.70	85.23	88.79	90.70	87.68	87.88
Chi-squared	81.71	88.11	85.91	87.91	84.25	87.72	89.58	86.74	86.49
SVM	84.17	88.67	87.86	87.69	86.45	88.86	89.32	87.27	87.54
Filter	83.84	88.81	88.24	88.46	84.76	88.63	91.00	87.42	87.65
Classification average	83.31	88.58	87.03	88.01	85.02	88.53	89.69	86.94	

Table 11. Average processing time with ranking on Full set.

Ranking Method	NB	J48	SMO	JRIP	DT	Rd. Frt	Mul. Pr	Kstar	F.S. Avg
Relief	0.02	0.06	2.67	0.22	0.34	0.19	24.90	0.00	3.55
GainRatio	0.00	0.04	1.25	0.18	0.32	0.18	24.93	0.00	3.36
InfoGain	0.01	0.04	1.39	0.16	0.33	0.18	24.99	0.00	3.39
OneR	0.01	0.04	1.05	0.17	0.37	0.17	25.02	0.00	3.35
SU	0.01	0.04	1.16	0.21	0.33	0.18	24.96	0.00	3.36
Chi-squared	0.01	0.04	1.15	0.22	0.36	0.17	24.97	0.00	3.37
SVM	0.01	0.03	0.41	0.18	0.17	0.11	5.61	0.00	0.82
Filter	0.00	0.04	0.88	0.19	0.35	0.17	24.87	0.00	3.31
Classification average	0.01	0.04	1.25	0.19	0.32	0.17	22.53	0.00	

Table 12. Average processing time with on selected features.

Ranking Method	**NB**	**J48**	**SMO**	**JRIP**	**DT**	**Rd. Frt**	**Mul. Pr**	**Kstar**	**F.S. Avg**
Relief	0.00	0.02	1.45	0.13	0.13	0.10	12.15	0.00	1.75
GainRatio	0.00	0.04	0.93	0.13	0.12	0.10	9.36	0.00	1.34
InfoGain	0.00	0.02	0.99	0.12	0.17	0.14	13.06	0.00	1.81
OneR	0.00	0.02	0.98	0.12	0.17	0.10	10.72	0.00	1.51
SU	0.00	0.02	1.23	0.11	0.14	0.13	13.04	0.00	1.83
Chi-squared	0.00	0.02	1.00	0.13	0.13	0.11	12.59	0.00	1.75
SVM	0.00	0.02	0.30	0.08	0.07	0.09	2.55	0.00	0.39
Filter	0.00	0.02	0.80	0.11	0.13	0.11	12.99	0.00	1.77
Classification average	0	0.0225	0.96	0.11625	0.1325	0.11	10.8075	0	

Table 13. Average Classification Accuracy and Processing Time for classification Algorithms.

Classification Algorithms	Without Ranking on Full set		With Ranking On Full set		With ranking On selected set	
	F.S Avg.	Processing Time(S)	F.S Avg.	Processing Time(S)	F.S Avg.	Processing Time(S)
NaiveBayes	84.61	00.01	84.34	0.01	83.31	00.00
J48	89.76	00.04	89.65	0.04	88.58	00.02
SMO	89.56	01.40	89.39	01.25	87.03	00.96
JRIP	89.37	00.16	89.05	00.19	88.01	00.11
Decision Tree	85.47	00.32	85.37	00.32	85.02	00.13
Random Forest	90.27	00.16	90.57	00.17	88.53	00.11
Multilayer Perceptron	90.28	25.05	88.03	22.53	89.69	10.80
Kstar	86.33	00.00	86.43	00.00	86.94	00.00

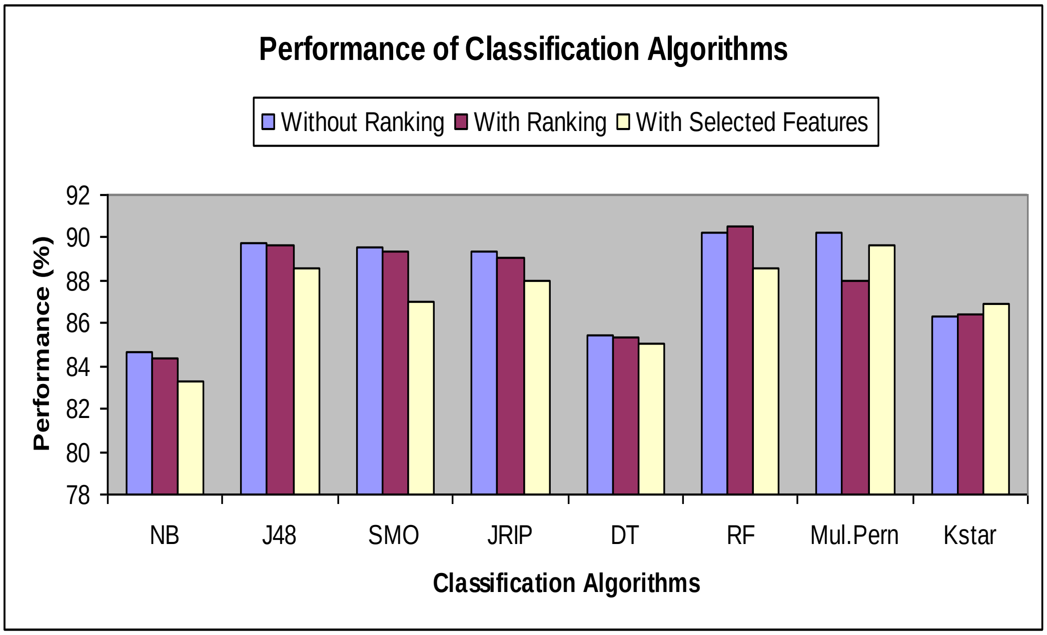


Figure 1. Performance of Classification Algorithms.

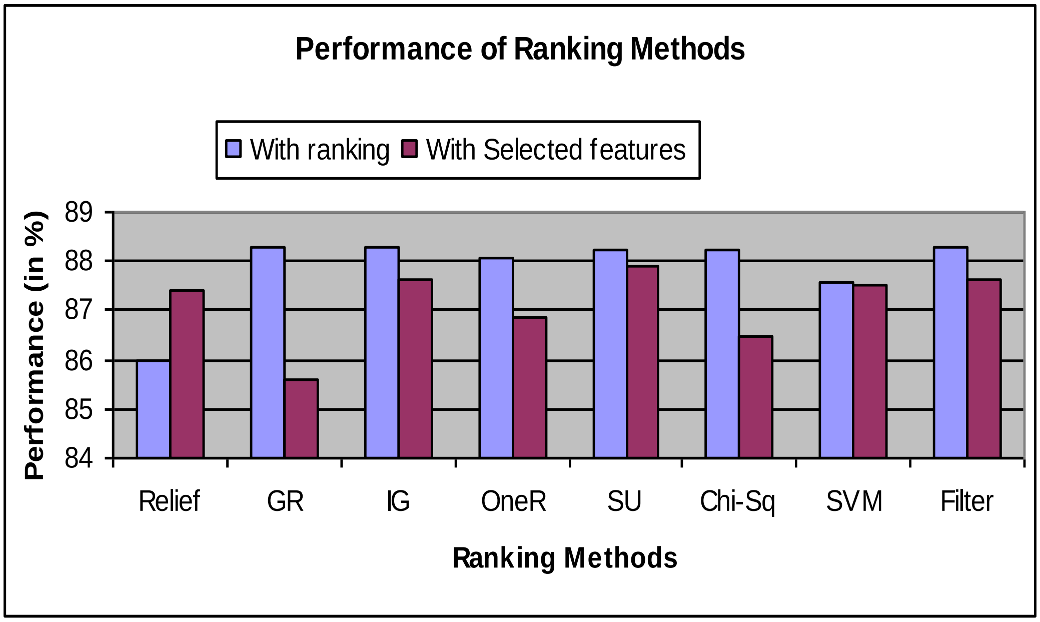


Figure 2. Performance of Ranking based on feature selection Algorithms

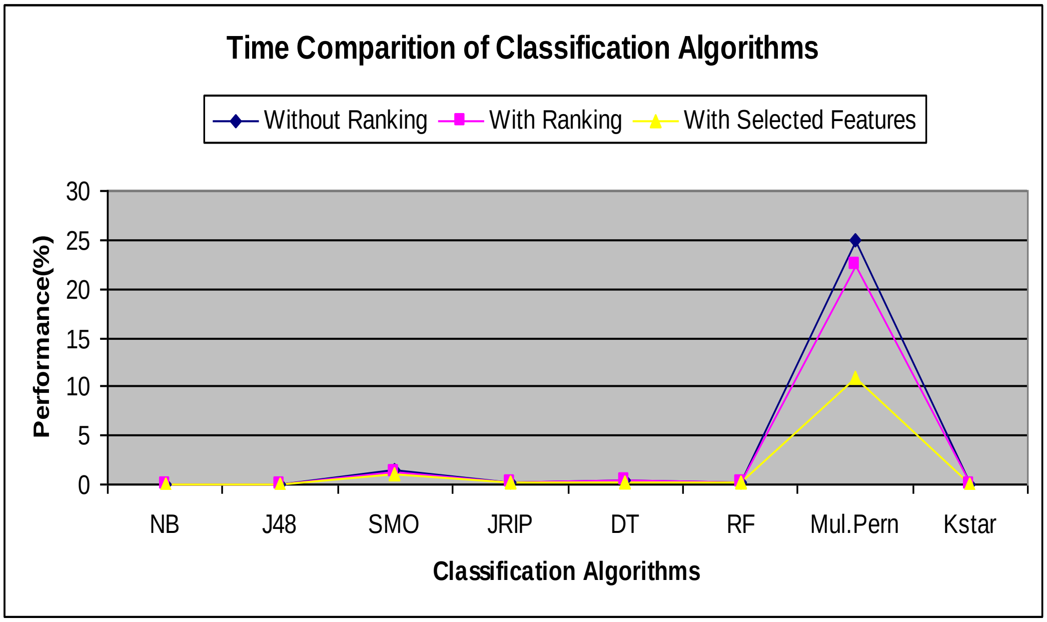


Figure 3. Processing Speed of Classification Algorithms.

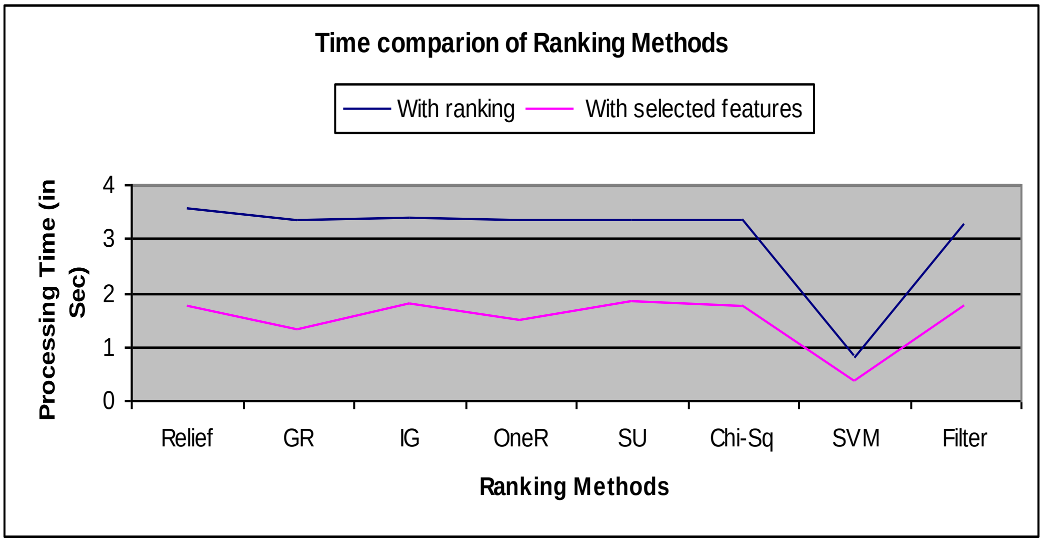


Figure 4. Processing Speed of Ranking Methods.

Results

Ranking from datasets is indeed a very important problem from both the algorithmic and performance perspective in data mining. Ranking methods with different classification algorithms gives different accuracy. Hence selection of ranking method is an important task for improving the classification accuracy. Not choosing the right ranking method for a dataset introduces bias towards selecting the best features. Furthermore predictive accuracy is not a useful measure when evolutionary classifies learned on datasets. In this study, out of eight ranking methods SVM scores the maximum accuracy for three datasets (vote, segment-challenge and diabetes) Chi-square scores for two datasets (ionosphere and soybean) and Filter, OneR, InfoGain scores for one datasets (ionosphere, diabetes). But it was found that Symmetrical Uncertainty (*SU*) which does not scores the maximum accuracy for any datasets give the maximum accuracy of 87.88 percentages comparing with other conventional ranking methods. The overall time taken by *SU* is higher when comparing with other ranking methods.

Conclusion

From this study, the following observations can be made:

1. Multilayer Perceptron, Random Forest, J48, SMO and JRIP perform better than other classification algorithms with and without ranking and also on selected features.
2. SVM ranking method will take a minimal processing time period with reasonable classification accuracy in comparison to other ranking methods.
3. The selected features by Relief ranking method provides better performance compared with ranking with full dataset.
4. With selected features, the performance of Gain Ratio is poorer than other ranking methods.
5. SU based ranking method reduces the number of initial attributes with maximum time period, and increases the classification performance, in comparison with other methods.

References

Amini, M.-R., Truong, T.-V., Goutte, C. (2008). A boosting algorithm for learning bipartite ranking functions with partially labeled data. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2008*, 99-106. doi:10.1145/1390334.1390354

Burges, C., Shaked, T., Renshaw, E., Lazier, A., Deeds, M., Hamilton, N., & Hullender, G. (2005). Learning to rank using gradient descent. In *Proceedings of the 22nd International Conference on Machine Learning*, 89-96. doi:10.1145/1102351.1102363

Cooper, W. S., Gey, F. C., & Dabney, D. P. (1992). Probabilistic retrieval based on staged logistic regression. In *Proceedings of the 15th annual international ACM SIGIR conference on research and development in information retrieval, SIGIR '92* (198-210). doi:10.1145/133160.133199

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*(3), 273-297. doi:10.1007/BF00994018

Fuhr, N. (1989). Optimum polynomial retrieval functions based on the probability ranking principle. *ACM Transactions on Information Systems, 7*(3), 183-204. doi:10.1145/65943.65944

Hall, M.A., & Smith, L.A. (1998). Practical feature subset selection for machine learning. In *Proceedings of the 21st Australian Computer Science Conference*, 181–191.

Herbrich, R., Graepel, T., & Obermayer, K. (2000). Large Margin Rank Boundaries for Ordinal Regression. *Advances in Large Margin Classifiers*, 115-132.

Holte, R.C. (1993). Very simple classification rules perform well on most commonly used datasets. *Machine Learning, 11*(1), 63-91. doi:10.1023/A:1022631118932

Kira, K., & Rendell, L. (1992). A practical approach to feature selection. In *ML92: Proceedings of the ninth international workshop on Machine learning* (249-256). Morgan Kaufmann Publishers Inc.

Kuo, J.-W., Cheng, P.-J., & Wang, H.-M. (2009). Learning to rank from bayesian decision inference. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management, CIKM'09*, doi:10.1145/1645953.1646058

Kuramochi, M., & Karypis, G. (2005). Gene classification using expression profiles: A feasibility study. *International Journal on Artificial Intelligence Tools*, *14*(4), 641-660. doi:10.1142/S0218213005002302

Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Liu, H., & Setiono, R. (1995). Chi2: Feature selection and discretization of numeric attributes. In *Tools with Artificial Intelligence, 1995. Proceedings, Seventh International Conference on*, 338-391. doi:10.1109/TAI.1995.479783

Moon, T., Smola, A., Chang, Y., & Zhen, Z. (2010). IntervalRank: Isotonic regression with listwise and pairwise constraint. In *Proceedings of the Third Acm International Conference on Web Search and Data Mining, WDSM "10*, 151-160. doi:10.1145/1718487.1718507

Rigutini, L., Papini, T., Maggini, M., & Scarselli, F. (2008). SortNet: learning to rank by a neural-based sorting algorithm. *SIGIR 2008 workshop: Learning to Rank for Information Retrieval*.

Robnik-Sikonja, M., & Kononenko, I. (2003). Theoretical and empirical analysis of ReliefF and RReliefF. *Machine Learning*, *53*(1/2), 23-69. doi:10.1023/A:1025667309714

Sculley, D. (2010) Combined regression and ranking. In *KDD '10: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 979-988. doi:10.1145/1835804.1835928

Valizadegan, H., Jin, R., Zhang, R., & Mao, J. (2009). Learning to rank by optimizing NDCG measure. *Advances in Neural Information Processing Systems*, *22*, 1883-1891.

Wu, Q., Burges, C. J. C., Svore, K. M., & Gao, J. (2008). Ranking, boosting, and model adaptation. *Microsoft Research Technical Report MSR-TR-2008-109*. Redmond, WA: Microsoft Research.

Yue, Y., Finley, T., Radlinski, F., & Joachims, T. (2007). A support vector method for optimizing average precision. In *SIGIR '07 : 30th annual International ACM SIGIR Conference on Research and Development in Information Retrieval : July 23-27, 2007, Amsterdam, the Netherlands*, 271. doi:10.1145/1277741.1277790

Journal of Modern Applied Statistical Methods
November 2015, Vol. 14, No. 2, 53-67.

ISSN 1538 − 9472

Two Stage Robust Ridge Method in a Linear Regression Model

Adewale Folaranmi Lukman
Ladoke Akintola
University of Technology
Ogbomoso, Nigeria

Oyedeji Isola Osowole
University of Ibadan
Ibadan, Nigeria

Kayode Ayinde
Ladoke Akintola
University of Technology
Ogbomoso, Nigeria

Two Stage Robust Ridge Estimators based on robust estimators M, MM, S, LTS are examined in the presence of autocorrelation, multicollinearity and outliers as alternative to Ordinary Least Square Estimator (OLS). The estimator based on S estimator performs better. Mean square error was used as a criterion for examining the performances of these estimators.

Keywords: Two Stage Least Square, Ridge Estimator, Ordinary Least Square, Robust Estimators, Two Stage Robust Ridge Estimator.

Introduction

Multiple regressions routinely assess the degree of relationship between one dependent variable and a set of independent variables. The Ordinary Least Squares (OLS) Estimator is most popularly used to estimate the parameters of regression model. Under certain assumptions, the estimator has some very attractive statistical properties which have made it one of the most powerful and popular estimators of regression model. A common violation in the assumption of classical linear regression model is the non-normal error terms. OLS estimator produces unstable prediction estimates when the assumption of normality of errors is not met (Ryan, 1996). Multiple regression methods also yield unstable results in the presence of outlier data points. When outliers occur in the data, the assumption of normally distributed errors is violated. An alternative strategy to deal with outliers is to accommodate them. Accommodation is accomplished by using any one of several robust regression estimation methods.

Adewale Folaranmi Lukman is a postgraduate student in the Department of Pure and Applied Statistics. Email at wale3005@yahoo.com. Dr. Oyedeji Isola Osowole is a Lecturer in the Department of Statistics. Email at dosowole@yahoo.com. Prof. Kayode Ayinde is a lecturer in the Department of Pure and Applied Mathematics. Email at: kayinde@lautech.edu.ng.

Also, the problem of autocorrelated error is another violation to the assumption of independence of error terms in classical linear regression model. The term autocorrelation may be defined as correlation between members of series of observations ordered in time as in time series data (Gujarati 1995). In the regression context, the classical linear regression model assumes that such autocorrelation does not exist in the disturbances ε_i. Symbolically

$$E\left(\varepsilon_i \varepsilon_j\right) = 0 \forall i \neq j \tag{1}$$

When this assumption breaks down, this is autocorrelation problem. A number of remedial procedures that rely on transformations of the variables have been developed. In order to correct for autocorrelation, one often uses Feasible Generalized Least Square (FGLS) procedures such as the Cochrane-Orcutt or Prais-Winsten two-step or the Maximum Likelihood Procedure or Two stage least Squares which are based on a particular estimator for the correlation coefficient (Green, 1993; Gujarati, 2003).

Another serious problem in regression estimation is multicollinearity. It is the term used to describe cases in which the explanatory variables are correlated. The regression coefficients possess large standard errors and some even have the wrong sign (Gujarati, 1995). In literature, there are various methods existing to solve this problem. Among them is the ridge regression estimator first introduced by Hoerl and Kennard (1970). Keijan (1993) proposed an estimator that is similar in form but different from the ridge regression estimator of Hoerl and Kennard. Ayinde and Lukman (2014) proposed some generalized linear estimator (CORC and ML) and principal components (PCs) estimator as alternative to multicollinearity estimation methods.

Inevitably, these problems can exist together in a data set. Holland (1973) proposed robust M-estimator for ridge regression to handle the problem of multicollinearity and outliers. Askin and Montgomery (1980) proposed ridge regression based on the M-estimates. Midi and Zahari (2007) proposed Ridge MM estimator (RMM) by combining the MM estimator and ridge regression. Samkar and Alpu (2010) proposed robust ridge regression methods based on M, S, MM and GM estimators. Maronna (2011) proposed robust MM estimator in ridge regression for high dimensional data. Eledum and Alkhaklifa (2012) proposed Generalized Two Stages Ridge Estimator (GTR) for the multiple linear model which suffers from both problem of autocorrelation AR (1) and multicollinearity.

The main objective of this study is to re-examine the study of Eledum and Alkhaklifa (2012). Efforts are made to correct the various assumptions violations of classical regression model which could have led into misleading conclusions. In this study, Two Stage Robust Ridge methods based on M, S, MM, LTS estimators are examined in the presence of outliers, autocorrelated errors and multicollinearity. A real life data considered in the study of Eledum and Alkhaklifa (2012) was used.

Outliers in least square regression

Barnett and Lewis (1994) define an outlier as an observation that appears inconsistent with the remainder of the data set. Outlier identification is important in OLS not only due to their impact on the OLS model, but also to provide insight into the process. These outlying cases may arise from a distribution different from the remaining data set. The distribution of the full dataset is contaminated in this instance. To statisticians, unusual observations are generally either outliers or 'influential' data points. In regression analysis, generally they categorize unusual observation (outliers) into three: outliers, high leverage points and influential observations. In other words, Hawkins (1980) pointed out that, an outlier is an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism.

Outliers are classified in three ways:

i. the change in the direction of response (Y) variable

ii. the deviation in the space of explanatory variable(s), deviated points in X-direction called leverage points and are also referred to as exterior X-space observation in this research, and

iii. The other is change in both directions (direction of the explanatory variable(s) and the response variable). According to Belsley, Kuh, and Welsch (1980), influential observations is one which either individual or together with several other observations have a demonstrably larger impact on the calculated values of various estimates than is the case for most of the other observations. Chatterjee and Hadi (1986) pointed out that, as with outliers, high leverage points need not be influential and influential observations are not necessarily high-leverage points. When an observation is considered to be both an outlier and influential, regression results are usually reported with and without the observation. When observations are not outliers but are influential, it is less clear what should be done.

Robustness ideas in regression

One idea to deal with this problem is to identify outliers, remove them, and then to proceed as before assuming we now have an appropriate data set for the standard methods. If the true coefficients were known, then outliers would not be hard to detect. Look for the points corresponding to the largest residuals. The field of regression diagnostics attempts to address the issue of how to identify influential points and outliers, in the general case when we do not know the true coefficient values. When there is only have one outlier, some diagnostic methods work very well by looking at the effect of one at a time deletion of data points. Unfortunately it is much more difficult to diagnose outliers when there are many of them, especially if the outliers appear in groups. In these situations, it is necessary to deal with the phenomena of outlier masking. Outlier masking occurs when a set of outliers goes undetected because of the presence of another set of outliers. Often when outliers are used to fit the parameter values, the estimates are badly biased, leaving residuals on the true outliers that do not indicate that they actually are outliers. Once there are several outliers, deletion methods are no longer computationally feasible. Then it is necessary to look at the deletion of all subsets of data points below a suitably chosen maximum number of outliers.

Another approach to dealing with outliers is robust regression, which tries to come up with estimators that are resistant or at least not strongly affected by the outliers. In studying the residuals of a robust regression, perhaps true outliers can be found. In this field many different ideas have been proposed, including Least Trimmed Squares (LTS), Least Median of Squares (LMS), M-estimators, and GM-estimators or bounded-influence estimators and S-estimators.

Robust regression and outlier diagnostic methods end up being very similar. They both involve trying to find outliers and trying to estimate coefficients in a manner that is not overly influenced by outliers. What is different is the order in which these two steps are performed. When using diagnostics, look for the outliers first and then once they have been removed use OLS on this clean data set for better estimates. Robust regression instead looks to find better robust estimates first and given these estimates, we can discover the outliers by analyzing the residuals.

Methodology

The data set was extracted from the study of Eledum and Alkhaklifa (2012); it represents the product in the manufacturing sector, the imported intermediate, the capital commodities and imported raw materials, in Iraq in the period from 1960 to 1990. An econometric model for this study is specified as follows:

$$Y = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon_t , t = 1, 2, \ldots, 31 \tag{2}$$

Where

Y = Product value in the manufacturing sector
X_1 = The value of the imported intermediate
X_2 = Imported capital commodities
X_3 = Value of imported raw materials
β_1, β_2, β_3 are the regression coefficients.

M-estimation procedure

The most common general method of robust regression is M-estimation, introduced by Huber (1964) that is nearly as efficient as OLS. Rather than minimize the sum of squared errors as the objective, the M-estimate minimizes a function ρ of the errors. The M-estimate objective function is,

$$\min \sum_{i=1}^{n} \rho\left(\frac{e_i}{s}\right) = \min \sum_{i=1}^{n} \rho\left(\frac{y_i - X'\hat{\beta}_i}{s}\right) \tag{3}$$

where s is an estimate of scale often formed from linear combination of the residuals. The function ρ gives the contribution of each residual to the objective function. A reasonable ρ should have the following properties:
$\rho(e) \geq 0, \rho(0) = 0, \rho(e) = \rho(-e),$ and $\rho(\mathrm{e}_i) \geq \rho(e_i')$ for $|e_i| \geq |e_i'|$
the system of normal equations to solve this minimization problem is found by taking partial derivatives with respect to β and setting them equal to 0, yielding,

$$\sum_{i=1}^{n} \psi\left(\frac{y_i - X'\hat{\beta}_i}{s}\right) X_i = 0 \tag{4}$$

where ψ is a derivative of ρ. The choice of the ψ function is based on the preference of how much weight to assign outliers. Newton-Raphson and iteratively reweighted Least Squares (IRLS) are the two methods to solve the M-estimates nonlinear normal equations. IRLS expresses the normal equations as,

$$X'WX\hat{\beta} = X'Wy \tag{5}$$

MM estimator

MM-estimation is special type of M-estimation developed by Yohai (1987). MM--estimators combine the high asymptotic relative efficiency of M-estimators with the high breakdown of class of estimators called S-estimators. It was among the first robust estimators to have these two properties simultaneously. The 'MM' refers to the fact that multiple M-estimation procedures are carried out in the computation of the estimator. Yohai (1987) describes the three stages that define an MM-estimator:

1. A high breakdown estimator is used to find an initial estimate, which we denote $\tilde{\beta}$ the estimator need to be efficient. Using this estimate the residuals, $r_i(\beta) = y_i - x_i^T\tilde{\beta}$ are computed.
2. Using these residuals from the robust fit and $\frac{1}{n}\sum_{i=1}^{n}\rho\left(\frac{r_i}{s}\right) = k$ where k is a constant and the objective function ρ, an M-estimate of scale with 50% BDP is computed. This $s\left(r_i\left(\tilde{\beta}\right),\ldots,r_n\left(\tilde{\beta}\right)\right)$ is denoted s_n. The objective function used in this stage is labeled ρ_0.
3. The MM-estimator is now defined as an M-estimator of β using a redescending score function, $\varphi_1(u) = \frac{\partial\rho_1(u)}{\partial u}$, and the scale estimate s_n obtained from stage 2. So an MM-estimator $\hat{\beta}$ defined as a solution to

$$\sum_{i=1}^{n} x_{ij}\varphi_1\left(\frac{y_i - x_i^T\tilde{\beta}}{s_n}\right) = 0,\ j = 1,\ldots,p. \tag{6}$$

S estimator

Rousseeuw and Yohai (1984) introduced *S* estimator, which is derived from a scale statistics in an implicit way, corresponding to $s(\theta)$ where $s(\theta)$ is a certain type of robust M-estimate of the scale of the residuals $e_1(\theta), \ldots, e_n(\theta)$. They are defined by minimization of the dispersion of the residuals: minimize $S\left(e_1(\theta), \ldots, e_n\left(\hat{\theta}\right)\right)$ with final scale estimate $\hat{\sigma} = S\left(e_1(\theta), \ldots, e_n\left(\hat{\theta}\right)\right)$. The dispersion $\left(e_1(\theta), \ldots, e_n\left(\hat{\theta}\right)\right)$ is defined as the solution of

$$\frac{1}{n}\sum_{i=1}^{n}\rho\left(\frac{e_i}{s}\right) = k \tag{7}$$

K is a constant and $\rho\left(\frac{e_i}{s}\right)$ is the residual function. Rousseeuw and Yohai (1984) suggest Tukey's biweight function given by:

$$\rho(x) = \begin{cases} \dfrac{x^2}{2} - \dfrac{x^4}{2c^2} + \dfrac{x^6}{6c^4} & \text{for } |x| \le c \\ \dfrac{c^2}{6} & \text{for } |x| > c \end{cases} \tag{8}$$

Setting $c = 1.5476$ and $K = 0.1995$ gives 50% breakdown point (Rousseeuw & Leroy, 1987).

LTS estimator

Rousseeuw (1984) developed the least trimmed squares estimation method. Extending from the trimmed mean, LTS regression minimizes the sum of trimmed squared residuals. This method is given by,

$$\hat{\beta}_{LTS} = \arg\min Q_{LTS}(\beta) \tag{9}$$

where $Q_{LTS}(\beta) = \sum_{i=1}^{h} e_i^2$ such that $e_{(1)}^2 \le e_{(2)}^2 \le e_{(3)}^2 \le \ldots \le e_{(n)}^2$ are the ordered squares residuals and *h* is defined in the range $\frac{n}{2} + 1 \le h \le \frac{3n + p + 1}{4}$, with *n* and *p* being sample size and number of parameters respectively. The largest squared residuals are excluded from the summation in this method, which allows those outlier data points to be excluded completely. Depending on the value of *h* and the outlier data configuration. LTS can be very efficient. In fact, if the exact numbers of outlying data points are trimmed, this method is computationally equivalent to OLS.

Two Stage Robust Ridge Estimator

Two Stage Ridge Regression approach used by Eledum and Alkhaklifa (2012) and Robust Ridge Regression Methods adopted by Samkar and Alpu (2010) are combined in this study to obtain Two Stage Robust Ridge Regression. This method is adopted to deal with the problem of autocorrelated error, outliers and, multicollinearity sequentially. Consider the Linear regression model:

$$Y = X\beta + u_t \tag{10}$$

X is an $n \times p$ matrix with full rank, Y is a $n \times 1$ vector of dependent variable, β is a $p \times 1$ vector of unknown parameters, and ε is the error term such that $E(\varepsilon) = 0$ and $E(\varepsilon\varepsilon') = \sigma^2 I$ and assume that the error term follows the AR(1) scheme, namely,

$$u_t = \rho u_{t-1} + \varepsilon_t, -1 < \rho < 1 \tag{11}$$

ε_t is a white noise error term such that $\varepsilon_t \sim N(0, \sigma^2 I)$
Premultiply equation (10) by P we obtain:

$$PY = PX\beta + PU \tag{12}$$

Equivalently, equation (12) becomes:

$$Y^* = X^*\beta + U^* \tag{13}$$

P is a non-singular matrix such that $P\Omega P' = I$ which implies $PP' = \Omega^{-1}$, $U^* \sim N(0, \sigma^2 I)$, $Y^* = PY$, $X^* = PX$, and $U^* = PU$.

Therefore, we can apply Robust Estimators to the transformed model (5) and obtain Two Stage Robust Estimator.

$$\hat{\beta}_{TRE} = \left(X^{*\prime} X^{*}\right)^{-1} X^{*\prime} Y^{*} = \left(X'P'PX\right)^{-1} X'P'PY$$

$$\hat{\beta}_{TRE} = \left(X'\Omega^{-1}X\right)^{-1} X'\Omega^{-1}Y \tag{14}$$

The variance-covariance matrix becomes:

$$V\left(\hat{\beta}_{TRE}\right) = \sigma^{2}\left(X'\Omega^{-1}X\right)^{-1}(3.6) \tag{15}$$

where

$$E\left(UU'\right) = \sigma^{2}\Omega = \sigma^{2}\begin{bmatrix} 1 & \rho & \rho^{2} & \cdots & \rho^{n-1} \\ \rho & 1 & \rho & \cdots & \rho^{n-2} \\ \rho^{2} & \rho & 1 & \cdots & \rho^{n-3} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \cdots & 1 \end{bmatrix}$$

$$\sigma^{2} = \frac{\sigma_{\varepsilon}^{2}}{1-\rho^{2}}$$

and the inverse of Ω is

$$\Omega^{-1} = \frac{1}{1-\rho^{2}}\begin{bmatrix} 1 & -\rho & 0 & \cdots & 0 \\ -\rho & 1+\rho^{2} & -\rho & \cdots & 0 \\ 0 & -\rho & 1+\rho^{2} & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}$$

Consider, $(n-1) \times n$ matrix P^{*} for transformation.

$$P^{*} = \begin{bmatrix} -\rho & 1 & 0 & \cdots & 0 \\ 0 & -\rho & 1 & \cdots & 0 \\ 0 & 0 & -\rho & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}$$

Therefore, $P^{*}{'}P^{*} = P$ by adding a new row with $\sqrt{1-\rho^{2}}$ in the first position and zero elsewhere.

$$P = \begin{bmatrix} \sqrt{1-\rho^{2}} & 0 & 0 & \cdots & 0 \\ -\rho & 1 & 0 & \cdots & 0 \\ 0 & -\rho & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & 0 \\ 0 & 1 & 1 & -\rho & 1 \end{bmatrix}$$

Then

$$Y^{*} = PY = \begin{bmatrix} \sqrt{1-\rho^{2}} & 0 & 0 & \cdots & 0 \\ -\rho & 1 & 0 & \cdots & 0 \\ 0 & -\rho & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix} \begin{bmatrix} Y_{1} \\ Y_{2} \\ Y_{3} \\ \vdots \\ Y_{n} \end{bmatrix}$$

$$X^{*} = PX = \begin{bmatrix} \sqrt{1-\rho^{2}} & 0 & 0 & \cdots & 0 \\ -\rho & 1 & 0 & \cdots & 0 \\ 0 & -\rho & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix} \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1p} \\ 1 & X_{21} & X_{22} & \cdots & X_{2p} \\ 1 & X_{31} & X_{32} & \cdots & X_{3p} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{np} \end{bmatrix}$$

$$\Omega = P'P = \begin{bmatrix} 1 & -\rho & 0 & \cdots & 0 \\ -\rho & 1+\rho^2 & -\rho & \cdots & 0 \\ 0 & -\rho & 1+\rho^2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}$$

However, the estimate obtained from applying Robust Estimators to the transformed model is used to obtain the ridge parameter K which is used in the Ridge Estimator since the estimates obtain from OLS will be inefficient when we have the problem of outliers or non-normal error term.

Results

From Table 1, it can be seen that estimation based on the OLS estimator produces residuals that reveals the problem of autocorrelation (DW p-value=0.0005) and multicollinearity (VIF>10) simultaneously. The problem of multicollinearity might be the reason for the wrong sign in the value of imported raw materials. We handle the problem of autocorrelation in Table 2 by transforming the data set. The original data set is transformed using $\hat{\rho} = 0.547$ (from Table 1) to correct the problem of autocorrelation by applying Two Stage Least Squares. Table 2 shows that the new data set obtain through transformation suffered the problem of non-normal error term using Jarque-Bera Statistic and Table 3 also shows the presence of bad leverages using robust diagnostics which might be the reason for the non-normality of the error term. The data set still suffered the problem of multicollinearity (VIF>10) as revealed in Table 2. Due to the presence of bad leverages OLS will not correctly estimate the parameters in the model. This prompts the use of the Two Stage Robust Estimators in Table 4. LTS and S estimators perform better than other estimators when we have leverages and outliers in y axis (bad leverages) in terms of the MSE (B). But the coefficient of LTS seems to be much different from the class of other estimators. We then prefer to consider S estimator in its stead. Due to the occurrence of both problem of multicollinearity and bad leverages in the new data set, we then use the Ridge combined with S estimator adopted from the concept of Samkar and Alpu (2010) to compute the ridge parameter. Geometric version of the ridge parameter proposed by Kibria (2003) was used $\hat{K}_{GM} = \frac{\hat{\sigma}^2}{\left(\prod_{i=1}^{p}\alpha_i\right)^{\frac{1}{p}}}$ where $\hat{\sigma}^2$ is the variance obtained from S estimator and α_i is the obtained coefficient.

Table 1. Ordinary Least Square (OLS)

Variable	Coefficient	Std. Error	*p*-value	VIF
*X*1	0.208	0.218	0.348	128.26
*X*2	0.921	0.196	0.000	103.43
*X*3	-1.34	0.162	0.415	70.87
R-squared	0.9896	DW	0.0005	
Jarque-Bera *p*-value	0.2493	σ^2	0.0111	
RHO	0.547			

Table 2. Two Stage Least Square (TS)

Variable	Coefficient	Std. Error	*p*-value	VIF
*X*1T	0.200	0.160	0.2211	26.839
*X*2T	0.963	0.191	0.0000	38.358
*X*3T	-0.1790	0.127	0.1687	16.904
R-squared	0.9735	DW *p*-value	0.2332	
Jarque-Bera *p*-value	0.0732	σ^2	0.028	
RHO	0.11			

Table 3. Robust Diagnostics

Observation	Mahalanobis	Robust MCD Distance	Leverage	Standardized Robust Residual	Outlier
12	1.5024	5.8641	*	4.7737	*
14	0.9716	3.0421		4.9055	*
15	4.6559	29.4708	*	9.1178	*
16	1.0615	8.2135	*	11.2653	*
17	1.6992	8.6846	*	1.4033	
18	2.2534	19.0971	*	-0.5591	
20	3.0865	24.3649	*	-2.4415	
21	3.8595	26.6181	*	0.4649	
22	1.2315	8.8886	*	0.4301	
30	3.421	3.0381		16.2649	*
31	1.2827	1.1007		-8.5191	*

Table 4. Two Stage Robust Estimators and OLS

Variables	OLS	TS	M	MM	S	LTS
*X*1T	0.208	0.200	0.329	0.328	0.346	0.032
*X*2T	0.921	0.963	0.976	0.976	0.963	1.723
*X*3T	-1.34	-0.1790	-0.228	-0.228	-0.221	-0.648
R-squared	0.9896	0.9735	0.7918	0.7939	0.8023	0.9951
σ^2	0.0111	0.028	0.0102	0.019	0.017	0.003
MSE(*B*)	0.1122	0.0782	0.0324	0.0303	0.0272	0.029

Table 5. Two Stage Robust Ridge Estimators

Variables	Coefficient	VIF
*X*1	0.3443	1.2972
*X*2	0.4278	1.0011
*X*3	0.1836	1.5526
MSE(β)	0.071687	
K	0.097	

Conclusion

OLS performs better than other estimators when there is no violation of assumptions in Classical Linear Regression Model. In this study the problem of autocorrelation was handled using Two Stage Least Square. The problem of multicollinearity and outlier are still presents. OLS will not be efficient because of the present of both problem therefore we apply Robust Methods to the transformed data. *S* and LTS estimators perform better than other Robust Methods in terms of the MSE. *S* estimator was chosen because LTS does not correctly estimate the model when compared with other estimators. Ridge parameter *K* is then obtained using the estimates obtain from *S* estimation. Robust ridge estimates was computed. Two stage robust ridge estimator performs better than the Generalized Two stage ridge regression proposed by Hussein et al (2012). This is because after the problem of autocorrelation was corrected in the study of Hussein et al (2012), the data sets still suffered the problem of multicollinearity and outlier. This was corrected in this study by obtaining the ridge parameter using a robust estimator instead of OLS.

Authors’ note

We acknowledge the assistance of Osowole O.I, Kayode Ayinde, and X Hussein.

References

Askin, G. R., & Montgomery, D. C. (1980). Augmented robust estimators. *Technometrics*, *22*(3), 333-341. doi:10.1080/00401706.1980.10486164

Ayinde, K. & Lukman, A. F. (2014). Combined estimators as alternative to multicollinearity estimation methods. *International Journal of Current Research*, *6*(1), 4505-4510.

Barnett, V. & Lewis, T. (1994). *Outliers in statistical data* (Vol. 3). New York: Wiley.

Belsley, D. A., Kuh, E., & Welsch, R. E. (1980). *Regression diagnostics: identifying influential data and sources of collinearity*. New York: Wiley.

Chatterjee, S., Hadi, A. S. (1986). Influential observations, high leverage points, and outliers in linear regression. *Statistical Science*, *1*(3), 379-416. doi:10.1214/ss/1177013622

Eledum, H. Y. A., Alkhalifa, A. A. (2012). Generalized two stages ridge regression estimator for multicollinearity and autocorrelated errors. *Canadian Journal on Science and Engineering Mathematics*, *3*(3), 79-85.

Green, W.H. (1993). *Econometric analysis* (2nd ed.). New York: MacMillan.

Gujarati, D. N. (1995). *Basic econometrics* (3rd ed.). New York: McGraw-Hill.

Gujarati, D. N. (2003). *Basic econometrics* (4th ed.) (pp. 748, 807). New Delhi: Tata McGraw-Hill.

Hawkins, D. M. (1980), *Identification of Outliers*, London: Chapman & Hall.

Hoerl, A. E. & Kennard, R.W. (1970). Ridge regression biased estimation for nonorthognal problems. *Technometrics*, *12*(1), 55-67. doi:10.1080/00401706.1970.10488634

Holland, P. W. (1973). *Weighted ridge regression: Combining ridge and robust regression methods* (NBER Working Paper No.11). Cambridge, MA: National Bureau of Economic Research. doi:10.3386/w0011

Huber, P. J. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, *35*(1), 73-101.doi:10.1214/aoms/1177703732

Keijan, L. (1993). A new class of biased estimate in linear regression. *Communications in Statistics-Theory and Methods*, *22*(2), 393-402. doi:10.1080/03610929308831027

Maronna, R. A. (2011). Robust ridge regression for high-dimensional data. *Technometrics*, *53*(1), 44-53. doi:10.1198/TECH.2010.09114

Midi, H., & Zahari, M. (2007). A simulation study on ridge regression estimators in the presence of outliers and multicollinearity. *Jurnal Teknologi*, *47*(C), 59-74. doi:10.11113/jt.v47.261

Rousseeuw, P.J. (1984). Least Median of Squares Regression. *Journal of the American Statistical Association*, 79, 871–880. doi:10.1080/01621459.1984.10477105

Rousseeuw P. J., & Leroy A. M. (1987). *Robust Regression and Outlier Detection*. New York: Wiley-Interscience.

Rousseeuw, P. J. & Van Driessen, K. (1998). Computing LTS regression for large data sets. *Technical Report*, University of Antwerp, submitted.

Rousseeuw P. J., & Yohai, V. (1984). Robust regression by means of S-estimators. In J. Franke, W. Härdle & D. Martin (Eds.), *Robust and Nonlinear Time Series Analysis* (Vol. 26, pp. 256-272). New York: Springer-Verlag. doi:10.1007/978-1-4615-7821-5_15

Ryan, T.P. (1996). *Modern Regression Method*. Canada: John Wiley & Sons.

Samkar, H. & Alpu, O. (2010). Ridge regression based on some robust estimators. *Journal of Modern Applied Statistical Methods*, *9*(2), 17.

Yohai, V. J. (1987). High breakdown point and high breakdown-point and high efficiency robust estimates for regression. *The Annals of Statistics*, *15*(2), 642-656. doi:10.1214/aos/1176350366

Journal of Modern Applied Statistical Methods
November 2015, Vol. 14, No. 2, 68-87.

ISSN 1538 − 9472

Semi-Parametric Non-Proportional Hazard Model with Time Varying Covariate

Kazeem A. Adeleke
Obafemi Awolowo University
Ile-Ife, Nigeria

Alfred A. Abiodun
University of Ilorin, Kwara State
Ilorin, Nigeria

R. A. Ipinyomi
University of Ilorin, Kwara State
Ilorin, Nigeria

The application of survival analysis has extended the importance of statistical methods for time to event data that incorporate time dependent covariates. The Cox proportional hazards model is one such method that is widely used. An extension of the Cox model with time-dependent covariates was adopted when proportionality assumption are violated. The purpose of this study is to validate the model assumption when hazard rate varies with time. This approach is applied to model data on duration of infertility subject to time varying covariate. Validity is assessed by a set of simulation experiments and results indicate that a non proportional hazard model performs well in the phase of violated assumptions of the Cox proportional hazards.

Keywords: Survival time, non-proportional hazards, time-dependent covariate, semi parametric model.

Introduction

In survival or life testing experiments, the assumption of Cox model (1972), may not hold. Example of this is when effect of a treatment on survival diminishes in the course of time to event. Different systems have different prognostic factors, some are time fixed although some are time varying. One advantage of Cox proportional regression models is the ability to incorporate time varying coefficients and time varying covariates (Cox, 1972, Therneau & Grambsch, 2000). The former refers to a variable that is measured at baseline and whose values remain fixed to a variable whose value remains fixed over the duration of follow-up. Although, its effects on hazards is allowed to change over the follow-up period. The later refers to a variable whose value itself varies over time of follow-up. Example of time varying covariate includes the exposure of a pharmaceutical agent to cumulative dosage of radiation, duration of relationship

Kazeem A. Adeleke is a lecturer in the Mathematics Department. Email him at: aadeleke@oauife.edu.ng. Alfred A. Abiodun is a lecturer in the Department of Statistics. R. A. Ipinyomi is an Professor of Statistics. Email him at: ipinyomira@yahoo.co.uk.

as a measure of duration of infertility in marriage, the receipt of an organ transplant. The natures of time varying covariate are very important and take major role of this work. In the above example, the first and second are continuous time variates whose value is non-decreasing over the time, the third example which is the receipt of an organ is also a time varying covariate but dichotomous in nature because the subject may be exposed or unexposed to the treatment.

Recently a number of studies have been directed towards modelling time varying covariates as well as stratification which are semi-parametric non-proportional hazard models (Austin, 2012, Lehr, 2004, Abrahamowicz, 2007, Bender, Augustin, & Blettner, 2005, Ata & Sozer, 2007, Austin, 2012, Zhou, 2001). A more advanced method of generating time varying covariate is the work of Zhou (2001) where the use of an exponential distribution was examined in conjunction with a transformation to the Cox model including time varying covariate. A piecewise exponential distribution was used to obtain a dichotomous or step function covariate which was in turn incorporated into the Cox model and analysed through a semi-parametric approach.

Bender et al. (2005) generated survival data that follows Cox proportional hazard model using three parametric distributions namely: exponential, Weibull and Gompertz and limited his study to only time fixed covariate. New extensions of Cox model with time varying covariate have been developed by Sylvestre and Abrahmowicz (2007) due to an undiscovered and complicated nature of longitudinal data structure where validation is made through simulation. They described and evaluated two alternatives for generation of survival times conditional on time varying covariate.

Applications of Cox model with time varying covariate are likely to continue to become increasingly important in medical research. The methods put forth by Sylvester and Abrahmowicz are however not presented in a close form. Leemis (1987), Leemis, Shih and Ryertson (1990), and Shih and Leemis, (1993) have offered different frameworks for generation of survival time that follow a Cox model with time varying following accelerated life and proportional hazards models where his procedures adopted one time varying covariate and no time fixed covariates. A recent study on Cox regression model in the presence of non-proportional hazards was carried out by Ata and Sozer (2007), where they worked on alternative different models in the violation of proportional assumption. Our study extend the work of Bender et al. (2005), and Zhou (2001), with an additional argument that allows for a fixed covariate, continuous time varying covariate and a step function covariate using exponential model see Austin (2012).

Non-proportional hazards models

Recall the Cox proportional hazards model with time fixed covariate *x*

$$h_i(t) = h_i(t, \underline{x}) = h_0(t)\exp\{\underline{\beta}'\underline{x}\} \tag{1}$$

where $h_0(t)$ is a non-parametric baseline hazard function $\beta' = (\beta_1, \beta_2, \ldots, \beta_p)$ is a vector of regression coefficients, and $x_i = x_1, x_2, \ldots, x_p$ is a vector of time fixed covariates for ith subject.

Although $h_0(t)$ is chosen arbitrarily with no distribution attached, the fact that $\exp(\underline{\beta}'\underline{x})$ is a parametric exponential function that assumes parametric forms of the predictors on hazards makes model in (1) a semi-parametric model.

Proportional hazard assumption

In linear regression modelling, the measure of effect is usually regression coefficient β, in logistic regression the measure of effect is an odds ratio, Walker and Duncan(1976), Hosmer and Lemeshow (2000), Agresti (2007), Adeleke and Adepoju (2010), the log of which is β, but in survival analysis, the measure of effect is the hazard ratio (Tableman and Kim, 2004). Proportional hazards assumption states that the hazard ratio is constant over time or the hazard for an individual is proportional to the hazard for any other individual (Kleinbaum and Klein, 2005). For example, if x and x^* are the covariates for two individual then

$$HR = \frac{h(t|x)}{h(t|x^*)} = \frac{h_0(t)\exp(\hat{\underline{\beta}}'\underline{x})}{h_0(t)\exp(\hat{\underline{\beta}}'\underline{x}_*)} = \exp^{\left(\sum(\underline{x}-\underline{x}_*)'\hat{\underline{\beta}}\right)} \tag{2}$$

The hazard ratio in (2) can also be expressed as $HR = \theta$, which implies that the hazard ration is time-independent.

Now let the effect of a time varying covariate on survival probability at a time $t(\beta_t)$ depend on the value of this variable at the same time, then an extended version of (1) by Cox (1972) can be given by

$$h(t, z(t), x) = h_0(t)\exp\sum_1^{p1}\beta_i x_i + \sum_1^{p2}\gamma_i z_i(t) \tag{3}$$

which can be written as $HR = \theta t$

Let the proportional hazard for a survival time T be given by

$$h\left(T_i / X\right) = \exp\left(\underline{\beta}'\underline{x}\right) h_0\left(t\right) \tag{4}$$

Then the cumulative distribution of T_i can be given as

$$F_T\left(t\right) = P\left(T_i \leq t\right) = P\left(\exp\left(\beta' x\right) h_0\left(t\right) \leq t\right) \tag{5}$$

$$\begin{aligned} H\left(t, z\left(t\right), x\right) &= \lambda \exp\left(\beta' x\right) t \\ F_T\left(t\right) &= P\left(T_i \leq t\right) \\ &= P\left(\exp\left(\underline{\beta}'\underline{x}\right) h_0\left(t\right) \leq t\right) \end{aligned} \tag{6}$$

$$\begin{aligned} P\left(h_0\left(t\right)\right) &\leq \frac{t}{\exp\left(\underline{\beta}'\underline{x}\right)} \\ &= F_{\exp(1)}\left(\frac{t}{\exp\left(\underline{\beta}'\underline{x}\right)}\right) \\ &= 1 - \exp\left(-\frac{t}{\exp\left(\underline{\beta}'\underline{x}\right)}\right) \end{aligned}$$

Now if $S_T(t) = 1 - F_T(t)$

$$T_i \approx S_T\left(t\right) = \exp\left(-\frac{1}{\exp\left(\underline{\beta}'\underline{x}\right)}\right) \tag{7}$$

Let Y_i be a uniform random variable with cumulative distribution function F and density function f, then

$$\begin{aligned} U &= F\left(Y_i\right) \sim U\left(0,1\right) \\ F\left(Y_i\right) &= \int_{-\infty}^{y} f\left(u\right) \sim U\left(0,1\right) \end{aligned}$$

Also

$$U = \exp\left[-H(T)\exp\left\{\underline{\beta}'\underline{x}\right\}\right] \sim U(0,1)$$

$$T = H^{-1}\left[-\log(U)\exp\left(\underline{\beta}'\underline{x}\right)\right] \tag{8}$$

where U is a uniform random variable (Bender et al, 2005). However, the survival time T does not involve time varying variable(s). By introducing the second covariate with time change when covariate is dichotomous, following the formulation of Zhou (2001) and Austin (2012), we define

$$Z_i(t) = \begin{cases} 0, & \text{for } t \le t_0 \\ 1, & \text{fot } t > t_0 \end{cases}$$

then the hazard function with dichotomous time changed covariate is

$$h_g(Y_i) = h_0(t)\exp\left\{\underline{\beta}'\underline{x} + \gamma'\underline{z}(t)\right\} \tag{9}$$

A natural problem is when time varying covariate is not dichotomous or step function but continuous. Zhou (2001) did not consider this, and Sylvestre and Abrahamowicz (2007) found the method was limited in applicability. For a case open to both time fixed and time varying covariate which is flexible for both step function and continuous system, see Austin (2012).

The cumulative hazard function and survival function H(.) and S(.) are:

$$H\left(t \middle| z(t), x\right) = \int_0^t h_0(s)\exp\left(\underline{\beta}'\underline{x} + \gamma'\underline{z}(s)\right)ds \tag{10}$$

$$S\left(t, z(t), x\right) = \exp\left[-H\left(t, z(t), x\right)\right]$$

Suppose the covariate follows a step function for $t \ge t_0$ i.e right censored data, then supposed the time is partitioned into two such that

$$Z_i(t) = \begin{cases} 0, & \text{for } t < t_0 \\ 1, & \text{fot } t \geq t_0 \end{cases}$$

Let D = domain and $D_1 = [0, t_0)$ and $D_2 = [t_0, \infty)$ then,
for $t < t_0$,

$$\begin{aligned} H(t, z(t), x) &= \int_0^t \lambda \exp\left(\underline{\beta}'\underline{x} + \gamma'\underline{z}(s)\right) ds \\ &= \lambda \exp\left(\underline{\beta}'\underline{x}\right) \int_0^t ds \\ &= \lambda \exp\left(\underline{\beta}'\underline{x}\right) [s]_0^t \\ &= \lambda \exp\left(\underline{\beta}'\underline{x}\right) t \\ H(t, z(t), x) &= \lambda \exp\left(\underline{\beta}'\underline{x}\right) t \end{aligned} \tag{11}$$

Using Bender et al. (2005), we obtain survival time

$$T = \frac{-\log U}{\lambda \exp\left(\underline{\beta}'\underline{x}\right)} \tag{12}$$

By Austin (2012), when $t \geq t_0$, using the condition above, the hazard function in (9) becomes

When $D_2 = t \geq t_0$, from 5, $Z(u) = 1$ then 6 becomes

$$\begin{aligned} &= \int_0^{t_0} \lambda \exp\left(\underline{\beta}'\underline{x}\right) du + \int_{t_0}^{0} \lambda \exp\left(\underline{\beta}'\underline{x} + \gamma\right) du \\ &= \lambda \exp\left(\underline{\beta}'\underline{x}\right) t_0 + \lambda \exp\left(\underline{\beta}'\underline{x} + \gamma\right)(t - t_0) \\ &= \lambda \exp\left(\underline{\beta}'\underline{x}\right) t_0 \left[1 - \exp(\gamma)\right] + \lambda \exp\left(\underline{\beta}'\underline{x} + \gamma\right) t \end{aligned}$$

by transformation

$$-\log(U) = \lambda \exp\left(\underline{\beta}'\underline{x}\right) t_0 \left[1 - \exp(\gamma)\right] + \lambda \exp\left(\underline{\beta}'\underline{x} + \gamma\right) T$$

The survival time obtained from the inverse cumulative hazards is

$$T=\left[\frac{-\log(U)-\lambda\exp\left(\underline{\beta}'\underline{x}\right)[t_0+\lambda\exp(\gamma_i)(t-t_0))}{\lambda\exp(\beta_i x_i+\gamma_i)}\right] \tag{13}$$

If however covariate is continuous the cumulative hazards is

$$H\left(t\middle|z(t),x\right)=\int_0^t h_0(s)\exp\left\{\underline{\beta}'\underline{x}+\gamma'\underline{z}(s)\right\}ds \tag{14}$$

Assume that $\underline{z}(s)$ is proportional to t such that $\underline{z}(s)=kt$ where $k > 0$. Hence the cumulative hazard from the above becomes

$$\begin{aligned} H\left(t\middle|z(t),x\right)&=\int_0^t \lambda\exp\left\{\underline{\beta}'\underline{x}+\gamma'\underline{z}(s)\right\}ds \\ &=\lambda\exp\left(\underline{\beta}'\underline{x}\right)\int_0^t \exp(\gamma ks)\,ds \\ (t,k,x)&=\frac{\lambda\exp\left(\underline{\beta}'\underline{x}\right)}{\lambda k}\left[\exp(\gamma kt-1)\right] \end{aligned} \tag{15}$$

Hence

$$\begin{aligned} &U=\exp\left(-H(T,k,x)\right) \\ &\left(-\log(U)\right)\gamma_i k=\lambda\exp\left(\underline{\beta}'\underline{x}\right)\left[\exp(\gamma_i kT-1)\right] \end{aligned}$$

so that

$$T=\frac{1}{\gamma_i k}\log\frac{\left[\lambda\exp\left(\underline{\beta}'\underline{x}\right)+\gamma_i k\left(-\log(U)\right)\right]}{\lambda\exp\left(\underline{\beta}'\underline{x}\right)} \tag{16}$$

Equations (12) and (13) and (16) will be used to obtain survival times for dicotonomous time varying covariate and continuous time varying; *U* can be obtained from *R*.

Non-parametric estimation

Follow the formulation of Kaplan and Meier (K-M) (1958) for estimating censored data. The method provides alternative way to life table approach where each interval contains only one observation.

The idea of K-M estimator is given by the conditional probability ($t \le t_0$) be the survival time of n randomly sampled individual study such that $t_1 \le t_2 \le ,.., \le t_n$ are of $T_1, T_2, ..., T_n$ where $S(t) \sim b(n, p)$ and $P = P(T \ge t)$ then, for $t \le t_{i+1}$

$$\begin{aligned} S(t_i) &\le P(T > t_i, T > t_{i-1}) \\ &= P(T > t_i | T > t_{i-1}) P(T > t_{i-1}) \\ &= P(T > t_i | T > t_{i-1}) P(T > t_{i-1} | T > t_{i-1}) \ldots P(T > t_0 = 0) \end{aligned}$$

Assume that at the beginning of the study all subjects were alive so, $P(T > t_0 = 0) = 1$, and

$$P(T > t_i | T > t_{i-1}) = \frac{n_i - d_i}{n_i}$$

The Kaplan Meier estimator is

$$S_{KM}(t) = \prod i : t_i \left(\frac{n_i - d_i}{n_i} \right) \text{ or } \prod i : t_i \le \left(1 - \frac{d_i}{n_i} \right)$$

For detail, see Greenwood (1926), Kaplan and Meier (1958), Adeleke (2012).

Semi-parametric estimation

For proportional hazard model of equation (1) where $h_0(t)$ is non-distributional and exp($\underline{\beta}'\underline{x}$) is a parametric function, we use partial likelihood estimate of Cox (1975)

$$\begin{aligned} L_i\left(\underline{\beta}\right) &= \prod_{j=1} L_i\left(\underline{\beta}\right) \\ &= \prod_{j=1} \frac{\exp\left(\underline{\beta}'\underline{x}_j + \gamma'\underline{z}_j\left(t\right)\right)}{\sum_{i\in R} \mathrm{e}\,xp\left(\underline{\beta}'\underline{x}_i + \gamma'\underline{z}_i\left(t\right)\right)} \end{aligned}$$

Application to a data of infertility

Data on period of infertility among women were obtained from a survey conducted in 2011 at Ijebu North Local Government (INLG) area of Ogun state. Information on the duration of infertility in years before a woman to get pregnant together with the causes of infertility were collected, along with covariates: duration of relationship (drelation) in years, respondent's age in years, marital status (married, cohabiting and single) and previous infertility treatment such as (ovulation induction, tubal surgery, antibiotic for infection, intercourse during fertile period and assisted conception).

Duration of infertility was measured as the time from marriage/first date of diagnose till fertile/date of first conception or the end of the study.

Let δ_i = 1 if a woman i = 1, 2, …, n become fertile at time t_i and δ_i = 0, if otherwise; let the survival time T = min (t_i, C_i), where t_i is the observed time and C_i is the censored time. Censored if either lost to follow-up or does not observe the event of interest (get pregnant) within the period of follow-up. First, consider the model of eqn (1) where age and duration of relationship and others were considered to be time fixed. The estimated regression coefficients are given in Table 1 together with associated p-values and Schoenfeld test result. As observed, intensity of being fertile is much higher for previous infertility treatment using ovulation induction and antibiotic for treatment of infections than when assisted with conception. Almost all the factors are negatively related with the hazards for the period of infertility. The aim is to know if model (1) is better used for the data or model 3 (i.e whether PH model assumption is satisfied or not). Age and duration of relationship were found to be significant.

Table 2 gives the estimates when age and duration of relationship are categorized as 1 if age less than 19 years, i.e (1-18), 2 if between (19-35) years inclusive and 3 if greater than 35 years. The result is not different much from what we had in Table 1. An indication of a significant variable implies the possibility of the variable varying with time and that implies violation of PH model assumption subject to some tests. The last column of the table is a report from Schoenfeld test with their respective *p*-values. The *p*-values for the correlation coefficient between time and covariates (duration of relationship) shows a significant relationship, supported by the Schoenfeld plot see fig 2. Another graphical test is log cumulative hazard plot. Log-cumulative hazard curves in fig 1 shows that only age of mothers is violating the assumption. Following the numerical test of the correlation coefficient between variable age of mothers and duration of relationship and time in Table 3, the *p*-values for both coefficients and Schoenfeld residual test for age of mothers and duration of relationship with time are indication that both age of mothers and duration of relationship are time varying.

Having detected this, an extended version of model (1) (i.e model 3) was introduced with age and duration of relationship categorized to see the effect within the age group (0-18, 19-34 and above 35) as shown in Table 4. Here the model is stable with the global test of Schoenfeld test showing a sign of proportionality.

Next, compare the two models, using Akaike's information criterion (AIC) or -2loglikelihood function (-2loglik). The values of AIC and -2loglik for Cox regression and Extended Cox are given in Table 5. According to the results, Extended Cox model gives most suitable result for modelling time to infertility data in the presence of non-proportional hazards followed by Cox model.

Results from infertility data

Table 1. Result from Cox model with Age, duration of relationship continuous

Variables	β (*p*-value)	Schoenfeld Test (rho))(*p*-value
Age	-0.086(1.4e-05)	0.169(0.198)
married	-1.67(0.108)	-0.024(0.840)
Cohabiting	-18.0(0.996)	-0.004(1.000)
drelation	-0.065(0.007)	0.287(0.028)
Ovulation	0.680(0.503)	0.066(0.591)
Tubla.S	-18.2 (0.998)	0.112(0.999)
Antibiotic	0.401 (0.697)	0.021(0.859)
Intercourse	-0.626 (0.659)	0.110(0.356)
		Global (0.0368)

Table 2. Result from Cox model with Age, duration of relationship categorized.

Variables	*β* (*p*-value)	Schoenfeld Test (rho)(*p*-value)
Age<=18	-0.777(0.460)	0.083(0.52)
Age>35	-1.225(4.30E-05)	-0.006(0.956)
Married	-1.69(0.103)	-0.011(0.93)
cohabiting	-17.827(1.00)	0.031(1.00)
dlv.cat1	0.447(0.110)	-0.201(0.0146)
Ovulation	0.862(0.400)	0.068(0.584)
Tubla.S	-17.448(1.00)	0.127(1.00)
Antibiotic	0.49(0.630)	0.026(0.584)
intercourse	-0.38(0.790)	0.087(0.479)
		Global (0.0506)

Table 3. Test for age and duration of relationship as time varying covariates

Variables	*β* (*p*-value)	Schoenfeld Test (rho)(*p*-value)
married	-1.0271(0.320)	-0.053(0.661)
cohabiting	-17.277(1.000)	-0.053(1.00)
Ovulation	0.94(0.360)	0.031(0. 802)
Tubla.S	-18.594(1.000)	0.086(1.00)
Antibiotic	0.617(0.550)	0.003(0. 980)
intercourse	-0.638(0.650)	0.130(0.283)
Age* time	-0.0187(0.000)	0.613(1.23E-09)
Drelation*time	-0.0055(0.021)	0.295(3.16E-03)
		Global(1.41E-06)

Table 4. Extended Cox model with age and duration of relationship as time varying.

Variables	*β* (*p*-value)	Schoenfeld Test (rho)(*p*-value)
married	-0.986(0.340)	0.0128(0.918)
cohabiting	-6.713(0.760)	0.0006(1.00)
Ovulation	1.384(0.180)	0.078(0.528)
Tubla.S	-8.640(0.940)	0.136(0.988)
Antibiotic	1.0257(0.320)	0.049(0.689)
intercourse	-0.275(0.840)	0.12908(0.296)
Age<=18*time	-4.612(1.70E-06)	0.194(0.548)
age.cat2*time	-4.717(1.0E06)	0.183(0.56)
Age>35*time	-4.713(9.70E-07)	0.198(0.544)
Time*dlv.cat1	0.001(0.980)	0.102(0.366)
		Global (0.982)

Table 5. AIC and -2loglik values.

	PHM	NPHM Extended Cox
AIC	525.813	311.6885
Loglik	509.813	291.688

Results from Simulation

Table 6. Mean values of the estimated regression coefficients for continuous time varying covariate model (16).

% cens	$\hat{\beta}$	$\hat{\gamma}$	AIC	loglik
C=0.0	-0.849(0.007)	0.724(0.151)	473.392	-309.929
C=0.5	-0.976(0.112)	2.016(0.0003)	158.962	-105.449
C=0.8	-0.770(0.261)	2.389(0.049)	62.032	-50.788

Table 7. Sample variances of the estimated regression coefficients for continuous time varying covariate model (16).

% cens	$\hat{\beta}$	$\hat{\gamma}$
C=0.0	0.0619	0.0552
C=0.5	0.1793	0.2073
C=0.8	0.4580	0.5744

Table 8. Mean values of the estimated regression coefficients for dicotonomous time varying ($t \geq t_0$); model 13.

% cens	$\hat{\beta}$	$\hat{\gamma}$	AIC	loglik
C=0.0	-0.363(0.211)	0.299(0.238)	625.857	-233.696
C=0.5	-0.348(0.363)	0.692(0.201)	240.578	-75.969
C=0.8	-0.184(0.411)	0.572(0.313)	107.576	-28.016

Table 9. Sample variances of the estimated regression coefficients for dicotonomous time varying ($t \geq t_0$); model 13.

% cens	$\hat{\beta}$	$\hat{\gamma}$
C=0.0	0.0537	0.0457
C=0.5	0.1271	0.1132
C=0.8	0.2664	0.2086

Table 10. Mean values of the estimated regression coefficients for time fixed covariate ($t \geq t_0$); model 12.

% cens	$\hat{\beta}$	$\hat{\gamma}$	AIC	loglik
C=0.0	-0.998 (2e-16)	0.043 (0.165)	11619.89	-5807.947
C=0.5	-1.058 (2e-16)	2.152 (2e-16)	5313.93	-2654.965
C=0.8	-8.060(2.4e-15)	-1.94(2e-16)	2585.184	-1290.592

Table 11. Sample variances of the estimated regression coefficients time fixed covariate ($t \geq t_0$); model 12.

% cens	$\hat{\beta}$	$\hat{\gamma}$
C=0.0	0.0047	0.00097
C=0.5	0.0088	0.0061
C=0.8	1.0365	0.0114

Table 12. Absolute Bias continuous TVC model 16**.**

% cens	$\underline{\beta}$	Abs Bias	MSE
$C = 0.0$	$\beta = -1$	0.150	0.069
	$\gamma = 0$	0.723	0.751
$C = 0.5$	$\beta = -1$	0.023	0.201
	$\gamma = 2$	0.015	0.257
$C = 0.8$	$\beta = -1$	0.229	0.659
	$\gamma = 3$	0.611	1.465

Table 13. Absolute Bias for dicotonomous time varying ($t \geq t_0$); model 13.

% cens	$\underline{\beta}$	Abs Bias	MSE
$C = 0.0$	$\beta = -1$	0.636	0.471
	$\gamma = 0$	0.298	0.143
$C = 0.5$	$\beta = -1$	0.651	0.611
	$\gamma = 2$	1.308	1.994
$C = 0.8$	$\beta = -1$	0.815	0.996
	$\gamma = 3$	2.428	6.143

Table 14. Absolute Bias for time fixed covariate ($t \geq t_0$); model 12.

% cens	$\underline{\beta}$	Abs Bias	MSE
$C = 0.0$	$\beta = -1$	0.002	0.918
	$\gamma = 0$	0.043	0.211
$C = 0.5$	$\beta = -1$	0.058	0.221
	$\gamma = 2$	0.152	1.133
$C = 0.8$	$\beta = -1$	7.06	1.110
	$\gamma = 3$	4.94	2.720

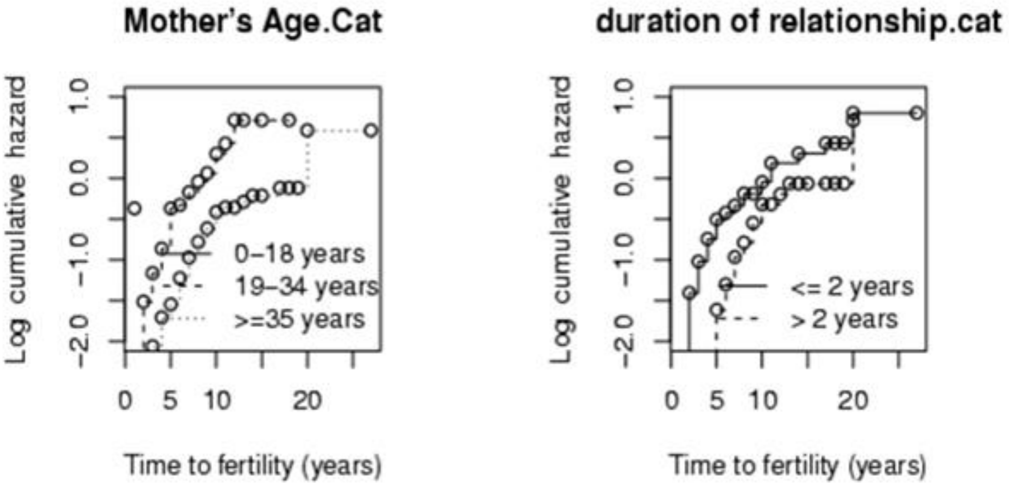


Figure 1. Log cumulative hazards for age and duration of relationship.

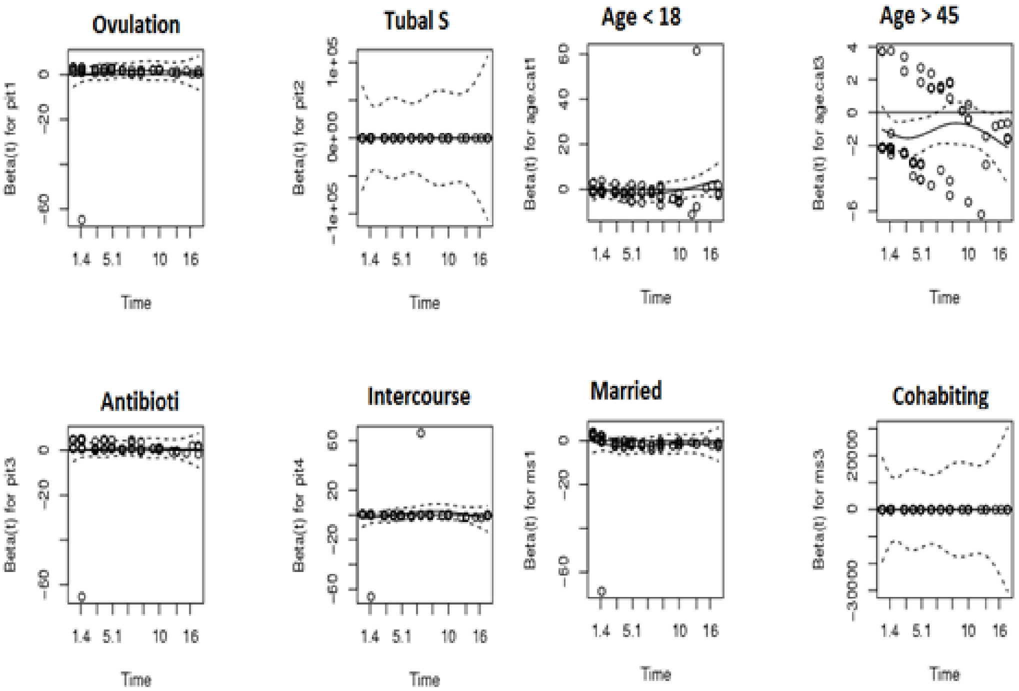


Figure 2. Schoenfeld Plots of residuals

In purpose of the simulation was to investigate the violation of the assumption and the use of Non-proportional hazard Model for different values of the true parameters β and γ, at different level of censoring. Hypothesis about the regression coefficients β and γ of the model 1.0 in various situations was tested. Each simulation consists of 80 replicates. The set-up of the simulated data resembles that of right censored and truncated data. For each sample, 1000 samples of survival times (months) were generated.

Given a time t_* , the time u were generated from a uniform $(0,t_*)$ distribution although the baseline survival time t_i were generated from an exponential distribution for fixed and time varying covariates in term of continuous and dichotomous covariates as define in eqn 12, 13 and 16. Two covariates; a time fixed and a binary with $P(z = 0) = P(z = 1)= ½$ and the other is distributed as normal and varies with time. Only the data that satisfy the condition $u_i + t_i \le t_*$ were kept in the sample given rise to right truncated data. The survival time is not only right truncated but also right censored. The simulation was carried out at three different percentage of censoring viz: 0%, 50% and 80%.

The true values of regression coefficients *β*, *γ* were taken to be either (-1, 0), (-1, 2), (-1, 3) in the simulation each at different level or percentage of censoring. Comparison were made using absolute bias Tables 6 to 11 showed the estimated mean values of $\hat{\beta}$ and $\hat{\gamma}$, *p*-values as well as the sample variances. The result in Tables 6 and 9 are from the analysis of (3) through the use of survival time obtained in (16) for fixed and continuous time varying covariates of (3). The estimated coefficients $\hat{\beta}$ is for the fixed covariate although $\hat{\gamma}$ is for the time varying (continuous or binary). The coefficients are significant at 50% and 80 % censoring and slightly overestimate its true value as percentage of censoring increases resulting in higher variance than the estimator of the other coefficient which appear to be more stable with lower variance than *γ*. Absolute Bias (AB) of Tables 12 to 14 showed the sensitivity of the model to change in percentage of censoring. At 0 percent censoring, model with time fixed covariate has the minimum AB followed by model with continuous time varying covariate. Also at 50% censoring, model with continuous time varying covariate has the minimum AB, followed by model with time fixed covariate. At 80% censoring, model with continuous time varying covariate has the minimum AB next is model with dicotonomous time varying covariate and least is time fixed model.

Checking the parameter of the time varying coefficient, as the values of the parameter *γ* increases from 0 to 3, At $\gamma = 0$, the AB of the parameter is minimum for model with time fixed covariate, followed by a model with dicotonomous time varying covariate and maximum for model with continuous time varying covariate. At $\gamma = 2$, AB is minimum for semi-parametric model via continuous time varying covariate (model 16), followed by a time fixed and maximum for semi-parametric model with dicotonomous time varying covariate. Lastly at $\gamma = 3$ AB increases from model with continuous time varying covariate to semi-parametric model with time fixed covariate. Hence, as parameter of time varying coefficient increase from 0-3, the semi-parametric model with continuous time varying covariate showed the minimum AB followed by dicotonomous time varying covariate and maximum with time fixed covariate model. This actually showed an evidence of time varying both in the coefficient and covariate.

For Mean Square Error (MSE), Semi-parametric model with continuous time varying covariate has being the best (with min MSE) among the three models as percentage of censoring increases from 0% to 80 percent. Also as parameter of time varying coefficient increases from 0 to 3, parameters of the semi-parametric model with continuous time varying coefficient showed the minimum MSE, and perform best. Followed by the parameters of time fixed covariate model and maximum MSE with model with dicotonomous time varying covariate.

Discussion

The result is more encouraging at 80% of censoring resulting from the outcome of the AIC and log-likelihood estimates of model selection criteria and generally accepted for all other results. Percentage of censoring contributes to the outcome and conclusion in that as the level of censoring increases from 0% through 50% to 80%. The coefficients of time varying covariates varying from zero to three (0-3). See Tables 6 and 10, the result also give a good sign of a well satisfactory size and power. The higher the percentage of censoring, the more closely the violation of PHM. It implies that at 80% censoring which is generally accepted from the results of our simulated data there exist an outright violation of the assumption of proportionality and this assume a semi-parametric non proportional hazard model.

In Tables 8, 9, 10, and 11 models 12 and 13 were used to generate survival time when both covariates are dichotomous and continuous, although time varying. The time varying covariate $Z(t)$ is zero when $t < t_0$ and 1 when $t \geq t_0$ as stated in the model, our t_0 is the maximum time it takes a woman to conceive (i.e 24 months), see Esther, Eunice , Kelly, CHESRenee, and Lee (2009), Ekwere, et al (2007) and Yusuff (2006). (When $t < t_0$, we obtain our survival time as we have in (12) and when $t \geq t_0$, it resulted in survival time of (13) as we notice from the estimated mean values and variances of Tables 8 and 9. None of the coefficients at any level of censoring is significant judging from the PH values of the coefficient. An indication of satisfying PH model assumption, but when $t \geq t_0$ (dicotonomous), the estimated mean values and sample variances of regression coefficient does not satisfy PH model assumption following parameters significant properties of the coefficients from the *p*-values.

The model with continuous time varying covariate (model 16) performed better (min AB and MSE) followed by model with dicotonomous time varying covariate and least with model with time fixed covariate see Tables 12 to 14. The same result follows when parameters of the time varying coefficient increase from 0-3.

References

Abrahamowicz, M., & MacKenzie, T. A. (2007). Joint estimation of time-dependent and non-linear effects of continuous covariates on survival. *Statistics in Medicine*, *26*(2), 392-408. doi:10.1002/sim.2519

Adeleke, K. A. (2012). A simulation study on kaplan meier non-parametric survival methods. *Journal of the Nigerian Mathematical Society, 31*, 243-254.

Adeleke, K.A. & Adepoju, A. A. (2010). Ordinal logistic regression model: An application to pregnancy outcomes. *Journal of Mathematics and Statistics*, *6*(3), 279-285. doi:10.3844/jmssp.2010.279.285

Agresti, A. (2007). *An introduction to categorical data analysis* (2nd ed.). New York: Wiley.

Ata, N. & Sozer, M. T. (2007). Cox regression models with non-proportional hazards applied to lung cancer survival data. *Hacettepe Journal of Mathematics and Statistics*, *36*(2), 157-167.

Austin, P. C. (2012). Generating survival times to simulate Cox proportional hazards models with time-varying covariates. *Statistics in Medicine, 31*(29), 3946–3958.doi:10.1002/sim.5452

Bender. R., Augustin, T., & Blettner, M. (2005). Generating survival times to simulate Cox proportional hazards models. *Statistics in Medicine*, *24*(11), 1713-1723. doi:10.1002/sim.2059

Cox, D. R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society. Series B (Methodological)*, *34*(2), 187-220.

Cox, D. R. (1975). Partial likelihood. *Biometrika*, *62*(2), 269-276. doi:10.2307/2335362

Ekwere, P. D., Archibong, E. I., Bassey, E. E., Ekabua, J. E., Ekanem, E. I. & Feyi-Waboso, P. (2007). Infertility among Nigerian couples as seen in Calabar. *Port Harcourt Medical Journal; 2*(1), 35-40. doi:10.4314/phmedj.v2i1.38890

Eisenberg, E., Shriver, E. K., Brumbaugh, K., Brown-Bryant, R & Warner, L. (2009). Frequently asked questions. *United State Department of Health and Human Services, Office on Women's Health*. Retrieved from http://www.womenshealth.gov

Greenwood, M. (1926). A report on the natural duration of cancer. *Reports on Public Health and Medical Subjects. Ministry of Health*, (33).

Hosmer, D. W. & Lemeshow, S. (2000). *Applied logistic regression* (2nd ed.). New York: Wiley.

Kaplan, E. L., & Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association, 53*(282), 457-481. doi:10.2307/2281868

Kleinbaum, D. G. & Klein, M. (2005). *Survival analysis: A self learning text* (2nd ed.). New York: Springer-Verlag.

Leemis, L. M. (1987). Techinical note-Variate generation for accelerated life and proportional hazards models. *Operations Research, 35*(6), 892-894. doi:10.1287/opre.35.6.892

Leemis, L. M., Shih, L. H., & Reynertson, K. (1990). Variate generation for accelerated life and proportional hazards models with time dependent covariates. *Statistics and Probability Letters; 10*(6), 335-339. doi:10.1016/0167-7152(90)90052-9

Lehr, S. (2004). *Over-fit in the analysis of time-dependent effects of prognostic factors*. Ph.D. Thesis, University of Vienna.

Shih, L. H. & Leemis, L. M. (1993). Variate generation for a nonhomogenous poisson process with time dependent covariates. *Journal of Statistical Computation and Simulation*; *44*(3-4), 165-186. doi:10.1080/00949659308811457

Sylvestre, M. P. & Abrahamowicz, M. (2007). Comparison of algorithms to generate event times conditional on time-dependent covariates. *Statistics in Medicine; 27*(14), 2618-2634.doi:10.1002/sim.3092

Tableman, M. & Kim, J. S. (2004). *Survival analysis using S: Analysis of time-to-event data*. New York: Chapman and Hall/CRC.

Therneau, T. M. & Grambsch, P. M. (2000). *Modeling survival data: Extending the Cox model*. New York: Springer-Verlag.

Walker, S. H & Duncan, D. B. (1967). Estimation of the probability of an event as a function of several independent variables. *Biometrika, 54*(1-2), 167-179. doi:10.1093/biomet/54.1-2.167

Yusuff, A. O. (2006). Infertility treatments and the conceptual dilemma of infertility: is there a right to reproduce? *Canadian Institutes of Health Research (CIHR) at the University of Toronto, ON, Canada.* Retrieved from http://www.nigerianlawguru.com/articles/human%20rights%20law/INFERTILITY%20TREATMENTS%20AND%20THE%20CONCEPTUAL%20DILEMMA%20OF%20INFERTILITY,%20IS%20THERE%20A%20RIGHT%20TO%20REPRODUCE.pdf.

Zhou, M. (2001). Understanding the Cox regression models with time-change covariates. *The American Statistician, 55*(2), 153-155. doi:10.1198/000313001750358491

Journal of Modern Applied Statistical Methods
November 2015, Vol. 14, No. 2, 88-109.

ISSN 1538 − 9472

New Entropy Estimators with Smaller Root Mean Squared Error

Amer Ibrahim Al-Omari
Al al-Bayt University
Mafraq, Jordan

New estimators of entropy of continuous random variable are suggested. The proposed estimators are investigated under simple random sampling (SRS), ranked set sampling (RSS), and double ranked set sampling (DRSS) methods. The estimators are compared with Vasicek (1976) and Al-Omari (2014) entropy estimators theoretically and by simulation in terms of the root mean squared error (RMSE) and bias values. The results indicate that the suggested estimators have less RMSE and bias values than their competing estimators introduced by Vasicek (1976) and Al-Omari (2014).

Keywords: Shannon entropy; simple random sampling, ranked set sampling; double ranked set sampling; root mean square error.

Introduction

The ranked set sampling was first suggested by McIntyre (1952) to estimate a mean of pasture and forage yields. It is a cost efficient sampling procedure alternative to the commonly used simple random sampling scheme. The RSS is useful in situations where the visual ordering of a set of units can be done easily, but the exact measurement of the units is difficult or expensive.

Let the variable of interest X has a probability density function (pdf) $g(x)$ and a cumulative distribution function (cdf) $G(x)$, with mean μ and variance σ^2. Let $g_{(i:n)}(x)$ and $G_{(i:n)}(x)$ be the pdf and cdf of the ith order statistic, $X_{(i:n)}$, $(1 \le i \le n)$ of a random sample of size n. The pdf and the cdf of $X_{(i:n)}$, respectively, are given by

$$g_{(i:n)}(x) = n\binom{n-1}{i-1} G^{i-1}(x)\left[1-G(x)\right]^{n-i} g(x), \ -\infty < x < \infty$$

Amer Ibrahim Al-Omari is Faculty of Science, in the Department of Mathematics. Email at: alomari_amer@yahoo.com.

and

$$G_{(i:n)}(x)=\sum_{j=1}^{n}\binom{n}{j}G^{j}(x)\left[1-G(x)\right]^{n-j},\ -\infty<x<\infty\ ,$$

with mean $\mu_{(i:n)}=\int_{-\infty}^{\infty}x\,g_{(i:n)}(x)\,dx$ and variance $\sigma_{(i:n)}^{2}=\int_{-\infty}^{\infty}\left(x-\mu_{(i:n)}\right)^{2}g_{(i:n)}(x)\,dx$.

The ranked set sampling method can be describes as follows:

Step 1. Randomly select n^2 units from the target population.
Step 2. Allocate the n^2 selected units randomly into n sets, each of size n.
Step 3. Without yet knowing any values for the variable of interest, rank the units within each set with respect to a variable of interest. This may be based on a personal professional judgment or based on a concomitant variable correlated with the variable of interest.
Step 4. The sample units are selected for actual measurement by including the ith smallest ranked unit of the ith sample (i = 1, 2, …, n).
Step 5. Repeat Steps 1 through 4 for r cycles to obtain a sample of size nr for actual measurement.

It is of interest to note here that even if n^2 units are selected from the population, but only n of them are measured for comparison with a simple random sampling of the same size n.

Let the measured RSS units are denoted by $X_{1(1:n)}$, $X_{2(2:n)}$, …, $X_{n(n:n)}$. The RSS estimator of the population mean is defined as $\bar{X}_{RSS}=\frac{1}{n}\sum_{i=1}^{n}X_{i(i:n)}$. Takahasi and Wakimoto (1968) provided the mathematical theory of the RSS and showed that

$$g(x)=\frac{1}{n}\sum_{i=1}^{n}g_{(i:n)}(x),\ \mu=\frac{1}{n}\sum_{i=1}^{n}\mu_{(i:n)},\ \mathrm{Var}\left(\bar{X}_{RSS}\right)=\frac{\sigma^{2}}{n}-\frac{1}{n^{2}}\sum_{i=1}^{n}\left(\mu_{i(i:n)}-\mu\right)^{2}.$$

Al-Saleh and Al-Kadiri (2000) suggested double ranked set sampling (DRSS) method for estimating the population mean to increase the efficiency of the estimators for fixed sample size. The DRSS method can be described as:

Step 1. Randomly choose n^2 samples of size n each from the target population.
Step 2. Apply the RSS method described above on the n^2 samples in Step 1. This step yields n samples of size n each.
Step 3. Reapply the RSS method again on the n samples obtained in Step 2 to obtain a sample of size n from the DRSS data. The cycle can be repeated r times if needed to obtain a sample of size rn units.

Let X be a continuous random variable with probability density function $g(x)$ and cumulative distribution function $G(x)$. The entropy $H[g(x)]$ of the random variable is defined by Shannon (1948a, 1948b) as

$$H\left[g(x)\right] = -\int_{-\infty}^{\infty} g(x)\log\left[g(x)\right]dx. \tag{1}$$

The problem of entropy estimation of a continuous random variable is considered by many authors. Vasicek's (1976) suggested an estimator of entropy based on spacing's as

$$H\left[g(x)\right] = \int_0^1 \log\left(\frac{dG^{-1}(p)}{dp}\right)dp, \tag{2}$$

where the estimation is found by replacing the distribution function $G(x)$ by the empirical distribution function $G_n(x)$, and using the difference operator instead of the differential operator. Then the derivative $\frac{d}{dp}G^{-1}(p)$ is estimated by a function of the order statistics.

Let $X_1, X_2, \ldots, X_n$ be a simple random sample of size n from $G(x)$ and $X_{(1)} < X_{(2)} < \ldots < X_{(n)}$ be the order statistics of the sample. Then Vasicek's (1976) estimator of $H[g(x)]$ is defined as

$$HV_{mn} = \frac{1}{n}\sum_{i=1}^{n}\log\left\{\frac{n}{2m}\left(X_{(i+m)} - X_{(i-m)}\right)\right\} \tag{3}$$

where $m < n / 2$ is a positive integer known as the window size, $X_{(i\text{ - m})} = X_{(1)}$ if $i \leq m$, and $X_{(i\text{ + m})} = X_{(n)}$ if $i \geq n - m$. He proved that $HV_{mn} \xrightarrow{P.} H\left[g\left(x\right)\right]$ as $n \to \infty,\ m \to \infty$, and $\frac{m}{n} \to 0$.

Van Es (1992) suggested an estimator of entropy based on spacings as

$$HVE_{mn} = \frac{1}{n-m}\sum_{i=1}^{n-m}\left\{\frac{n+1}{m}\left(X_{(i+m)} - X_{(i)}\right)\right\} + \sum_{k=m}^{n}\frac{1}{k} + \log\left(m\right) - \log\left(n+1\right) \tag{4}$$

and proved the consistency and the asymptotic normality of the estimator under some conditions.

Ebrahimi, Pflughoeft, and Soofi (1994) adjusted the weights of Vasicek (1976) estimator to have a smaller weights and proposed an entropy estimator given by

$$HE_{mn} = \frac{1}{n}\sum_{i=1}^{n}\log\left\{\frac{n}{\theta_i m}\left(X_{(i+m)} - X_{(i-m)}\right)\right\} \tag{5}$$

where

$$\theta_i = \begin{cases} 1 + \dfrac{i-1}{m}, & 1 \leq i \leq m, \\ 2, & m+1 \leq i \leq n-m, \\ 1 + \dfrac{n-i}{m}, & n-m+1 \leq i \leq n, \end{cases}$$

where $X_{(i\text{-m})} = X_{(1)}$ for $i \leq m$ and $X_{(i\text{+m})} = X_{(n)}$ for $i \geq n - m$. Ebrahimi et al. (1994) showed by simulation that their estimator has a smaller bias and mean squared error than Vasicek (1976) estimator. Also, they proved that

$$HE_{mn} \xrightarrow{P.} H\left[g\left(x\right)\right] \text{ as } n \to \infty,\ m \to \infty,\ m/n \to 0.$$

Noughabi and Noughabi (2013) suggested a new estimator of entropy of an unknown continuous probability density function as

$$HNN_{mn} = -\frac{1}{n}\sum_{i=1}^{n}\log\left\{s_i\left(n,m\right)\right\}, \tag{6}$$

where

$$s_i\left(n,m\right) = \begin{cases} \hat{g}\left(X_{(i)}\right), & 1 \le i \le m, \\ \dfrac{2m/n}{X_{(i+m)} - X_{(i-m)}}, & m+1 \le i \le n-m, \\ \hat{g}\left(X_{(i)}\right), & n-m+1 \le i \le n, \end{cases}$$

and $\hat{g}\left(X_i\right) = \frac{1}{nh}\sum_{j=1}^{n} k\left(\frac{X_i - X_j}{h}\right)$, where h is bandwidth and k is a kernel function satisfies $\int_{-\infty}^{\infty} k\left(x\right)dx = 1$. They proved that $HNN_{mn} \xrightarrow{P} H\left[g\left(x\right)\right]$ as $n \to \infty$, $m \to \infty$, $m/n \to 0$. Note that the kernel function in Noughabi and Noughabi (2013) is selected to be the standard normal distribution and the bandwidth h is chosen to be $h = 1.06sn^{-1/5}$, where s is the sample standard deviation.

To estimate the entropy $H\,[g(x)]$ of an unknown continuous probability density function $g(x)$, Noughabi and Arghami (2010) suggested an entropy estimator given by

$$HN_{mn} = \frac{1}{n}\sum_{i=1}^{n}\log\left\{\frac{n}{c_i m}\left(X_{(i+m)} - X_{(i-m)}\right)\right\} \tag{7}$$

where

$$c_i = \begin{cases} 1, & 1 \le i \le m, \\ 2, & m+1 \le i \le n-m, \\ 1, & n-m+1 \le i \le n, \end{cases}$$

and $X_{(i\text{-m})} = X_{(1)}$ if $i \le m$ and $X_{(i\text{+m})} = X_{(n)}$ for $i \ge n - m$.

Correa (1995) suggested a modified entropy estimator to have smaller mean squared error in the form

$$HC_{mn}=-\frac{1}{n}\sum_{i=1}^{n}\log\left(\frac{\sum_{j=i-m}^{i+m}(j-i)\left(X_{(j)}-\bar{X}_{(i)}\right)}{n\sum_{j=i-m}^{i+m}\left(X_{(j)}-\bar{X}_{(i)}\right)^{2}}\right), \tag{8}$$

where $\bar{X}_{(i)}=\frac{1}{2m+1}\sum_{j=i-m}^{i+m}X_{(j)}$.

Al-Omari (2014) suggested three estimators of entropy of an unknown continuous probability density function $g(x)$ using SRS, RSS, and DRSS methods. Based on SRS his first suggested estimator is defined as

$$AHESRS_{mn}=\frac{1}{n}\sum_{i=1}^{n}\log\left\{\frac{n}{\omega_i m}\left(X_{(i+m)}-X_{(i-m)}\right)\right\} \tag{9}$$

where $X_{(i\text{-m})}=X_{(1)}$ for $i\leq m$, $X_{(i\text{+m})}=X_{(n)}$ for $i\geq n-m$, and

$$\omega_i=\begin{cases}1+\frac{1}{2}, & 1\leq i\leq m,\\ 2, & m+1\leq i\leq n-m,\\ 1+\frac{1}{2}, & n-m+1\leq i\leq n,\end{cases} \tag{10}$$

The second and third estimators suggested by Al-Omari (2014), based on RSS and DRSS respectively, are given by

$$AHERSS_{mn}=\frac{1}{n}\sum_{i=1}^{n}\log\left\{\frac{n}{\omega_i m}\left(X^{*}_{(i+m)}-X^{*}_{(i-m)}\right)\right\} \tag{11}$$

and

$$AHEDRSS_{mn}=\frac{1}{n}\sum_{i=1}^{n}\log\left\{\frac{n}{\omega_i m}\left(X^{**}_{(i+m)}-X^{**}_{(i-m)}\right)\right\} \tag{12}$$

where $X^{*}_{(i-m)}=X^{*}_{(1)}$ for $i\leq m$ and $X^{*}_{(i+m)}=X^{*}_{(n)}$ for $i\geq n-m$, and

$X^{**}_{(i-m)} = X^{**}_{(1)}$ for $i \leq m$ and $X^{**}_{(i+m)} = X^{**}_{(n)}$ for $i \geq n-m$.

For more about entropy estimators, see Choi, Kim, and Song (2004), Park, Park (2003), Goria, Leonenko, Mergel, and Novi Inverardi (2005) and Choi (2008).

The remaining part of this paper is organized as follows. The suggested entropy estimators are given in the section, "Proposed Estimators". Next, a simulation study is conducted to compare the new estimators with their counterparts suggested by Vasicek (1976) and Al-Omari (2014). Finally, some conclusions and suggestions for further works.

The proposed estimators

The coefficient of the entropy estimators in Ebrahimi et al. (1994), Noughabi and Arghami (2010), and Al-Omari (2014) are adjusted. Let $X_1^{(0)}, X_2^{(0)}, \ldots, X_n^{(0)}$ be a simple random sample of size n from $G(x)$. Based on SRS the first suggested estimator is given by

$$SHESRS_{mn} = \frac{1}{n}\sum_{i=1}^{n} \log\left\{ \frac{n}{\varepsilon_i m}\left(X^{(0)}_{(i+m)} - X^{(0)}_{(i-m)} \right) \right\} \tag{13}$$

where

$$\varepsilon_i = \begin{cases} 1+\dfrac{1}{4}, & 1 \leq i \leq m, \\ 2, & m+1 \leq i \leq n-m, \\ 1+\dfrac{1}{4}, & n-m+1 \leq i \leq n, \end{cases} \tag{14}$$

$X^{(0)}_{(i-m)} = X^{(0)}_{(1)}$ for $i \leq m$ and $X^{(0)}_{(i+m)} = X^{(0)}_{(n)}$ for $i \geq n-m$. Comparing (3) with (13), we have

$$\begin{aligned} SHESRS_{mn} &= \frac{1}{n}\sum_{i=1}^{n}\log\left\{\frac{n}{\varepsilon_i m}\left(X_{(i+m)}^{(0)} - X_{(i-m)}^{(0)}\right)\right\} \\ &= HVSRS_{mn} + \frac{1}{n}\sum_{i=1}^{n}\log\frac{2}{\varepsilon_i} \qquad (15) \\ &= HVSRS_{mn} + \frac{2m}{n}\log\frac{8}{5} \end{aligned}$$

Let $X_{(1:n)}^{(1)}, X_{(2:n)}^{(1)}, ..., X_{(n:n)}^{(1)}$ be a RSS of size n, Vasicek (1976) entropy estimator using RSS as considered by Mahdizadeh (2012) is given by

$$HVRSS_{mn} = \frac{1}{n}\sum_{i=1}^{n}\log\left\{\frac{n}{2m}\left(X_{(i+m)}^{(1)} - X_{(i-m)}^{(1)}\right)\right\} \qquad (16)$$

Based on the RSS units $X_{(1:n)}^{(1)}$, $X_{(2:n)}^{(1)}$, …, $X_{(n:n)}^{(1)}$, the second suggested entropy estimator is

$$SHERSS_{mn} = \frac{1}{n}\sum_{i=1}^{n}\log\left\{\frac{n}{\varepsilon_i m}\left(X_{(i+m)}^{(1)} - X_{(i-m)}^{(1)}\right)\right\} \qquad (17)$$

where ε_i is defined as in (14), and $X_{(i-m)}^{(1)} = X_{(1)}^{(1)}$ for $i \le m$ and $X_{(i+m)}^{(1)} = X_{(n)}^{(1)}$ for $i \ge n - m$. Comparing (16) with (17) to have

$$\begin{aligned} SHERSS_{mn} &= \frac{1}{n}\sum_{i=1}^{n}\log\left\{\frac{n}{\varepsilon_i m}\left(X_{(i+m)}^{(1)} - X_{(i-m)}^{(1)}\right)\right\} \\ &= HVRSS_{mn} + \frac{1}{n}\sum_{i=1}^{n}\mathrm{Log}\frac{2}{\varepsilon_i} \qquad (18) \\ &= HVRSS_{mn} + \frac{2m}{n}\log\frac{8}{5} \end{aligned}$$

Assume that $X_{(1:n)}^{(2)}, X_{(2:n)}^{(2)}, ..., X_{(n:n)}^{(2)}$ is a DRSS sample of size n. The third suggested entropy estimator has the form

$$SHEDRSS_{mn} = \frac{1}{n}\sum_{i=1}^{n}\log\left\{\frac{n}{\varepsilon_i m}\left(X_{(i+m)}^{(2)} - X_{(i-m)}^{(2)}\right)\right\} \tag{19}$$

where ε_i is defined as in (14), and $X_{(i-m)}^{(2)} = X_{(1)}^{(2)}$ for $i \le m$ and $X_{(i+m)}^{(2)} = X_{(n)}^{(2)}$ for $i \ge n-m$. Based on DRSS method Mahdizadeh (2012) showed that Vasicek (1976) estimator will be

$$SHEDRSS_{mn} = \frac{1}{n}\sum_{i=1}^{n}\log\left\{\frac{n}{2m}\left(X_{(i+m)}^{(2)} - X_{(i-m)}^{(2)}\right)\right\} \tag{20}$$

Comparing (19) with (20) to get

$$\begin{aligned} SHEDRSS_{mn} &= \frac{1}{n}\sum_{i=1}^{n}\log\left\{\frac{n}{\varepsilon_i m}\left(X_{(i+m)}^{(2)} - X_{(i-m)}^{(2)}\right)\right\} \\ &= HVDRSS_{mn} + \frac{1}{n}\sum_{i=1}^{n}\log\frac{2}{\varepsilon_i} \\ &= HVDRSS_{mn} + \frac{2m}{n}\log\frac{8}{5} \end{aligned} \tag{21}$$

Remark 1: The entropy $H\left(f_n^{ME}\right)$ of an empirical maximum entropy density f_n^{ME} which is related to *HVSRS*$_{1n}$ and SHESRS$_{1n}$ can be computed following Theil (1980) as:

$$\begin{aligned} H\left(f_n^{ME}\right) &= HVSRS_{1n} + \frac{2-2\log 2}{n} \\ &= SHESRS_{1n} - \frac{2}{n}\log\frac{8}{5} + \frac{2-2\log 2}{n} \\ &= SHESRS_{1n} + \frac{2}{n}\left(1-\log\frac{4}{5}\right) \end{aligned} \tag{22}$$

Remark 2: If $n \to \infty$ in (22), then $H\left(f_n^{ME}\right) = SHESRS_{1n}$.

In the following two theorems, we compared the suggested estimators with Vasicek (1967) and Al-Omari (2014).

Theorem 1: The suggested estimators have the following properties:

a) Let $X_1^{(0)}, X_2^{(0)}, ..., X_n^{(0)}$ be SRS of size n, then $SHESRS_{mn} > HVSRS_{mn}$.

b) Let $X_{(1)}^{(1)}, X_{(2)}^{(1)}, ..., X_{(n)}^{(1)}$ be a RSS of size n, then $SHERSS_{mn} > HVRSS_{mn}$.

c) Let $X_{(1)}^{(2)}, X_{(2)}^{(2)}, ..., X_{(n)}^{(2)}$ be a DRSS of size n, then $SHEDRSS_{mn} > HVDRSS_{mn}$.

Proof: The proof of (a), (b), (c), is straightforward by using (15), (18), (21), respectively, where $\frac{2m}{n}\log\frac{8}{5} > 0$.

In the following theorem, we compare our suggested entropy estimators with their competitors in Al-Omari (2014).

Theorem 2: Based on the suggested estimators and Al-Omari (2014) entropy respectively, we have

$$SHEj_{mn} > AHEj_{mn},\ j = \text{SRS, RSS, DRSS}.$$

Proof: Compare (9) with (13) based on SRS to obtain

$$SHESRS_{mn} - AHESRS_{mn} = \frac{2m}{n}\log\frac{6}{5},$$

and since $\frac{2m}{n}\log\frac{6}{5} > 0$, then the case of SRS holds. Also, compare (11) with (17) based on RSS, and (12) with (19) using DRSS to complete the proof of this theorem.

The following theorem proves the consistency of the suggested estimators $SHESRS_{mn}$, $SHERSS_{mn}$, and $SHEDRSS_{mn}$.

Theorem 3: Let Ω be the class of continuous densities with finite entropies and let $X_1, X_2, \ldots, X_n$ be a random sample from $g \in \Omega$. If $n \to \infty$, $m \to \infty$, $m/n \to 0$, then $SHEj_{mn}$, (j = SRS, RSS, DRSS) converges in probability to $H[g(x)]$.

Proof: Based on the simple random sampling, from (15) we have

$$SHESRS_{mn} = HVSRS_{mn} + \frac{2m}{n}\log\frac{8}{5},$$

and Vasicek (1976) showed that $HVSRS_{mn}$ converges in probability to $H\,[g(x)]$ and since $\frac{2m}{n}\log\frac{8}{5}$ converges to zero as n goes to infinity, then we proved the case of the SRS. Follow the same approach and use (18) and (21) to prove the theorem for RSS and DRSS estimators, respectively.

Methodology

Simulation study

A simulation was conducted to investigate the performance of the suggested entropy estimators with Vasicek (1976) and Al-Omari (2014) entropy estimators using sampling methods considered in this study. The comparison is based on the root mean squared errors (RMSEs) and bias values of the estimators for 10000 samples generated from the uniform, exponential and the standard normal distributions using SRS, RSS and DRSS methods. The selection of the optimal values of the window size of m for a given value n is as yet an open problem in the entropy estimation. Therefore, we used the heuristic formula $m = \sqrt{n} + 0.5$ suggested by Wieczorkowski and Grzegorzewski (1999) to select m and to compute the RMSEs of entropy estimators. In this study, we considered the sample and window sizes as given in Table 1.

Table 1. The sample and window sizes considered in this simulation

Sample size	$n = 10$	$n = 20$	$n = 30$
Window size	$1 \le m \le 5$	$1 \le m \le 10$	$1 \le m \le 15$

Also, the performance of the RMSE of the suggested estimators for samples generated from the uniform, exponential and standard normal distributions is evaluated based on the quantity

$$Q_N = \frac{HVj_{mn} - N}{HVj_{mn}} \times 100,\ N = SHEj_{mn}, AHEj_{mn},\ \ j = SRS, RSS, DRSS\,.$$

The results are summarized in Tables 2-6. Also, we compared the suggested estimators of entropy with their competitors suggested by Al-Omari (2014) and the results presented in Table 7 are taken from Al-Omari (2014).

Based on these results observe the following.

- The suggested entropy estimators using SRS, RSS and DRSS methods are more efficient than their competitors HV_{mn} based on the same method for all cases considered in this study. As an example, from Table 3, with $n = 10$ and $m = 3$ for the exponential distribution with $H[g(x)] = 1$ using RSS method, the RMSE and bias value of $SHERSS_{mn}$ are 0.230412 and -0.052759 compared to 0.401125 and -0.332760 the RMSE and bias of $HVRSS_{mn}$.
- The $SHEDRSS_{mn}$ is superior to the other suggested estimators, $SHERSS_{mn}$ and $SHESRS_{mn}$ under the uniform, exponential and normal distributions. From Table 1, consider the case of $n = 20$ and $m = 4$ under the uniform distribution when $H[g(x)] = 0$, it can be noted that the RMSE values of $SHEDRSS_{mn}$, $SHERSS_{mn}$, and $SHESRS_{mn}$ are 0.052373, 0.068747 and 0.114983, respectively.
- The nature of the underlying distribution as well as the value of $H[g(x)]$ affect on the efficiency of the estimator using the same method. As an example, the $Q_{SHERSS_{mn}}$ values with $n = 30$ and $m = 3$ for the uniform, exponential, and the standard normal distributions are 95.39025, 31.76442 and 32.75544, respectively. However, the values of $Q_{SHE_{mn}}$ for the uniform distribution with $H[g(x)] = 0$ are superior to their counterparts for the exponential and normal distributions.
- Finally, the suggested entropy estimators are found to be more efficient than their competitors in Al-Omari (2014) entropy estimators using SRS, RSS and DRSS schemes for the same window and sample sizes. For illustration, assume that $n = 30$ and $m = 8$ when the underlying distribution is the standard normal, from Table 4, the RMSE of $SHERSS_{mn}$ is 0.120242 compared to 0.157726 which is the RMSE of $AHERSS_{mn}$ as shown in Table 7.

Table 2. The Monte Carlo RMSEs and bias values of HV_{mn} and SHE_{mn} for the uniform distribution with $H\,[g(x)] = 0$.

		SRS					**RSS**				
		HV_{mn}		SHE_{mn}		$Q_{SHE_{mn}}$	HV_{mn}		SHE_{mn}		$Q_{SHE_{mn}}$
n	***m***	**Bias**	**RMSE**	**Bias**	**RMSE**		**Bias**	**RMSE**	**Bias**	**RMSE**	
10	1	-0.519826	0.569537	-0.430151	0.490404	13.89427	-0.396308	0.443439	-0.303703	0.361606	22.63043
	2	-0.415135	0.452358	-0.226627	0.290240	35.83843	-0.304078	0.329233	-0.116915	0.172961	90.35100
	3	-0.422613	0.453818	-0.135797	0.213148	53.03227	-0.327681	0.343991	-0.045891	0.114159	201.3262
	4	-0.458940	0.487054	-0.080015	0.179669	63.11107	-0.371538	0.383103	0.004574	0.093383	310.24920
	5	-0.502063	0.527918	-0.032713	0.167982	68.18029	-0.425903	0.436521	0.042936	0.105150	315.14120
20	1	-0.393900	0.418346	-0.349192	0.376728	9.94822	-0.343340	0.365754	-0.294874	0.320679	14.05611
	2	-0.271880	0.290818	-0.177492	0.204940	29.52981	-0.217937	0.233026	-0.125116	0.150017	55.33306
	3	-0.253931	0.270200	-0.112786	0.145519	46.14397	-0.205321	0.216879	-0.063859	0.093348	132.33380
	4	-0.260596	0.274678	-0.074069	0.114983	58.13898	-0.214042	0.222524	-0.026611	0.068747	223.68540
	5	-0.276800	0.288985	-0.043624	0.095299	67.02286	-0.235141	0.242179	0.000439	0.052744	359.15930
	6	-0.299321	0.310256	-0.017934	0.085705	72.37604	-0.258899	0.264554	0.022973	0.059480	344.77810
	7	-0.322084	0.332301	0.005663	0.082331	75.22397	-0.285310	0.290156	0.043299	0.067712	328.51490
	8	-0.348254	0.357901	0.028228	0.087902	75.43958	-0.314138	0.318471	0.061191	0.081194	292.23460
	9	-0.374620	0.383864	0.048022	0.097710	74.54567	-0.343410	0.347711	0.079914	0.096721	259.49900
	10	-0.402840	0.411741	0.066866	0.108377	73.67836	-0.371780	0.375737	0.097578	0.112133	235.08160
30	1	-0.352853	0.368369	-0.323835	0.340961	7.44037	-0.319230	0.333509	0.288992	0.305176	9.28415
	2	-0.223356	0.235685	-0.161288	0.178121	24.42412	-0.190866	0.201625	-0.127419	0.142794	41.19991
	3	-0.197719	0.208362	-0.104892	0.124359	40.31589	-0.165182	0.173360	-0.070574	0.088725	95.39025
	4	-0.196240	0.205882	-0.071025	0.093814	54.43312	-0.162899	0.169841	-0.038020	0.061304	177.04720
	5	-0.202003	0.210395	-0.046135	0.075603	64.06616	-0.172441	0.178293	-0.014997	0.046725	281.57950
	6	-0.213804	0.221385	-0.024700	0.063205	71.45019	-0.185622	0.190458	0.002250	0.043550	337.33180
	7	-0.226688	0.233521	-0.007941	0.057695	75.29344	-0.200036	0.204048	0.018588	0.045106	352.37440
	8	-0.242599	0.248992	0.007775	0.057090	77.07155	-0.217704	0.221309	0.033174	0.051831	326.98190
	9	-0.259471	0.265356	0.022036	0.060359	77.25358	-0.235661	0.238850	0.046793	0.060639	293.88840

Table 2 continued on next page

	10	-0.276934	0.282548	0.036215	0.067383	76.15167	-0.254437	0.257257	0.058627	0.069646	269.37800
	11	-0.295302	0.300725	0.049094	0.074862	75.10616	-0.273700	0.276336	0.072000	0.081003	241.14290
	12	-0.313803	0.319255	0.062218	0.085295	73.28311	-0.293398	0.295911	0.083363	0.091704	222.68060
	13	-0.332279	0.337432	0.075374	0.095536	71.68733	-0.311978	0.341101	0.095165	0.102770	231.90720
	14	-0.351090	0.356205	0.087783	0.106535	70.09166	-0.332096	0.334518	0.106272	0.113446	194.86980
	15	-0.370555	0.375518	0.099545	0.116477	68.98231	-0.352077	0.354327	0.118516	0.125081	183.27800

Table 3. The Monte Carlo RMSEs and bias values of HV_{mn} and SHE_{mn} for the exponential distribution with $H\,[g(x)] = 1$.

		SRS					RSS				
		HV_{mn}		SHE_{mn}		$Q_{SHE_{mn}}$	HV_{mn}		SHE_{mn}		$Q_{SHE_{mn}}$
n	***m***	**Bias**	**RMSE**	**Bias**	**RMSE**		**Bias**	**RMSE**	**Bias**	**RMSE**	
10	1	-0.552032	0.677001	-0.457584	0.600041	11.36778	-0.430553	0.505229	-0.342184	0.432785	14.33884
	2	-0.442683	0.571820	-0.253108	0.442568	22.60362	-0.337494	0.404667	-0.148595	0.269907	33.30146
	3	-0.435444	0.561640	-0.154607	0.391369	30.31675	-0.332760	0.401125	-0.052759	0.230412	42.55855
	4	-0.451545	0.575390	-0.076188	0.371210	35.48550	-0.348029	0.420617	0.025378	0.233566	44.47062
	5	-0.469437	0.597761	0.005489	0.372418	37.69784	-0.366628	0.445977	0.101893	0.270512	39.34396
20	1	-0.414064	0.490107	-0.360711	0.445976	9.00436	-0.357765	0.398661	-0.312513	0.358752	10.01076
	2	-0.285717	0.376086	-0.193143	0.310495	17.44043	-0.234959	0.280262	-0.140851	0.207405	25.99603
	3	-0.260773	0.351341	-0.122104	0.272095	22.55530	-0.213397	0.261261	-0.072871	0.165700	36.57683
	4	-0.256116	0.352810	-0.067569	0.251502	28.71461	-0.210620	0.259248	-0.017564	0.152350	41.23388
	5	-0.262412	0.358638	-0.022414	0.244018	31.95980	-0.214122	0.265246	0.022190	0.156584	40.96650
	6	-0.265650	0.360325	0.016823	0.248330	31.08166	-0.218028	0.272315	0.061287	0.174543	35.90401
	7	-0.266934	0.365008	0.055461	0.256349	29.76894	-0.224596	0.282196	0.103601	0.200858	28.82323
	8	-0.273952	0.377519	0.100674	0.274582	27.26671	-0.232629	0.293062	0.145963	0.231970	20.84610
	9	-0.280123	0.381968	0.143573	0.293999	23.03046	-0.236125	0.302083	0.188596	0.267430	11.47135
	10	-0.285183	0.391290	0.179545	0.322338	17.62171	-0.238413	0.310922	0.231203	0.303760	2.30347
30	1	-0.367058	0.423423	-0.332016	0.394742	6.77360	-0.332526	0.361491	-0.303272	0.334033	7.59576

Table 3 continued on next page

2	-0.233677	0.306086	-0.173511	0.262016	14.39791	-0.203455	0.236001	-0.137679	0.182964	22.47321
3	-0.202277	0.281503	-0.108684	0.223191	20.71452	-0.170859	0.207468	-0.078000	0.141567	31.76442
4	-0.194424	0.275072	-0.067472	0.207505	24.56339	-0.160246	0.199410	-0.036059	0.123278	38.17863
5	-0.191705	0.272356	-0.033792	0.197718	27.40457	-0.159714	0.200465	-0.002510	0.122595	38.84469
6	-0.186870	0.272196	0.000772	0.195841	28.05148	-0.158702	0.202869	0.027994	0.128086	36.86270
7	-0.191094	0.275374	0.029066	0.198154	28.04186	-0.161705	0.206226	0.059517	0.141042	31.60804
8	-0.195662	0.280589	0.056849	0.208607	25.65389	-0.164468	0.212265	0.085540	0.160732	24.27767
9	-0.196983	0.282040	0.088082	0.220610	21.78060	-0.165511	0.217222	0.115128	0.182796	15.84830
10	-0.197171	0.283394	0.115949	0.235447	16.91885	-0.167152	0.220237	0.144441	0.205632	6.63149
11	-0.198853	0.286241	0.142656	0.253233	11.53154	-0.173076	0.229318	0.172966	0.220033	4.04896
12	-0.204089	0.293653	0.171742	0.274080	6.66535	-0.171555	0.232740	0.200259	0.214615	7.78766
13	-0.202908	0.298108	0.204980	0.228389	23.38717	-0.176996	0.240454	0.231487	0.232102	3.47343
14	-0.205700	0.300842	0.232277	0.290007	3.60156	-0.176922	0.244541	0.262425	0.211142	13.65780
15	-0.210699	0.305809	0.258234	0.300011	1.89595	-0.177959	0.248760	0.291253	0.239115	3.87723

Table 4. The Monte Carlo RMSEs and bias values of HV_{mn} and SHE_{mn} for the standard normal distribution and $H[g(x)] = 1.419$.

		SRS					RSS				
		HV_{mn}		SHE_{mn}		$Q_{SHE_{mn}}$	HV_{mn}		SHE_{mn}		$Q_{SHE_{mn}}$
n	***m***	**Bias**	**RMSE**	**Bias**	**RMSE**		**Bias**	**RMSE**	**Bias**	**RMSE**	
10	1	-0.598925	0.676499	-0.499469	0.434171	35.8208955	-0.484489	0.549750	-0.388446	0.466743	15.09905
	2	-0.521455	0.591007	-0.335907	0.436633	26.1205028	-0.422169	0.471157	-0.238609	0.320258	32.02733
	3	-0.563002	0.623188	-0.275063	0.382983	38.5445484	-0.462240	0.504378	-0.181597	0.269765	46.51531
	4	-0.610651	0.663364	-0.236072	0.351842	46.9609445	-0.523019	0.557792	-0.149270	0.244690	56.13239
	5	-0.671777	0.719069	-0.200702	0.325688	54.7069892	-0.584483	0.614209	-0.111978	0.218489	64.42758
20	1	-0.435480	0.483459	-0.380981	0.434171	10.1948666	-0.382986	0.420310	-0.335512	0.377639	10.15227
	2	-0.327145	0.375798	-0.231087	0.296133	21.1988888	-0.275716	0.313472	-0.182040	0.234712	25.12505
	3	-0.317948	0.364927	-0.175301	0.251511	31.0790925	-0.268657	0.304811	-0.125104	0.189103	37.96057

Table 4 continued on next page

	4	-0.327070	0.372436	-0.143556	0.230357	38.1485678	-0.285331	0.318855	-0.098619	0.172598	45.86944
	5	-0.352658	0.395796	-0.117332	0.215233	45.6202185	-0.305555	0.337744	-0.073404	0.160748	52.40537
	6	0.375996	0.416964	-0.098719	0.204234	51.0187930	-0.335066	0.365185	-0.051912	0.152608	58.21077
	7	-0.404050	0.442997	-0.083445	0.199295	55.0121107	-0.363782	0.391748	-0.036080	0.148138	62.18538
	8	-0.439618	0.475094	-0.061765	0.187822	60.4663498	-0.395221	0.421583	-0.020165	0.147835	64.93336
	9	-0.467134	0.500777	-0.043230	0.186628	62.7323140	-0.428042	0.451680	-0.006860	0.144519	68.00412
	10	-0.496926	0.527456	-0.029603	0.178984	66.0665534	-0.454818	0.477152	0.009882	0.145955	69.41121
30	1	-0.378860	0.413455	-0.346828	0.384885	6.91006276	-0.343626	0.370512	-0.313688	0.342854	7.464805
	2	-0.259105	0.299687	-0.196988	0.246877	17.6217187	-0.226914	0.255947	-0.163491	0.201857	21.13328
	3	-0.236758	0.277238	-0.145212	0.203905	26.4512801	-0.204698	0.234358	-0.108571	0.157593	32.75544
	4	-0.234369	0.275867	-0.108651	0.179817	34.8175026	-0.204765	0.234413	-0.081230	0.140863	39.90820
	5	-0.244288	0.283027	-0.088572	0.166051	41.3303324	-0.214434	0.243683	-0.056181	0.127184	47.80760
	6	-0.255248	0.293332	-0.068084	0.157937	46.1575962	-0.227340	0.255901	-0.038603	0.122294	52.21043
	7	-0.269724	0.305134	-0.048333	0.151084	50.4860160	-0.241325	0.268228	-0.021655	0.120957	54.90516
	8	-0.285713	0.321039	-0.036608	0.151194	52.9047873	-0.254983	0.282376	-0.008427	0.120242	57.41777
	9	-0.304064	0.337563	-0.020683	0.147718	56.2398723	-0.274697	0.301420	0.010331	0.123468	59.03789
	10	-0.320051	0.352764	-0.009717	0.148068	58.0263292	-0.295057	0.319933	0.018501	0.125482	60.77866
	11	-0.339131	0.369866	0.005731	0.147483	60.1252886	-0.314201	0.339141	0.030498	0.129224	61.89667
	12	-0.361226	0.392070	0.016315	0.149674	61.8246742	-0.333173	0.356224	0.042772	0.133458	62.53537
	13	-0.382347	0.410463	0.027129	0.152493	62.8485393	-0.353582	0.375170	0.053690	0.138200	63.16337
	14	-0.400618	0.428008	0.039711	0.155154	63.7497430	-0.375752	0.397462	0.064272	0.140967	64.53321
	15	-0.423597	0.449968	0.048426	0.156576	65.2028590	-0.394363	0.414605	0.072957	0.147206	64.49488

Table 5. The Monte Carlo RMSEs and bias values of HV_{mn} and SHE_{mn} for the uniform distribution with $H[g(x)] = 0$ and exponential distribution with $H[g(x)] = 1$ using DRSS.

		SRS					RSS				
		HV_{mn}		SHE_{mn}		$Q_{SHE_{mn}}$	HV_{mn}		SHE_{mn}		$Q_{SHE_{mn}}$
n	***m***	**Bias**	**RMSE**	**Bias**	**RMSE**		**Bias**	**RMSE**	**Bias**	**RMSE**	

Table 5 continued on next page

10	1	-0.327408	0.369593	-0.230787	0.285326	22.7999448	-0.365854	0.425279	-0.267318	0.345821	18.68373
	2	-0.260621	0.278731	-0.071592	0.121826	56.2926262	-0.288898	0.340618	-0.101687	0.207273	39.14796
	3	-0.296104	0.306116	-0.014117	0.078474	74.3646199	-0.300393	0.351750	-0.018027	0.181245	48.47335
	4	-0.346305	0.352712	0.029482	0.073990	79.0225453	-0.322839	0.377437	0.056521	0.201436	46.63056
	5	-0.404121	0.409902	0.065862	0.095121	76.7942093	-0.335248	0.399189	0.134718	0.252269	36.80462
20	1	-0.308453	0.329353	-0.260588	0.285042	13.4539537	-0.329105	0.363241	-0.278366	0.317530	12.58421
	2	-0.189231	0.202666	-0.095093	0.119561	41.0058915	-0.204908	0.240316	-0.112945	0.168444	29.90729
	3	-0.182095	0.191163	-0.041993	0.071976	62.3483624	-0.191216	0.228320	-0.050530	0.133863	41.37044
	4	-0.197693	0.204342	-0.010391	0.052373	74.3699288	-0.190904	0.229986	-0.003685	0.126728	44.89752
	5	-0.220876	0.225845	0.012711	0.049477	78.0924971	-0.197900	0.239789	0.036502	0.139896	41.65871
	6	-0.247733	0.251580	0.035133	0.056178	77.6699261	-0.207032	0.251002	0.078413	0.161731	35.56585
	7	-0.275808	0.278919	0.053697	0.068101	75.5839509	-0.209883	0.258152	0.118656	0.192217	25.54115
	8	-0.303823	0.306608	0.071232	0.082285	73.1628007	-0.218701	0.271560	0.158069	0.224230	17.42893
	9	-0.333903	0.336495	0.089491	0.098489	70.7309172	-0.223692	0.278728	0.200103	0.262984	5.648518
	10	-0.363272	0.365731	0.106408	0.114566	68.6747910	-0.228126	0.290431	0.244783	0.283888	2.252859
30	1	-0.298092	0.312767	-0.267592	0.283216	9.44824742	-0.308011	0.331033	-0.278838	0.304383	8.050557
	2	-0.170745	0.180210	-0.107748	0.122162	32.2113090	-0.182416	0.207785	-0.118447	0.154790	25.50473
	3	-0.146113	0.153646	-0.052193	0.070391	54.1862463	-0.152039	0.180708	-0.059074	0.114805	36.46933
	4	-0.149143	0.154886	-0.023125	0.047458	69.3593998	-0.145325	0.176699	-0.019990	0.102139	42.19605
	5	-0.159888	0.164564	-0.003052	0.038571	76.5617024	-0.146632	0.179028	0.009230	0.105307	41.17847
	6	-0.174419	0.178204	0.013102	0.038421	78.4398779	-0.149443	0.184598	0.038407	0.115953	37.18621
	7	-0.191854	0.194940	0.027534	0.046606	76.0921309	-0.150245	0.188158	0.068588	0.133307	29.15156
	8	-0.209886	0.212509	0.040817	0.052754	75.1756396	-0.153441	0.194332	0.095598	0.152215	21.67270
	9	-0.229010	0.231261	0.052824	0.061955	73.2099230	-0.157250	0.199936	0.123844	0.175122	12.41097
	10	-0.248006	0.249993	0.065446	0.072283	71.0859904	-0.162854	0.208891	0.151295	0.198703	4.877185
	11	-0.267506	0.269188	0.077163	0.082922	69.1955065	-0.163540	0.213175	0.182129	0.207543	2.641961
	12	-0.287408	0.289018	0.088169	0.093391	67.6867877	-0.167660	0.221482	0.207757	0.202062	8.768207
	13	-0.307160	0.308699	0.100118	0.104801	66.0507485	-0.171024	0.225764	0.239466	0.211883	6.148456
	14	-0.327370	0.328890	0.111085	0.115458	64.8946456	-0.170880	0.232977	0.268159	0.210502	9.646875
	15	-0.346997	0.348439	0.122960	0.126985	63.5560313	-0.169873	0.235173	0.299068	0.210721	10.397450

Table 6. The Monte Carlo RMSEs and bias values of HV_{mn} and SHE_{mn} for the standard normal distribution and $H\,[g(x)] = 1.419$.

		HV_{mn}		SHE_{mn}		$Q_{SHE_{mn}}$
n	**m**	**Bias**	**RMSE**	**Bias**	**RMSE**	
10	1	-0.415021	0.472162	-0.316672	0.385139	18.43075
	2	-0.373395	0.412666	-0.186378	0.256423	37.86185
	3	-0.427401	0.459119	-0.143329	0.218981	52.30409
	4	-0.492911	0.518275	-0.115918	0.202153	60.99503
	5	-0.554351	0.577281	-0.084253	0.181100	68.62880
20	1	-0.350703	0.383160	-0.303014	0.340790	11.05804
	2	-0.245907	0.277809	-0.152363	0.200155	27.95230
	3	-0.246496	0.276941	-0.104439	0.162172	41.44168
	4	-0.262789	0.290545	-0.078826	0.147712	49.16037
	5	-0.291340	0.317967	-0.055774	0.138687	56.38321
	6	-0.316105	0.341597	-0.037661	0.134214	60.70984
	7	-0.349246	0.373132	-0.021199	0.132559	64.47397
	8	-0.384526	0.406764	-0.008681	0.134158	67.01822
	9	-0.416151	0.436696	0.006082	0.132054	69.76066
	10	-0.445901	0.465518	0.023744	0.134764	71.05074
30	1	-0.321940	0.345223	-0.292331	0.318084	7.861300
	2	-0.206709	0.231560	-0.143028	0.177006	23.55934
	3	-0.187163	0.212774	-0.094482	0.138090	35.10015
	4	-0.190073	0.215577	-0.066854	0.122350	43.24534
	5	-0.199843	0.224569	-0.044224	0.111818	50.20773
	6	-0.214636	0.239021	-0.025579	0.108667	54.53663
	7	-0.231613	0.255278	-0.012061	0.108224	57.60543
	8	-0.247340	0.271084	0.001734	0.109348	59.66269
	9	-0.268298	0.291044	0.014961	0.113895	60.86674
	10	-0.286538	0.308661	0.027278	0.118811	61.50761
	11	-0.305310	0.326485	0.040250	0.123778	62.08769
	12	-0.324892	0.346062	0.051274	0.129747	62.50759
	13	-0.343097	0.363236	0.061548	0.135452	62.70964
	14	-0.369990	0.388586	0.070900	0.140756	63.77739
	15	-0.387740	0.406081	0.080947	0.145418	64.18990

Table 7. The Monte Carlo RMSEs and bias values of $AHEj_{mn}$, j = SRS, RSS, DRSS (Al-Omari, 2014).

		$AHESRS_{mn}$		Q_{AHESRS}	$AHERSS_{mn}$		Q_{AHERSS}	$AHEDRSS_{mn}$		$Q_{AHEDRSS}$
n	***m***	**Bias**	**RMSE**		**Bias**	**RMSE**		**Bias**	**RMSE**	
Uniform distribution with $H\,[g(x)] = 0$										
10	2	-0.298609	0.350332	22.554260	-0.189664	0.228762	30.516686	-0.145388	0.176159	36.799638
	3	-0.249056	0.298944	34.126897	-0.154894	0.186380	45.818350	-0.122180	0.144286	52.865580
20	4	-0.144016	0.167779	38.917933	-0.100304	0.118284	46.844385	-0.082268	0.096978	52.541328
	5	-0.133179	0.157805	45.393360	-0.091608	0.108584	55.163743	-0.077708	0.091093	59.665700
30	7	-0.092957	0.109089	53.285144	-0.066053	0.077716	61.912883	-0.058041	0.067650	65.297015
	8	-0.089259	0.105818	57.501446	-0.064713	0.076188	65.573926	-0.056421	0.065369	69.239420
Exponential distribution with $H\,[g(x)] = 1$										
10	2	-0.323532	0.483573	15.432654	-0.220406	0.315220	22.103853	-0.173991	0.251460	26.175364
	3	-0.265713	0.443276	21.074710	-0.159787	0.276197	31.144406	-0.128545	0.223802	36.374698
20	4	0.141143	0.279706	20.720501	-0.098056	0.179990	30.572271	-0.075338	0.179771	21.833938
	5	0.118697	0.271887	24.189015	-0.072456	0.172661	34.905333	-0.052175	0.145269	39.417988
30	7	-0.058550	0.205261	25.461009	-0.027194	0.130283	36.825134	-0.046556	0.115023	38.868929
	8	-0.036080	0.200329	28.604115	-0.010631	0.136358	35.760488	-0.001239	0.120306	38.092543
Standard normal distribution with $H\,[g(x)] = 1.419$										
10	2	-0.409842	0.496627	15.969354	-0.308706	0.375690	20.262250	-0.262149	0.316029	23.417728
	3	-0.386562	0.468471	24.826698	-0.291133	0.353844	29.845470	-0.254450	0.303820	33.825435
20	4	-0.214227	0.279269	25.015573	-0.168035	0.219922	31.027583	-0.148107	0.194728	32.978368
	5	-0.205782	0.272804	31.074594	-0.160392	0.213700	36.727225	-0.145734	0.191755	39.693427
30	7	-0.132038	0.196792	35.506368	-0.105796	0.158654	40.851067	-0.095517	0.143483	43.793433
	8	-0.129915	0.193509	39.724146	-0.102504	0.157726	44.143270	-0.094560	0.145579	46.297458

Conclusion

Three entropy estimators are suggested using SRS, RSS, and DRSS methods. The consistency of these estimators is proved as well as some properties are reported. Based on theoretical and numerical comparisons the suggested entropy estimators are more efficient than Vasicek (1976) and Al-Omari (2014) entropy estimators. However, the suggested estimators of entropy in this paper can be extended by considering other sampling methods such as the multistage RSS and median RSS methods.

Acknowledgements

The author thanks the referees for their helpful and valuable comments that substantially improved this paper.

References

Al-Omari, A. I. (2014). Estimation of entropy using random sampling. *Journal of Computation and Applied Mathematics*, *261*, 95-102. doi:10.1016/j.cam.2013.10.047

Al-Saleh, M. F. & Al-Kadiri, M. A. (2000). Double ranked set sampling. *Statistics and Probability Letters*, *48*(2), 205-212. doi:10.1016/S0167-7152(99)00206-0

Choi, B. (2008). Improvement of goodness of fit test for normal distribution based on entropy and power comparison. *Journal of Statistical Computation and Simulation*, *78*(9), 781-788. doi:10.1080/00949650701299451

Choi, B., Kim, K., & Song, S. H. (2004). Goodness of fit test for exponentiality based on Kullback-Leibler information. *Communication in Statistics-Simulation and Computation*, *33*(2), 525–536. doi:10.1081/SAC-120037250

Goria, M. N., Leonenko, N. N., Mergel, V. V., & Novi Inverardi, P. L. (2005). A new class of random vector entropy estimators and its applications in testing statistical hypotheses. *Journal of Nonparametric Statistics*, *17*(3), 277-297. doi:10.1080/104852504200026815

Correa, J. C. (1995). A new estimator of entropy. *Communication in Statistics-Theory Methods*, *24*(10), 2439-2449. doi:10.1080/03610929508831626

Ebrahimi, N., Pflughoeft, K., & Soofi, E. S. (1994). Two measures of sample entropy. *Statistics & Probability Letters*, *20*(3), 225-234. doi:10.1016/0167-7152(94)90046-9

Mahdizadeh, M. (2012). On the use of ranked set samples in entropy based test of fit for the Laplace distribution. *Revista Colombiana de Estadística*, *35*(3), 443-455.

McIntyre, G. A. (1952). A method for unbiased selective sampling using ranked sets. *Australian Journal of Agricultural Research*, *3*(4), 385-390. doi:10.1071/AR9520385

Noughabi, H. A. & Noughabi, R. A. (2013). On the entropy estimators. *Journal of Statistical Computation and Simulation*, *83*(4), 784-792. doi:10.1080/00949655.2011.637039

Noughabi, H. A. & Arghami, N. R. (2010). A new estimator of entropy. *Journal of the Iranian Statistical Society*, *9*(1), 53-64.

Park, S. & Park, D. (2003). Correcting moments for goodness of fit tests based on two entropy estimates. *Journal of Statistical Computation and Simulation*, *73*(9), 685-694. doi:10.1080/0094965031000070367

Shannon, C. E. (1948a). A mathematical theory of communications. *Bell System Technical Journal 27*(3), 379-423. doi:10.1002/j.1538-7305.1948.tb01338.x

Shannon, C. E. (1948b). A mathematical theory of communications. *Bell System Technical Journal 27*(4), 623-656. doi:10.1002/j.1538-7305.1948.tb00917.x

Takahasi, K. & Wakimoto, K. (1968). On the unbiased estimates of the population mean based on the sample stratified by means of ordering. *Annals of the Institute of Statistical Mathematics*, *20*(1), 1-31. doi:10.1007/BF02911622

Theil, J. (1980). The entropy of maximum entropy distribution. *Economics Letters*, *5*(2), 145–148. doi:10.1016/0165-1765(80)90089-0

Van Es, B. (1992). Estimating functionals related to a density by class of statistics based on spacings. *Scandinavian Journal of Statistics*, *19*(1), 61-72.

Vasicek, O. (1976). A test for normality based on sample entropy. *Journal of the Royal Statistical Society, B*, *38*, 54-59.

Wieczorkowski, R. & Grzegorzewsky, P. (1999). Entropy estimators - improvements and comparisons. *Communication in Statistics-Simulation and Computation*, *28*(2), 541-567. doi:10.1080/03610919908813564

Journal of Modern Applied Statistical Methods
November 2015, Vol. 14, No. 2, 110-122.

ISSN 1538 − 9472

Bayesian Analysis Under Progressively Censored Rayleigh Data

Gyan Prakash
Department of Community Medicine
S. N. Medical College, Agra, U. P., India

The one-parameter Rayleigh model is considered as an underlying model for evaluating the properties of Bayes estimator under Progressive Type-II right censored data. The One-Sample Bayes prediction bound length (OSBPBL) is also measured. Based on two different asymmetric loss functions a comparative study presented for Bayes estimation. A simulation study was used to evaluate their comparative properties.

Keywords: Rayleigh model, Bayes estimator, Progressive Type-II right censoring scheme, ISELF, LLF, OSBPBL.

Introduction

The Rayleigh distribution is considered as a useful life distribution. It plays an important role in statistics and operations research. Rayleigh model is applied in several areas such as health, agriculture, biology and physics. It often used in physics, related fields to model processes such as sound and light radiation, wave heights, as well as in communication theory to describe hourly median and instantaneous peak power of received radio signals. The model for frequency of different wind speeds over a year at wind turbine sites and daily average wind speed are considered under the Rayleigh model.

The probability density function and distribution function of Rayleigh distribution are

$$f\left(x;\sigma\right)=\frac{x}{\sigma^2}\exp\left(-\frac{x^2}{2\sigma^2}\right); x\geq 0,\sigma>0 \tag{1}$$

and

Gyan Prakash is an Assistant Professor in the department of Community Medicine at S. N. Medical College. Email him at ggyanji@yahoo.com.

$$F(x;\sigma) = 1 - \exp\left(-\frac{x^2}{2\sigma^2}\right); x \geq 0, \sigma > 0. \tag{2}$$

Here, the parameter σ is known as location parameter. The considered model is useful in life testing experiments, in which age with time as its failure rate and is a linear function of time. The present distribution also plays an important role in communication engineering and electro-vacuum device.

The focus is on measurement of One-Sample Bayes prediction bound length based on Progressive Type-II right censored data. A comparative study of Bayes estimation under two different asymmetric loss functions is presented. For evaluation of performances of the proposed procedures, a simulation study carries out also.

A great deal of literature is available on Rayleigh model under different criterions, such as Sinha (1990), Bhattacharya & Tyagi (1990), Fernandez (2000), Hisada & Arizino (2002), Ali-Mousa & Al–Sagheer (2005), Wu, Chen, and Chen (2006), Kim & Han (2009), Prakash & Prasad (2010), Prakash & Singh (2013). Soliman, Amin, and Abd-El Aziz (2010) presented results on estimation and prediction of inverse Rayleigh distribution based on lower record values. Recently, Prakash (2013) presented Bayes estimators for inverse Rayleigh model. Bayesian analysis for Rayleigh distribution was also discussed by Ahmed, Ahmad, and Reshi (2013).

The progressive Type-II right censoring

The progressive censoring appears to be a great importance in planned duration experiments in reliability studies. In many industrial experiments involving lifetimes of machines or units, experiments have to be terminated early and the number of failures must be limited for various reasons. In addition, some life tests require removal of functioning test specimens to collect degradation related information to failure time data.

Progressive censored sampling is an important method of obtaining data in lifetime studies. Live units removed early on can be readily used in others tests, thereby saving cost to experimenter and a compromise can be achieved between time consumption and the observation of some extreme values. The Progressive Type-II right censoring scheme is describes as follows.

Suppose an experiment in which n independent and identical units $X_1, X_2, \ldots, X_n$ are placed on a life test at the beginning time and first r; $(1 \leq r \leq n)$ failure times are observed. At time of each failure occurring prior to the termination point, one or more surviving units removed from the test. The experiment is terminated at time of r^{th} failure, and all remaining surviving units are removed from the test.

Let $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(r)}$ be the lifetimes of completely observed units to fail and $R_1, R_2, \ldots, R_r$; $(r \leq n)$ are the numbers of units withdrawn at these failure times. Here, $R_1, R_2, \ldots, R_r$; $(r \leq n)$ all are predefined integers follows the relation $R_1 + R_2 + \ldots + R_r + r = n$.

At the first failure time $x_{(1)}$, withdraw R_1 units randomly from remaining n - 1 surviving units. Immediately after second observed failure time $x_{(2)}$, R_2 units are withdrawn from remaining $n - 2 - R_1$ surviving units at random, and so on. The experiments continue until at r^{th} failure time x_r, remaining units $R_r = n - r - \sum_{j=1}^{r-1} R_j$ are withdrawn. Here, $X_{1:r:n}^{(R_1,R_2,\ldots,R_r)}, X_{2:r:n}^{(R_1,R_2,\ldots,R_r)}, \ldots, X_{r:r:n}^{(R_1,R_2,\ldots,R_r)}$ be the r ordered failure items and $(R_1, R_2, \ldots, R_r)$ be progressive censoring scheme.

Progressively Type-II right censoring scheme reduces to conventional Type-II censoring scheme when

$$R_i = 0 \forall i = 1, 2, \ldots, r-1 \Rightarrow R_r = n - r$$

and for complete sample case when

$$R_i = 0 \forall i = 1, 2, \ldots, r \Rightarrow n = r.$$

Based on progressively Type-II censoring scheme the joint probability density function of order statistics $X_{1:r:n}^{(R_1,R_2,\ldots,R_r)}, X_{2:r:n}^{(R_1,R_2,\ldots,R_r)}, \ldots, X_{r:r:n}^{(R_1,R_2,\ldots,R_r)}$ is defined as

$$f_{\left(X_{1:r:n}, X_{2:r:n}, \ldots, X_{r:r:n}\right)}\left(\sigma \middle| \underline{x}\right) = K_p \prod_{i=1}^{r} f\left(x_{(i)}; \sigma\right)\left(1 - F\left(x_{(i)}; \sigma\right)\right)^{R_i}. \tag{3}$$

Here, K_p is called as progressive normalizing constant and is defined as

$$K_p = n\left(n - R_1 - 1\right)\left(n - R_1 - R_2 - 2\right)\ldots\left(n + 1 - \sum_{j=1}^{r-1} R_j - r\right)$$

Progressive Type-II censored sample is denoted by $x \equiv (x_{(1)}, x_{(2)}, \ldots, x_{(r)})$ and $(R_1, R_2, \ldots, R_r)$ being Progressive censoring scheme for the considered model. Simplifying (3)

$$\Rightarrow f_{(X_{1:r:n}, X_{2:r:n}, \ldots, X_{r:r:n})}(\sigma | \underline{x}) = K_p A_r(\underline{x}) \sigma^{-2r} \exp\left(-\frac{T_r(\underline{x})}{\sigma^2}\right); \tag{4}$$

where $A_r(\underline{x}) = \prod_{i=1}^{r} x_{(i)}$ and $T_r(\underline{x}) = \frac{1}{2} \sum_{i=1}^{r} (1 + R_i) x_{(i)}^2$.

The Bayes estimation

There is no clear-cut way to determine if one prior probability estimate is better than the other. It is more frequently the case that attention is restricted to a given flexible family of priors, and one is chosen from that family that matches best with personal beliefs. However, there is adequate information about the parameter it should be used; otherwise it is preferable to use the non-informative prior. In present study, the extended Jeffrey's prior proposed by Al-Kutubi & Ibrahim (2009) is considered:

$$\pi(\sigma) \propto (I(\sigma))^c; c \in R^+, I(\sigma) = -nE\left[\frac{\partial^2 \log f(x, \sigma)}{\partial \sigma^2}\right] \tag{5}$$

Thus, the extended Jeffrey's prior for present model is

$$\pi(\sigma) \propto \left(\frac{n}{\sigma^2}\right)^c; c \in R^+. \tag{6}$$

Based on Bayes theorem, the posterior density is defined as

$$\pi^*(\sigma | \underline{x}) = \frac{f_{(X_{1:r:n}, X_{2:r:n}, \ldots, X_{r:r:n})}(\sigma | \underline{x}) \cdot \pi(\sigma)}{\int_{\sigma} f_{(X_{1:r:n}, X_{2:r:n}, \ldots, X_{r:r:n})}(\sigma | \underline{x}) \cdot \pi(\sigma) d\sigma}. \tag{7}$$

Using (4) and (6) in (7), the posterior density is obtain as

$$\pi^*\left(\sigma|\underline{x}\right)\propto\frac{K_pA_r\left(\underline{x}\right)\sigma^{-2r}\exp\left(-\frac{T_r\left(\underline{x}\right)}{\sigma^2}\right)\cdot\left(\frac{n}{\sigma^2}\right)^c}{\int\limits_\sigma K_pA_r\left(\underline{x}\right)\sigma^{-2r}\exp\left(-\frac{T_r\left(\underline{x}\right)}{\sigma^2}\right)\cdot\left(\frac{n}{\sigma^2}\right)^c d\sigma}$$

$$\Rightarrow\pi^*\left(\sigma|\underline{x}\right)=\frac{2\left(T_r\left(\underline{x}\right)\right)^{r+c-\frac{1}{2}}}{\Gamma\left(r+c-\frac{1}{2}\right)}\sigma^{-2(r+c)}\exp\left(-\frac{T_r\left(\underline{x}\right)}{\sigma^2}\right). \tag{8}$$

The selection of loss function may be crucial in Bayesian analysis. If most commonly used loss function, squared error loss function (SELF) is taken as a measure of inaccuracy, and then the resulting risk is often too sensitive to assumptions about behavior of tail of probability distribution. In Bayesian point of view, SELF is inappropriate in many situations. To overcome this difficulty, a useful asymmetric loss function based on SELF has selected. This asymmetric loss function is known as invariant squared error loss function (ISELF) and is defined for any estimate $\hat{\sigma}$ corresponding to the parameter σ as

$$L\left(\hat{\sigma},\sigma\right)=\left(\sigma^{-1}\partial\right)^2;\partial=\hat{\sigma}-\sigma. \tag{9}$$

The Bayes estimator $\hat{\sigma}_I$ for location parameter σ under ISELF is obtained as

$$\begin{aligned}\hat{\sigma}_1&=\left[E\left(\sigma^{-1}\right)\right]\left[E\left(\sigma^{-2}\right)\right]^{-1}\\&=\left[\int\limits_\sigma\sigma^{-1}\pi^*\left(\sigma|\underline{x}\right)d\sigma\right]\left[\int\limits_\sigma\sigma^{-2}\pi^*\left(\sigma|\underline{x}\right)d\sigma\right]^{-1}\end{aligned}$$

$$\Rightarrow\hat{\sigma}_1=\frac{\Gamma\left(r+c\right)}{\Gamma\left(r+c+\frac{1}{2}\right)}\sqrt{T_r\left(\underline{x}\right)}. \tag{10}$$

Some estimation problems overestimation is more serious than the underestimation, or vice-versa. In addition, there are some cases when the positive and negative errors have different consequences. In such cases, a useful and flexible class of asymmetric loss function (LINEX loss function (LLF)) is defined as

$$L\left(\partial^*\right)=e^{a\partial^*}-a\partial^*-1;a\neq 0,\partial^*=\left(\sigma^{-1}\partial\right). \tag{11}$$

The shape parameter of LLF is denoted by 'a'. Negative (positive) value of 'a' gives more weight to overestimation (underestimation) and its magnitude reflect the degree of asymmetry. It is seen that, for $a = 1$ the function is quite asymmetric with overestimation being more costly than underestimation. For small values of $| a |$, LLF is almost symmetric and is not far from SELF.

The Bayes estimator $\hat{\sigma}_L$ of location parameter under LLF is obtain by simplifying following equality

$$E\left\{\frac{1}{\sigma}e^{-a\frac{\hat{\sigma}_L}{\sigma}}\right\}=e^a E\left\{\frac{1}{\sigma}\right\}$$

$$\Rightarrow\int_\sigma \sigma^{-1}e^{-a\frac{\hat{\sigma}_L}{\sigma}}\pi^*\left(\sigma|\underline{x}\right)d\sigma=e^a\int_\sigma \sigma^{-1}\pi^*\left(\sigma|\underline{x}\right)d\sigma$$

$$\Rightarrow\int_\sigma \sigma^{-(2r+2c+1)}\exp\left(-\left(\frac{T_r\left(\underline{x}\right)}{\sigma^2}+a\frac{\hat{\sigma}_L}{\sigma}\right)\right)d\sigma=\frac{e^a}{2}\Gamma\left(r+c\right)T_r\left(\underline{x}\right)^{-(r+c)}. \tag{12}$$

A closed form of Bayes estimator $\hat{\sigma}_L$ does not exist. A numerical technique is applied here for obtaining the risk for the Bayes estimator corresponding to their loss.

One-sample Bayes prediction bound length

Consider the nature of future behavior of the observation when sufficient information about the past and the present behavior of an event or an observation is known or given. The Bayesian statistical analysis to predict the future statistic from the considered model is based on the Progressive Type-II right ordered data.

Let $x_{(1)}, x_{(2)},\ldots, x_{(r)}$ be the first r observed failure units from a sample of size n under the Progressive Type-II right censoring scheme from underlying model (1). If $y \equiv (y_{(1)}, y_{(2)},\ldots, y_{(s)})$ be the second independent random sample of future observations from same model. Then Bayes predicative density of future observation Y is denoted by $h\left(Y|\underline{x}\right)$ and obtained by simplifying the following relation

$$h\left(Y|\underline{x}\right)=\int_{\sigma} f\left(y;\sigma\right)\cdot\pi^{*}\left(\sigma|\underline{x}\right)d\sigma$$

$$=\frac{2y\left(T_{r}\left(\underline{x}\right)\right)^{r+c-\frac{1}{2}}}{\Gamma\left(r+c-\frac{1}{2}\right)}\int_{\sigma}\exp\left(-\frac{1}{\sigma^{2}}\left(\frac{y^{2}}{2}+T_{r}\left(\underline{x}\right)\right)\right)\cdot\sigma^{-2(r+c+1)}d\sigma$$

$$\Rightarrow h\left(Y|\underline{x}\right)=\left(r+c-\frac{1}{2}\right)y\left(T_{r}\left(\underline{x}\right)\right)^{r+c-\frac{1}{2}}\left(T_{r}\left(\underline{x}\right)+\frac{y^{2}}{2}\right)^{-r-c-\frac{1}{2}}. \qquad (13)$$

Let l_1 and l_2 are the lower and upper Bayes prediction limits for the random variable Y and $1 - \vartheta$ is called the confidence prediction coefficient. Then (l_1, l_2) be the 100(1 - ϑ) % prediction limits for future random variable Y, if

$$\Pr\left(l_{1}\le Y\le l_{2}\right)=1-\vartheta. \qquad (14)$$

Now, the Central Coverage Bayes Prediction lower and upper limits are obtain by solving following equality

$$\Pr\left(Y\le l_{1}\right)=\frac{1-\vartheta}{2}=\Pr\left(Y\ge l_{2}\right). \qquad (15)$$

Solving (15), the lower and upper Bayes prediction limits for the future random observation Y are obtain as

$$l_{1}=\sqrt{\left(2T_{r}\left(\underline{x}\right)\left(\vartheta^{*}-1\right)\right)} \text{ and } l_{2}=\sqrt{\left(2T_{r}\left(\underline{x}\right)\left(\vartheta^{**}-1\right)\right)};$$

where

$$\vartheta^{*}=\left(\frac{1+\vartheta}{2}\right)^{-\lambda}, \vartheta^{**}=\left(\frac{1-\vartheta}{2}\right)^{-\lambda} \text{ and } \lambda=\left(r+c-\frac{1}{2}\right)^{-1}.$$

The One-Sample Bayes Prediction bound length under the Central Coverage is obtained as

$$I = l_2 - l_1. \tag{16}$$

Numerical illustration

The procedure is illustrated by presenting a complete analysis under a simulated data set in present section. A comparative study of Bayes estimators based on simulation in terms of risk ratios under Progressively Type-II right censored data is presented as follows:

1) Random values of parameter σ are generated from prior density (6) for selected parametric values of c (= 0, 0.50, 1.50, 2.00, 5.00) and $n = 20$.
2) The value of $c = 0$ is used for Uniform distribution. For the values of $c = 0.50$ and $c = 1.50$ the analysis corresponding to the Jeffrey's prior and Hartigan's prior (Hartigan (1964)) respectively.
3) Using generated values of σ obtained in step (1), generate a Progressively Type-II censored sample of size m form given values of censoring scheme R_i ; $i = 1, 2, \ldots, m$, for considered model, according to an algorithm proposed by Balakrishnan and Aggarwala (2000).
4) The censoring scheme for different values of m is presented in Table 1.
5) The risk ratio of the Bayes estimators are calculated form 1,00,000 generated future ordered samples each of size $n = 20$ of Rayleigh model.
6) For selected values of shape parameter a (= 0.25, 0.50, 1.00, 1.50) of LLF, a risk ratio between the Bayes estimator $\hat{\sigma}_L$ and $\hat{\sigma}_I$ are obtained for considered parametric values and presented in Tables 2-3 under ISELF and LLF respectively.
7) From both tables, note the risk ratios are smaller than unity. This shows that the magnitude of risk with respect to LLF is smaller than the ISELF, when other parameters values considered to be fixed.

8) A decreasing trend has been seen for risk ratio when *c* increases in both cases. Similar behavior also seen when censoring scheme *m* changed.

9) Further, it is noted also that the risk ratios tend to be wider as shape parameter '*a*' increases when other parametric values are consider to be fixed.

10) The magnitude of risk ratio will be wider for ISELF as compared to LLF when other parametric values considered to be fixed.

11) Further, the magnitude of the risk ratio for both case are robust.

The random samples are generated for One-Sample Bayes Prediction Central Coverage bound length. The procedure and results are as follows.

1) A set of 1,00,000 random samples of size $n = 20$ was drawn from the model for similar set of parametric values as consider earlier in step (1) to (5).

2) For the selected values of level of significance $\vartheta = 99\%, 95\%, 90\%$; the central coverage Bayes prediction lengths of bounds were obtained and presented them in Table 4.

3) It is observed from Table 4 that the Central Coverage Bayes prediction bounds lengths under One–Sample plan tend to be wider as *c* increases when other parametric values are fixed (except for $c = 5.00$).

4) The bound length expended also, when progressive censoring plan *m* changed.

5) Note the length of bounds tends to be closer when level of significance ϑ decreases when other parametric values are fixed.

6) The magnitudes of lengths are smaller or nominal. This shows that the central Coverage Bayes prediction criterion is robust.

Table 1. Censoring scheme for different values of m

Case	m	R_i ; $i = 1, 2, \ldots, r$
1	10	1 2 1 0 0 1 2 0 0 0
2	10	1 0 0 3 0 0 1 0 0 1
3	20	1 0 2 0 0 1 0 2 0 0 0 1 0 0 0 1 0 0 1 0

Table 2. Risk ratio between $\hat{\sigma}_L$ and $\hat{\sigma}_I$ under ISELF

$m \downarrow$	$c \downarrow a \rightarrow$	**0.25**	**0.5**	**1**	**1.5**
10	0	0.7765	0.7842	0.7915	0.7988
	0.5	0.7583	0.7659	0.773	0.7802
	1.5	0.7148	0.722	0.7287	0.7354
	2	0.6124	0.6186	0.6243	0.63
	5	0.385	0.3889	0.3925	0.3961
10	0	0.7522	0.7597	0.7668	0.7738
	0.5	0.7346	0.742	0.7488	0.7556
	1.5	0.6924	0.6993	0.7059	0.7123
	2	0.5933	0.5992	0.6049	0.6104
	5	0.373	0.3767	0.3802	0.3837
20	0	0.7288	0.7359	0.7429	0.7496
	0.5	0.7117	0.7187	0.7255	0.7322
	1.5	0.6707	0.6774	0.6838	0.6901
	2	0.5747	0.5803	0.5857	0.5912
	5	0.3613	0.3649	0.3682	0.3717

Table 3. Risk ratio between $\hat{\sigma}_L$ and $\hat{\sigma}_I$ under LLF

$m\downarrow$	$c\downarrow a\rightarrow$	**0.25**	**0.5**	**1**	**1.5**
10	0	0.7741	0.7819	0.7891	0.7964
	0.5	0.7561	0.7636	0.7707	0.7776
	1.5	0.7125	0.7198	0.7265	0.7332
	2	0.6105	0.6166	0.6225	0.6281
	5	0.3838	0.3878	0.3913	0.3948
10	0	0.6748	0.6815	0.6879	0.6941
	0.5	0.659	0.6655	0.6717	0.6778
	1.5	0.6211	0.6273	0.6332	0.6389
	2	0.5321	0.5375	0.5426	0.5475
	5	0.3346	0.3378	0.3411	0.3441
20	0	0.5898	0.5957	0.6013	0.6068
	0.5	0.5759	0.5817	0.5871	0.5926
	1.5	0.5429	0.5483	0.5534	0.5585
	2	0.4651	0.4698	0.4742	0.4785
	5	0.2924	0.2952	0.2981	0.3008

Table 4. One-Sample Central Coverage Bayes Prediction Bound Length

$m\downarrow$	$c\downarrow \vartheta\rightarrow$	**99%**	**95%**	**90%**
10	0	0.4195	0.3246	0.2711
	0.5	0.6243	0.4796	0.4021
	1.5	0.7737	0.5961	0.4988
	2	1.0101	0.7785	0.6516
	5	0.385	0.3839	0.3825
10	0	0.441	0.3409	0.2853
	0.5	0.637	0.4905	0.4115
	1.5	0.7859	0.6062	0.507
	2	1.0193	0.7864	0.6578
	5	0.373	0.3707	0.3682
20	0	0.45	0.3465	0.2901
	0.5	0.6436	0.4958	0.4149
	1.5	0.7899	0.609	0.51
	2	1.0231	0.7885	0.6602
	5	0.3713	0.3699	0.3678

References

Ahmed, A., Ahmad, S. P. & Reshi, J. A. (2013). Bayesian analysis of Rayleigh distribution. *International Journal of Scientific and Research Publications, 3* (10), 1-9.

Al-Kutubi, H. S. & Ibrahim, N. A. (2009). Bayes estimator for exponential distribution with extension of Jeffrey prior information. *Malaysian Journal of Mathematical Science, 3*, 297-313.

Ali-Mousa, M. A. M. & Al-Sagheer, S. A. (2005). Bayesian prediction for progressively type-II censored data from Rayleigh model. *Communication in Statistics ‑Theory and Methods, 34*, 2353-2361. doi:10.1080/03610920500313767

Balakrishnan, N. & Aggarwala, R. (2000). *Progressive Censoring: Theory, Methods and Applications*. Birkhauser Publishers, Boston.

Bhattacharya, S. K. & Tyagi, R. K. (1990). Bayesian survival analysis based on the Rayleigh model. *Trabajos de Estadistica, 5* (1), 81-92. doi:10.1007/BF02863677

Fernandez, A. J. (2000). Bayesian inference from type II doubly censored Rayleigh data. *Statistical Probability Letters, 48*, 393‑399. doi:10.1016/S0167-7152(00)00021-3

Hartigan, J. (1964). Invariant prior distribution. *Annals of Mathematical Statistics, 35*(2), 836-845. doi:10.1214/aoms/1177703583

Hisada, K. & Arizino, I. (2002). Reliability tests for Weibull distribution with varying shape parameter based on complete data. *Reliability, IEEE Transactions on, 51*(3), 331-336. doi:10.1109/TR.2002.801845

Kim, C. & Han, K. (2009). Estimation of the scale parameter of the Rayleigh distribution under general progressive censoring. *Journal of the Korean Statistical Society*, *38*(3), 239-246. doi:10.1016/j.jkss.2008.10.005

Prakash, G. (2013). Bayes estimation in the inverse Rayleigh model. *Electronic Journal of Applied Statistical Analysis, 6*(1), 67-83. doi:10.1285/i20705948v6n1p67

Prakash, G. & Prasad, B. (2010). Bayes prediction intervals for the Rayleigh model. *Model Assisted Statistics and Applications, 5*(1), 43-50. doi:10.3233/MAS-2010-0128

Prakash, G. & Singh, D. C. (2013). Bayes prediction intervals for the Pareto model. *Journal of Probability and Statistical Science*, *11*(1), 109-122.

Sinha, S. K. (1990). On the prediction limits for Rayleigh life distribution. *Calcutta Statistical Association Bulletin, 39*, 105‑109.

Soliman, A., Amin, E. A., & Abd-El Aziz, A. A. (2010). Estimation and prediction from inverse Rayleigh distribution based on lower record values. *Applied Mathematical Sciences, 4*(62), 3057-3066

Wu, S. J., Chen, D. H. & Chen, S. T. (2006). Bayesian inference for Rayleigh distribution under progressive censored sample. *Applied Stochastic Models in Business & Industry, 22*(3), 269‑279. doi:10.1002/asmb.615

Journal of Modern Applied Statistical Methods
November 2015, Vol. 14, No. 2, 123-140.

ISSN 1538 − 9472

Monte Carlo Comparison of the Parameter Estimation Methods for the Two-Parameter Gumbel Distribution

Demet Aydin
Sinop University
Sinop, Turkey

Birdal Şenoğlu
Ankara University
Ankara, Turkey

The performances of the seven different parameter estimation methods for the Gumbel distribution are compared with numerical simulations. Estimation methods used in this study are the method of moments (ME), the method of maximum likelihood (ML), the method of modified maximum likelihood (MML), the method of least squares (LS), the method of weighted least squares (WLS), the method of percentile (PE) and the method of probability weighted moments (PWM). Performance of the estimators is compared with respect to their biases, MSE and deficiency (Def) values via Monte-Carlo simulation. A Monte Carlo Simulation study showed that the method of PWM was the best performance the other methods of bias criterion and the method of ML outperforms the other methods in terms of Def criterion. A real life example taken from the hydrology literature is given at the end of the paper.

Keywords: Gumbel distribution, estimation methods, Monte Carlo simulation, efficiency

Introduction

The Gumbel distribution was first proposed by E. J. Gumbel in 1941. It is a special case of the Generalized Extreme Value (GEV) distribution and is sometimes referred to as Extreme value type I distribution or just the log-Weibull distribution. It is widely used for modeling extreme events, or extreme order statistics. It has two forms, one for "minimum order statistics" and the other for "maximum order statistics." In this study, we focus on the second form.

The Gumbel distribution has many applications in practice, such as annual maximum flow of river, floods, rainfalls, earthquake magnitudes, annual sea-level prediction and so on. It is of considerable importance in many areas of

Demet Aydın is Assistant Professor in the Department of Statistics. Email at demethanaydin@gmail.com. Birdal Şenoğlu is Professor in the Department of Statistics. Email at senoglu@science.ankara.edu.tr.

environmental sciences, e.g., hydrology, see Wallis and Wood (1985). Mathematical modeling of natural phenomena is becoming more and more important in this age of global warming, especially for public safety and economic issues. Therefore, estimating the model parameters precisely and efficiently is very important. There are various different estimation methods in the literature for estimating the parameters of the Gumbel distribution. The method of moments and the method of maximum likelihood (ML) are the most well known among them. There exist various studies in the literature identifying the most efficient method of estimation for the Gumbel distribution via Monte Carlo simulation study, see for example Landwehr et al. (1979) and Mahdi and Cenac (2004).

In the present work, these studies were extended by including four other estimation methods, namely, modified maximum likelihood (MML), least squares (LS), weighted least squares (WLS) and method of percentile. This is the first study comparing these seven different methods of estimation in the same study.

Gumbel distribution

The probability density function (PDF) and the cumulative density function (CDF) of the two-parameter Gumbel distribution with the location parameter μ and the scale parameter σ are defined as follows:

$$f\left(x\right)=\frac{1}{\sigma}\exp\left(-\frac{\left(x-\mu\right)}{\sigma}\right)\exp\left(-\exp\left(-\frac{\left(x-\mu\right)}{\sigma}\right)\right),\ -\infty<x<\infty,\ \mu\in\Re,\ \sigma\in\Re^{+} \tag{1}$$

and

$$F\left(x\right)=\exp\left(-\exp\left(-\frac{\left(x-\mu\right)}{\sigma}\right)\right) \tag{2}$$

respectively.

To understand the basic characteristics of the Gumbel distribution, the mean, the variance, the skewness and the kurtosis values are given as follows:

$$E\left(x\right)=\mu+\sigma\gamma,\ V\left(x\right)=\frac{\pi^{2}}{6}\sigma^{2},\ \beta_{1}:1.14\ \text{and}\ \beta_{2}:5.40 \tag{3}$$

respectively. Here, γ is the Euler's constant, with approximate value 0.5772.

It is seen that Gumbel distribution is positively skewed and moderately long tailed. See Figure 1 for the plot of the Gumbel distribution.

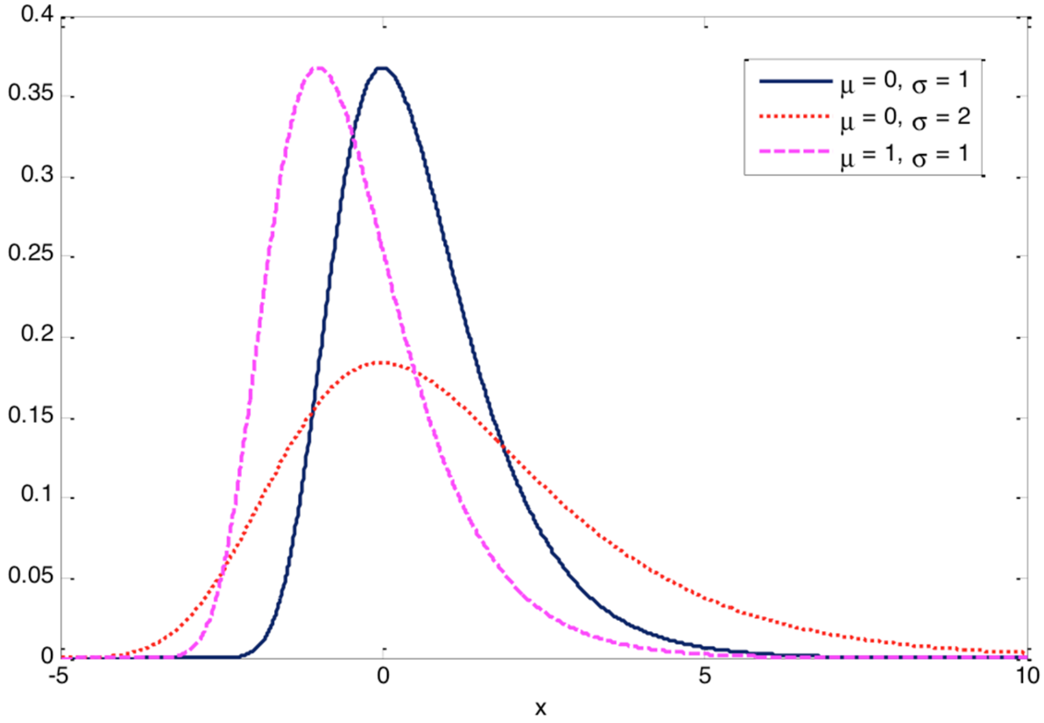


Figure 1. Plot of the Gumbel distribution for various μ and σ values.

The methods of estimation

In this section, we briefly describe the methods of estimation for the Gumbel distribution used in this study.

The method of moments

Moment estimators of the location parameter μ and the scale parameter σ of the Gumbel distribution are found by equating the sample moments to the corresponding theoretical moments.

In other words, they are the solutions of the following equalities

$$\bar{X} = \mu + \gamma\sigma \text{ and } S^2 = \frac{\pi^2}{6}\sigma^2 \tag{4}$$

ME of μ and σ are then obtained as

$$\tilde{\mu}_{ME} = \bar{X} - \gamma\tilde{\sigma}_{ME} \text{ and } \tilde{\sigma}_{ME} = \frac{\sqrt{6}}{\pi}S \tag{5}$$

respectively.

The method of Maximum Likelihood

ML estimators of the two-parameter Gumbel distribution in (1) are found by maximizing the following log-likelihood function with respect to the parameters of interest (i.e., with respect to μ and σ),

$$\ln L = -n\ln\sigma - \sum_{i=1}^{n} z_i - \sum_{i=1}^{n} g\left(z_i\right) \tag{6}$$

where

$$g\left(z_i\right) = \exp\left(-z_i\right),\ z_i = \frac{\left(x_i - \mu\right)}{\sigma}$$

First, we obtain the likelihood functions given below:

$$\frac{\partial \ln L}{\partial \mu} = \sum_{i=1}^{n}\frac{1}{\sigma} - \frac{1}{\sigma}\sum_{i=1}^{n} g\left(z_i\right) = 0 \tag{7}$$

$$\frac{\partial \ln L}{\partial \sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma}\sum_{i=1}^{n} z_i - \frac{1}{\sigma}\sum_{i=1}^{n} z_i g\left(z_i\right) = 0 \tag{8}$$

It is clear that likelihood equations do not have explicit solutions. Therefore, we apply numerical methods to solve the equations (7) and (8). Iterative solutions of these equations are the ML estimates of the location parameter μ and the scale parameter σ.

The method of Modified Maximum Likelihood

MML methodology was first introduced by Tiku (1967, 1968). It is used as an alternative to the well known ML methodology when the estimators of the parameters can not be obtained explicitly. Idea behind the MML methodology is based on the linearization of the nonlinear terms in the likelihood equations.

MML methodology is based on the following steps:

i) Likelihood equations given in (7) and (8) are written in terms of the order statistics, since complete sums are invariant to ordering, i.e., $\sum_{i=1}^{n} z_i = \sum_{i=1}^{n} z_{(i)}$

$$\frac{\partial \ln L}{\partial \mu} = \sum_{i=1}^{n} \frac{1}{\sigma} - \frac{1}{\sigma} \sum_{i=1}^{n} g\left(z_{(i)}\right) = 0 \tag{9}$$

and

$$\frac{\partial \ln L}{\partial \sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma} \sum_{i=1}^{n} z_{(i)} - \frac{1}{\sigma} \sum_{i=1}^{n} g\left(z_{(i)}\right) z_{(i)} = 0 \tag{10}$$

where

$$g\left(z_{(i)}\right) = \exp\left(-z_{(i)}\right) \text{ and } z_{(i)} = \frac{\left(x_{(i)} - \mu\right)}{\sigma},\ i = 1, 2, \ldots, n$$

ii) Linearize the nonlinear term in (9) and (10) by using the first two terms of the Taylor series expansion around the expected values of the order statistics

$$g\left(z_{(i)}\right) \cong g\left(t_{(i)}\right) + \left(z_{(i)} - t_{(i)}\right) \left\{ \frac{d}{dz} g\left(z_{(i)}\right) \right\}_{z_{(i)} = t_{(i)}}$$

or equivalently

$$g\left(z_{(i)}\right) \cong \alpha_i + \beta_i z_{(i)} \tag{11}$$

where

$$\alpha_i = \exp\left(-t_{(i)}\right) + \exp\left(-t_{(i)}\right) t_{(i)} \text{ and } \beta_i = -\exp\left(-t_{(i)}\right)$$

Here, $t_{(i)}$'s (i = 1, 2, …, n) are the expected values of the standardized order statistics $z_{(i)}$, i.e., $t_{(i)} = E(z_{(i)})$, and are obtained from the following equality:

$$\int_{-\infty}^{t_{(i)}} \exp(-z)\exp\left(-\exp(-z)\right) dz = \frac{i}{n+1} \tag{12}$$

Equation (12) gives

$$t_{(i)} = -\ln\left(-\ln\left(\frac{i}{n+1}\right)\right),\ i = 1, 2, \ldots, n$$

iii) By incorporating (11) into (9) and (10), we obtain the modified likelihood equations given below

$$\frac{\partial \ln L}{\partial \mu} \cong \frac{\partial \ln L^*}{\partial \mu} = \sum_{i=1}^{n} \frac{1}{\sigma} - \frac{1}{\sigma}\sum_{i=1}^{n}\left(\alpha_i + \beta_i z_{(i)}\right) = 0 \tag{13}$$

and

$$\frac{\partial \ln L}{\partial \sigma} \cong \frac{\partial \ln L^*}{\partial \sigma} = \frac{n}{\sigma} + \frac{1}{\sigma}\sum_{i=1}^{n} z_{(i)} - \frac{1}{\sigma}\sum_{i=1}^{n}\left(\alpha_i + \beta_i z_{(i)}\right) z_{(i)} = 0 \tag{14}$$

iv) Solutions of the modified likelihood equations in (13) and (14) with respect to the unknown parameters are the following MML estimators

$$\hat{\mu}_{MML} = K_{MML} + L_{MML}\hat{\sigma}_{MML} \text{ and } \hat{\sigma}_{MML} = \frac{-B + \sqrt{B^2 - 4nC}}{2\sqrt{n(n-1)}} \tag{15}$$

where

$$K_{MML}=\frac{1}{m}\sum_{i=1}^{n}\beta_i x_{(i)},\ L_{MML}=\frac{\Delta}{m},=\sum_{i=1}^{n}\Delta_i,\ \Delta_i=\left(\alpha_i-1\right),$$
$$m=\sum_{i=1}^{n}\beta_i,\ B=\sum_{i=1}^{n}\Delta_i\left(x_{(i)}-\hat{\mu}\right)\text{ and }C=\sum_{i=1}^{n}\beta_i\left(x_{(i)}-\hat{\mu}\right)^2$$

MML estimators are asymptotically equivalent to the ML estimators. Therefore, they are asymptotically unbiased and minimum variance bound (MVB) estimators under the regularity conditions. However, in contrast to ML estimators, they are the explicit functions of the sample observations and avoid the computational difficulties encountered in the numerical solutions, such as multiple roots, nonconvergence of iterations or convergence to wrong values, see for example Barnett (1966). It should be noted that MML estimators are nearly unbiased and MVB estimators even for small samples.

The method of Least Squares

Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from the distribution function $F(.)$. LS estimators of the unknown parameters of $F(.)$ are obtained by minimizing the following equation:

$$\sum_{i=1}^{n}\left\{F\left(X_{(i)}\right)-\frac{i}{n+1}\right\}^2 \tag{16}$$

with respect to the parameters of interest. It is known that $X_{(1)} < X_{(2)} < \ldots < X_{(n)}$ are the ordered random variables.

Then the LS estimators of the parameters of the two-parameter Gumbel distribution are obtained by minimizing the function

$$G\left(\mu,\sigma\right)=\sum_{i=1}^{n}\left\{\exp\left(-\exp\left(-z_{(i)}\right)\right)-\frac{i}{n+1}\right\}^2 \tag{17}$$

with respect to the parameters μ and σ.

The method of Weighted Least Squares

Let, $X_1, X_2, \ldots, X_n$ be a random sample of size n from the distribution function $F(.)$ and $X_{(1)} < X_{(2)} < \ldots < X_{(n)}$ be the ordered random variables.

WLS estimators of the unknown parameters are obtained by minimizing the function

$$\sum_{i=1}^{n} w_i \left\{ F\left(X_{(i)}\right) - \frac{i}{n+1} \right\}^2 \tag{18}$$

with respect to the parameters of interest.

In case of the Gumbel distribution, the WLS estimators of the model parameters are obtained by minimizing the following function

$$G(\mu,\sigma) = \sum_{i=1}^{n} w_i \left\{ \exp\left(-\exp\left(-z_{(i)}\right)\right) - \frac{i}{n+1} \right\}^2 \tag{19}$$

with respect to the parameters μ and σ. Here,

$$z_{(i)} = \frac{x_{(i)} - \mu}{\sigma},\ w_i = \frac{1}{Var\left(F\left(z_{(i)}\right)\right)},\ \text{and } Var\left(F\left(z_{(i)}\right)\right) = \frac{(n+1)^2(n+2)}{i(n-i+1)}$$

The method of percentile

Percentile estimators of the unknown parameters of the distribution function $F\left(\dfrac{x_{(i)} - \mu}{\sigma}\right)$ are found by minimizing the equation

$$\sum_{i=1}^{n} \left\{ x_{(i)} - F^{-1}\left(\frac{i}{n+1}\right) \right\}^2 \tag{20}$$

with respect to the unknown parameters. Here, $X_{(i)}$'s are defined as the i^{th} order statistics. For the Gumbel distribution, equation (20) reduces to

$$PE(\mu,\sigma) = \sum_{i=1}^{n} \left\{ x_{(i)} - \mu + \sigma \ln\left(-\ln\left(\frac{i}{n+1}\right)\right) \right\}^2 \tag{21}$$

Solutions of the equation (21) are the following percentile estimators of the location parameter μ and the scale parameter σ

$$\tilde{\mu}_{PE} = K_{PE} + L_{PE}\tilde{\sigma}_{PE} \text{ and } \tilde{\sigma}_{PE} = \frac{\sum_{i=1}^{n} \delta_i \left(\tilde{\mu}_{PE} - x_{(i)} \right)}{\sum_{i=1}^{n} \delta_i^2} \tag{22}$$

where

$$K_{PE} = \frac{1}{n}\sum_{i=1}^{n} x_{(i)},\ L_{PE} = \frac{\delta}{n},\ \delta_i = \ln\left(-\ln\left(\frac{i}{n+1} \right) \right) \text{ and } \delta = \sum_{i=1}^{n} \delta_i$$

The method of Probability Weighted Moments

The method of probability weighted moments has been defined by Greenwood et al. (1979). Similar to the traditional method of moments, parameter estimates are obtained by equating the analytical expressions for PWM to sample estimates.

They defined the PWM as follows

$$M_{i,j,k} = E\left[X^i F\left(X\right)^j \left(1 - F\left(X\right)\right)^k \right] = \int_0^1 [x(F)]^i F^j \left(1-F\right)^k dF,\ i, j, k \in R \tag{23}$$

where $F(X)$ is the cdf of the random variable X and $x(F)$ is the inverse distribution function.

By adopting the convention $M_{1,0,k} = M_{(k)}$, the PWM estimators of μ and σ are obtained as

$$\hat{\mu}_{PWM} = \hat{M}_{(0)} - \gamma\hat{\sigma}_{PWM} \text{ and } \hat{\sigma}_{PWM} = \frac{\hat{M}_{(0)} - 2\hat{M}_{(1)}}{\ln 2} \tag{24}$$

respectively. $\hat{M}_{(k)}$ in (24) is an unbiased estimate of $M_{(k)}$ and is given by

$$\hat{M}_{(k)} = \frac{1}{n}\sum_{i=1}^{n} x_{(i)} \frac{C\left(n-i,k\right)}{C\left(n-1,k\right)} \tag{25}$$

where $x_{(i)}$ are the ordered observations and k is a nonnegative integer. See Landwehr et al. (1979) for more detailed information about the method of PWM.

Methodology 1

Monte Carlo simulation study

An extensive Monte Carlo simulation study was conducted to compare the performance of the different estimators proposed in the previous section. Performances of the different estimators are compared with respect to their biases, MSE and Def values. Def is the natural measure of the joint efficiency of the pair ($\hat{\mu}, \hat{\sigma}$), see Tiku and Akkaya (2004). It is defined as given below.

Definition: Let $\hat{\theta}_1$ and $\hat{\theta}_2$ be the estimators of the parameters θ_1 and θ_2, respectively. Def is a MSE based measure of the joint efficiency of estimators of a set of parameters of a probability distribution. Then, the Def of the estimators $\hat{\theta}_1$ and $\hat{\theta}_2$ is defined as

$$Def\left(\hat{\theta}_1, \hat{\theta}_2\right) = MSE\left(\hat{\theta}_1\right) + MSE\left(\hat{\theta}_2\right) \tag{26}$$

where

$$MSE\left(\hat{\theta}\right) = Var\left(\hat{\theta}\right) + Bias^2\left(\hat{\theta}\right) \text{ and } Bias\left(\hat{\theta}\right) = E\left(\hat{\theta}\right) - \theta.$$

Results 1

The Mean, MSE and Def values of the parameter estimators were computed based on ⟦100000/n⟧ Monte Carlo runs for various sample sizes ranging from 5 to 1000 (i.e., n = 5, 10, 50, 100 and 1000). Here, ⟦.⟧ shows the integer value function. The location parameter μ and the scale parameter σ are taken to be 0 and 1 without loss of generality throughout the study, since all the estimators are invariant under the linear transformations of the data. All the computations were conducted in MATLAB R2010a. Simulation results are presented in Table 1.

Table 1. Simulated Means, Variance, MSE and Def values for the different parameter estimators of *μ* and *σ*; *μ* = 0, *σ* = 1

		μ			***σ***			
n		**Mean**	**Variance**	**MSE**	**Mean**	**Variance**	**MSE**	**Def**
5	*ML*	0.0876	0.2365	0.2441	0.8491	0.1221	0.1449	0.3890
	MML	0.1965	0.2508	0.2894	0.9785	0.1989	0.1994	0.4888
	LS	-0.0127	0.2869	0.2871	1.2366	0.5363	0.5923	0.8794
	WLS	-0.0238	0.5399	0.5404	1.2688	2.0432	2.1155	2.6559
	PE	0.0033	0.2394	0.2395	1.2715	0.3541	0.4278	0.6673
	ME	0.0569	0.2369	0.2401	0.9178	0.1747	0.1815	0.4216
	PWM	0.0057	0.2324	0.2324	1.0066	0.1976	0.1977	0.4301
10	*ML*	0.0358	0.1133	0.1146	0.9197	0.0611	0.0675	0.1821
	MML	0.0957	0.1168	0.1260	0.9741	0.0691	0.0697	0.1957
	LS	-0.0088	0.1205	0.1206	1.1031	0.1234	0.1341	0.2547
	WLS	-0.0170	0.1249	0.1252	1.1259	0.1432	0.1590	0.2842
	PE	-0.0069	0.1140	0.1140	1.1698	0.1395	0.1683	0.2823
	ME	0.0233	0.1154	0.1159	0.9512	0.0920	0.0944	0.2103
	PWM	-0.0031	0.1120	0.1120	0.9970	0.0872	0.0872	0.1992
50	*ML*	0.0097	0.0224	0.0224	0.9839	0.0122	0.0124	0.0349
	MML	0.0229	0.0226	0.0231	0.9915	0.0124	0.0125	0.0356
	LS	0.0011	0.0248	0.0248	1.0195	0.0190	0.0194	0.0442
	WLS	-0.0015	0.0261	0.0261	1.0273	0.0227	0.0235	0.0496
	PE	-0.0035	0.0231	0.0231	1.0617	0.0235	0.0273	0.0504
	ME	0.0084	0.0233	0.0233	0.9884	0.0207	0.0208	0.0442
	PWM	0.0022	0.0224	0.0225	0.9991	0.0164	0.0164	0.0389
100	*ML*	0.0037	0.0110	0.0110	0.9930	0.0060	0.0060	0.0170
	MML	0.0106	0.0110	0.0111	0.9962	0.0061	0.0061	0.0172
	LS	-0.0001	0.0121	0.0121	1.0119	0.0092	0.0093	0.0215
	WLS	-0.0016	0.0128	0.0128	1.0164	0.0111	0.0113	0.0242
	PE	-0.0044	0.0114	0.0114	1.0388	0.0112	0.0127	0.0241
	ME	0.0035	0.0115	0.0115	0.9941	0.0105	0.0105	0.0220
	PWM	0.0001	0.0110	0.0110	0.9999	0.0079	0.0079	0.0190
1000	*ML*	0.0000	0.0011	0.0011	0.9990	0.0006	0.0006	0.0017
	MML	0.0007	0.0011	0.0011	0.9992	0.0006	0.0006	0.0017
	LS	-0.0003	0.0012	0.0012	1.0008	0.0008	0.0008	0.0021
	WLS	-0.0004	0.0013	0.0013	1.0012	0.0011	0.0011	0.0024
	PE	-0.0019	0.0012	0.0012	1.0072	0.0011	0.0011	0.0023
	ME	-0.0002	0.0012	0.0012	0.9995	0.0011	0.0011	0.0022
	PWM	-0.0003	0.0011	0.0011	0.9997	0.0007	0.0007	0.0019

The following conclusions are drawn from the results of the Monte Carlo simulation study.

i) According to the bias comparisons of the estimators:

As far as the location parameter μ is concerned, MML did not perform well especially for small n values ($n = 5$ and 10). PE and PWM estimators show the best performance among the others, since they are more or less unbiased even for small sample sizes. It is observed in Table 1 that biases of the different estimators considered in this study decrease as the sample size n increases.

If our concern is the scale parameter σ, all the scale estimators (except PWM and MML) have substantial bias in cases where a small number of data samples ($n = 5$ and 10) are available. For these sample sizes, LS, WLS and PE overestimate σ while ML and ME underestimate. PWM shows the best performance and followed by the MML estimator for all the sample sizes. Similar to the comments made about the location estimators, bias of the scale estimators decreases as the sample size n increases.

ii) According to the efficiency comparisons of the estimators:

Simulation results show that the method of ML outperforms the other methods for estimating the location parameter μ in all cases except $n = 5$ and 10. For these sample sizes, the method of PWM shows the best performance among the other methods with the smallest MSE.

For estimating the scale parameter σ, it is observed that ML works the best for all sample sizes.

It should be noted that there is not much difference in the performances between ML and MML estimators especially for moderate ($n = 50$ and 100) and large ($n = 1000$) sample sizes as mentioned in the section on MML.

iii) According to the joint efficiency (Def) comparisons of the estimators:

It is clear from the simulation results presented in Table 1 that the method of ML provides the smallest Def values in all cases, therefore it is the best method for jointly estimating the location parameter μ and the scale parameter σ of the Gumbel distribution. Second best performance is shown by the method of MML for all values of n except $n = 5$. For $n = 5$, ME is the second most efficient method of the seven. Third place (in terms of the joint efficiency) was taken by the method of PWM.

Note that the simulation results presented in this study are in accordance with those of the Landwehr et al. (1979) who compared the methods of PWM, ME and ML.

Methodology 2

Asymptotic variances

In this part, obtain the exact variances of the ML estimators as

$$V\left(\hat{\mu}\right) \cong \frac{6}{\pi^2}\frac{\sigma^2}{n}\left(\frac{\pi^2}{6}+\left(1-\gamma\right)^2\right) \text{ and } V\left(\hat{\sigma}\right) \cong \frac{6}{\pi^2}\frac{\sigma^2}{n}$$

by using the diagonal elements of I^{-1} (where $I=\left[I_{ij}\right]_{i,j=1,2}$ is the Fisher information matrix), see Panjer (2006). These variances are also known as the Rao-Cramer Lower Bounds (RCLBs) for the parameters μ and σ. Elements of the symmetric matrix **I** are given by

$$I_{11} = -E\left(\frac{\partial^2 \ln L}{\partial \mu^2}\right) = \frac{n}{\sigma^2}$$

$$I_{12} = -E\left(\frac{\partial^2 \ln L}{\partial \mu \partial \sigma}\right) = -\frac{n}{\sigma^2}\left(1-\gamma\right)$$

$$I_{22} = -E\left(\frac{\partial^2 \ln L}{\partial \sigma^2}\right) = \frac{n}{\sigma^2}\left(\frac{\pi^2}{6}+\left(1-\gamma\right)^2\right)$$

Results 2

Table 2 shows that the RCLBs for the parameters and for various different sample sizes.

Table 2. RCLBs for the parameters *μ* and *σ*

n	$V(\hat{\mu})$	$V(\hat{\sigma})$
5	0.2217	0.1215
10	0.1108	0.0607
50	0.0221	0.0121
100	0.0110	0.0060
1000	0.0011	0.0006

It is seen that simulated variances of the ML estimators given in Table 1 are very close to the RCLBs even for small sample sizes. This is another indication of the fact that the ML estimators show the best performance for estimating the parameters of the Gumbel distribution.

A real life example

Meriç (Maritsa or Evros) is the longest river of the Balkan Peninsula and the second longest river of in South-Eastern Europe. Its length is 530 km with a catchments area of more than 53,000 square kilometers, see Sezen et al. (2007). It is a highly industrialized, highly agricultural and highly populated area with approximately 2 million inhabitants. The Meriç River basin is distributed over the territories of three countries, namely, Bulgaria (66%), Turkey (28%) and Greece (6%). The Meriç River has four main tributaries known as Ardas (Bulgaria and Greece), Tundzha (Bulgaria and Turkey), Erythropotamos (mostly in Greece) and Ergene (in Turkey), see Skiyas and Kallioras (2007).

The main reason for analyzing the data belonging to the Meriç River is its high risk of flooding. It is known that one or two flooding events have occurred annually during the last decade. They have caused severe economic, socioeconomic and environmental impacts, see Skiyas and Kallioras (2007).

The maximum daily flood discharge (annual) is measured in cubic meters per second (m^3/s) for the Meriç River at Turkey, recorded during the period 1982-2006. These measurements have been taken from the Kirişhane station, Edirne (Turkey), see Sezen et al. (2007).

Discharge is defined as the volume of the water flowing through a specified point of a stream in a given interval of time. Therefore, especially in flood periods, identifying the distributional characteristics (such as mean and variance) of the maximum daily discharge data is extremely important for flood control, water resources planning, design of hydraulic structures, management and decision making (Chen & Chiu, 2004).

The aim is to fit a distribution to the maximum daily discharge (annual) data by using the Methods of Estimation described. To have an idea about the underlying distribution of the data, we use the Kolmogorov-Simirnov (KS) test. According to the KS test, we do not reject the null hypothesis

H_0: Distribution of the maximum daily discharge (annual) data is Gumbel since $KS_{cal} = 1.1349 < KS_{tab} = 0.2376$.

For the maximum daily discharge (annual) data, estimates of the parameters of the Gumbel distribution are obtained as reported in Table 3.

Table 3. Parameter estimates of the Gumbel distribution for the Meriç River during 1982-2006.

Estimator	$\hat{\mu}$	$\hat{\sigma}$
ML	539.8018	302.2066
MML	545.5504	303.8097
LS	504.1084	314.3558
WLS	497.1617	323.3498
PE	509.0342	430.4036
ME	509.4286	395.1687
PWM	527.4405	363.9631

See Figure 2 for the plots of the fitted densities based on these estimate values. It can be seen from the figure that the fitted densities based on the ML and the MML estimates provide better fit than the fitted densities based on the other estimates for the Meriç River data.

Conclusion

Seven estimation methods for estimating the parameters of the two-parameter Gumbel distribution were compared. Performance of the estimators is compared with respect to their biases, MSE and Def values.

Comparing all the seven methods, it is clear that as far as bias is concerned, the method of PWM outperforms the other methods for all sample sizes. It can also be seen from the simulation results that all the estimators of the location parameter μ and the scale parameter σ are asymptotically unbiased. In terms of the joint efficiency, the method of ML works the best for all sample sizes. However,

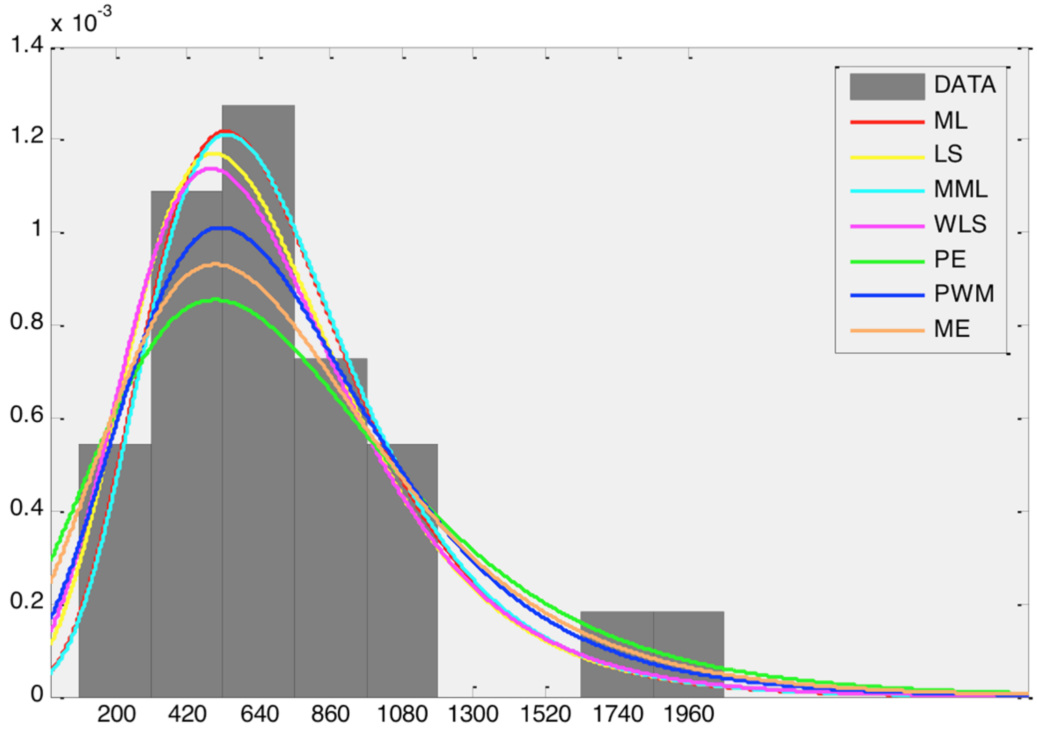


Figure 2. Histogram of the maximum daily flood discharges (annual) for the Meriç River data and the fitted densities.

the Def values of the MML estimators $(\hat{\mu}, \hat{\sigma})$ are quite close to that of ML estimators especially for moderate and large sample sizes as expected. As far as computation is concerned, MML estimators are easy to compute and do not have the computational complexities of ML estimators. Therefore, their computation takes very little CPU time, see Kantar and Şenoğlu (2008). If our consideration is both efficiency and the CPU time, then we recommend to use the MML estimators for estimating the pair (μ, σ) for moderate and large sample sizes.

References

Barnett, V. D. (1966). Order statistics estimators of the location of the Cauchy Distribution. *Journal of the American Statistical Association*, *61*(316), 1205–1218. Correction in *63*, 383–385. doi:10.1080/01621459.1966.10482205

Chen, Y. C., & Chiu C. L. (2004). A fast method of flood discharge estimation. *Hydrological Processes*, *18*(9), 1671–1684. doi:10.1002/hyp.1476

Greenwood, J. A., Landwehr, J. M., Matalas, N. C., & Wallis, J. R. (1979). Probability weighted moments: Definition and relation to parameters of several distributions expressible in inverse form. *Water Resources Research, 15*(5), 1049–1054.

Gumbel, E. J. (1941). The return period of flood flows. *The Annals of Mathematical Statistics*, *12*(2), 163–190.

Kantar, Y. M., & Şenoğlu, B. (2008). A comparative study for the location and scale parameters of the Weibull distribution with given shape parameter. *Computers and Geosciences, 34*, 1900–1909. doıː10.1016/j.cageo.2008.04.004

Landwehr, J. M., Matalas, N. C., & Wallis J. R. (1979). Probability weighted moments compared with some traditional techniques in estimating Gumbel parameters and quantiles. *Water Resources Research*, *15*(5), 1055–1064.

Mahdi, S., & Cenac, M. (2005). Estimating parameters of Gumbel distribution using the methods of moments, probability weighted moments and maximum likelihood. *Revista de Matemática: Teoría y Aplicaciones*, *12*(1-2), 151–156.

Panjer, H. H. (2006). *Operational risk modeling analytics*. Hoboken, NJ: John Wiley and Sons.

Sezen, N., Gündüz, N., & Malkaralı, S. (2007). *Meriç River Floods And Turkish–Bulgarian Cooperations*. Paper presented at the International Congress on River Basin Management, Antalya, Turkey.

Skiyas, S., & Kallioras, A. (2007). Cross border co-operation and the problem of flooding in the Evros delta. In J. Verwijmeren & M. Wiering (Eds), *Many rivers to cross: cross border co-operation in river management* (pp. 119–144). Delft, The Netherlands: Eburon Academic Publishers.

Tiku, M. L. (1967). Estimating the mean and standard deviation from censored normal samples. *Biometrika*, *54*(1-2), 155–165. doi:10.1093/biomet/54.1-2.155

Tiku, M. L. (1968). Estimating the parameters of log-normal distribution from censored samples. *Journal of the American Statistical Association*, *63*(321), 134–140. doi:10.1080/01621459.1968.11009228

Tiku, M. L., & Akkaya, A. D. (2004). *Robust estimation and hypothesis testing*. New Delhi: New Age International (P) Ltd. Publishers.

Wallis, J. R., & Wood, E. F. (1985). Relative accuracy of log Pearson III procedures. *Journal of Hydraulic Engineering, 111*(7), 1043-1056. doi:10.1061/(ASCE)0733-9429(1985)111:7(1043)

Journal of Modern Applied Statistical Methods
November 2015, Vol. 14, No. 2, 141-158.

ISSN 1538 − 9472

Structural Properties of Transmuted Weibull Distribution

Kaisar Ahmad
University of Kashmir
Srinagar, India

S. P. Ahmad
University of Kashmir
Srinagar, India

A. Ahmed
Aligarh Muslim University
Aligarh, India

The transmuted Weibull distribution, and a related special case, is introduced. Estimates of parameters are obtained by using a new method of moments.

Keywords: Transmuted Weibull distribution, moment generating function, sample coefficient of variation, Standard deviation, Skewness and kurtosis

Introduction

The Weibull distribution was introduced by the Swedish Physicist Waloddi Weibull in 1939. He applied this distribution to analyze the breaking strength of materials. This distribution has been extensively used in lifetime and reliability problem. The Weibull family is a generalization of the exponential family and can model data exhibiting monotone hazard rate behavior, i.e., it can accommodate three types of failure rates, namely increasing, decreasing and constant. Its application in connection with lifetimes of many types of manufactured items has been widely advocated (e.g., Weibull, 1951; Berrettoni, 1964), and it has been used as a model with diverse types of items such as vacuum tubes (Kao, 1959), ball bearings (Lieblein & Zelen, 1956), and electrical insulation. It is also widely used in biomedical applications.

A simple explanation of the Weibull distribution and its applications can be found in Franck (1988). A comprehensive review of this model is available in Johnson, Kotz, and Balakrishnan (1995). A generalization of the Weibull distribution with application to the analysis of survival data is given by Mudholkar, Srivastava, and Kollia (1996). Inferences from grouped data in the three-parameter Weibull models is introduced by Hirose and Lai (1997). Lawless

K. Ahmad and S. P. Ahmad are in the Department of Statistics. Email at ahmadkaisar31@gmail.com and sprvz@yahoo.com, respectively. A. Ahmed is in the Department of Statistics and O.R.

(2002) provided statistical models and methods for lifetime data. Al-Athari (2011) and Hossain and Zimmer (2003) did some comparative studies on the estimation of Weibull parameters using complete and censored samples. Nadarajah and Kotz (2005) presented a procedure on some recent modifications of Weibull distribution.

For deriving new moment estimators of three parameters transmuted Weibull distribution, a similar approach to that of Huang and Hwang (2006) was used. Nadarajah and Kotz (2005) discussed products and ratios of Weibull random variables. Gokarna and Tsokos (2009) proposed a method on the transmuted extreme value distribution with application. Ahmad and Ahmad (2013) presented a procedure of Bayesian analysis of Weibull distribution.

A random variable x is said to have a Weibull distribution with parameters $\alpha > 0$ and $\beta > 0$ if its pdf is given by

$$g(x) = \frac{\beta}{\alpha} x^{\beta-1} \exp\left(-\frac{x^{\beta}}{\alpha}\right) \quad x \geq 0, \alpha > 0, \beta > 0$$

The cdf of Weibull distribution is given by

$$G(x) = \int_0^x g(x)\,dx$$

$$G(x) = \int_0^x \frac{\beta}{\alpha} x^{\beta-1} \exp\left(-\frac{x^{\beta}}{\alpha}\right) dx$$

$$\Rightarrow \qquad G(x) = 1 - \exp\left(-\frac{x^{\beta}}{\alpha}\right) \tag{1}$$

Transmuted Weibull distribution

In order to obtain the pdf of transmuted Weibull distribution, use the following cdf which is given by

$$F(x) = (1+\lambda)G(x) - \lambda G(x)^2 \tag{2}$$

where $G(x)$ is the cdf of base distribution. If $\lambda = 0$, we have the distribution of base random variable.

Now using equation (1) in equation (2),

$$F(x) = (1+\lambda)\left(1-\exp\left(-\frac{x^{\beta}}{\alpha}\right)\right) - \lambda\left(1-\exp\left(-\frac{x^{\beta}}{\alpha}\right)\right)^{2}$$

$$\Rightarrow \qquad F(x) = (1+\lambda)k - \lambda k^{2}$$

where

$$k = 1-\exp\left(-\frac{x^{\beta}}{\alpha}\right)$$

$$\Rightarrow \qquad F(x) = k(1+\lambda-\lambda k)$$

$$\Rightarrow \qquad F(x) = k\{1+\lambda(1-k)\}$$

$$\Rightarrow \qquad F(x) = \left\{1-\exp\left(-\frac{x^{\beta}}{\alpha}\right)\right\}\left\{1+\lambda\exp\left(-\frac{x^{\beta}}{\alpha}\right)\right\} \tag{3}$$

This is the required cdf of Transmuted Weibull distribution.

In order to find the pdf of Transmuted Weibull distribution, first differentiate equation (3) w.r.t. x which is given by

$$f(x) = \frac{d}{dx}\{F(x)\}$$

$$\Rightarrow \qquad f(x) = \frac{d}{dx}\left[\left\{1-\exp\left(-\frac{x^{\beta}}{\alpha}\right)\right\}\left\{1+\lambda\exp\left(-\frac{x^{\beta}}{\alpha}\right)\right\}\right]$$

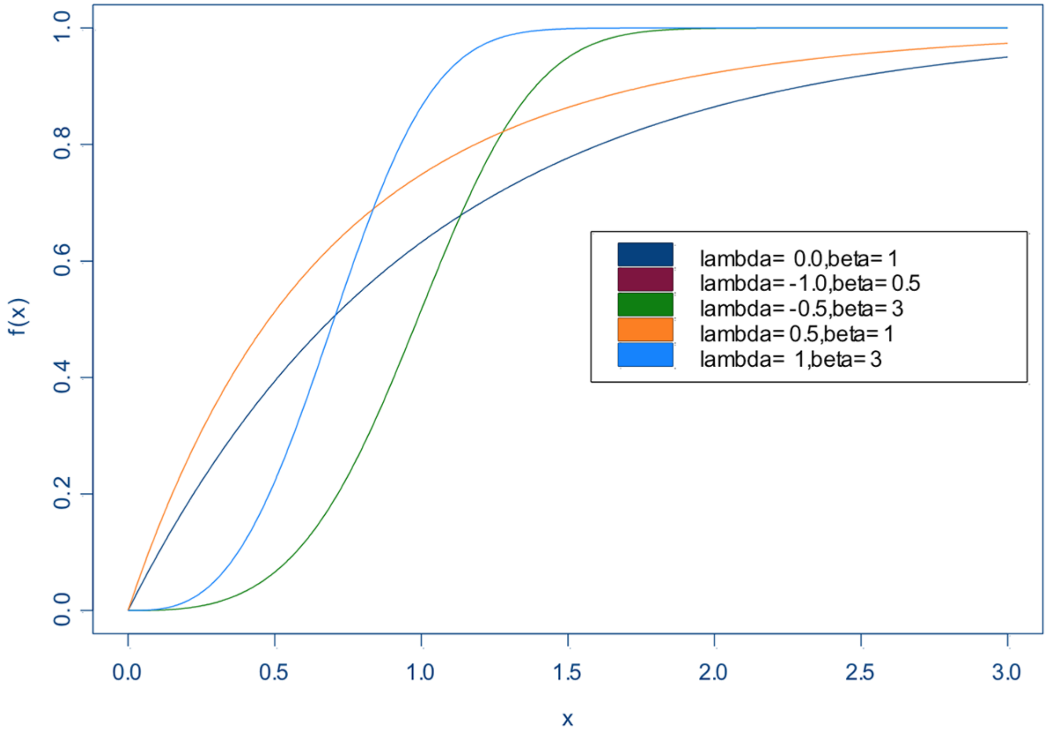


Figure 1. The cdfs of various transmuted Weibull distributions.

After differentiating the above equation w.r.t. *x*,

$$f(x)=\frac{\beta}{\alpha}x^{\beta-1}\exp\left(-\frac{x^{\beta}}{\alpha}\right)\left\{1-\lambda+2\lambda\exp\left(-\frac{x^{\beta}}{\alpha}\right)\right\} \tag{4}$$

which is the required pdf of Transmuted Weibull distribution with parameters α, β and λ.

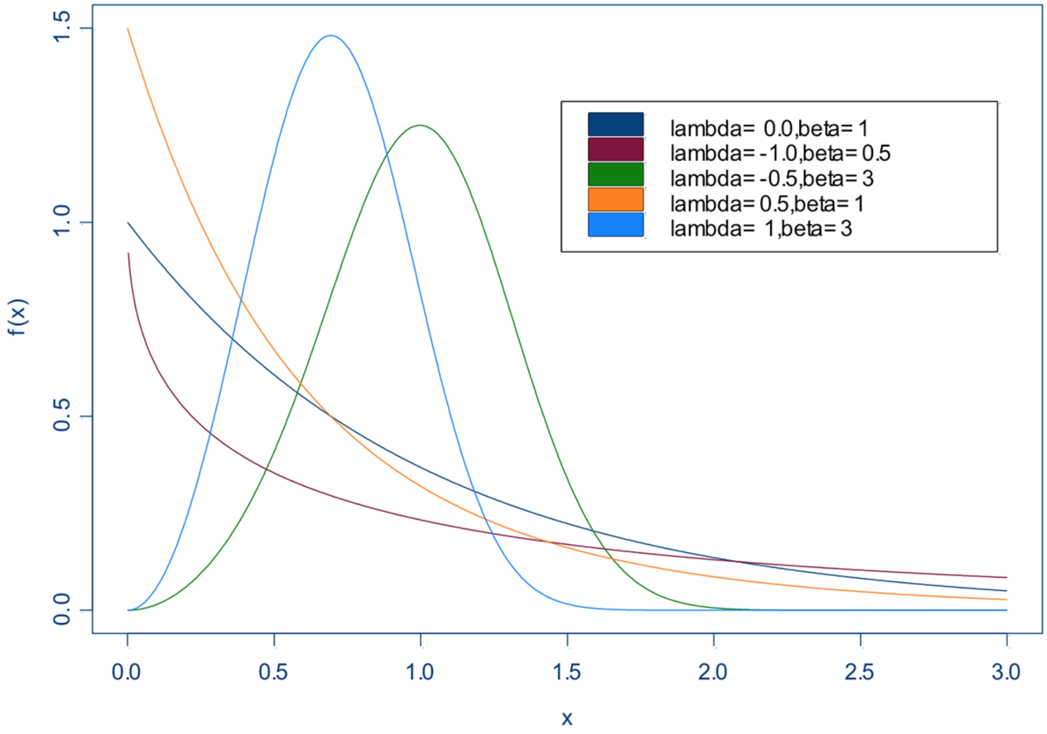


Figure 2. The pdfs of various Transmuted Weibull distributions.

Special cases

1) If $\lambda = 0$, then Transmuted Weibull distribution reduced to two parameter Weibull distribution with parameters α and β.

$$f(x;\alpha,\beta) = \frac{\beta}{\alpha} x^{\beta-1} \exp\left(-\frac{x^{\beta}}{\alpha}\right) \quad x \geq 0 \quad \alpha,\beta > 0$$

2) If $\lambda = 0$ and $\beta = 1$, then Transmuted Weibull distribution reduced to exponential distribution with parameter $\left(\frac{1}{\alpha}\right)$, i.e.

$$f(x) = \frac{1}{\alpha} \exp\left(-\frac{x}{\alpha}\right) \quad x > 0, \alpha > 0$$

3) If $\lambda = 0$ and $\alpha = \beta = 1$, then Transmuted Weibull distribution reduced to standard exponential distribution, i.e.

$$f(x) = \exp(-x) \;\; x > 0$$

Moments of Transmuted Weibull distribution

Moments are the expected values of certain functions of a random variable. They serve to numerically describe the variable with respect to given characteristics for location, variation, skewness and kurtosis, to name a few. The expected value of x^r is termed as r^{th} moment about origin of the random variable x which is given by

$$\mu_r' = E(x)^r$$

Thus the r^{th} moment of Transmuted Weibull distribution is given by

$$\mu_r' = \int_0^\infty x^r f(x;\alpha,\beta,\lambda)\,dx$$

$$\mu_r' = \int_0^\infty x^r \frac{\beta}{\alpha} x^{\beta-1} \exp\left(-\frac{x^\beta}{\alpha}\right)\left\{1-\lambda+2\lambda\exp\left(-\frac{x^\beta}{\alpha}\right)\right\}dx$$

After solving the above equation,

$$\mu_r' = \alpha^{\frac{r}{\beta}}\Gamma\left(\frac{r}{\beta}+1\right)\left(1-\lambda+\lambda 2^{-\frac{r}{\beta}}\right) \tag{5}$$

Mean of the Transmuted Weibull distribution

Setting $r = 1$ in equation (5) leads to the mean of the Transmuted Weibull distribution, which is given by

$$\mu_1' = \alpha^{\frac{1}{\beta}}\Gamma\left(\frac{1}{\beta}+1\right)\left(1-\lambda+\lambda 2^{\frac{-1}{\beta}}\right) \tag{6}$$

Second moment of the Transmuted Weibull distribution

Setting r = 2 in equation (5),

$$\mu_2' = \alpha^{\frac{2}{\beta}}\Gamma\left(\frac{2}{\beta}+1\right)\left(1-\lambda+\lambda 2^{\frac{-2}{\beta}}\right) \tag{7}$$

Variance of Transmuted Weibull distribution

The variance of Transmuted Weibull distribution is given by

$$\mu_2 = \alpha^{\frac{2}{\beta}}\left\{\Gamma\left(\frac{2}{\beta}+1\right)\left(1-\lambda+\lambda 2^{\frac{-2}{\beta}}\right)-\Gamma^2\left(\frac{1}{\beta}+1\right)\left(1-\lambda+\lambda 2^{\frac{-1}{\beta}}\right)^2\right\} \tag{8}$$

Third and fourth moments of Transmuted Weibull distribution

Setting r = 3 in equation (5),

$$\mu_3' = \alpha^{\frac{3}{\beta}}\Gamma\left(\frac{3}{\beta}+1\right)\left(1-\lambda+\lambda 2^{\frac{-3}{\beta}}\right)$$

and

$$\mu_3 = \alpha^{\frac{3}{\beta}}\left[\begin{array}{l}\Gamma\left(\frac{3}{\beta}+1\right)\left(1-\lambda+\lambda 2^{\frac{-3}{\beta}}\right)\\ -\Gamma\left(\frac{1}{\beta}+1\right)\left(1-\lambda+\lambda 2^{\frac{-1}{\beta}}\right)\left\{\begin{array}{l}3\Gamma\left(\frac{2}{\beta}+1\right)\left(1-\lambda+\lambda 2^{\frac{-2}{\beta}}\right)\\ -2\Gamma^2\left(\frac{1}{\beta}+1\right)\left(1-\lambda+\lambda 2^{\frac{-1}{\beta}}\right)^2\end{array}\right\}\end{array}\right] \tag{9}$$

If r = 4 in equation (5),

$$\mu_4' = \alpha^{\frac{4}{\beta}}\Gamma\left(\frac{4}{\beta}+1\right)\left(1-\lambda+\lambda 2^{\frac{-4}{\beta}}\right)$$

thus

$$\mu_4 = \alpha^{\frac{4}{\beta}} \left[\begin{array}{l} \Gamma\left(\frac{4}{\beta}+1\right)\left(1-\lambda+\lambda 2^{\frac{-4}{\beta}}\right) \\ -\Gamma\left(\frac{1}{\beta}+1\right)\left(1-\lambda+\lambda 2^{\frac{-1}{\beta}}\right) \left\{ \begin{array}{l} 4\Gamma\left(\frac{3}{\beta}+1\right)\left(1-\lambda+\lambda 2^{\frac{-3}{\beta}}\right) \\ -6\Gamma^2\left(\frac{1}{\beta}+1\right)\left(1-\lambda+\lambda 2^{\frac{-1}{\beta}}\right) \\ \Gamma\left(\frac{2}{\beta}+1\right)\left(1-\lambda+\lambda 2^{\frac{-2}{\beta}}\right) \\ +3\Gamma^3\left(\frac{1}{\beta}+1\right)\left(1-\lambda+\lambda 2^{\frac{-1}{\beta}}\right)^3 \end{array} \right\} \end{array} \right] \tag{10}$$

MGF of Transmuted Weibull distribution

The mgf of Transmuted Weibull distribution is given by

$$M_x(t) = \int_0^\infty e^{tx} f(x)\,dx$$

$$M_x(t) = \int_0^\infty \left\{1 + tx + \frac{(tx)^2}{2!} + \frac{(tx)^3}{3!} + \cdots + \frac{(tx)^n}{n!} + \cdots\right\} f(x)\,dx$$

$$M_x(t) = \int_0^\infty \sum_{r=0}^\infty \frac{t^r x^r}{r!} f(x)\,dx$$

$$M_x(t) = \sum_{r=0}^\infty \frac{t^r}{r!} \int_0^\infty x^r f(x)\,dx$$

$$M_x(t) = \sum_{r=0}^\infty \frac{t^r}{r!} \mu_r'$$

Now by using the equation (5) in the above equation, we have

$$M_x(t)=\sum_{r=0}^{\infty}\frac{t^r}{r!}\alpha^{\frac{r}{\beta}}\Gamma\left(\frac{r}{\beta}+1\right)\left(1-\lambda+\lambda 2^{-\frac{r}{\beta}}\right) \tag{11}$$

This is the required mgf of Transmuted Weibull distribution.

Standard deviation of Transmuted Weibull distribution

The positive square root of the variance is called standard deviation. Symbolically, $\sigma = \sqrt{\sigma^2}$. From equation (8), the variance of Transmuted Weibull distribution is given as

$$\sigma^2=\alpha^{\frac{2}{\beta}}\left\{\Gamma\left(\frac{2}{\beta}+1\right)\left(1-\lambda+\lambda 2^{\frac{-2}{\beta}}\right)-\Gamma^2\left(\frac{1}{\beta}+1\right)\left(1-\lambda+\lambda 2^{\frac{-1}{\beta}}\right)^2\right\}$$

$$\Rightarrow\ \sigma=\alpha^{\frac{1}{\beta}}\sqrt{\left\{\Gamma\left(\frac{2}{\beta}+1\right)\left(1-\lambda+\lambda 2^{\frac{-2}{\beta}}\right)-\Gamma^2\left(\frac{1}{\beta}+1\right)\left(1-\lambda+\lambda 2^{\frac{-1}{\beta}}\right)^2\right\}}$$

$$\Rightarrow \qquad \sigma=\alpha^{\frac{1}{\beta}}\sqrt{\sigma_2-\sigma_1^2}$$

where

$$\sigma_k=\Gamma\left(\frac{k}{\beta}+1\right)\left(1-\lambda+\lambda 2^{\frac{-k}{\beta}}\right) \tag{12}$$

Coefficient of variation of Transmuted Weibull distribution

This is the ratio of standard deviation and mean. Usually, it is denoted by C.V. and is given by

$$C.V.=\frac{\sigma}{\mu}$$

$$\Rightarrow\ C.V.=\frac{\alpha^{\frac{1}{\beta}}\sqrt{\left\{\Gamma\left(\frac{2}{\beta}+1\right)\left(1-\lambda+\lambda 2^{\frac{-2}{\beta}}\right)-\Gamma^{2}\left(\frac{1}{\beta}+1\right)\left(1-\lambda+\lambda 2^{\frac{-1}{\beta}}\right)^{2}\right\}}}{\alpha^{\frac{1}{\beta}}\Gamma\left(\frac{1}{\beta}+1\right)\left(1-\lambda+\lambda 2^{\frac{-1}{\beta}}\right)}$$

$$\Rightarrow\ C.V.=\frac{\sqrt{\sigma_2-\sigma_1^2}}{\sigma_1} \tag{13}$$

where $\sigma_k=\Gamma\left(\frac{k}{\beta}+1\right)\left(1-\lambda+\lambda 2^{\frac{-1}{\beta}}\right)$

Skewness and kurtosis of Transmuted Weibull distribution

The most popular way to measure the skewness and kurtosis of a distribution function rests upon ratios of moments. Lack of symmetry of tails (about mean) of frequency distribution curve is known as skewness. The formula for measure of skewness given by Karl Pearson in terms of moments of frequency distribution is given by

$$\beta_1=\frac{\mu_3^2}{\mu_2^3}$$

After using equation (8) and equation (9) in the above equation, we have

$$\beta_1=\frac{\left[\begin{array}{l}\Gamma\left(\frac{3}{\beta}+1\right)\left(1-\lambda+\lambda 2^{\frac{-3}{\beta}}\right)\\ -\Gamma\left(\frac{1}{\beta}+1\right)\left(1-\lambda+\lambda 2^{\frac{-1}{\beta}}\right)\left\{\begin{array}{l}3\Gamma\left(\frac{2}{\beta}+1\right)\left(1-\lambda+\lambda 2^{\frac{-2}{\beta}}\right)\\ -2\Gamma^{2}\left(\frac{1}{\beta}+1\right)\left(1-\lambda+\lambda 2^{\frac{-1}{\beta}}\right)^{2}\end{array}\right\}\end{array}\right]^{2}}{\left[\Gamma\left(\frac{2}{\beta}+1\right)\left(1-\lambda+\lambda 2^{\frac{-2}{\beta}}\right)-\Gamma^{2}\left(\frac{1}{\beta}+1\right)\left(1-\lambda+\lambda 2^{\frac{-1}{\beta}}\right)^{2}\right]^{3}}$$

$$\Rightarrow \ \beta_1 = \frac{\left\{\sigma_3 - \sigma_1\left(3\sigma_2 - \sigma_1^2\right)\right\}^2}{\left(\sigma_2 - \sigma_1^2\right)^3}$$

where

$$\sigma_k = \Gamma\left(\frac{k}{\beta} + 1\right)\left(1 - \lambda + \lambda 2^{\frac{-1}{\beta}}\right)$$

Therefore

$$\gamma_1 = \sqrt{\beta_1}$$

$$\Rightarrow \ \gamma_1 = \frac{\left\{\sigma_3 - \sigma_1\left(3\sigma_2 - \sigma_1^2\right)\right\}}{\left(\sigma_2 - \sigma_1^2\right)^{\frac{3}{2}}}$$

If $\gamma_1 < 0$, then the frequency curve is negatively skewed. If $\gamma_1 > 0$, then the frequency curve is positively skewed.

Kurtosis

The formula for measure of kurtosis is given by

$$\beta_2 = \frac{\mu_4}{\mu_2^2}$$

After using equation (8) and equation (10) in the above equation,

$$\beta_2 = \frac{\begin{array}{l} \left[\Gamma\left(\frac{4}{\beta}+1\right)\left(1-\lambda+\lambda 2^{\frac{-4}{\beta}}\right)\right. \\ -\Gamma\left(\frac{1}{\beta}+1\right)\left(1-\lambda+\lambda 2^{\frac{-1}{\beta}}\right)\left\{4\Gamma\left(\frac{3}{\beta}+1\right)\left(1-\lambda+\lambda 2^{\frac{-3}{\beta}}\right)\right. \\ \quad -6\Gamma\left(\frac{1}{\beta}+1\right)\left(1-\lambda+\lambda 2^{\frac{-1}{\beta}}\right)\Gamma\left(\frac{2}{\beta}+1\right)\left(1-\lambda+\lambda 2^{\frac{-2}{\beta}}\right) \\ \quad \left. +3\Gamma^3\left(\frac{1}{\beta}+1\right)\left(1-\lambda+\lambda 2^{\frac{-1}{\beta}}\right)^3 \right\} \end{array}}{\left[\Gamma\left(\frac{2}{\beta}+1\right)\left(1-\lambda+\lambda 2^{\frac{-2}{\beta}}\right)-\Gamma^2\left(\frac{1}{\beta}+1\right)\left(1-\lambda+\lambda 2^{\frac{-1}{\beta}}\right)^2\right]^2}$$

$$\Rightarrow \quad \beta_2 = \frac{\left\{\sigma_4-\sigma_1\left(4\sigma_3-6\sigma_1\sigma_2+3\sigma_1\right)\right\}^3}{\left(\sigma_2-\sigma_1\right)^2}$$

where

$$\sigma_k = \Gamma\left(\frac{k}{\beta}+1\right)\left(1-\lambda+\lambda 2^{\frac{-1}{\beta}}\right)$$

and

$$\gamma_2 = \beta_2 - 3$$

$$\Rightarrow \quad \gamma_2 = \frac{\left\{\sigma_4-\sigma_1\left(4\sigma_3-6\sigma_1\sigma_2+3\sigma_1\right)\right\}^3}{\left(\sigma_2-\sigma_1\right)^2} - 3$$

If $\gamma_2 > 0$, then the frequency curve is leptokurtic. If $\gamma_2 < 0$, then the frequency curve is platykurtic. If $\gamma_2 = 0$, then the frequency curve is mesokurtic, or we can say that there is no kurtosis.

Harmonic mean of Transmuted Weibull distribution

$$\frac{1}{H}=\int_0^\infty f\left(x;\alpha,\beta,\lambda\right)dx$$

$$\frac{1}{H}=\int_0^\infty \frac{1}{x}\frac{\beta}{\alpha}x^{\beta-1}\exp\left(-\frac{x^\beta}{\alpha}\right)\left\{1-\lambda+2\lambda\exp\left(-\frac{x^\beta}{\alpha}\right)\right\}dx$$

$$\frac{1}{H}=\left(1-\lambda\right)\int_0^\infty \frac{1}{x}\frac{\beta}{\alpha}x^{\beta-1}\exp\left(-\frac{x^\beta}{\alpha}\right)dx+2\lambda\int_0^\infty \frac{1}{x}\frac{\beta}{\alpha}x^{\beta-1}\exp\left(-\frac{2x^\beta}{\alpha}\right)dx$$

After substitution,

$$\frac{1}{H}=\frac{\left(1-\lambda\right)}{\alpha}\int_0^\infty \frac{1}{z^{\frac{1}{\beta}}}\exp\left(-\frac{z}{\alpha}\right)dz+2\frac{\lambda}{\alpha}\int_0^\infty \frac{1}{z^{\frac{1}{\beta}}}\exp\left(-\frac{2z}{\alpha}\right)dz$$

After solving the above equation

$$\frac{1}{H}=\alpha^{\frac{1}{\beta}}\Gamma\left(1-\frac{1}{\beta}\right)\left(1-\lambda+\lambda 2^{\frac{-1}{\beta}}\right)$$

$$\Rightarrow\ H=\frac{1}{\alpha^{\frac{1}{\beta}}\Gamma\left(1-\frac{1}{\beta}\right)\left(1-\lambda+\lambda 2^{\frac{-1}{\beta}}\right)} \qquad (14)$$

New moment estimator of the Transmuted Weibull distribution

For deriving new moment estimators of three parameters transmuted Weibull distribution, we need the following theorem obtained by using the similar approach of Huang and Hwang (2006).

Theorem 1. Let $n\geq 3$ and let $X_1, X_2, X_3, \ldots, X_n$ be n positive identical independently random variables having probability density function $f(x)$. Then the independence of the sample mean $\bar{X}_n$ and the sample coefficient of variance

$V_n = \dfrac{S_n}{\bar{X}_n}$ is equivalent to that $f(x)$ is a Transmuted Weibull density where S_n is the sample standard deviation.

The next theorem requires the derivation of the expectation and the variance of $V_n^{\,2} = \left(\dfrac{S_n}{\bar{X}_n}\right)^2$, where $\bar{X}_n$ and S_n are respectively the sample mean and the sample standard deviation.

Theorem 2. Let $X_1, X_2, X_3, \ldots, X_n$ be n positive identical independently distributed random samples drawn from a population having Transmuted Weibull density

$$f\left(x;\alpha,\beta,\lambda\right) = \frac{\beta}{\alpha} x^{\beta-1} \exp\left(-\frac{x^{\beta}}{\alpha}\right)\left\{1-\lambda+2\lambda\exp\left(-\frac{x^{\beta}}{\alpha}\right)\right\}$$

then

$$E\left(S_n^2\right) = \alpha^{\frac{2}{\beta}}\left\{\Gamma\left(\frac{2}{\beta}+1\right)\left(1-\lambda+\lambda 2^{\frac{-2}{\beta}}\right) - \Gamma^2\left(\frac{1}{\beta}+1\right)\left(1-\lambda+\lambda 2^{\frac{-1}{\beta}}\right)^2\right\}$$

Proof: Because the r^{th} moment of a random variable x about origin is given by

$$\mu_r' = \int_0^{\infty} x^r f\left(x;\alpha,\beta,\lambda\right) dx$$

$$\mu_r' = \int_0^{\infty} x^r \frac{\beta}{\alpha} x^{\beta-1} \exp\left(-\frac{x^{\beta}}{\alpha}\right)\left\{1-\lambda+2\lambda\exp\left(-\frac{x^{\beta}}{\alpha}\right)\right\} dx$$

After solving the above equation,

$$\mu_r' = \alpha^{\frac{r}{\beta}}\Gamma\left(\frac{r}{\beta}+1\right)\left(1-\lambda+\lambda 2^{-\frac{r}{\beta}}\right)$$

If $r = 1$ in the above equation,

$$E\left(\bar{X}_n\right) = \alpha^{\frac{1}{\beta}}\Gamma\left(\frac{1}{\beta}+1\right)\left(1-\lambda+\lambda 2^{-\frac{1}{\beta}}\right)$$

Also if $r = 2$ in the above equation,

$$E\left(\bar{X}_n^2\right) = \frac{\alpha^{\frac{2}{\beta}}\left[\begin{array}{l}\Gamma\left(\frac{2}{\beta}+1\right)\left(1-\lambda+\lambda 2^{-\frac{2}{\beta}}\right)\\ +\left(n-1\right)\Gamma^2\left(\frac{1}{\beta}+1\right)\left(1-\lambda+\lambda 2^{\frac{-1}{\beta}}\right)^2\end{array}\right]}{n} \tag{15}$$

and $$V\left(\bar{X}_n\right) = \frac{\alpha^{\frac{2}{\beta}}\left\{\begin{array}{l}\Gamma\left(\frac{2}{\beta}+1\right)\left(1-\lambda+\lambda 2^{\frac{-2}{\beta}}\right)\\ -\Gamma^2\left(\frac{1}{\beta}+1\right)\left(1-\lambda+\lambda 2^{\frac{-1}{\beta}}\right)^2\end{array}\right\}}{n}\ .$$

Thus

$$E\left(S_n^2\right) = nV\left(\bar{X}_n\right)$$

$$E\left(S_n^2\right) = \alpha^{\frac{2}{\beta}}\left\{\begin{array}{l}\Gamma\left(\frac{2}{\beta}+1\right)\left(1-\lambda+\lambda 2^{\frac{-2}{\beta}}\right)\\ -\Gamma^2\left(\frac{1}{\beta}+1\right)\left(1-\lambda+\lambda 2^{\frac{-1}{\beta}}\right)^2\end{array}\right\} \tag{16}$$

where $\bar{X}_n$ and S_n^2 are respectively the sample mean and the sample variance.

Theorem 3. Let $X_1, X_2, X_3, \ldots, X_n$ be n positive identical independently distributed random samples drawn from a population having Transmuted Weibull density

$$f\left(x;\alpha,\beta,\lambda\right)=\frac{\beta}{\alpha}x^{\beta-1}\exp\left(-\frac{x^{\beta}}{\alpha}\right)\left\{1-\lambda+2\lambda\exp\left(-\frac{x^{\beta}}{\alpha}\right)\right\}$$

then

$$E\left(\frac{S_n^2}{\bar{X}_n^2}\right)=\frac{n\begin{Bmatrix}\Gamma\left(\frac{2}{\beta}+1\right)\left(1-\lambda+\lambda 2^{\frac{-2}{\beta}}\right)\\ -\Gamma^2\left(\frac{1}{\beta}+1\right)\left(1-\lambda+\lambda 2^{\frac{-1}{\beta}}\right)^2\end{Bmatrix}}{\begin{bmatrix}\Gamma\left(\frac{2}{\beta}+1\right)\left(1-\lambda+\lambda 2^{\frac{-2}{\beta}}\right)\\ +\left(n-1\right)\Gamma^2\left(\frac{1}{\beta}+1\right)\left(1-\lambda+\lambda 2^{\frac{-1}{\beta}}\right)^2\end{bmatrix}}$$

where $\bar{X}_n$ and S_n^2 are respectively the sample mean and the sample variance.

Proof: By using the theorem (1), we have

$$E\left(S_n^2\right)=E\left(\frac{S_n^2}{\bar{X}_n^2}\bar{X}_n^2\right)=E\left(\frac{S_n^2}{\bar{X}_n^2}\right)E\left(\bar{X}_n^2\right)$$

$$\Rightarrow \qquad E\left(\frac{S_n^2}{\bar{X}_n^2}\right)=\frac{E\left(S_n^2\right)}{E\left(\bar{X}_n^2\right)} \tag{17}$$

Now using equations (15) and (16) in equation (17), we have

$$E\left(\frac{S_n^2}{\bar{X}_n^2}\right)=\frac{n\left\{\begin{array}{l}\Gamma\left(\frac{2}{\beta}+1\right)\left(1-\lambda+\lambda 2^{\frac{-2}{\beta}}\right)\\ -\Gamma^2\left(\frac{1}{\beta}+1\right)\left(1-\lambda+\lambda 2^{\frac{-1}{\beta}}\right)^2\end{array}\right\}}{\left[\begin{array}{l}\Gamma\left(\frac{2}{\beta}+1\right)\left(1-\lambda+\lambda 2^{\frac{-2}{\beta}}\right)\\ +(n-1)\Gamma^2\left(\frac{1}{\beta}+1\right)\left(1-\lambda+\lambda 2^{\frac{-1}{\beta}}\right)^2\end{array}\right]}$$

$$E\left(\frac{S_n^2}{\bar{X}_n^2}\right)\to\frac{\left\{\Gamma\left(\frac{2}{\beta}+1\right)\left(1-\lambda+\lambda 2^{\frac{-2}{\beta}}\right)\right\}}{\left[\Gamma^2\left(\frac{1}{\beta}+1\right)\left(1-\lambda+\lambda 2^{\frac{-1}{\beta}}\right)^2\right]}-1$$

as $n\to\infty$ and that this limit is the square of the population coefficient of variation. Thus, $\frac{S_n^2}{\bar{X}_n^2}$ is an asymptotically unbiased estimator of the square of the population coefficient of variation.

References

Ahmad, S. P. & Ahmad, K. (2013). Bayesian analysis of Weibull distribution using R software. *Australian Journal of Basic and Applied Sciences, 7*(9), 156-164. http://ajbasweb.com/old/ajbas/2013/July/156-164.pdf

Al-Athari, F. M. (2011). Parameter estimation for the double-Pareto distribution. *Journal of Mathematics and Statistics*, *7*(4), 289–294. doi:10.3844/jmssp.2011.289.294

Berrettoni, J. N. (1964). Practical applications of the Weibull distribution, *Industrial Quality Control*, *21*(2), 71-79.

Franck, J. R. (1988). A simple explanation of the Weibull distribution and its applications. *Reliability Review, 8*(3), 93-116.

Gokarna R. A. & Tsokos, C. P. (2009). On the transmuted extreme value distribution with application. *Nonlinear Analysis: Theory, Methods and Applications, 71*(12), e1401–e1407. doi:10.1016/j.na.2009.01.168

Hirose, H. & Lai, T. L. (1997). Inference from grouped data in three-parameter Weibull models with applications to breakdown-voltage experiments, *Technometrics, 39*(2), 199-210. doi:10.1080/00401706.1997.10485085

Hossain, A. & Zimmer, W. (2003). Comparison of estimation methods for Weibull parameters: complete and censored samples. *Journal of Statistical Computation and Simulation, 73*(2), 145–153. doi:10.1080/0094965021000033486

Huang, P. H. & Hwang, T. Y. (2006). On new moment estimation of parameters of the generalized gamma distribution using it's characterization. [sic] *Taiwanese Journal of Mathematics, 10*(4), 1083 -1093.

Johnson, N. L., Kotz, S., & Balakrishnan, N. (1995). *Continuous univariate distributions* (Vol. 2). New York: John Wiley & Sons.

Kao, J. H. K. (1959). A graphical estimation of mixed Weibull parameters in life testing of electron tubes. *Technometrics*, *1*(4), 389-407. doi:10.1080/00401706.1959.10489870

Lawless, J. F. (2002). *Statistical models and methods for lifetime data* (2nd Ed.). New York: John Wiley & Sons.

Lieblein, J. & Zelen, M. (1956). Statistical investigation of the fatigue life of deep groove ball bearing. *Journal of Research of the National Bureau of Standards, 57*(5), 273-316. doi:10.6028/jres.057.033

Mudholkar, G. S., Srivastava, D. K. & Kollia, G. D. (1996). A generalization of the Weibull distribution with application to the analysis of survival data. *Journal of the American Statistical Association, 91*(436), 1575-1583. doi:10.1080/01621459.1996.10476725

Nadarajah, S. & Kotz, S. (2005). On some recent modifications of Weibull distribution, *IEEE Transactions on Reliability, 54*(4), 561-562. doi:10.1109/TR.2005.858811

Weibull, W. (1939). A statistical theory of strength of materials. *Ingeniörs Vetenskapsakademien Handlingar, 151*, 1-45.

Weibull, W. (1951). A statistical distribution function of wide applicability. *Journal of Applied Mechanics, 18*, 293-297.

Journal of Modern Applied Statistical Methods
November 2015, Vol. 14, No. 2, 159-171.

ISSN 1538 − 9472

A Robust Panel Unit Root Test in the Presence of Cross Sectional Dependence

Nurul Sima Mohamad Shariff
Universiti Sains Islam Malaysia
Negeri Sembilan, Malaysia

Nor Aishah Hamzah
University of Malaya
Kuala Lumpur, Malaysia

Problems arise in testing the stationarity of the panel in the presence of cross sectional dependence and outliers. The currently available panel unit root tests are very much affected by the presence of outliers. As such, this article introduces an alternative test which is robust to outliers and cross sectional dependence. The performance and robustness of the proposed test is discussed and comparisons are made to the existing tests via simulation studies.

Keywords: Cross sectional dependence, outliers, unit root, robust test, panel model.

Introduction

The investigation of the stationary in panel data has received great attention in panel analysis for the past few decades. It is an important issue in modeling the panel with the involvement of times series dimension in this study. This investigation can be done via unit root test. The panel unit root tests can be found in Im et al. (2003), Levin and Lin (1992, 1993), Levin et al. (2002), Bai and Ng (2004), Philips and Sul (2003), Moon and Perron (2004), Pesaran (2007) and Choi (2001, 2002). Hurlin (2010) distinguished two generations of unit root tests on which the first generation tests relied on the assumption that all cross sectional units are independent. The first generation of unit root tests were those proposed by Quah (1994), Breitung and Meyer (1994) and Levin and Lin (1992, 1993).

For the second generation of panel unit root tests, the presence of cross sectional dependence (hereafter CD) among the residuals is allowed within the panel. The assumption of CD is due to the evidence obtained on the strong co-movements among the economic variables (Barbieri, 2009). The assumption that the individual time series in the panel are cross sectional independent is not practical in the context of cross country regressions. As argued by O'Connell,

Dr. Shariff is a Lecturer with the Faculty of Science of Technology. Email her at nurulsima@usim.edu.my.

(1998), the presence of such CD may affect the finite sample behaviour of the panel unit root test which subsequently results to the incorrect decision in a unit root test. Those who proposed the tests which incorporated the CD were: Pesaran (2007), Philips and Sul (2003), Bai and Ng (2004), Moon and Perron (2004) and Choi (2002).

The existence of outliers implies that some shocks will only have temporary effects and thus, providing that they are sufficiently large or sufficiently frequent indicated that the series is stationary (Franses & Haldrup, 1994). Martin and Yohai (1986) showed via the simulation experiment that an additive outliers biases Ordinarily Least Squares (OLS) estimator downward for the parameter in a stationary first order autoregressive process. Hence, in some situations it could be expected that the additive outliers will establish the wrong impression that a time series is stationary when it is actually non-stationary. In addition, the presence of a cross sectional dependence may deteriorate the asymptotic distribution of the standard unit root test which is normally distributed (Philips & Sul, 2003; Banerjee, 1999). Due to such interest, a robust unit root test in the panel data model is proposed which aims at reducing the effects of outliers in the presence of the CD. Specifically, the presence of the unit root will be tested when both the CD and outliers exist in the panel. The finite sample behaviour of the proposed test is studied and its performance is evaluated through the Monte Carlo simulation study.

Model and Tests

Pesaran Unit Root Test

Specifically, in the presence of CD, the following model was considered by Pesaran (2007) to test the presence of the unit root:

$$\Delta y_{it} = \alpha_i + b_i y_{it-1} + \gamma_i f_t + \varepsilon_{it}; \quad i = 1, 2, \ldots, N. \quad t = 1, 2, \ldots, T \tag{1}$$

where $\Delta y_{it} = y_{it} - y_{it-1}$; y_{it} is an i^{th} observation observed at a particular time t, α_i is the intercept, and b_i is a parameter for the variable of y_{it-1}. The presence of CD is represented by $\gamma_i f_t$ where f_t is the latent factor and γ_i is factor loadings that is common across cross sectional units i and ε_{it} is the random error. This model can be employed for a larger and complicated set of time series. In the absence of the unit root, negative values for b_i are expected. Specifically, the hypothesis test for a unit root is defined as follows:

$$\begin{aligned} &H_0: \; b_i = 0; \quad \text{for all } i = 1, 2, ..., N \\ &H_1: \; b_i < 0; \quad \text{for some } i = 1, 2, ..., N \end{aligned} \tag{2}$$

Rejecting the null explains that the panel is stationary (no unit root). Model (1) can be expressed as cross sectional Augmented Dickey-Fuller (CADF) model:

$$\Delta y_{it} = \alpha_i + b_i y_{it-1} + c_i \bar{y}_{t-1} + d_i \Delta \bar{y}_t + e_{it}; \; i = 1, 2, \ldots, N. \;\; t = 1, 2, \ldots, T \tag{3}$$

where the standard of Augmented Dickey-Fuller (ADF) model is improved up to more variables in independent variables in model (3), that are; cross section averages of lagged levels ($\bar{y}_{t-1}$) and first differences of the individual series ($\Delta \bar{y}_t$), *i* in the model. Pesaran has shown that the effect of CD can be eliminated by using model (3). Thus, let CADF$_i$ be the ADF statistics for the i^{th} cross sectional unit given by the t-ratio of the OLS estimate $\hat{b}_i$ of b_i in the CADF regression (3). Then, the Pesaran unit root test is given by

$$\text{CIPS} = \frac{\sum_{i=1}^{N} \text{CADF}_i}{N} \tag{4}$$

where CIPS stands for cross sectional augmented IPS (Im et al. (1997) unit root test) . This CADF$_i$ is given by

$$\text{CADF}_i = t_i(N, T) = \frac{\left(\mathbf{y}_{i,-1}^T \bar{\mathbf{M}} \mathbf{y}_{i,-1}\right)^{-1} \left(\mathbf{y}_{i,-1}^T \bar{\mathbf{M}} \Delta \mathbf{y}_i\right)}{\sqrt{\sigma_i^2 \left(\mathbf{y}_{i,-1}^T \bar{\mathbf{M}} \mathbf{y}_{i,-1}\right)^{-1}}} \tag{5}$$

where $\mathbf{y}_{i,-1} = \left(y_{i1}, \ldots, y_{iT-1}\right)^T$, $\Delta \mathbf{y}_i = \left(\Delta y_{i2}, \Delta y_{i3}, \ldots, \Delta y_{iT}\right)^T$; $\sigma_i^2 = \frac{\sum_{t=1}^{T} \hat{e}_{it}^2}{T-4}$,with $\hat{e}_{it} = \Delta y_{it} - \Delta \hat{y}_{it}$ and $\bar{\mathbf{M}}$ is defined as $\bar{\mathbf{M}} = \mathbf{I}_t - \bar{\mathbf{H}}\left(\bar{\mathbf{H}}^T \bar{\mathbf{H}}\right)^{-1} \bar{\mathbf{H}}^T$ and $\bar{\mathbf{H}} = (\mathbf{1}, \Delta \bar{\mathbf{y}}_t, \bar{\mathbf{y}}_{t-1})$. $\mathbf{I}_t$ is a unit matrix of order $T \times T$ and $\bar{\mathbf{H}}$ is the combination of the dummy variables, average of cross section of the first difference of y_{it} and its first lagged value y_{it-1}. The asymptotic distribution of this distribution is more skewed compared to the ADF (asymptotically normal) distribution in the presence of CD (Philips and Sul, 2003). The critical value of the test statistics in (5) is given in Table 1 and those are obtained from the simulation experiment based on the CADF model.

Proposed Unit Root Test

The Pesaran's unit root test uses the OLS procedure that is non-robust. It has been known in the literature that the OLS is sensitive to the influence of outliers in the data. Hence, to limit the influence of outliers in the data in investigating the presence of the unit root in the model, the Generalized M-estimator is applied and it is obtained by solving the following equation:

$$\sum_{t=1}^{T} u_i\left(y_{it-1}\right) v_i\left(y_{it-1}\right) \psi_i\left(\frac{\hat{e}_{it}\left(b_i\right)}{\hat{\sigma}_i v_i\left(y_{it-1}\right)}\right) y_{it-1} = 0; \quad \text{for } i = 1, 2, \ldots, N \tag{6}$$

where $u_i\left(y_{it-1}\right) = 1$ and $v_i\left(y_{it-1}\right) = \frac{1}{d\left(y_{it-1}\right)}$. The $d\left(y_{it-1}\right)$ is given as a measure of the outlying the $y_{it\text{-}1}$ in the X-space from its mean value. Here, ψ_i (.) is the derivative of ρ_i (.), where ρ_i (.) is a differential convex function (with minimum at 0) and is known as the robustifying criterion function while $\hat{e}_{it}\left(b_i\right)$ is the estimated residuals and $\hat{\sigma}_i$ is the robust scale obtained from the first iteration of M-estimation.

To test for a unit root, a similar hypothesis statement as in (2) is considered. Under H_0 of no unit root, the generalization of the test is given by:

$$t_i^* = \frac{\hat{b}_i^* - b_i}{\sqrt{Var(\hat{b}_i^*)}} \tag{7}$$

where $\hat{b}_i^*$ is the Generalized M-estimator where it is computed as follows:

$$\hat{b}_i^* = \left(\mathbf{y}_{i,-1}^T \mathbf{G}_i \mathbf{y}_{i,-1}\right)^{-1} \left(\mathbf{y}_{i,-1}^T \mathbf{G}_i \Delta \mathbf{y}_i\right) \tag{8}$$

where $\mathbf{y}_{i,-1} = \left(y_{i1}, \ldots, y_{iT-1}\right)^T$, $\Delta\mathbf{y}_i = \left(\Delta y_{i2}, \Delta y_{i3}, \ldots, \Delta y_{iT}\right)^T$ and $\mathbf{G}_i = \bar{\mathbf{M}}^* \mathbf{W}_i\left(z_{it}\right)$ with $z_{it} = d_i\left(y_{it-1}\right)\frac{\hat{e}_{it}\left(b_i\right)}{\hat{\sigma}_i}$. The $Var(\hat{b}_i^*)$ is given by

$$Var(\hat{b}_i^*) = \left(\mathbf{y}_{i,-1}^T \bar{\mathbf{M}}^* \mathbf{y}_{i,-1}\right)^{-1} \hat{\sigma}_i^2 \frac{E\left(\psi_i\left(\frac{\hat{e}_{it}\left(b_i\right)}{\hat{\sigma}_i v_i\left(y_{it-1}\right)}\right)\right)^2}{\left(E\left(\psi_i'\left(\frac{\hat{e}_{it}\left(b_i\right)}{\hat{\sigma}_i v_i\left(y_{it-1}\right)}\right)\right)\right)^2}. \tag{9}$$

where $E\left(\psi_i\left(.\right)\right)$ and $E\left(\psi_i'\left(.\right)\right)$ are the expected values of robustifying criterion function $\psi_i\left(.\right)$ and derivative of $\psi_i\left(.\right)$, respectively. The $\bar{\mathbf{M}}^*$ is computed as $\bar{\mathbf{M}}^* = \mathbf{I}_t - \bar{\mathbf{H}}^*\left(\bar{\mathbf{H}}^{*T}\bar{\mathbf{H}}^*\right)^{-1}\bar{\mathbf{H}}^{*T}$; $\mathbf{I}_t$ is an identity T by T matrix and $\bar{\mathbf{H}}^* = \left(\mathbf{1}, \psi\left(\bar{y}_{t-1}\right), \psi\left(\Delta\bar{y}_t\right)\right)$. The value of $\psi\left(.\right)$ in $\bar{\mathbf{H}}^*$ takes the form

$$\psi\left(\bar{y}_{t-1}\right) = \begin{cases} \bar{y}_{t-1} & \text{,if } \left|\bar{y}_{t-1}\right| \le c \\ sign\left(\bar{y}_{t-1}\right) \times \left|\underset{i}{\text{median}}\left(y_{1t-1}, \ldots, y_{Nt-1}\right)\right| & \text{,elsewhere} \end{cases}$$

and (10)

$$\psi\left(\Delta\bar{y}_t\right) = \begin{cases} \Delta\bar{y}_t & \text{,if } \left|\Delta\bar{y}_t\right| \le d \\ sign\left(\Delta\bar{y}_t\right) \times \left|\underset{i}{\text{median}}\left(\Delta y_{1t}, \ldots, \Delta y_{Nt}\right)\right| & \text{,elsewhere} \end{cases}$$

where c and d are the critical values and computed as $3\hat{\sigma}_{\bar{y}_{t-1}}$ and $3\hat{\sigma}_{\Delta\bar{y}_t}$, respectively. The $\hat{\sigma}_{\bar{y}_{t-1}}$ and $\hat{\sigma}_{\Delta\bar{y}_t}$ are robust scale with $\hat{\sigma}_{\bar{y}_{t-1}} = 1.4825 \ \underset{t}{\text{median}}\left|\bar{y}_{t-1} - \underset{t}{\text{median}}\left(\bar{y}_{t-1}\right)\right|$, $\hat{\sigma}_{\Delta\bar{y}_t} = 1.4825 \ \underset{t}{\text{median}}\left|\Delta\bar{y}_t - \underset{t}{\text{median}}\left(\Delta\bar{y}_t\right)\right|$, respectively. These robust scales are chosen to achieve specified level of efficiency and are called as the Median Absolute Deviation (MAD) with the tuning constant 1.4825 where $\hat{\sigma}_{\bar{y}_{t-1}}$ and $\hat{\sigma}_{\Delta\bar{y}_t}$ are consistent for σ at the normal distribution.

The proposed unit root test is the average of t_i^* which is given by

$$\text{RCIPS} = \overline{t}_i^* = \frac{\sum_{i=1}^{N} t_i^*}{N} \tag{11}$$

where t_i^* is given in (7).

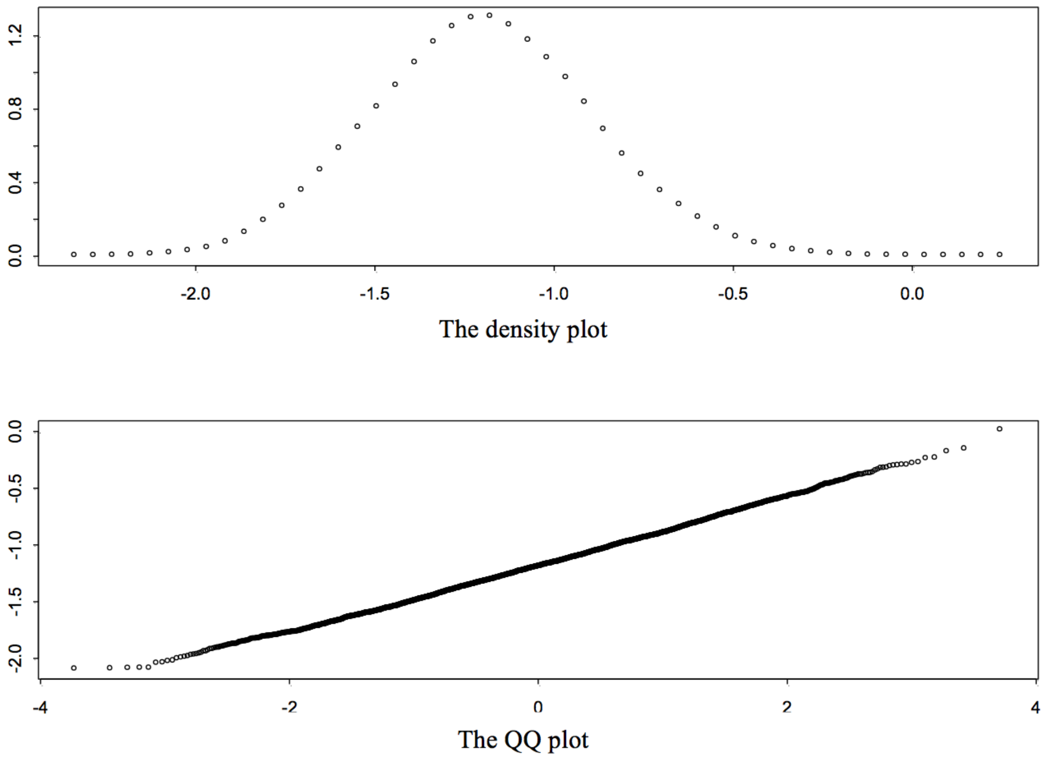
The asymptotic distribution of the test statistics given in (7) is obtained through the extensive simulation experiment. Based on Figure 1, the RCIPS unit root test tends to have an approximate t-distribution with a mean μ and a standard deviation, σ. As the sample size increase, it is believed that the RCIPS will approach to a standard normal distribution. This result is comparable with Pesaran (2007) under conditions where e_{it} is normally distributed.

To investigate the performance of the RCIPS, the critical region of test statistics is required. Therefore, the critical region of RCIPS test is obtained through simulation experiment at the 0.05 level of significance and it is given in Table 2. The data generating process (DGP) and results are given in the next section.

Finite Sample Behavior of the Tests

Following Pesaran (2007), the following DGP is considered: $y_{it} = \mu_i(1-\varphi_i) + \varphi_i y_{it-1} + e_{it}$; $e_{it} = \gamma_i^T f_t + \varepsilon_{it}$; $\mu_i \sim iidN(0,1)$; $\varepsilon_{it} \sim iidN(0,\sigma_i^2)$; $\sigma_i^2 \sim iidU\left[0.5,1.5\right]$. The presence of CD is characterized by the latent factor $f_t \sim iidN(0, 1)$ and strong CD, $\gamma_i \sim iidU\,(0.5, 1.5)$. The performance of the tests is measured by setting: 1) $\varphi_i = 1$ and 2) $\varphi_i \sim U\,[0.75, 0.95]$ for computing the size (incorrect detection) and power (correct detection) of the test, respectively.

A panel contaminated by outliers is represented by $y^*_{it} = y_{it} + \xi\left(L\right)\omega I_{it}\left(\tau\right)$ for $i = 1, 2, \ldots, N$. $t = 1, 2, \ldots, T$, where y^*_{it} is the observed contaminated series, y_{it} is the uncontaminated series, $\xi\,(L)$ is the dynamic pattern of the outliers, ω is the magnitude of outliers. $I_{it}\,(\tau)$ is the indicator function of the presence of outlier and will takes the value of 1 at time $t = \tau$ (chosen at random) and 0 elsewhere.



The density plot

The QQ plot

Figure 1. The Density and QQ plots of t-statistics (RCIPS unit root test)

***Note.** Figure 1 provides results of the test statistic of the proposed unit root test (RCIPS) which is based on 5,000 runs for a sample size (*N*, *T*) = (200, 200). Based on this figure, the RCIPS tends to have a approximate t-distribution with mean μ and a standard deviation σ.

Two types of outliers are considered in this study; additive outliers (AO) and temporary change (TC). The AO only affect the level but leave the variance unaffected. The TC will produce an abrupt step and dies out gradually in time. Hence, in y^*_{it} , $\xi(L)=1$ in the presence of AO and TC takes the form of $\xi(L)=\dfrac{1}{1-\delta L}$ where δ represents the velocity of the dynamic effect and is bounded by [0, 1] (Tsay, 1998). The performance of the tests is investigated at the 5% level of significance using the sample sizes $N = (20, 30, 50)$ and $T = (20, 30, 50, 100, 200)$ with 1,000 replications.

Results and Discussion

The size and power of the unit root tests are investigated for the uncontaminated panel, the panel with AO and the panel with TC. These are tabulated in Tables 3 to 4 for the size and power of the tests, respectively. The results of the tests are reported by rows: 1) CIPS and 2) RCIPS with three columns of the number of cross sectional units, $N = (20, 30, 50)$. For each column of $N = (20, 30, 50)$, results of the size and power of the unit roots tests are reported when the panel is 1) uncontaminated, 2) contaminated with AO, and 3) contaminated with TC.

In the uncontaminated panel, the CIPS unit root test gives a smaller size for a small sample but attains a reasonable size as T increases whereas the RCIPS is slightly oversized even when N and T are large. In the presence of the AO and TC, the sizes for the CIPS test are all zeros for all sample sizes. The RCIPS has smaller size in the presence of AO but achieves a good size of the test in the presence of TC compared to CIPS. These results are comparable when the panel is free from the outliers effect (see column of “no cont” of Table 3).

In investigating the power of the test in the uncontaminated panel, the CIPS gives slightly lower correct detection (power) of a unit root for $T \leq 50$. The probability of correctly detect the presence of unit root however increasing (good power) as T increases and the result is comparable to those obtained in Pesaran (2007). The RCIPS outperforms the CIPS even for small sample. In the presence of the AO and TC the panel, the powers for the CIPS test are poor when $T \leq 50$. The power however increases for $T \geq 100$ with an increasing N. The powers for both tests are good as N increases in the presence of TC in the panel. The RCIPS provides a sensible power when $T \leq 30$ in the presence of the AO but outperforms the CIPS in the presence of the TC. Based on these results, the RCIPS provides a good reasonable size (close to 0.05) and power (greater than 0.95) in the presence of the AO and TC relative to CIPS especially when N and T are small.

Conclusion

An alternative approach to Pesaran unit root test is proposed in order to investigate the stationarity of the data when outliers occur in the panel. The proposed test is robust to the effect of spurious observation in data. The finite sample behaviour of the tests is studied and compared via the Monte Carlo experiments. The results show that the proposed unit root test provide comparable size and power of the test in uncontaminated panel and yield better results than Pesaran unit root test in the presence of outliers in the panel especially for the small pair of sample size.

Table 1. Critical Values of CIPS

N	20			30			50		
Level of significance / T	1%	5%	10%	1%	5%	10%	1%	5%	10%
20	-2.40	-2.21	-2.10	-2.32	-2.15	-2.07	-2.25	-2.11	-2.03
30	-2.38	-2.20	-2.11	-2.30	-2.15	-2.07	-2.23	-2.11	-2.04
50	-2.36	-2.20	-2.11	-2.30	-2.16	-2.08	-2.23	-2.11	-2.05
100	-2.36	-2.20	-2.11	-2.30	-2.16	-2.08	-2.23	-2.12	-2.05
200	-2.36	-2.20	-2.11	-2.30	-2.16	-2.08	-2.23	-2.12	-2.05

These results are quoted from Pesaran (2007). The critical values are obtained from the estimates of $\Delta Y_{it} = \alpha_i + b_i Y_{it-1} + c_i \Delta \overline{Y}_t + d_i \overline{Y}_{t-1} + e_{it}$ with the test statistic is given by regression based on 10,000 runs. The test statistic is given by $\overline{t}_i = \sum_{i=1}^{N} t_i \, / \, N$ (the details of this expression can be referred in equation (5)) and the results of the test statistics are reported at 1%, 5% and 10% level of significance.

Table 2. Critical Values of RCIPS

N	20			30			50		
Level of significance / T	1%	5%	10%	1%	5%	10%	1%	5%	10%
20	-1.6240	-1.3834	-1.2711	-1.5179	-1.3423	-1.2458	-1.4291	-1.2888	-1.2124
30	-1.6565	-1.4592	-1.3637	-1.6139	-1.4300	-1.3319	-1.5264	-1.3843	-1.2931
50	-1.7569	-1.5555	-1.4484	-1.6979	-1.4987	-1.4138	-1.6126	-1.4483	-1.3692
100	-1.8267	-1.6090	-1.5238	-1.7662	-1.5894	-1.6866	-1.6866	-1.5242	-1.4575
200	-1.8983	-1.6946	-1.5992	-1.8397	-1.6613	-1.5646	-1.7706	-1.6182	-1.5319

Following the work of Im et al. (2003), the DGP computing critical values for RCIPS test is given by $y_{it} = y_{it-1} + e_{it}$, with $e_{it} \sim iidN(0,1)$; for $i = 1, 2, \ldots, N$. $t = 1, 2, \ldots, T$ based on 5,000 runs. The test statistic is given by $\overline{t}_i^* = \sum_{i=1}^{N} t_i^* \, / \, N$ (the details of this expression can be referred in equation (11)) and the results of the test statistics are reported at 1%, 5% and 10% level of significance.

Table 3. The size of the unit root tests

CIPS

	no cont	AO	TC	no cont	AO	TC	no cont	AO	TC
T/N		**20**			**30**			**50**	
20	0.006	0.000	0.000	0.003	0.000	0.000	0.006	0.000	0.000
30	0.011	0.000	0.000	0.004	0.000	0.000	0.002	0.000	0.000
50	0.012	0.000	0.000	0.014	0.000	0.000	0.009	0.000	0.000
100	0.047	0.000	0.000	0.022	0.000	0.000	0.028	0.000	0.000
200	0.034	0.000	0.008	0.035	0.000	0.000	0.025	0.000	0.000

RCIPS

	no cont	AO	TC	no cont	AO	TC	no cont	AO	TC
T/N		**20**			**30**			**50**	
20	0.041	0.008	0.039	0.058	0.006	0.038	0.056	0.004	0.056
30	0.074	0.013	0.042	0.042	0.011	0.023	0.062	0.002	0.045
50	0.053	0.004	0.030	0.049	0.026	0.032	0.059	0.021	0.054
100	0.076	0.048	0.051	0.078	0.073	0.045	0.074	0.052	0.039
200	0.069	0.081	0.042	0.057	0.076	0.052	0.080	0.073	0.044

The values are the probability of rejecting the null of a unit root based on 1000 replications in uncontaminated panel (column no cont), contaminated with AO (column AO) and contaminated with TC (column TC). The size (probability of rejecting the null of a unit root when the unit root is present in the data) of the test is computed for $\varphi_i = 1$. The H_0 is rejected if the respective test statistics is greater than theirs critical values (tabulated in Tables 1 and 2) at 5% level of significance.

Table 4. The Power of the unit root tests

CIPS

	no cont	AO	TC	no cont	AO	TC	no cont	AO	TC
T/N		**20**			**30**			**50**	
20	0.002	0.000	0.000	0.022	0.000	0.000	0.018	0.000	0.000
30	0.207	0.000	0.000	0.241	0.000	0.000	0.283	0.000	0.001
50	0.862	0.011	0.026	0.952	0.005	0.023	0.999	0.007	0.022
100	1.000	0.918	0.836	1.000	0.282	0.955	1.000	0.355	0.977
200	1.000	0.981	1.000	1.000	0.993	1.000	1.000	1.000	1.000

RCIPS

	no cont	AO	TC	no cont	AO	TC	no cont	AO	TC
T/N		**20**			**30**			**50**	
20	0.793	0.422	0.788	0.912	0.481	0.833	0.952	0.683	0.961
30	0.920	0.617	0.865	0.964	0.755	0.965	0.981	0.804	0.980
50	0.994	0.834	0.968	1.000	0.922	0.988	1.000	0.986	1.000
100	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
200	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

The values are the probability of rejecting the null of a unit root based on 1000 replications in uncontaminated panel (column no cont), contaminated with AO (column AO) and contaminated with TC (column TC). The power (probability of correctly rejecting the null of a unit root when the unit root is absence in the data) of the test is computed for $\varphi_i \sim U\,[1.75, 0.95]$. The H_0 is rejected if the respective test statistics is greater than theirs critical values (tabulated in Tables 1 and 2) at 5% level of significance.

Acknowledgments

The authors gratefully acknowledge the helpful comments from the participant at the International Conference on Computing, Mathematics and Statistics 2013 held in Bay View Beach Resort, Penang, Malaysia and the grant support USIM/ERGS-FST-33-50312 from the Universiti Sains Islam Malaysia (USIM), Malaysia.

References

Bai, J. & Ng, S. (2004). A PANIC attack on unit roots and cointegration. *Econometrica*, *72*(4), 1127-1177.

Banerjee, A. (1999). Panel data unit root and cointegration: An overview. *Oxford Bulletin of Economics & Statistics*, *61*(S1), 607-629. doi:10.1111/1468-0084.0610s1607

Barbieri, L. (2009). Panel unit root tests under cross-sectional dependence: An overview. *Journal of Statistics: Advances in Theory and Applications*, *1*(2), 117–158.

Breitung, J. & Meyer, W. (1994). Testing for unit roots in panel data: Are wages on different bargaining levels cointegrated? *Applied Economics*, *26*(4), 353-361. doi:10.1080/00036849400000081

Choi, I. (2001). Unit root tests for panel data. *Journal of International Money and Finance*, *20*(2), 249-272. doi:10.1016/S0261-5606(00)00048-6

Choi, I. (2002). *Combination Unit Root Tests for Cross-Sectionally Correlated Panels*. Mimeo, Hong Kong University of Science and Technology.

Franses, P. H. & Haldrup, N. (1994). The effects of additive outliers on tests for unit roots and cointegration. *Journal of Business and Economic Statistics*, *12*(4), 471-478. doi:10.1080/07350015.1994.10524569

Hurlin, C. (2010). What would Nelson and Plosser find had they used panel unit root tests? *Journal of Applied Economics*, *42*(12), 1515-1531. doi:10.1080/00036840701721539

Im, K. S., Pesaran, M. H., & Shin, Y. (1997). *Testing for Unit Roots in Heterogenous Panels*. DAE, Working Paper 9526, University of Cambridge.

Im, K. S., Pesaran, M. H. & Shin, Y. (2003). Testing for Unit Roots in Heterogeneous Panels. *Journal of Econometrics*, *115*(1), 53-74. doi:10.1016/S0304-4076(03)00092-7

Levin, A. & Lin, C. F. (1992). *Unit Root Test in Panel Data: Asymptotic and Finite Sample Properties*. University of California at San Diego, Discussion Paper. 92-93.

Levin, A. & Lin, C. F. (1993). *Unit Root Test in Panel Data: New Results. Discussion Paper No 93-56*. Department of Economics, University of California at San Diego.

Levin, A., Lin, C. F. & Chu, C. S. J. (2002). Unit root tests in panel data: Asymptotic and finite-sample properties. *Journal of Econometrics*, *108*(1), 1-24. doi:10.1016/S0304-4076(01)00098-7

Martin, R. D. & Yohai, V. J. (1986). Influence functionals or time series. *The Annals of Statistics*, *14*(3), 781-818. doi:10.1214/aos/1176350027

Moon, H. R. & Perron, B. (2004). Testing for a unit root in panels with dynamic factors. *Journal of Econometrics*, *122*(1), 81-12. doi:10.1016/j.jeconom.2003.10.020

O'Connell, P. (1998). The overvaluation of purchasing power parity. *Journal of International Economics*, *44*(1), 1-19. doi:10.1016/S0022-1996(97)00017-2

Pesaran, M. H. (2007). A simple panel unit root test in the presence of cross section dependence. *Journal of Applied Economics*, *22*(2), 265-312. doi:10.1002/jae.951

Philips, P. C. B. & Sul, D. (2003). Dynamic panel estimation and homogeneity testing under cross section dependence. *Econometrics Journal*, *6*(1), 217-259. doi:10.1111/1368-423X.00108

Quah, D. (1994). Exploiting cross-section variation for unit root inference in dynamic data. *Economics Letters*, *44*(1-2), 9-19. doi:10.1016/0165-1765(93)00302-5

Tsay, R. S. (1998). Outliers, level shift, and variance changes in time series. *Journal of Forecasting*, *7*(1), 1-20. doi:10.1002/for.3980070102

Journal of Modern Applied Statistical Methods
November 2015, Vol. 14, No. 2, 172-200.

ISSN 1538 − 9472

The Distribution of the Inverse Square Root Transformed Error Component of the Multiplicative Time Series Model

Bright F. Ajibade
Petroleum Training Institute
Efurum, Warri, Nigeria

Chinwe R. Nwosu
Nnamdi Azikiwe University
Awka, Nigeria

J. I. Mbegbu
University of Benin
Benin City, Nigeria

The probability density function, mean and variance of the inverse square-root transformed left-truncated $N\left(1,\sigma^2\right)$ error component $e_t^*\left(=\frac{1}{\sqrt{e_t}}\right)$ of the multiplicative time series model were established. A comparison of key-statistical properties of e_t^* and e_t confirmed normality with mean 1 but with $Var\left(e_t^*\right)\approx\frac{1}{4}Var\left(e_t\right)$ when $\sigma\leq 0.14$. Hence $\sigma\leq 0.14$ is the required condition for successful transformation.

Keywords: Multiplicative time series model, Error component, Left truncated normal distribution, Inverse square root transformation, Successful transformation, Moments

Introduction

The general multiplicative time series model for descriptive time series analysis is

$$X_t = T_t\, S_t\, C_t\, e_t\,,\quad t=1,2,\ldots n \tag{1}$$

where for time t, X_t denotes the observed value of the series, T_t is the trend, S_t, the seasonal component, C_t the cyclical term and e_t is the random or irregular component of the series. Model (1) is regarded as adequate when the irregular component is purely random. For a short period of time, the cyclical component is

Mr. Ajibade, B. F. is a Deputy Chief Officer in the Department of General Studies. Email him at: equalright_bright@yahoo.com. C. R. Nwosu is an Associate Professor in the Department of Statistics. J. I. Mbegbu is a Professor in the Department of Maths and Statistics.

superimposed into the trend (Chatfield, 2004) to yield a trend-cycle component denoted by M_t and hence

$$X_t = M_t \, S_t \, e_t \tag{2}$$

where e_t are independent identically distributed normal errors with mean 1 and variance $\sigma^2 > 0$ $\left(e_t \sim N\left(1, \sigma^2\right)\right)$

According to Uche (2003), the left truncated normal distribution $\left(N\left(\mu, \sigma^2\right)\right)$ for X is

$$f^*(x) = \begin{cases} 0 & -\infty < x \le 0 \\ \dfrac{ke^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}}{\sigma\sqrt{2\pi}} & 0 < x < \infty \end{cases} \tag{3}$$

Using Equation 3, Iwueze (2007) obtained the left truncated normal distribution $\left(N\left(1, \sigma^2\right)\right)$ for $e_t\,(=X)$ as

$$f_{LTN}(x) = \begin{cases} 0 & -\infty < x \le 0 \\ \dfrac{e^{-\frac{1}{2}\left(\frac{x-1}{\sigma}\right)^2}}{\sigma\sqrt{2\pi}\left[1-\Phi\left(-\frac{1}{\sigma}\right)\right]} & 0 < x < \infty \end{cases} \tag{4}$$

with mean

$$E_{LTN}(X) = 1 + \frac{\sigma e^{-\frac{1}{2\sigma^2}}}{\sqrt{2\pi}\left(1-\Phi\left(-\frac{1}{\sigma}\right)\right)} \tag{5}$$

and

$$Var_{LTN}\left(X\right) = \frac{\sigma^2}{2\left(1-\Phi\left(-\frac{1}{\sigma}\right)\right)}\left[1+\Pr\left(\chi^2_{(1)} < \frac{1}{\sigma^2}\right)\right] - \frac{\sigma e^{-\frac{1}{2\sigma^2}}}{\sqrt{2\pi}\left(1-\Phi\left(-\frac{1}{\sigma}\right)\right)} - \left[1+\frac{\sigma e^{-\frac{1}{2\sigma^2}}}{\sqrt{2\pi}\left(1-\Phi\left(-\frac{1}{\sigma}\right)\right)}\right]^2 \qquad (6)$$

Iwueze (2007) also showed that $f_{LTN}\left(x\right) > 0$ provided $\sigma < 0.30$.

Data transformations are the application of mathematical modifications to values of a variable. There are a great variety of possible data transformations, including $\log_e\left(X_t\right), \sqrt{X_t}$, $\frac{1}{X_t}$, $\frac{1}{\sqrt{X_t}}$, X_t^2, and $\frac{1}{X_t^2}$. In practice many multiplicative time series data do not meet the assumptions of a parametric statistical analysis; they are not normally distributed, the variances are not homogenous or both. In analyzing such data, there are two choices:

i. Adjusting the data to fit the assumptions by making a transformation, or

ii. Developing new methods of analysis with assumptions which fit the data in its "original" form.

If a satisfactory transformation can be found, it will almost always be easier and simpler to use it rather than developing new methods of analysis (Turkey, 1957). Hence the need for this work which aims at finding conditions for satisfactory inverse square root transformation with respect to the error component of the multiplicative time series model from a study of its distribution. A transformation is considered satisfactory or successful, if the basic assumptions of the model are not violated after transformation. (Iwueze et al., 2008)The basic assumptions of a multiplicative time series model placed on the error component are: (i) unit mean (ii) constant variance (iii) Normality. According to Roberts (2008), transforming data made it much easier to work with - It was like sharpening a knife. For more information on choice of appropriate transformations see Osborne (2002), Osborne (2010) and Watthanacheewakul (2012).

Data Classification

For a time series data to be classified appropriate for inverse square root transformation,

i. the data must be amenable to the multiplicative time series model. The appropriateness of the multiplicative model is accessed by (a) displaying the data in the Buy's-Ballot Table. (b) Plotting the periodic (yearly) means (μ_i) and standard deviations σ_i against the period (year) i. If there is a dependency relationship between μ_i and σ_i, then the multiplicative model is appropriate.

ii. the variance must be unstable. The stability of the variance of the time series is ascertained by observing both the row and column means and standard deviations. If the variance is not stable the appropriate transformation is determined using Bartlett (1947) as was applied by Akpanta and Iwueze (2009);

$$Y = \begin{cases} \log_e X & , \quad \beta = 1 \\ X^{1-\beta} & , \quad \beta \neq 1 \end{cases} \tag{7}$$

The linear relationship between the natural log of periodic standard deviations ($\log_e \sigma_i$) and natural log of the periodic means ($\log_e \mu_i$) is given as

$$\log_e \sigma_i = \alpha + \beta \log_e \mu_i \tag{8}$$

The value of slope β according to Bartlett (1947) should be approximately 1.5 for the inverse square root transformation (see Table 1).

Table 1. Bartlett's transformations for some values of β

β	**0**	$\frac{1}{2}$	**1**	$\frac{3}{2}$	**2**	**3**	**-1**
Transformation	No transformation	$\sqrt{X}$	$\log_e X$	$\frac{1}{\sqrt{X}}$	$\frac{1}{X}$	$\frac{1}{X^2}$	X^2

Background of the Study

Since Iwueze (2007) investigated the effect of the logarithmic transformation on the error component, ($e_t \sim N(1, \sigma^2)$) of the multiplicative time series model, a number of studies investigating the effects of data transformation on the various components of the multiplicative time series model have been carried out. (See Iwueze et al., 2008; Iwu et al., 2009; Otuonye et al., 2011; Nwosu et al., 2013; and Ohakwe et al., 2013). The overall aim of such studies is to determine the conditions for successful transformation. That is, to establish the conditions where:

a. the required basic assumptions of the model are not violated after transformation, with respect to (i) the error term (ii) the seasonal component.

b. with respect to the trend component, there is no alteration in the form of the trend curve. In other words the form of the trend curve in the original series is maintained in the transformed series.

Iwueze (2007) found that the logarithmic transformation of the error component e_t $\left(e_t \sim N\left(1, \sigma^2\right)\right)$ to $e_t^* \left(= \log_e e_t\right)$ is normal with mean 0 and variance σ_1^2 provided $\sigma < 0.1$, in which case $\sigma_1 = \sigma$. It was established that the assumption for the error term e_t^*, for the additive model obtained after the logarithmic transformation, is valid if and only if $\sigma_1 < 0.10$. Observe from Table 1 that $\beta \approx 1$ for a time series data to be classified fit for logarithmic transformation.

Otuonye et al. (2011) investigated the distribution and properties of the error component of the multiplicative time series model under square root transformation, and found that the square root transformed error component $e_t^* \left(= \sqrt{e_t}\right)$ is normally distributed with mean ≈ 1 and variance $\approx \frac{1}{4}$ times that of the untransformed error component. That is $Var\left(e_t^*\right) = \frac{1}{4}\left[Var\left(e_t\right)\right]$ when $0 < \sigma \leq 0.3$. Thus $0 < \sigma \leq 0.3$ is the recommended condition for successful square root transformation. Only time series data with $\beta \approx \frac{1}{2}$ are classified fit for square root transformation. Similarly, Nwosu et al. (2013), while investigating the distribution of the inverse transformed error component of the multiplicative time series model $e_t^*\left(=\frac{1}{e_t}\right)$, obtained that the desirable statistical properties of e_t and e_t^* were found to be approximately the same and normally distributed with unit mean for $\sigma \leq 0.10$. Hence, $\sigma \leq 0.10$ is the recommended condition for successful inverse transformation of the multiplicative time series model. Time series data classified fit for inverse transformation must have $\beta \approx 2$. Also, Ohakwe et al. (2013) found that for the square transformation $e_t^*\left(=e_t^2\right)$ that $e_t^* \sim N(1,1)$ in the interval $0 < \sigma \leq 0.027$. Hence, $0 < \sigma \leq 0.027$is the condition for successful square transformation. Observe that a time series data is classified fit for square transformation when $\beta \approx$ -1.

Note that the overall aim of these works is to establish conditions for successful transformation, hence provide better choice of right transformation. According to Roberts (2008), choosing a good transformation improved his analyses in three ways: (i) increase in visual clarity as graphs were made more informative (ii). Reduction or elimination of outliers (iii). Increase in statistical clarity; his statistical test became more sensitive, *F* and *t* values increased making it more likely to detect differences when they exist.

Justification for this Study

The value of the slope β, categorized time series data into mutually exclusive groups, in the sense that any time series data belongs exclusively to one and only one group hence can only be appropriately transformed by only one of the six transformations listed in Table 1. Thus despite the fact that Iwueze (2007), Otuonye et al., (2011), Nwosu et al. (2013), and Ohakwe et al. (2013) carried out similar studies with respect to the logarithmic, square root, inverse and square transformations respectively, this work on inverse square root transformation is still very necessary since results established for the above listed four transformations cannot be applied in the analysis of time series data requiring inverse square root transformation.

Inverse Square Root Transformation

When $\beta \approx \frac{3}{2}$, adopt inverse square root transformation on the multiplicative time series model given in Equation 2 to obtain

$$Y_t = \frac{1}{\sqrt{X_t}} = \frac{1}{\sqrt{M_t}}\frac{1}{\sqrt{S_t}}\frac{1}{\sqrt{e_t}} = M_t^* S_t^* e_t^* \qquad (9)$$

where $M_t^* = \frac{1}{\sqrt{M_t}}$, $S_t^* = \frac{1}{\sqrt{S_t}}$ *and* $e_t^* = \frac{1}{\sqrt{e_t}}$, $e_t > 0$

Because e_t does not admit negative or zero values, the use of the left truncated normal distribution as the pdf of e_t shall be exploited.

Thus, it will be of interest to find what the distribution of e_t^* is. Is e_t^* ~ iid $N\left(1, \sigma_1^2\right)$. What is the relationship between σ_1^2 and σ^2?

Aim and Objectives

The aim of this work is to obtain the distribution of the inverse square root transformed error component of the multiplicative time series model and the objectives are:

i. to examine the nature of the distribution.

ii. to verify the satisfaction of the assumption on the mean of the error terms; $\mu = 1$.

iii. to determine the relationship between σ_1^2 and σ^2.

Methodology

To achieve the above stated objectives the following were conducted:

Let $X = e_t$ and $Y = e_t^* = \frac{1}{\sqrt{e_t}} = \frac{1}{\sqrt{X}}$

1. Obtain the pdf of e_t^*, $g(y)$.
2. Plot the curves of the two pdfs, $g(y)$ and $f_{LTN}(x)$ for various values of σ.
3. Obtain the region where $g(i)$ satisfies the following normality conditions (Bell-shaped conditions).

   i. Mode $\approx 1 \approx$ Mean.

   ii. Median $\approx$ Mean $\approx$ 1.

iii. Approved normality test, Anderson Darling’s test statistic (AD) was used to confirm the normality of the simulated error terms e_t and the inverse square root transformed error term.

$Y = e_t^* = \dfrac{1}{\sqrt{e_t}} = \dfrac{1}{\sqrt{X}}$ for some values of σ

iv. Obtain and use the functional expressions for the mean and variance of e_t^* to validate some of the results obtained using simulated data.

The probability density function of $Y = \frac{1}{\sqrt{x}}$, $g(y)$

Given the pdf of X in Equation 4 and the transformation

$$Y = \frac{1}{\sqrt{x}}$$

then

$$X = \frac{1}{y^2} \text{ and} \frac{dx}{dy} = -\frac{2}{y^3}$$

using the transformation of variable technique

$$g(y) = f_{LTN}(x)\left|\frac{dx}{dy}\right|$$

(see Freund & Walpole, 1986). Hence

$$g(y) = \begin{cases} \dfrac{2}{y^3\sigma\sqrt{2\pi}\left[1-\Phi\left(\dfrac{-1}{\sigma}\right)\right]} e^{\frac{-1}{2\sigma^2}\left(\frac{1}{y^2}-1\right)^2} & ,0 < y < \infty \\ 0 & -\infty < y < 0 \end{cases} \qquad (10)$$

Plot of the Probability density curves $f_{LTN}(x) = f^*(x)$ and *g*(*y*)

Using the pdf of the two variables given in Equation 4 and Equation 10, the curves $f^*(x)$ and $g(y)$ were plotted for some values of $\sigma \in (0, 0.4]$. For want of space only five are shown in Figures 1 to 5.

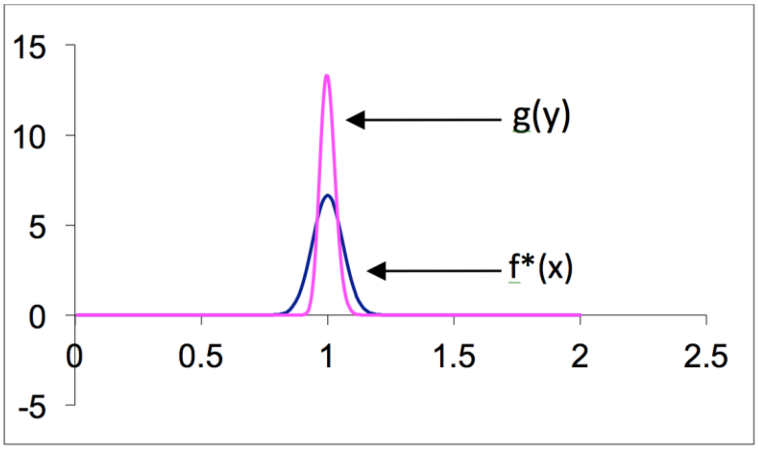


Figure 1. Curve Shapes for σ = 0.06

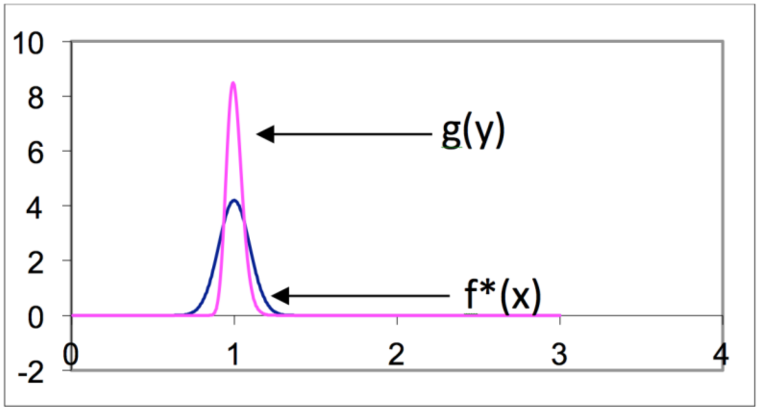


Figure 2. Curve Shapes for σ = 0.095

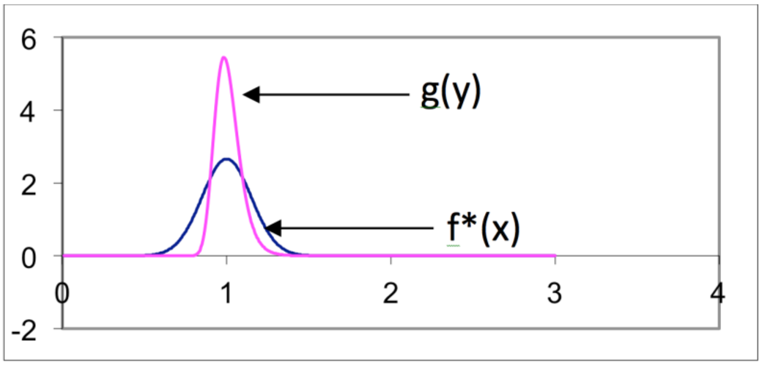


Figure 3. Curve Shapes for σ = 0.15

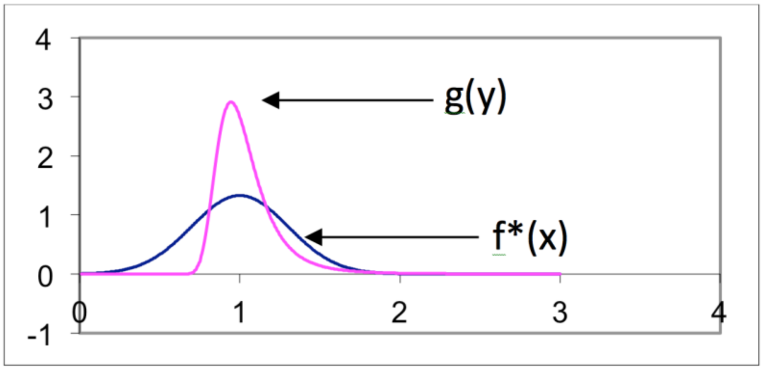


Figure 4. Curve Shapes for σ = 0.3

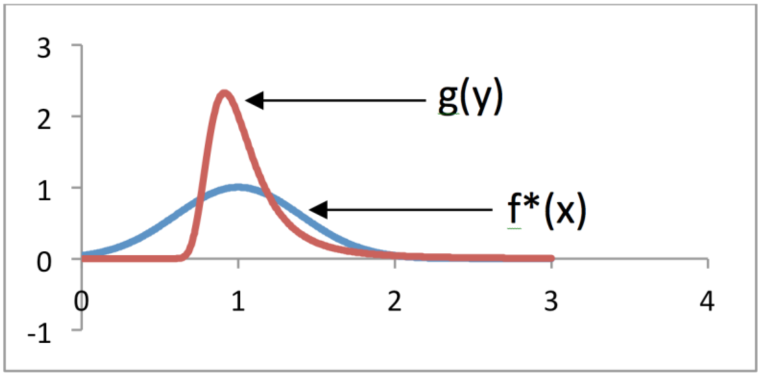


Figure 5. Curve Shapes for σ = 0.4

Observations:

i. The curve $g(y)$ is positively skewed for $\sigma > 0.15$ (see Figures 3-5).

ii. $f^*(x)$ is positively skewed for $\sigma > 0.30$ (see Figure 5) as reported in Iwueze (2007).

Normality Region for *g*(*y*)

From Figures 1 to 5, it is clear that the curve $g(y)$ has one maximum point, y_{max} (mode), and one maximum value, $g(y_{\text{max}})$, for all values of σ. To obtain the values of σ that satisfy the symmetric and bell-shaped condition of mode = mean, we invoke Rolle's Theorem and proceed to obtain the maximum point (mode) for a given value of σ.

Differentiating $g(y)$ in Equation 10 gives

$$g'(y) = \frac{2}{\sigma\sqrt{2\pi\left[1-\Phi\left(\frac{-1}{\sigma}\right)\right]}}\left[-3y^{-4}e^{\frac{-1}{2\sigma^2}\left(\frac{1}{y^2}-1\right)^2} + y^{-3}\left(\frac{-1}{2\sigma^2}\right)\left(\frac{-4}{y^3}\right)\left(\frac{1}{y^2}-1\right)e^{\frac{-1}{2\sigma^2}\left(\frac{1}{y^2}-1\right)^2}\right] \tag{11}$$

$$= \frac{2e^{\frac{-2}{2\sigma^2}\left(\frac{1}{y^2}-1\right)^2}}{\sigma\sqrt{2\pi\left[1-\Phi\left(\frac{-1}{\sigma}\right)\right]}}\left[\frac{2(1-y^2)}{\sigma^2 y^8}-\frac{3}{y^4}\right]$$

Equating $g`(y) = 0$, gives

$$\frac{2(1-y^2)}{\sigma^2 y^8}-\frac{3}{y^4}=0$$

$$3\sigma^2 y^4 + 2y^2 - 2 = 0 \tag{12}$$

Putting $w = y^2$ in Equation 12, gives

$$3\sigma^2 w^2 + 2w - 2 = 0 \tag{13}$$

Solving Equation 13, gives

$$w = \frac{-1 \pm \sqrt{1+6\sigma^2}}{3\sigma^2}$$

Because $y_{\max}$ is positive
then

$$w = \frac{-1 + \sqrt{1+6\sigma^2}}{3\sigma^2}$$

hence

$$y = \pm\sqrt{\frac{-1 + \sqrt{1+6\sigma^2}}{3\sigma^2}}$$

and

$$y_{\max} = \sqrt{\frac{-1 + \sqrt{1+6\sigma^2}}{3\sigma^2}}$$

The bell-shaped condition would imply $y_{\max} \approx 1$, see Table 2 for the numerical computation of

$$y_{\max} = \sqrt{\frac{-1 + \sqrt{1+6\sigma^2}}{3\sigma^2}}$$

Table 2. Computation of $y_{\max} = \sqrt{\frac{-1+\sqrt{1+6\sigma^2}}{3\sigma^2}}$, for $\sigma \; \varepsilon \; [0.01, 0.3]$

σ	$y_{\max} = \sqrt{\frac{-1+\sqrt{1+6\sigma^2}}{3\sigma^2}}$	$1-y_{\max}$
0.010	0.99992502	0.000075
0.015	0.99970031	0.000300
0.020	0.99932659	0.000673
0.025	0.99880501	0.001195
0.030	0.99813720	0.001863
0.035	0.99732519	0.002675
0.040	0.99637147	0.003629
0.045	**0.99527886**	0.004721
0.050	0.99405059	0.005949
0.055	0.99269018	0.007310
0.060	0.99120149	0.008799
0.065	0.98958860	0.010411
0.070	0.98785584	0.012144
0.075	0.98600775	0.013992
0.080	0.98404899	0.015951
0.085	0.98198438	0.018016
0.090	0.97981881	0.020181
0.095	0.97755725	0.022443
0.100	0.97520469	0.024795
0.105	0.97276613	0.027234
0.110	0.97024653	0.029753
0.115	0.96765082	0.032349
0.120	0.96498387	0.035016
0.125	0.96225045	0.037750
0.130	0.95945523	0.040545
0.135	0.95660279	0.043397
0.140	0.95369754	0.046302
0.145	**0.95074378**	0.049256
0.150	0.94774567	0.052254
0.155	0.94470721	0.055293
0.160	0.94163225	0.058368
0.165	0.93852446	0.061476
0.170	0.93538739	0.064613
0.175	0.93222440	0.067776
0.180	0.92903869	0.070961
0.185	0.92583333	0.074167
0.190	0.92261120	0.077389
0.195	0.91937505	0.080625
0.200	0.91612748	0.083873
0.205	0.91287093	0.087129
0.210	0.90960772	0.090392
0.215	0.90634001	0.093660
0.220	0.90306986	0.096930
0.225	0.89979918	0.100201
0.230	0.89652976	0.103470
0.235	0.89326328	0.106737
0.240	0.89000132	0.109999
0.245	0.88674534	0.113255
0.250	0.88349669	0.116503
0.255	0.88025665	0.119743
0.260	0.87702640	0.122974
0.265	0.87380702	0.126193
0.270	0.87059952	0.129400
0.275	0.86740484	0.132595
0.280	0.86422383	0.135776
0.285	0.86105729	0.138943
0.290	0.85790594	0.142094
0.295	0.85477043	0.145230
0.300	0.85165139	0.148349

Thus $g(y)$ is symmetrical about 1 with Mode ≈ 1 ≈ Mean correct to two decimal places when $0 < \sigma < 0.045$ and correct to one decimal place when $0 < \sigma < 0.045$.

Use of simulated error terms

To find the region where the bell-shaped conditions (ii-iii) listed in methodology are satisfied, we made use of artificial data generated from $N\left(1,\sigma^2\right)$ for e_t, subsequently transformed to obtain $e_t^* = \frac{1}{\sqrt{e_t}}$ for $0.05 \le \sigma \le 0.20$. Values of the required statistical characteristics were obtained for each variable e_t and e_t^* as shown in Tables 3 to 6. For each configuration of (n = 100, $0.05 \le \sigma \le 0.15$), 1000 replications were performed for values of σ in steps of 0.01. For want of space the results of the first 25 replications are shown for the configurations, (n = 100, σ = 0.06), (n = 100, σ = 0.1), (n = 100, σ = 0.15), and (n = 100, σ = 0.2).

Functional expressions for the mean and variance of *g*(*y*)

By definition, the mean of Y, $E(Y)$ is given by:

$$E(Y)=\int_0^\infty yg(y)dy=\frac{2}{\sigma\sqrt{2\pi}\left[1-\Phi\left(\frac{1}{\sigma}\right)\right]}\int_0^\infty \frac{1}{y^2}e^{\frac{-1}{2\sigma^2}\left(\frac{1}{y^2}-1\right)^2}dy \tag{14}$$

let $u=\frac{1}{y^2}$, then $y=\frac{1}{u^{\frac{1}{2}}}$ and $dy=\frac{-du}{2u^{\frac{3}{2}}}$, for $\infty<u<0$

$$\therefore E(Y)=k\int_\infty^0 ue^{\frac{-1}{2}\left(\frac{u-1}{\sigma}\right)^2}\frac{-du}{2u^{\frac{3}{2}}}\cdot=\frac{k}{2}\int_0^\infty u^{\frac{-1}{2}}e^{\frac{-1}{2}\left(\frac{u-1}{\sigma}\right)^2}du\cdots \tag{15}$$

$$\text{where } k=\frac{2}{\sigma\sqrt{2\pi}\left[1-\Phi\left(\frac{-1}{\sigma}\right)\right]}$$

let $z=\frac{u-1}{\sigma}$, then $z\sigma+1=u$ and $du=\sigma dz$ for $\frac{-1}{\sigma}<z<\infty$

$$\therefore E(Y)=\frac{k}{2}\int_{\frac{-1}{\sigma}}^\infty(1+z\sigma)^{\frac{-1}{2}}e^{\frac{-1}{2}z^2}\sigma dz=\frac{\sigma k}{2}\int_{\frac{-1}{\sigma}}^\infty(1+z\sigma)^{\frac{-1}{2}}e^{\frac{-z^2}{2}}dz \tag{16}$$

Table 3. Simulation Results when $\sigma = 0.06$

$$X = e_t \sim N\left(1, \sigma^2\right), \sigma = 0.06$$

$$Y = e_t^* = \frac{1}{\sqrt{e_t}}, e_t \sim N\left(1, \sigma^2\right), \sigma = 0.06$$

Mean	StD	Variance	Median	AD	p-value	Mean	StD	Variance	Median	AD	p-value
1	0.06	0.0036	0.9927	.235	.788	1.0013	0.0303	0.000918	1.0037	.206	.867
1	0.06	0.0036	1.0009	.183	.908	1.0013	0.0302	0.000914	0.9995	.298	.580
1	0.06	0.0036	1.0002	.195	.889	1.0013	0.0303	0.000916	0.9999	.275	.654
1	0.06	0.0036	1.0029	.234	.790	1.0013	0.0303	0.000917	0.9985	.334	.505
1	0.06	0.0036	1.0037	.178	.918	1.0013	0.0302	0.000915	0.9982	.312	.546
1	0.06	0.0036	1.0045	.435	.294	1.0013	0.0301	0.000908	0.9978	.364	.433
1	0.06	0.0036	1.0037	.178	.918	1.0013	0.0302	0.000915	0.9982	.312	.546
1	0.06	0.0036	1.0013	.137	.976	1.0013	0.0302	0.000910	0.9993	.213	.851
1	0.06	0.0036	0.9941	.196	.888	1.0013	0.0302	0.000911	1.0030	.302	.569
1	0.06	0.0036	1.0017	.250	.739	1.0014	0.0304	0.000924	0.9991	.453	.266
1	0.06	0.0036	1.0004	.200	.880	1.0013	0.0302	0.000915	0.9998	.314	.540
1	0.06	0.0036	1.0045	.435	.294	1.0013	0.0301	0.000908	0.9978	.364	.433
1	0.06	0.0036	0.9991	.183	.908	1.0013	0.0303	0.000916	1.0005	.214	.846
1	0.06	0.0036	0.9983	.250	.739	1.0013	0.0301	0.000908	1.0009	.206	.866
1	0.06	0.0036	1.0010	.209	.859	1.0013	0.0300	0.000901	0.9995	.241	.767
1	0.06	0.0036	1.0028	.195	.889	1.0013	0.0302	0.000913	0.9986	.284	.625
1	0.06	0.0036	1.0031	.141	.972	1.0013	0.0302	0.000911	0.9985	.208	.862
1	0.06	0.0036	0.9975	.310	.550	1.0013	0.0299	0.000894	1.0012	.232	.795
1	0.06	0.0036	1.0006	.262	.699	1.0014	0.0304	0.000924	0.9997	.385	.387
1	0.06	0.0036	0.9983	.182	.911	1.0013	0.0302	0.000913	1.0009	.318	.531
1	0.06	0.0036	0.9958	.150	.962	1.0013	0.0303	0.000916	1.0021	.218	.835
1	0.06	0.0036	0.9938	.290	.606	1.0013	0.0299	0.000896	1.0031	.185	.906
1	0.06	0.0036	0.9931	.450	.270	1.0013	0.0300	0.000903	1.0035	.336	.503
1	0.06	0.0036	0.9950	.199	.882	1.0013	0.0301	0.000907	1.0025	.390	.376
1	0.06	0.0036	0.9987	.216	.841	1.0013	0.0302	0.000914	1.0006	.315	.538
1	0.06	0.0036	0.9942	.311	.546	1.0013	0.0300	0.000899	1.0029	.165	.940

*Note. For each row, $\frac{Var\left(e_t\right)}{Var\left(e_t^*\right)}$ equals 4.

Table 4. Simulation Results when σ = 0.1

$X = e_t \sim N\left(1, \sigma^2\right), \sigma = 0.1$						$Y = e_t^* = \dfrac{1}{\sqrt{e_t}}, e_t \sim N\left(1, \sigma^2\right), \sigma = 0.1$					
Mean	**StD**	**Variance**	**Median**	**AD**	**p-value**	**Mean**	**StD**	**Variance**	**Median**	**AD**	**p-value**
1	0.1	0.01	0.9878	.235	0.788	1.0038	0.0514	0.00265	1.0061	.298	.582
1	0.1	0.01	1.0016	.183	0.908	1.0038	0.0511	0.00262	0.9992	.457	.260
1	0.1	0.01	1.0003	.195	0.889	1.0038	0.0513	0.00263	0.9998	.428	.306
1	0.1	0.01	1.0049	.234	0.790	1.0038	0.0513	0.00264	0.9976	.502	.201
1	0.1	0.01	1.0062	.178	0.918	1.0038	0.0512	0.00262	0.9969	.495	.211
1	0.1	0.01	1.0074	.435	0.294	1.0038	0.0509	0.00259	0.9963	.424	.313
1	0.1	0.01	1.0062	.178	0.918	1.0038	0.0512	0.00262	0.9969	.495	.211
1	0.1	0.01	1.0022	.137	0.976	1.0038	0.0509	0.00259	0.9989	.357	.450
1	0.1	0.01	0.9902	.196	0.888	1.0038	0.0510	0.00260	1.0050	.464	.251
1	0.1	0.01	1.0029	.250	0.739	1.0038	0.0516	0.00267	0.9986	.685	.071
1	0.1	0.01	1.0007	.200	0.880	1.0038	0.0512	0.00262	0.9997	.495	.210
1	0.1	0.01	1.0074	.435	0.294	1.0038	0.0509	0.00259	0.9963	.424	.313
1	0.1	0.01	0.9984	.183	0.908	1.0038	0.0513	0.00263	1.0008	.326	.516
1	0.1	0.01	0.9971	.250	0.739	1.0038	0.0509	0.00259	1.0014	.272	.664
1	0.1	0.01	1.0016	.209	0.859	1.0037	0.0505	0.00255	0.9992	.359	.445
1	0.1	0.01	1.0047	.195	0.889	1.0038	0.0511	0.00261	0.9977	.446	.277
1	0.1	0.01	1.0052	.141	0.972	1.0038	0.0510	0.00260	0.9974	.346	.477
1	0.1	0.01	0.9959	.310	0.550	1.0037	0.0502	0.00252	1.0021	.278	.642
1	0.1	0.01	1.0011	.262	0.699	1.0038	0.0516	0.00266	0.9995	.554	.150
1	0.1	0.01	0.9971	.182	0.911	1.0038	0.0511	0.00261	1.0014	.499	.205
1	0.1	0.01	0.9931	.150	0.962	1.0038	0.0513	0.00263	1.0035	.368	.424
1	0.1	0.01	0.9897	.290	0.606	1.0037	0.0503	0.00253	1.0052	.221	.827
1	0.1	0.01	0.9884	.450	0.270	1.0037	0.0506	0.00256	1.0058	.366	.428
1	0.1	0.01	0.9917	.306	0.559	1.0038	0.0508	0.00258	1.0042	.547	.156
1	0.1	0.01	0.9979	.199	0.882	1.0038	0.0511	0.00261	1.0011	.497	.207
1	0.1	0.01	0.9904	.216	0.841	1.0037	0.0504	0.00254	1.0048	.226	.815

*Note. For each row, $\dfrac{Var\left(e_t\right)}{Var\left(e_t^*\right)}$ equals 4.

Table 5. Simulation Results when $\sigma = 0.15$

$$X = e_t \sim N\left(1, \sigma^2\right), \sigma = 0.15 \qquad Y = e_t^* = \frac{1}{\sqrt{e_t}}, e_t \sim N\left(1, \sigma^2\right), \sigma = 0.15$$

Mean	StD	Variance	Median	AD	p-value		Mean	StD	Variance	Median	AD	p-value
1	0.15	0.0225	0.9818	.235	.788	*	1.0089	0.0803	0.00645	1.0092	.582	.126
1	0.15	0.0225	1.0024	.183	.908		1.0088	0.0791	0.00626	0.9988	.761	.046
1	0.15	0.0225	1.0005	.195	.889		1.0088	0.0798	0.00637	0.9997	.756	.047
1	0.15	0.0225	1.0073	.234	.790		1.0088	0.0798	0.00636	0.9964	.857	.027
1	0.15	0.0225	1.0093	.178	.918		1.0088	0.0792	0.00628	0.9954	.842	.029
1	0.15	0.0225	1.0111	.435	.294		1.0087	0.0788	0.00620	0.9945	.646	.089
1	0.15	0.0225	1.0093	.178	.918		1.0088	0.0792	0.00628	0.9954	.842	.029
1	0.15	0.0225	1.0034	.137	.976		1.0087	0.0786	0.00618	0.9983	.656	.085
1	0.15	0.0225	0.9853	.196	.888	*	1.0087	0.0788	0.00621	1.0075	.785	.040
1	0.15	0.0225	1.0043	.250	.739		1.0089	0.0804	0.00646	0.9979	1.109	.005
1	0.15	0.0225	1.0010	.200	.880		1.0088	0.0793	0.00628	0.9995	.860	.026
1	0.15	0.0225	1.0111	.435	.294		1.0087	0.0788	0.00620	0.9945	.646	.089
1	0.15	0.0225	0.9976	.183	.908		1.0088	0.0796	0.00633	1.0012	.596	.119
1	0.15	0.0225	0.9957	.250	.739		1.0087	0.0788	0.00621	1.0022	.486	.221
1	0.15	0.0225	1.0025	.209	.859		1.0086	0.0775	0.00601	0.9988	.620	.104
1	0.15	0.0225	1.0070	195	889		1.0088	0.0791	0.00626	0.9965	.779	.042
1	0.15	0.0225	1.0077	141	.972		1.0087	0.0787	0.00619	0.9962	.635	.095
1	0.15	0.0225	0.9938	.310	.550		1.0085	0.0770	0.00593	1.0031	.450	.271
1	0.15	0.0225	1.0016	.262	.699		1.0089	0.0799	0.00639	0.9992	.880	.023
1	0.15	0.0225	0.9957	.182	.911		1.0087	0.0789	0.00622	1.0022	.838	.030
1	0.15	0.0225	0.9896	.500	.962		1.0088	0.0798	0.00636	1.0052	.701	.065
1	0.15	0.0225	0.9846	.290	.606		1.0085	0.0770	0.00593	1.0078	.398	.361
1	0.15	0.0225	0.9826	.450	.270		1.0086	0.0781	0.00609	1.0088	.545	.157
1	0.15	0.0225	0.9876	.306	.559		1.0087	0.0782	0.00611	1.0063	.868	.025
1	0.15	0.0225	0.9968	.199	.882		1.0088	0.0790	0.00624	1.0016	.860	.026
1	0.15	0.0225	0.9856	.216	.841		1.0085	0.0772	0.00596	1.0073	.419	.322

*Note. For each row, $\frac{Var\left(e_t\right)}{Var\left(e_t^*\right)}$ equals 4 except where indicated by *. For those rows, $\frac{Var\left(e_t\right)}{Var\left(e_t^*\right)}$ equals 3.

Table 6. Simulation Results when σ = 0.2

$$X = e_t \sim N\left(1, \sigma^2\right), \sigma = 0.2$$

$$Y = e_t^* = \frac{1}{\sqrt{e_t}}, e_t \sim N\left(1, \sigma^2\right), \sigma = 0.2$$

Mean	StD	Variance	Median	AD	p-value	Mean	StD	Variance	Median	AD	p-value
1	0.2	0.04	0.9757	.235	0.788	1.0167	0.1147	0.0132	1.0124	1.176	<0.005
1	0.2	0.04	1.0032	.183	0.908	1.0162	0.1107	0.0123	0.9984	1.220	<0.005
1	0.2	0.04	1.0007	.195	0.889	1.0165	0.1127	0.0127	0.9997	1.315	<0.005
1	0.2	0.04	1.0097	.234	0.790	1.0164	0.1124	0.0126	0.9952	1.435	<0.005
1	0.2	0.04	1.0124	.178	0.918	1.0163	0.1109	0.0123	0.9939	1.353	<0.005
1	0.2	0.04	1.0148	.435	0.294	1.0161	0.1105	0.0122	0.9927	1.097	0.007
1	0.2	0.04	1.0124	.178	0.918	1.0163	0.1109	0.0123	0.9939	1.353	<0.005
1	0.2	0.04	1.0045	.137	0.976	1.0161	0.1095	0.0120	0.9978	1.117	0.006
1	0.2	0.04	0.9803	.196	0.888	1.0161	0.1100	0.0121	1.0100	1.276	<0.005
1	0.2	0.04	1.0057	.250	0.739	1.0166	0.1133	0.0128	0.9971	1.734	<0.005
1	0.2	0.04	1.0013	.200	0.880	1.0163	0.1110	0.0123	0.9994	1.418	<0.005
1	0.2	0.04	1.0149	.435	0.294	1.0161	0.1105	0.0122	0.9927	1.097	0.007
1	0.2	0.04	0.9968	.183	0.908	1.0164	0.1120	0.0125	1.0016	1.072	0.008
1	0.2	0.04	0.9943	.250	0.739	1.0162	0.1107	0.0123	1.0029	0.915	0.019
1	0.2	0.04	1.0033	.209	0.859	1.0157	0.1072	0.0115	0.9984	1.026	0.010
1	0.2	0.04	1.0094	.195	0.889	1.0162	0.1109	0.0123	0.9953	1.293	<0.005
1	0.2	0.04	1.0103	.141	0.972	1.0161	0.1097	0.0120	0.9949	1.084	0.007
1	0.2	0.04	0.9917	.310	0.550	1.0156	0.1066	0.0114	1.0042	0.768	0.045
1	0.2	0.04	1.0021	.260	0.699	1.0165	0.1119	0.0125	0.9989	1.371	<0.005
1	0.2	0.04	0.9942	.182	0.911	1.0162	0.1100	0.0121	1.0029	1.331	<0.005
1	0.2	0.04	0.9862	.150	0.962	1.0165	0.1128	0.0127	1.007	1.267	<0.005
1	0.2	0.04	0.9795	.290	0.606	1.0156	0.1064	0.0113	1.0104	0.745	0.051
1	0.2	0.04	0.9768	.450	0.270	1.0159	0.109	0.0119	1.0118	0.933	0.017
1	0.2	0.04	0.9835	.306	0.559	1.0159	0.1084	0.0118	1.0084	1.348	<0.005
1	0.2	0.04	0.9958	.199	0.882	1.0162	0.1101	0.0121	1.0021	1.402	<0.005
1	0.2	0.04	0.9808	.216	0.841	1.0156	0.1066	0.0114	1.0097	0.766	0.045

*Note. For each row, $\frac{Var\left(e_t\right)}{Var\left(e_t^*\right)}$ equals 3.

Using the binomial expansion ,

$$(1+x)^n = 1+\frac{nx}{1!}+\frac{n(n-1)x^2}{2!}+\frac{n(n-1)(n-2)x^3}{3!}+\dots \tag{17}$$

(Smith and Minton, 2008).

$$\therefore (1+z\sigma)^{\frac{-1}{2}} = \cdot 1+\frac{\left(\frac{-1}{2}\right)z\sigma}{1!}+\frac{\frac{-1}{2}\left(\frac{-1}{2}-1\right)(z\sigma)^2}{2!}+\frac{\frac{-1}{2}\left(\frac{-1}{2}-1\right)\left(\frac{-1}{2}-2\right)(z\sigma)^3}{3!}+\dots$$

$$= 1-\frac{z\sigma}{2}+\frac{3(z\sigma)^2}{8}-\frac{15(z\sigma)^3}{48}+\dots \tag{18}$$

$$\therefore E(Y) = \frac{\sigma k}{2}\int_{\frac{-1}{\sigma}}^{\infty}\left[1-\frac{z\sigma}{2}+\frac{3(z\sigma)^2}{8}-\frac{15(z\sigma)^3}{48}+\dots\right]e^{\frac{-z^2}{2}}\,d \tag{19}$$

$$= \frac{1}{\sqrt{2\pi}\left[1-\Phi\left(\frac{-1}{\sigma}\right)\right]}\left[\int_{\frac{-1}{\sigma}}^{\infty} e^{\frac{-z^2}{2}}dz-\int_{\frac{-1}{\sigma}}^{\infty}\frac{z\sigma}{2}e^{\frac{-z^2}{2}}dz+\int_{\frac{-1}{\sigma}}^{\infty}\frac{3(z\sigma)^2}{8}e^{\frac{-z^2}{2}}dz-\int_{\frac{-1}{\sigma}}^{\infty}\frac{15(z\sigma)^3}{48}e^{\frac{-z^2}{2}}dz+\dots\right]$$

$$E(Y) = \frac{1}{\left[1-\Phi\left(\frac{-1}{\sigma}\right)\right]}\left[\begin{aligned}&\frac{1}{\sqrt{2\pi}}\int_{\frac{-1}{\sigma}}^{\infty} e^{\frac{-z^2}{2}}dz-\frac{1}{\sqrt{2\pi}}\int_{\frac{-1}{\sigma}}^{\infty}\frac{z\sigma}{2}e^{\frac{-z^2}{2}}dz\\&+\frac{1}{\sqrt{2\pi}}\int_{\frac{-1}{\sigma}}^{\infty}\frac{3(z\sigma)^2}{8}e^{\frac{-z^2}{2}}dz-\frac{1}{\sqrt{2\pi}}\int_{\frac{-1}{\sigma}}^{\infty}\frac{15(z\sigma)^3}{48}e^{\frac{-z^2}{2}}dz+\dots\end{aligned}\right]$$

$$E(Y)=\frac{1}{\left[1-\Phi\left(\frac{-1}{\sigma}\right)\right]}\left[\begin{array}{l}\frac{1}{\sqrt{2\pi}}\int_{\frac{-1}{\sigma}}^{\infty}e^{\frac{-z^2}{2}}dz-\frac{\sigma}{\sqrt{2\pi}}\int_{\frac{-1}{\sigma}}^{\infty}\frac{z}{2}e^{\frac{-z^2}{2}}dz\\ +\frac{\sigma^2}{\sqrt{2\pi}}\int_{\frac{-1}{\sigma}}^{\infty}\frac{3z^2}{8}e^{\frac{-z^2}{2}}dz-\frac{\sigma^3}{\sqrt{2\pi}}\int_{\frac{-1}{\sigma}}^{\infty}\frac{15z^3}{48}e^{\frac{-z^2}{2}}dz+\dots\end{array}\right]$$

$$=\frac{1}{\left[1-\Phi\left(\frac{1}{\sigma}\right)\right]}\left[\begin{array}{l}\Pr\left(z>\frac{-1}{\sigma}\right)-\frac{\sigma}{2\sqrt{2\pi}}e^{\frac{-1}{2\sigma^2}}+\frac{3}{8}\left(\frac{\sigma^2}{2}-\frac{\sigma e^{\frac{-1}{2\sigma^2}}}{\sqrt{2\pi}}+\frac{\sigma^2}{2}\Pr\left(x^2{}_{(1)}<\frac{1}{\sigma^2}\right)\right)\\ -\frac{5}{16}\left(\frac{\sigma}{\sqrt{2\pi}}e^{\frac{-1}{2\sigma^2}}-\frac{2\sigma^3}{\sqrt{2\pi}}e^{\frac{-1}{2\sigma^2}}\right)+\dots\end{array}\right]$$

$$=\frac{1}{\left[1-\Phi\left(\frac{-1}{\sigma}\right)\right]}\left[\begin{array}{l}\left[1-\Phi\left(\frac{-1}{\sigma}\right)\right]-\frac{\sigma}{2\sqrt{2\pi}}.e^{\frac{-1}{2\sigma^2}}+\frac{3\sigma}{8\sqrt{2\pi}}e^{\frac{-1}{2\sigma^2}}-\frac{5\sigma}{16\sqrt{2\pi}}e^{\frac{-1}{2\sigma^2}}\\ +\frac{3\sigma^2}{16}\left[1+\Pr\left(\chi^2{}_{(1)}<\frac{1}{\sigma^2}\right)\right]-\frac{10\sigma^3}{26\sqrt{2\pi}}e^{\frac{-1}{2\sigma^2}}+\dots\end{array}\right]$$

$$E(Y)=\frac{1}{\left[1-\Phi\left(\frac{-1}{\sigma}\right)\right]}\left[\begin{array}{l}\left[1-\Phi\left(\frac{-1}{\sigma}\right)\right]-\frac{\sigma\left(8e^{\frac{-1}{2\sigma^2}}+6e^{\frac{-1}{2\sigma^2}}+5e^{\frac{-1}{2\sigma^2}}\right)}{16\sqrt{2\pi}}\\ +\frac{3\sigma^2}{16}\left[1+\Pr\left(\chi^2{}_{(1)}<\frac{1}{\sigma^2}\right)\right]-\frac{10\sigma^3}{16\sqrt{2\pi}}e^{\frac{-1}{2\sigma^2}}+\dots\end{array}\right]$$

$$\begin{aligned}\therefore E(Y)=1&-\frac{19\sigma e^{\frac{-1}{2\sigma^2}}}{16\sqrt{2\pi}\left[1-\Phi\left(\frac{1}{\sigma}\right)\right]}+\frac{3\sigma^2}{16\left[1-\Phi\left(\frac{1}{\sigma}\right)\right]}\left[1+\Pr\left(\chi^2{}_{(1)}<\frac{1}{\sigma^2}\right)\right]\\ &-\frac{10\sigma^3 e\frac{-1}{e^{2\sigma^2}}}{16\sqrt{2\pi}\left[1-\Phi\left(\frac{1}{\sigma}\right)\right]}+\dots\end{aligned}\qquad(20)$$

To find the variance, first obtain the second moment;

$$E(Y^2) = \int_0^\infty y^2 g(y)dy$$

$$= \frac{2}{\sigma\sqrt{2\pi}\left[1-\Phi\left(\frac{-1}{\sigma}\right)\right]} \int_0^\infty \frac{1}{y} e^{\frac{-1}{2\sigma^2}\left(\frac{1}{y^2}-1\right)^2} dy$$

let $u = \frac{1}{y^2}$ then $du = \frac{-2}{y^3}dy,$ and $dy = \frac{-du}{2u^{\frac{3}{2}}}$ for $\infty < u < 0$

$$\therefore E(Y^2) = -k\int_\infty^0 u^{\frac{1}{2}} e^{\frac{-1}{2\sigma^2}(u-1)^2} \frac{du}{2u^{\frac{3}{2}}} = \frac{k}{2}\int_0^\infty u^{-1} e^{\frac{-1}{2}\left(\frac{u-1}{\sigma}\right)^2} du \qquad (21)$$

where $k = \frac{2}{\sigma\sqrt{2\pi}\left[1-\Phi\left(\frac{-1}{\sigma}\right)\right]}$

let $z = \frac{u-1}{\sigma}$ then $u = z\sigma + 1$ and $du = \sigma dz$ for $\frac{-1}{\sigma} < z < \infty$

$$\therefore E(Y^2) = .\frac{k}{2}\int_{\frac{-1}{\sigma}}^\infty (1+z\sigma)^{-1} e^{\frac{-z^2}{2}} \sigma dz$$

Using the binomial expansion on $(1+z\sigma)^{-1}$, given in Equation 16 we have

$$(1+z\sigma)^{-1} = 1 - 1z\sigma + 1(z\sigma)^2 - 1(z\sigma)^3 + \ldots$$

$$\therefore E(Y^2) = \frac{\sigma k}{2}\int_{\frac{-1}{\sigma}}^\infty [1 - z\sigma + (z\sigma)^2 - (z\sigma)^3 + \ldots]\, e^{\frac{-z^2}{2}} dz \qquad (22)$$

$$E(Y^2)=\frac{1}{\left[1-\Phi\left(\frac{-1}{\sigma}\right)\right]}\left[\begin{array}{l}\frac{1}{\sqrt{2\pi}}\int\limits_{\frac{-1}{\sigma}}^{\infty}e^{\frac{-z^2}{2}}dz-\frac{\sigma}{\sqrt{2\pi}}\int\limits_{\frac{-1}{\sigma}}^{\infty}ze^{\frac{-z^2}{2}}dz\\+\frac{\sigma^2}{\sqrt{2\pi}}\int\limits_{\frac{-1}{\sigma}}^{\infty}z^2e^{\frac{-z^2}{2}}dz-\frac{\sigma^3}{\sqrt{2\pi}}\int\limits_{\frac{-1}{\sigma}}^{\infty}z^3e^{\frac{-z^2}{2}}dz+\dots\end{array}\right]$$

$$=\frac{1}{\left[1-\Phi\left(\frac{1}{\sigma}\right)\right]}\left[\begin{array}{l}\left[\Pr\left(z>\frac{-1}{\sigma}\right)-\frac{\sigma}{\sqrt{2\pi}}e^{\frac{-1}{2\sigma^2}}+\left(\frac{\sigma^2}{2}-\frac{\sigma e^{\frac{-1}{2\sigma^2}}}{\sqrt{2\pi}}+\frac{\sigma^2}{2}\Pr\left(x^2{}_{(1)}<\frac{1}{\sigma^2}\right)\right)\right]\\-\left(\frac{\sigma e^{\frac{-1}{2\sigma^2}}}{\sqrt{2\pi}}-\frac{2\sigma^3e^{\frac{-1}{2\sigma^2}}}{\sqrt{2\pi}}\right)+\dots\end{array}\right]$$

$$=\frac{1}{\left[1-\Phi\left(\frac{-1}{\sigma}\right)\right]}\left[\begin{array}{l}\left(1-\Phi\left(\frac{-1}{\sigma}\right)\right)-\frac{3\sigma}{\sqrt{2\pi}}e^{\frac{-1}{2\sigma^2}}\\+\frac{\sigma^2}{2}\left[1+\Pr\left(x^2{}_{(1)}<\frac{1}{\sigma^2}\right)\right]-\frac{2\sigma^3e^{\frac{-1}{2\sigma^2}}}{\sqrt{2\pi}}+\dots\end{array}\right]$$

$$\begin{aligned}E(Y^2)=1&-\frac{3\sigma}{\sqrt{2\pi}\left[1-\Phi\left(\frac{-1}{\sigma}\right)\right]}e^{\frac{-1}{2\sigma^2}}+\frac{\sigma^2}{2\left[1-\Phi\left(\frac{-1}{\sigma}\right)\right]}\left[1+\Pr\left(\chi^2{}_{(1)}<\frac{1}{\sigma^2}\right)\right]\\&-\frac{2\sigma^3e^{\frac{-1}{2\sigma^2}}}{\sqrt{2\pi}\left[1-\Phi\left(\frac{-1}{\sigma}\right)\right]}\end{aligned}\tag{23}$$

Observe the following:

1. Subsequent terms in series (20) and (23) for $E(Y)$ and $E(Y^2)$ respectively all have $e^{\frac{-1}{2\sigma^2}}$ as a factor.

2. $e^{\frac{-1}{2\sigma^2}} = 0$ for $\sigma \le 0.22$ correct to 4 decimal places. (See Table 7, column 3)
3. Conditions (1) and (2) imply that all subsequent terms for $E(Y)$ and $E(Y^2)$ are all zeros for $\sigma \le 0.22$.

Thus, without loss of generality

$$E(Y) = 1 + \frac{3\sigma^2}{16\left[1-\Phi\left(\frac{-1}{\sigma}\right)\right]}\left[1+\Pr\left(\chi^2_{(1)} < \frac{1}{\sigma^2}\right)\right] \text{ for } \sigma \le 0.22 \qquad (24)$$

and

$$E(Y^2) = 1 + \frac{\sigma^2}{2\left[1-\Phi\left(\frac{-1}{\sigma}\right)\right]}\left[1+\Pr\left(\chi^2_{(1)} < \frac{1}{\sigma^2}\right)\right] \text{ for } \sigma \le 0.2 \qquad (25)$$

thus

$$Var(Y) = E(Y^2) - [E(Y)]^2$$

$$Var(Y) = 1 + \frac{\sigma^2}{2\left[1-\Phi\left(\frac{-1}{\sigma}\right)\right]}\left[1+\Pr\left(\chi^2_{(1)} < \frac{1}{\sigma^2}\right)\right] - \left[1 + \frac{3\sigma^2}{16\left[1-\Phi\left(\frac{-1}{\sigma}\right)\right]}\left[1+\Pr\left(\chi^2_{(1)} < \frac{1}{\sigma^2}\right)\right]\right]^2$$

$$= \frac{\sigma^2}{8\left[1-\Phi\left(\frac{-1}{\sigma}\right)\right]}\left[1+\Pr\left(\chi^2_{(1)} < \frac{1}{\sigma^2}\right)\right] - \left[\frac{3\sigma^2}{16\left[1-\Phi\left(\frac{-1}{\sigma}\right)\right]}\left[1+\Pr\left(\chi^2_{(1)} < \frac{1}{\sigma^2}\right)\right]\right]^2 \tag{26}$$

Numerical computations of mean and variance of $Y\left(=e_t^*\right)$

Now compute the values of *E*(*Y*) and Var(*Y*) for $\sigma \in [0.01, 0.22]$ using the functional expressions obtained in Equations 24 and 26, respectively. Table 7 shows the computations of *E*(*Y*) and Var(*Y*). For these computations we write

$$E\left(Y\right) = 1 + \frac{3\sigma^2 B}{8\left(2A\right)} \qquad \sigma < 0.22$$

and

$$Var\left(Y\right) = \frac{\sigma^2 B}{8A} - \left(\frac{3\sigma^2 B}{16A}\right)^2$$

where $\mathrm{A} = 1 - \Phi\left(-\frac{1}{\sigma}\right)$ and $\mathrm{B} = 1 + \Pr\left(\chi^2_{(1)} < \frac{1}{\sigma^2}\right)$

From Table 7, columns 4 and 5, A = 1 and B = 2 for σ <0.22

$$\therefore E\left(Y\right) = 1 + \frac{3\sigma^2}{8} \qquad \sigma < 0.22 \tag{27}$$

and

$$Var\left(Y\right) = \frac{\sigma^2}{4} - \left(\frac{3\sigma^2}{8}\right)^2 \qquad \sigma < 0.22 \tag{28}$$

Equation 27 is the relationship observed with simulated data in Tables 3-6.

Results

The following results were obtained from the investigations carried out on the pdf of $e_t^*\left(=\frac{1}{\sqrt{e_t}}\right)$, $g(y)$ where $e_t \sim N\left(1,\sigma^2\right)$, left truncated at 0.

i. The curve shapes are bell-shaped, with mode $\approx$ mean ≈ 1 when $0 < \sigma \leq 0.145$ correct to 1 decimal place.

Using simulated data, whenever $\sigma < 0.15$

ii. Median $\approx$ Mean ≈ 1

iii. $E\left(e_t^*\right)=1+\frac{3}{8}\sigma^2$

iv. $\frac{Var\left(e_t\right)}{Var\left(e_t^*\right)}=4$, thus $\text{var}(e_t^*) \approx \frac{1}{4}Var(e_t)$

v. e_t^* is normally distributed when $\sigma \leq 0.14$. It was observed that the normality of a pdf curve at a point b implied normality at points $0 < a \leq b \in \Re$.

Using the functional expressions for mean and variance of e_t^*

vi. $E\left(e_t^*\right)=1+\frac{3}{8}\sigma^2 \qquad \sigma \leq 0.22$

≈ 1 correct to 2 decimal places (dp) when $\sigma \leq 0.11$

correct to 1 dp when $\sigma \leq 0.22$

vii. $Var(e_t^*)=\frac{\sigma^2}{4}-\left(\frac{3\sigma^2}{8}\right)^2 \quad \sigma \leq 0.22$

viii. $\frac{Var\left(e_t\right)}{Var\left(e_t^*\right)} \approx 4$

correct to 2 dp when $\sigma \leq 0.04$

correct to 1 dp when $\sigma \leq 0.14$

Table 7. Computations of $E(Y)$ & $Var(Y)$ for $\sigma \in [0.01, 0.3]$

σ	σ^2	$e^{\frac{-1}{2\sigma^2}}$	A	B	$E(Y)$	$Var(Y)$	$VarX\,/\,Var\ Y$
0.01	0.0001	0.0000000	1.00000	2.00000	1.00004	0.0000250	4.00023
0.02	0.0004	0.0000000	1.00000	2.00000	1.00015	0.0001000	4.00090
0.03	0.0009	0.0000000	1.00000	2.00000	1.00034	0.0002249	4.00203
0.04	0.0016	0.0000000	1.00000	2.00000	1.00060	0.0003996	4.00360
0.05	0.0025	0.0000000	1.00000	2.00000	1.00094	0.0006241	4.00563
0.06	0.0036	0.0000000	1.00000	2.00000	1.00135	0.0008982	4.00812
0.07	0.0049	0.0000000	1.00000	2.00000	1.00184	0.0012216	4.01106
0.08	0.0064	0.0000000	1.00000	2.00000	1.00240	0.0015942	4.01445
0.09	0.0081	0.0000000	1.00000	2.00000	1.00304	0.0020158	4.01831
0.10	0.0100	0.0000000	1.00000	2.00000	1.00375	0.0024859	4.02263
0.11	0.0121	0.0000000	1.00000	2.00000	1.00454	0.0030044	4.02741
0.12	0.0144	0.0000000	1.00000	2.00000	1.00540	0.0035708	4.03266
0.13	0.0169	0.0000000	1.00000	2.00000	1.00634	0.0041848	4.03839
0.14	0.0196	0.0000000	1.00000	2.00000	1.00735	0.0048460	4.04459
0.15	0.0225	0.0000000	1.00000	2.00000	1.00844	0.0055538	4.05127
0.16	0.0256	0.0000000	1.00000	2.00000	1.00960	0.0063078	4.05844
0.17	0.0289	0.0000000	1.00000	2.00000	1.01084	0.0071075	4.06610
0.18	0.0324	0.0000002	1.00000	2.00000	1.01215	0.0079524	4.07425
0.19	0.0361	0.0000010	1.00000	2.00000	1.01354	0.0088417	4.08291
0.20	0.0400	0.0000037	1.00000	2.00000	1.01500	0.0097750	4.09207
0.21	0.0441	0.0000119	1.00000	2.00000	1.01654	0.0107515	4.10175
0.22	0.0484	0.0000326	1.00000	1.99999	1.01815	0.0117706	4.11195
0.23	0.0529	0.0000785	0.99999	1.99999	1.01984	0.0128315	4.12268
0.24	0.0576	0.0001699	0.99998	1.99997	1.02160	0.0139334	4.13394
0.25	0.0625	0.0003355	0.99997	1.99994	1.02344	0.0150757	4.14575
0.26	0.0676	0.0006134	0.99994	1.99988	1.02535	0.0162574	4.15811
0.27	0.0729	0.0010503	0.99989	1.99979	1.02734	0.0174777	4.17104
0.28	0.0784	0.0016993	0.99982	1.99964	1.02940	0.0187356	4.18454
0.29	0.0841	0.0026181	0.99972	1.99944	1.03154	0.0200304	4.19862
0.30	0.0900	0.0038659	0.99957	1.99914	1.03375	0.0213609	4.21330

From the probability density curves, the results obtained from simulated data and the functional expressions for the mean and variance, $\sigma \leq 0.14$ (intersecting region) is the recommended condition for successful inverse square root transformation.

The results of this investigation together with findings from similar investigations with respect to the error term $e_t \sim N\left(1, \sigma^2\right)$ under other types of

Table 8. Summary of this and similar findings with respect to the error term e_t ~ $N\left(1,\sigma^2\right)$ under different transformations

e_t^*	Distribution of e_t^*	Condition for successful transformation	Relationship between σ and σ_1
$\log_e e_t$	$e_t^* \sim N\left(0,\sigma_1^2\right)$	$\sigma < 0.1$	$\sigma_1 \approx \sigma$
$\frac{1}{e_t}$	$e_t^* \sim N\left(1,\sigma_1^2\right)$	$\sigma \le 0.1$	$\sigma_1 \approx \sigma$
$\sqrt{e_t}$	$e_t^* \sim N\left(1,\sigma_1^2\right)$	$\sigma \le 0.59$	$\sigma_1 \approx \frac{1}{2}\sigma$
e_t^2	$e_t^* \sim N\left(1,\sigma_1^2\right), \sigma_1^2 = 1$	$\sigma \le 0.027$	$\sigma_1 > \sigma$
$\frac{1}{\sqrt{e_t}}$	$e_t^* \sim N\left(1,\sigma_1^2\right)$	$\sigma \le 0.14$	$\sigma_1 \approx \frac{1}{2}\sigma$

Conclusion

From the results of the investigations of the distributions of the error term $\left(e_t\right)$ of the multiplicative time series model and its inverse square root transformed error term $\left(e_t^*\right)$, it is clear that the condition for successful inverse square root transformation is $\sigma < 0.14$. This is because the two stochastic processes e_t and e_t^* are normally distributed with mean 1, but with the variance of inverse square root transformed error term being one quarter of the variance of the untransformed error component whenever $\sigma < 0.14$, outside this region transformation is not advisable since the basic assumption on the error term are violated after the transformation. This relationship between the two variances, $Var\left(e_t^*\right) \approx \frac{1}{4}\ Var\left(e_t\right)$, agrees with findings of Otuonye et al. (2011) under square root transformation, however the region of successful transformation obtained is closer to the region obtained for the logarithmic and inverse transformations by Iwueze (2007) and Nwosu et al. (2013).

References

Akpanta, A. C. & Iwueze I. S. (2009). On applying the Bartlett transformation method to time series data. *Journal of Mathematical Sciences, 20*(3), 227-243.

Bartlett, M. S. (1947). The use of transformations. *Biometrica, 3*(1), 39-52. doi:10.2307/3001536

Chatfield, C. (2004). *The analysis of time series: An introduction* (6th ed.). London: Chapman & Hall/CRC.

Freund, J. E & Walpole, R. E (1986). *Mathematical statistics* (4th ed.). Upper Saddle River, NJ: Prentice-Hall, Inc.

Iwu, H., Iwueze, I. S., & Nwogu, E. C. (2009). Trend analysis of transformations of the multiplicative time series model. *Journal of Nigerian Statistical Association*, *21*, 40-54.

Iwueze, I. S. (2007). Some implications of truncating the $N\left(1,\sigma^2\right)$ distribution to the left at zero. *Journal of Applied Science, 7*(2), 189-195.

Iwueze, I. S., Akpanta, A. C., & Iwu, H. C. (2008). Seasonal analysis of transformations of the multiplicative time series model. *Asian Journal of Mathematics and Statistics 1*(2), 80-89. doi:10.3923/ajms.2008.80.89

Nwosu, C.R., Iwueze, I.S., & Ohakwe J. (2013). Condition for successful inverse transformation of the error component of the multiplicative time series model. *Asian Journal of Applied Science 6*(1), 1-15. doi:10.3923/ajaps.2013.1.15

Ohakwe, J., Iwuoha, O., & Otuonye, E. L. (2013). Condition for successful square transformation in time series modeling. *Applied Mathematics*, *4*(4), 680-687. doi:10.4236/am.2013.44093

Osborne, J. (2002). Notes on the use of data transformations. *Practical Assessment Research and Evaluation*, *8*(6). Available online: http://PAREonline.net/getvn.asp?v=8&n=6

Osborne, J. W. (2010). Improving your data transformation. *Practical Assessment Research & Evaluation*, *15*(12). Available online: http://pareonline.net/getvn.asp?v=15&n=12

Otuonye, E. L., Iwueze, I. S., & Ohakwe, J. (2011). The effect of square root transformation on the error component of the multiplicative time series model. *International Journal of Statistics and Systems*, *6*(4), 461-476.

Roberts, S. (2008). Transform your data. *Nutrition*, *24*, 492-494. doi:10.1016/j.nut.2008.01.004

Smith, R. T., & Minton, R. B. (2008) *Calculus* (3rd ed.). NY: McGraw Hill.

Turkey, J. W. (1957). On the comparative anatomy of transformations. *The Annals of Mathematical Statistics, 28*(3), 602-632. doi:10.1214/aoms/1177706875

Uche, P. I (2003). *Probability: Theory and applications*. Ikega, Lagos: Longman Nigeria PLC.

Watthanacheewakul, L. (2012). Transformation with right skew data. In S. I. Ao, L. Gelmen, D. W. L. Hukins, A. Hunter and A. M. Korsunsky, Eds. *Proceedings of the World Congress on Engineering Vol. 1*. London, UK: Newswood Limited.

Journal of Modern Applied Statistical Methods
November 2015, Vol. 14, No. 2, 201-218.

ISSN 1538 − 9472

The Bayes Factor for Case-Control Studies with Misclassified Data

Tzesan Lee
Centers for Disease Control & Prevention
Atlanta, GA

The question of how to test if collected data for a case-control study are misclassified was investigated. A mixed approach was employed to calculate the Bayes factor to assess the validity of the null hypothesis of no-misclassification. A real-world data set on the association between lung cancer and smoking status was used as an example to illustrate the proposed method.

Keywords: Bayes factor, Misclassification, *p*-value.

Introduction

Misclassification is a ubiquitous problem in epidemiologic studies. Particularly, it often occurs if the data are obtained from the proxy or surrogate (Nelson, Longstreth, Koesell, and van Belle 1990). Methods for dealing with misclassified data from case-control studies have been widely studied. See, for example, Kleinbaum, Kupper & Morgenstern (1982), Fleiss, Levin & Paik (2003), and Rothman, Greenland & Lash (2008). Almost all studies make an assumption in the beginning that the collected data are misclassified. Yet how to test the validity of this assumption has not been addressed.

These issues can also be considered from a Bayesian perspective. First, the misclassification probabilities are included in both the null and alternative hypothesis. Second, bias-adjusted estimators for the proportion of exposure in cases or controls are presented. Third, the uniform and the Beta distributions are adopted respectively as the prior distribution for the misclassification probability and population proportion parameter in cases or controls. Finally, the lower-bound for the Bayes factor is calculated. A real-world data set was used as an example to illustrate the proposed method. A comparison between the *p*-value and the Bayes factor is made.

Tzesan Lee was retired from the Centers for Disease Control and Prevention and is currently working as President, Applied Math Press, LLC. Email at leetzesan@gmail.com.

Methodology

Consider the data for case-control studies given in Table 1. The random variable E^* denotes the classified surrogate for the true exposure variable E, while the variable D indicates the disease status of the subjects with $D = 1$ and $D = 0$ representing cases and controls respectively. Suppose that E^* is misclassified, but D is not misclassified.

Table 1. Case-control studies with misclassified data

Classified exposure status	**Group of subjects**	
	***D* = 1 (cases)**	***D* = 0 (controls)**
E^* = 1 (exposed)	n_{11}	n_{10}
E^* = 0 (unexposed)	n_{01}	n_{00}
Sample size	$n_{[1]}$	$n_{[0]}$

It is well known that the traditional sample proportion estimator of the exposed group given by

$$\hat{p}_i = n_{ji} \big/ n_{[i]}, \hat{q}_i = 1 - \hat{p}_i \tag{1}$$

In terms of the sensitivity and specificity defined by

$$\varphi_i = \Pr\left(E^* = 1 \middle| E = 1, D = i\right), \bar{\varphi}_i = 1 - \varphi_i \tag{2}$$

$$\psi_i = \Pr\left(E^* = 1 \middle| E = 0, D = i\right), \bar{\psi}_i = 1 - \psi_i \tag{3}$$

it was shown (Lee, 2009) that

$$E\left(\hat{p}_i\right) = \varphi_i p_i + \left(1 - \psi_i\right) q_i = p_i \cdot \Delta_i + 1 - \psi_i \tag{4}$$

$$E\left(\hat{q}_i\right) = \left(1 - \varphi_i\right) p_i + \psi_i q_i = q_i \cdot \Delta_i + 1 - \varphi_i \tag{5}$$

From Equations 4 and 5 it is seen that the traditional sample proportion estimators, $\hat{p}_i$ and $\hat{q}_i$, are no longer unbiased. By solving Equations 4 and 5 with the left-side $E(\hat{p}_i)$ or $E(\hat{q}_i)$ being replaced by $\hat{p}_i$ or $\hat{q}_i$, it follows

$$\breve{p}_i = (\psi_i - \hat{q}_i)/\Delta_i, \tag{6}$$

$$\breve{q}_i = (\varphi_i - \hat{p}_i)/\Delta_i, \tag{7}$$

where

$$\Delta_i = \varphi_i + \psi_i - 1, \; i \; = \; 0, \, 1. \tag{8}$$

Equations 6 and 7 are called the bias-adjusted proportion (BAP) estimators of p_i and q_i. The BAP estimators are said to be admissible if they are greater than zero but less than one plus their sum equals to one. Evidently, the following constraints are required to be imposed on the sensitivity and specificity in order for Equations 6 and 7 to be admissible (Lee, 2009):

$$\begin{aligned} \varphi_i &> \hat{p}_i, \\ \psi_i &> \hat{q}_i, \\ \varphi_i + \psi_i &> 1. \end{aligned} \tag{9}$$

A concern is aimed at testing whether the given data in Table 1 are misclassified - whether the exposure rates for cases and control are the same. This can be tested through the hypothesis testing which is formulated as follows:

$$H_0 : \varepsilon_{RD} = 0 \quad \text{versus} \quad H_1 : \varepsilon_{RD} \neq 0, \tag{10}$$

where $\varepsilon_{RD} = p_1 - p_0$, the subscript "*RD*" means the rate difference. However, Equation 10 can't be used to test whether the observed data of Table 1 are misclassified. In order to test if the data are misclassified, the hypotheses of Equation 10 has to be enlarged by including the misclassification probabilities associated with both cases and controls given as follows:

$$H_0 : \varepsilon_{RD} = 0, \bar{\varphi}_i = \bar{\psi}_i = 0 \text{ versus } H_1 : \varepsilon_{RD} \neq 0, \bar{\varphi}_i \neq 0, \bar{\psi}_i \neq 0, \; i \; = \; 0, \, 1, \tag{11}$$

To test the hypotheses of Equation 11, a mixed Bayesian approach is taken to tackle this problem (Kass & Raftery, 1995).

Let

$$\breve{\varepsilon}_{RD} = \breve{p}_1 - \breve{p}_0 - \varepsilon_{RD} \tag{12}$$

It can be shown

$$E\left(\breve{\varepsilon}_{RD}\right) = 0, \tag{13}$$

$$\begin{aligned} Var\left(\breve{\varepsilon}_{RD}\right) &= Var\left(\breve{p}_1\right) + Var\left(\breve{p}_0 + \varepsilon_{RD}\right) \\ &= \sum_{i=0}^{1}\left(p_i \cdot \Delta_i + 1 - \psi_i\right)\left(q_i \cdot \Delta_i + 1 - \varphi_i\right) \cdot n_{[i]}^{-1} \end{aligned} \tag{14}$$

Define

$$\breve{x}_{RD} = \breve{\varepsilon}_{RD}^{2} \big/ Var\left(\breve{\varepsilon}_{RD}\right) \tag{15}$$

To assess the evidence in favor of supporting the null against the alternative hypothesis of Equation 11, the Bayes factor for favoring H_0 relative H_1 from using Equation 15 can be calculated as follows:

$$B^{g}\left(\breve{x}_{RD}\right) = \frac{f\left(\breve{x}_{RD} \middle| H_0\right)}{m_g\left(\breve{x}_{RD}\right)} \tag{16}$$

where

$$m_g\left(\breve{x}_{RD}\right) = \iint_{R\times\Omega} f\left(\breve{x}_{RD} \middle| H_1\right) \prod_{i=0}^{1} h_0\left(\varphi_i, \psi_i\right) g\left(p_i, q_i\right) d\varphi_i d\psi_i dp_i dq_i \tag{17}$$

$f\left(\breve{x}_{RD} \mid H_1\right)$ is the central chi-square distribution with one degree of freedom, $g\left(p_i, q_i\right) = \Gamma\left(\eta + \tau\right) p_i^{\eta-1} q_i^{\tau-1} \big/ \left[\Gamma\left(\eta\right)\Gamma\left(\tau\right)\right]$, the beta distribution with the parameters η and τ over [0, 1], and $h_0\left(\varphi_i, \psi_i\right) = \left[\bar{\varphi}_i \bar{\psi}_i\right]^{-1}$ is the uniform distribution over $\Omega_i = [a_i ,1] \times [b_i,1]$, where a_i and b_i are specified in the Appendix. Although the posterior marginal probability density function of m_g (Equation 17) depends on two hyper-parameters η and τ, a Bayes/non-Bayes compromise rather than a type III hyper-distribution for η and τ is adopted to estimate η and τ (Good & Crook, 1974). As a result, the parameters η and τ are estimated by employing the likelihood method. The maximum likelihood estimators for η and τ and the relative maximum value of m_g of Equation 17 are denoted respectively by $\left(\eta_{\max}, \tau_{\max}\right)$ and $m_g^{\max} = m_g\left(\eta_{\max}, \tau_{\max}\right)$. Thus, define the lower bound of the Bayes factor (Equation 16) as follows:

$$\underline{B}^g = f\left(\breve{x}_{RD} \middle| H_0\right) \big/ m_g^{\max} \tag{18}$$

The details of calculating Equation 18 are given in the Appendix.

Example

Although there is some evidence of a greater than average risk in some occupations to have the lung cancer, these occupations could not account for the general increase in pulmonary cancer. It is thought of interest to select a particular population group, homogeneous economically, with little occupational exposure to respiratory irritants and with equal access to diagnostic facilities. Physicians are believed to represent such a group. Wynder and Cornfield (1953) reported a study on the exposure to tobacco and other possible respiratory irritants of 63 physicians with lung cancer and 133 physicians with cancers in areas where respiratory irritants are not believed to play a part. Among these 133 physicians, 43 cases were cancer of stomach and kidney, 45 cases cancer of colon and lymphoma, and 45 cases cancer of bladder, leukemia and sarcoma. The data in Table 2 is taken from Cornfield (1956) who only used 43 cases from cancer of stomach and kidney as a control group. The non-smoker is defined to be those who smoked the equivalent of less than 1 cigarette a day. Here it is of interest to test whether the data concerning the smoking status in Table 2 for both cases and controls are misclassified.

Table 2. The data of physicians with and without lung cancer by smoking status

Smoking status	Lung cancer patients	Controls
Smoker	60	32
Nonsmoker	3	11
Total	63	43

Before calculating the Bayes factor, the data in Table 2 are first to be checked if the two required conditions are satisfied before using the formula derived in the Appendix. Because $\hat{p}_1 = 0.952381 > \hat{p}_0 = 0.744186$ and $\hat{\sigma}_{\hat{p}_1} = \sqrt{n_{[1]}^{-1}\hat{p}_1\hat{q}_1} = 0.027 > \hat{\sigma}_{\hat{p}_0} = \sqrt{n_{[0]}^{-1}\hat{p}_0\hat{q}_0} = 0.067$, where $n_{[1]} = 63$, $n_{[0]} = 43$, the two required conditions are indeed being satisfied; hence it was free to use the formula in the Appendix. Let $a_i = \hat{p}_i + 0.005$ and $b_i = \hat{q}_i + 0.005$, $i = 0,1$, be substituted into Equations A17 to A11, it follows that $\dot{M}_{[1,1,0,0]} = 1.1011$, $M_{[1,0,1,0]} = 0.0828$, $M_{[1,0,0,1]} = -0.0037$, $\dot{M}_{[1,1,0,1]} = 0.0513$, $M_{[1,0,1,1]} = 1.2369$, $\dot{M}_{[0,1,0,0]} = 1.1169$, $M_{[0,0,1,0]} = 0.6287$, $M_{[0,0,0,1]} = -0.0567$, $\dot{M}_{[0,1,0,1]} = 0.4819$, and $M_{[0,0,1,1]} = 4.8652$. Then, substituting the above information into Equations A12 and A14, this leads to that $N_0 = 0.1957$, $N_1 = 5.4652$, $N_2 = -31.4597$, $R_0 = 0.0016$, $R_1 = 0.1967$, $R_2 = -0.0041$, $R_3 = 0.0704$, $R_4 = 0.234$, $R_5 = -0.0252$, $R_6 = -0.1988$, and $a = 133.5876$. Again, by substituting the above information into Equations A13 and A16, it follows that

$$m_g^{(1)}(\eta,\tau) \equiv \frac{-400.8(\eta+\tau)\left\{\begin{array}{l}\eta\tau(\eta+\tau)\left[0.003\eta(\eta+\tau)-0.002\right] \\ +0.017\eta\tau(\eta+\tau)+0.002\eta\left.\right]+0.009\tau\end{array}\right\}+5.97\tau}{2\sqrt{\eta\tau}\eta^2(\eta+\tau)^3} \tag{19}$$

and

$$m_g^{(2)}(\eta,\tau) \equiv \frac{2.33\eta\tau(\eta+\tau)^2 + 2.23\tau(\eta+\tau) - 3.82}{2\left[\sqrt{\eta\tau}(\eta+\tau)\right]^3} \tag{20}$$

Consequently, m_g (η, τ) was readily obtained from substituting Equations 19 and 20 into Equation A17.

To find the relative maximum of $m_g(\eta, \tau)$, the 2-dimensional unit square $[0,1] \times [0,1]$ was partitioned into 100 lattice points $(0.1, 0.1), (0.1, 0.2), \ldots, (1.0, 0.9), (1.0, 1.0)$ and then evaluated the function value of $m_g(\eta, \tau)$ at these lattice points. After identifying the proximity of the relative maximum a finer neighborhood was then searched to locate it. Equation A17 was found to have a unique relative maxima: $m_g^{\max}(0.14,1.0) = 2.15$. The value of $f(\breve{x}_{RD} \mid H_0)$ was evaluated directly from the probability density function of the central chi-square distribution with one degree of freedom; hence we have $f(\breve{x}_{RD} \mid H_0) = 6.4 \times 10^{-6}$. After dividing the value of $f(\breve{x}_{RD} \mid H_0) = 6.4 \times 10^{-6}$ by $m_g^{\max} = 2.15$, we thus obtained the lower bound of the Bayes factor given by $\underline{B}^g(\breve{x}_{RD}) = 3.0 \times 10^{-6}$.

Since $\breve{x}_{RD} \big| H_0 = \hat{x}_{RD} = \hat{p}_D^2 \big/ Var(\hat{p}_D) = 19.1$ (p-value = 1.2×10^{-5}), where $\hat{p}_D \equiv \hat{p}_1 - \hat{p}_0$, the null hypothesis H_0 was rejected for Table 2. Yet, the evidence from the lower bound of the Bayes factor ($\underline{B}^g(\breve{x}_{RD})$ = 3.0×10^{-6}) was in favor of supporting H_1 (Equation 11) by at most a factor of "3.3×10^5 to 1". Hence the data in Table 2 are likely to be misclassified.

Discussion

Although both the p-value and the Bayes factor rejected the null hypothesis H_0 with respect to the data in Table 2, the p-value seemed much inclined to reject the null hypothesis H_0 in Equation 10 rather than that in Equation 11. In other words, the p-value is inadequate to reject the null hypothesis in Equation 11. This study provides another example to corroborate the p-value fallacy (Goodman 1999a, Goodman 1999b).

Because the Beta distribution which is the conjugate family of the binomial distribution was used as the prior distributions, the Bayes factor could of course change accordingly if other family of distributions is used as the prior distribution (Delampady & Berger, 1990).

The derivation of the formula provided in the Appendix was based on the two assumptions: (i) $p_1 > p_0$, and (ii) $\sigma_{\hat{p}_1} = \sqrt{n_{[1]}^{-1} p_1 q_1} < \sigma_{\hat{p}_0} = \sqrt{n_{[0]}^{-1} p_0 q_0}$. These two assumptions can be verified if it is valid by substituting the crude prevalence estimator ($\hat{p}_i$, $i = 0, 1$) into the inequality. Should the both of the two assumptions fail to be satisfied, all we need to do is to switch the index accordingly for cases and controls before using the formula provided in the Appendix. However, if only one of the assumptions is violated, Equation A4 has to be revised accordingly.

References

Askey, R. A. & Roy, R. (2010). Gamma function, *NIST handbook of mathematical functions* (pp.135). F. W. Olver, D. W. Lozier, R. F. Boisvert, C. W. Clark (Eds.), United Kingdom: National Institute of Standards and Technology, Cambridge University Press.

Cornfield, J. (1956). A statistical problem arising from retrospective studies. In *Proceedings of the third Berkeley symposium on mathematical statistics and probability* (Vol. 4, pp. 135-148). Berkeley: University of California Press.

Delampady, M. & Berger, J. O. (1990). Lower bounds on Bayes factors for multinomial distributions with application to chi-squared tests of fit. *The Annals of Statistics*, *18*(3), 1295-1316. doi:10.1214/aos/1176347750

Fleiss, J., Levin, B. & Paik, M. C. (2003). *Statistical methods for rates and proportions* (3rd ed.). New York: John Wiley & Sons.

Good, I. J. & Crook, J. F. (1974). The Bayes/non-Bayes compromise and the multinomial distribution. *Journal of American Statistical Association*, *69*(347), 711-720. doi:10.2307/2286006

Goodman, S. N. (1999a). Toward evidence-based medical statistics. 1: The P-value fallacy, *Annals of Internal Medicine*, *130*(12), 995-1004. doi:10.7326/0003-4819-130-12-199906150-00008

Goodman, S. N. (1999b). Toward evidence-based medical statistics. 2: The Bayes factor, *Annals of Internal Medicine*, *130*(12), 1005-1013. doi:10.7326/0003-4819-130-12-199906150-00019

Kass, R. E. and Raftery, A. E. (1995). Bayes factor, *Journal of American Statistical Association*, *90*(430), 773-795. doi:10.1080/01621459.1995.10476572

Kleinbaum, D. G., Kupper, L. L. & Morgenstern, H. (1982). *Epidemiologic research: Principles and quantitative methods*. Belmont, CA: Lifetime Learning.

Lee, T-S. (2009). Bias-adjusted exposure odds ratio for misclassified data, *The Internet Journal of Epidemiology*, *6*(2), 1-19.

Mietinen, O. & Nurminen, M. (1985). Comparative analysis of two rates. *Statistics in Medicine*, *4*(2), 213-226. doi:10.1002/sim.4780040211

Nelson, L. M., Longstreth, W. T., Koesell, T. D., & van Belle, G. (1990). Proxy respondents in epidemiologic research. *Epidemiologic Reviews*, *12*(1), 71-86.

Rothman, K. J., Greenland, S. & Lash, T. L. (2008). *Modern epidemiology* (3rd ed.). Philadelphia, PA: Lippincott Williams & Wilkins.

Wynder, E. L. & Cornfield, J. (1953). Cancer of the lung in physicians, *New England Journal of Medicine*, *248*(11), 441-444. doi:10.1056/NEJM195303122481101

Appendix

By applying the quadratic approximation to the probability density function of the central chi-square distribution with one degree of freedom in Equation 17, we have

$$\begin{aligned} f\left(\breve{x}_{RD} \middle| \varepsilon_{RD}, \varphi_0, \psi_0, \varphi_1, \psi_1\right) &= \frac{1}{\sqrt{2\pi}} \breve{x}_{RD}^{-\frac{1}{2}} e^{-\frac{1}{2}\breve{x}_{RD}} \\ &\approx \frac{1}{\sqrt{2\pi}} \cdot \frac{1}{\sqrt{\breve{x}_{RD}}} \left(1 - \tfrac{1}{2}\breve{x}_{RD} + \tfrac{1}{8}\breve{x}_{RD}^2\right) \\ &= \frac{1}{\sqrt{2\pi}} \left[\frac{\sqrt{Var\left(\breve{\varepsilon}_{RD}\right)}}{\breve{\varepsilon}_{RD}} - \tfrac{1}{2} \frac{\breve{\varepsilon}_{RD}}{\sqrt{Var\left(\breve{\varepsilon}_{RD}\right)}} + \tfrac{1}{8} \left(\frac{\breve{\varepsilon}_{RD}}{\sqrt{Var\left(\breve{\varepsilon}_{RD}\right)}} \right)^3 \right], \end{aligned}$$

(A1)

where $\breve{\varepsilon}_{RD}$ and $Var(\breve{\varepsilon}_{RD})$ are given by Equations 12 and 14, respectively.

By using the linear approximation:

$$\left[\left(1 - \varepsilon_{RD}^{-1}\left(\breve{p}_1 - \breve{p}_0\right)\right]^{-1} \approx 1 + \varepsilon_{Rd}^{-1}\left(\breve{p}_1 - \breve{p}_0\right),$$

it follows that

$$\begin{aligned}
\frac{\sqrt{Var\left(\breve{\varepsilon}_{Rd}\right)}}{\breve{\varepsilon}_{Rd}} &= \frac{\sqrt{\begin{array}{l}\Delta_1^{-2} n_{[1]}^{-1}\left(p_1\Delta_1+1-\psi_1\right)\left(q_1\Delta_1+1-\varphi_1\right)\\ +\Delta_0^{-2} n_{[0]}^{-1}\left(p_0\Delta_0+1-\psi_0\right)\left(q_0\Delta_0+1-\varphi_0\right)\end{array}}}{\breve{p}_1-\breve{p}_0-\varepsilon_{RD}} \\
&= -\frac{\sqrt{A}}{\varepsilon_{RD}}\cdot\frac{\sqrt{1+A^{-1}\left\{\sum_{i=0}^{1} n_{[i]}^{-1}\Delta_i^{-1}\left[\left(1-p_i\varphi_i-q_i\psi_i\right)+\Delta_i^{-1}\overline{\varphi}_i\overline{\psi}_i\right]\right\}}}{1-\varepsilon_{RD}^{-1}\left(\breve{p}_1-\breve{p}_0\right)} \\
&= -I^{-1}\cdot\varepsilon_{RD}^{-1}\cdot\sqrt{1+I^2J}\cdot\left[1-\varepsilon_{RD}^{-1}\left(\breve{p}_1-\breve{p}_0\right)\right]^{-1} \\
&\approx -I^{-1}\cdot\varepsilon_{RD}^{-1}\left(1+\tfrac{1}{2}I^2J\right)\left[1+\varepsilon_{RD}^{-1}\left(\breve{p}_1-\breve{p}_0\right)\right] \\
&= -I^{-1}\cdot\varepsilon_{RD}^{-1}\left[1+\varepsilon_{RD}^{-1}\left(\breve{p}_1-\breve{p}_0\right)+\tfrac{1}{2}I^2J+\tfrac{1}{2}\varepsilon_{RD}^{-1}I^2J\left(\breve{p}_1-\breve{p}_0\right)\right] \\
&= -I^{-1}\cdot\varepsilon_{RD}^{-1}\left\{\begin{array}{l}1+\varepsilon_{RD}^{-1}\left[\Delta_1^{-1}u\left(\varphi_1\right)-\Delta_0^{-1}u\left(\varphi_0\right)\right]+\tfrac{1}{2}I^2J\\ +\tfrac{1}{2}\varepsilon_{RD}^{-1}I^2J\left[\Delta_1^{-1}u\left(\varphi_1\right)-\Delta_0^{-1}u\left(\varphi_0\right)\right]\end{array}\right\}
\end{aligned} \tag{A2}$$

where

$$\begin{aligned}
A &= n_{[1]}^{-1}p_1q_1+n_{[0]}^{-1}p_0q_0 \\
I &= A^{-\frac{1}{2}} \\
J &= \sum_{i=0}^{1}K_i \\
K_i &= n_{[1]}^{-1}\left[\begin{array}{l}\Delta_1^{-1}\left(1-p_i\varphi_i-q_i\psi_i\right)\\ +\Delta_i^{-2}\overline{\varphi}_i\overline{\psi}_i\right)\end{array}\right] = n_{[1]}^{-1}\left[\begin{array}{l}-q_i+\Delta_i^{-1}s\left(\varphi_i\right)\\ +\Delta_i^{-2}t\left(\varphi_i\right)\end{array}\right] \\
s\left(\varphi_i\right) &= q_i\left(2\varphi_i-1\right) \\
t\left(\varphi_i\right) &= \varphi_i\left(1-\varphi_i\right) \\
u\left(\varphi_i\right) &= \hat{p}_i-\varphi_i
\end{aligned} \tag{A3}$$

By using the quadratic approximation on ε_{RD}^{-1}, I^{-1} and I, we have by assuming that $p_1 > p_0$ and $n_{[1]}^{-1}p_1q_1 < n_{[0]}^{-1}p_0q_0$

$$
\begin{aligned}
&\varepsilon_{RD}^{-1} \approx p_1^{-1} + p_0 p_1^{-2} + p_0^2 p_1^{-3} \\
&I^{-1} \approx \frac{1}{\sqrt{n_{[0]}}}\left[\sqrt{p_0 q_0} + \tfrac{1}{2}\frac{n_{[0]}}{n_{[1]}} \cdot \frac{p_1 q_1}{\sqrt{p_0 q_0}} - \tfrac{1}{8}\frac{n_{[0]}^2}{n_{[1]}^2} \cdot \frac{\left(p_1 q_1\right)^2}{\left(p_0 q_0\right)^{\frac{3}{2}}}\right] \\
&I \equiv \frac{1}{\sqrt{A}} \approx \sqrt{\frac{n_{[0]}}{p_0 q_0}}\left[1 - \tfrac{1}{2} \cdot \frac{n_{[0]} p_1 q_1}{n_{[1]} p_0 q_0} + \tfrac{3}{8}\left(\frac{n_{[0]} p_1 q_1}{n_{[1]} p_0 q_0}\right)^2\right]
\end{aligned}
\tag{A4}
$$

For fixed i = 0, 1 let

$$
M_{[i,j,k,l]} \equiv \int_{a_i}^{1}\int_{b_i}^{1}\left[s^j\left(\varphi_i\right)t^k\left(\varphi_i\right)u^l\left(\varphi_i\right)\right]\Big/\Delta_i^{j+2k+l}\,d\psi_i\,d\varphi_i \tag{A5}
$$

where $a_i = \hat{p}_i + 0.005$, $b_i = \hat{q}_i + 0.005$, $s(\varphi_i)$, $t(\varphi_i)$ and $u(\varphi_i)$ are all defined in Equation A3. Let us calculate some of Equation A5 which will be needed later. For j = 1, k = l = 0 we have

$$
\begin{aligned}
M_{[i,1,0,0]} &\equiv \int_{a_i}^{1}\int_{b_i}^{1}\left[s\left(\varphi_i\right)/\Delta_i\right]d\psi_i\,d\varphi_i = \int_{a_i}^{1} s\left(\varphi_i\right)\left[\ln\varphi_i - \ln\left(\varphi_i - \overline{b}_i\right)\right]d\varphi_i \\
&= \int_{a_i}^{1}\left\{\Big[\Big(s'(0)\varphi_i + s(0)\Big]\ln\varphi_i - \begin{bmatrix} s'\left(\overline{b}_i\right)\left(\varphi_i - \overline{b}_i\right) \\ +s\left(\overline{b}_i\right) \end{bmatrix}\ln\left(\varphi_i - \overline{b}_i\right)\right\}d\varphi_i \\
&= q_i \dot{M}_{[i,1,0,0]}
\end{aligned}
\tag{A6}
$$

where $\delta_i = a_i + b_i - 1$, $\overline{b}_i = 1 - b_i$, and

$$
\begin{aligned}
\dot{M}_{[i,1,0,0]} \equiv\ & \delta_i^2 \ln\delta_i - a_i^2 \ln a_i - b_i^2 \ln b_i + a_i \ln a_i \\
& + \left(2\overline{b}_i - 1\right)\left(\delta_i \ln\delta_i - b_i \ln b_i + b_i - \delta_i\right) + \tfrac{1}{2}\left(1 + a_i^2 + b_i^2 - \delta_i^2\right)
\end{aligned}
$$

For j = l = 0, k = 1 we have

$$
\begin{aligned}
M_{[i,0,1,0]} &= \int_{a_i}^{1}\int_{b_i}^{1}\Big[t\left(\varphi_i\right)\big/\Delta_i^2 d\psi_i d\varphi_i \\
&= \int_{a_i}^{1}\left[\begin{array}{l}\dfrac{\frac{1}{2}t''\left(\bar{b}_i\right)\left(\varphi_i-\bar{b}_i\right)^2+t'\left(\bar{b}_i\right)\left(\varphi_i-\bar{b}_i\right)+t\left(\bar{b}_i\right)}{\varphi_i-\bar{b}_i} \\ -\dfrac{\frac{1}{2}t''\left(0\right)\varphi_i^2+t'\left(0\right)\varphi_i+t\left(0\right)}{\varphi_i}\end{array}\right]d\varphi_i \qquad \text{(A7)} \\
&= \sum_{m=1}^{3}\left(d_{m[i,0,1,0]}+e_{m[i,0,1,0]}\right)=b_i\bar{b}_i\ln\left(b_i/\delta_i\right)
\end{aligned}
$$

where $\bar{a}_i=1-a_i$, and

$$
\begin{aligned}
&d_{1[i,0,1,0]}=-\tfrac{1}{2}\bar{a}_i\left(b_i+\delta_i\right), d_{2[i,0,1,0]}=\bar{a}_i\left(1-2\bar{b}_i\right), d_{3[i,0,1,0]}=b_i\bar{b}_i\ln\left(b_i/\delta_i\right), \\
&e_{1[i,0,1,0]}=\tfrac{1}{2}\bar{a}_i\left(1+a_i\right), e_{2[i,0,1,0]}=-\bar{a}_i, e_{3[i,0,1,0]}=0.
\end{aligned}
$$

For $j = k = 0$, $l = 1$ we have

$$
\begin{aligned}
M_{[i,0,0,1]} &\equiv \int_{a_i}^{1}\int_{b_i}^{1}\frac{u\left(\varphi_i\right)}{\Delta_i}d\psi_i d\varphi_i \\
&= -\hat{p}_i a_i\ln a_i+\left(\hat{q}_i-b_i\right)\left(b_i\ln b_i-\delta_i\ln\delta_i\right) \qquad \text{(A8)} \\
&\quad +\tfrac{1}{2}\left(a_i^2\ln a_i+b_i^2\ln b_i-\delta_i^2\ln\delta_i\right) \\
&\quad -\tfrac{1}{4}\left(a_i^2+b_i^2-\delta_i^2-1\right)-\bar{a}_i\bar{b}_i
\end{aligned}
$$

For $j = l = 1$, $k = 0$ we have

$$\begin{aligned}
M_{[i,1,0,1]} &= \int_{a_i}^{1}\int_{b_i}^{1} \frac{s(\varphi_i)}{\Delta_i}\cdot\frac{u(\varphi_i)}{\Delta_i}\, d\psi_i d\varphi_i \\
&\quad -\int_{a_i}^{1}\left[\begin{array}{l} \dfrac{\frac{1}{2}v_1''(\bar{b}_i)(\varphi_i-\bar{b}_i)^2 + v_i'(\bar{b}_i)(\varphi_i-\bar{b}_i)+v(\bar{b}_i)}{\varphi_i-\bar{b}_i} \\ -\dfrac{\frac{1}{2}v''(0)\varphi_i^2 + v'(0)\varphi_i + v(0)}{\varphi_i} \end{array}\right] d\varphi_i \qquad \text{(A9)} \\
&= \sum_{m=1}^{3}\left(d_{m[i,1,0,1]} + e_{m[i,1,0,1]}\right) = q_i \dot{M}_{[i,1,0,1]}
\end{aligned}$$

where

$$\begin{aligned}
v_i(\varphi_i) &= s(\varphi_i)u(\varphi_i) = q_i(2\varphi_i - 1)(\hat{p}_i - \varphi_i), \\
d_{1[i,1,0,1]} &= -q_i\bar{a}_i(b_i+\delta_i), d_{2[i,1,0,1]} = q_i\bar{a}_i\left[1+2(\hat{p}_i-\bar{b}_i)\right], \\
d_{3[i,1,0,1]} &= q_i\bar{b}_i\left[1+2(\hat{p}_i-\bar{b}_i)-\hat{p}_i\right]\ln(b_i/\delta_i), \\
e_{1[i,1,0,1]} &= q_i\bar{a}_i(1+a_i), e_{2[i,1,0,1]} = -q_i(1+2\hat{p}_i)\bar{a}_i, e_{3[i,1,0,1]} = \hat{p}_i q_i \ln a_i, \\
\dot{M}_{[i,1,0,1]} &\equiv \bar{b}_i\left[(1+\hat{p}_i - 2\bar{b}_i)\ln(b_i/\delta_i) + \hat{p}_i \ln a_i\right]
\end{aligned}$$

For $j = 0$, $k = l = 1$ we have

$$\begin{aligned}
M_{[i,0,1,1]} &= \int_{a_i}^{1}\int_{b_i}^{1} \frac{t(\varphi_i)}{\Delta_i^2}\cdot\frac{u(\varphi_i)}{\Delta_i}\, d\psi_i d\varphi_i \\
&= \tfrac{1}{2}\int_{a_i}^{1}\left[\begin{array}{l} \dfrac{\begin{array}{l}\frac{1}{6}v_2'''\left(\bar{b}_i\right)\left(\varphi_i-\bar{b}_i\right)^3+\frac{1}{2}v_2''\left(\bar{b}_i\right)\left(\varphi_i-\bar{b}_i\right)^2\\ +v_2'\left(\bar{b}_i\right)\left(\varphi_i-\bar{b}_i\right)+v_2\left(\bar{b}\right)_i\end{array}}{\left(\varphi_i+\bar{b}_i\right)^2} \\ -\dfrac{\frac{1}{6}v_2'''(0)\varphi_1^3+\frac{1}{2}v_2''(0)\varphi_1^2+v_2'(0)\varphi_1+v_2(0)}{\varphi_1^2}\end{array}\right] d\varphi_i \\
&= \sum_{m=1}^{4}\left(d_{m[i,0,1,1]}+e_{m[i,0,1,1]}\right) \\
&= 2\bar{a}_i\bar{b}_i+\tfrac{1}{2}\left\{\begin{array}{l}\left[3\bar{b}_i^2-2(1+\hat{p}_i)\bar{b}_i+\hat{p}_i\right]\ln\left(b_i/\delta_i\right)\\ +b_i\bar{a}_i\bar{b}_i\left(\hat{p}_i-\bar{b}_i\right)\left(b_i+\delta_i\right)\big/\left(b_i\delta_i\right)^2\end{array}\right\}
\end{aligned} \tag{A10}$$

where

$$\begin{aligned}
&v_2\left(\varphi_i\right)=t\left(\varphi_i\right)u\left(\varphi_i\right)=\varphi\left(1-\varphi_i\right)\left(\hat{p}_i-\varphi_i\right),\\
&d_{1[i,0,1,1]}=\tfrac{1}{4}\bar{a}_i\left(b_i+\delta_i\right), d_{1[i,0,1,1]}=\tfrac{1}{2}\bar{a}_i\left[3\bar{b}_i-\left(1+\hat{p}_i\right)\right],\\
&d_{3[i,0,1,1]}=\tfrac{1}{2}\left[3\bar{b}_i^2-2\left(1+\hat{p}_i\right)\bar{b}_i+\hat{p}_i\right]\ln\left(b_i/\delta_i\right),\\
&d_{4[i,0,1,1]}=\tfrac{1}{2}b_i\bar{b}_i\left(\hat{p}_i-\bar{b}_i\right)\bar{a}_i\left(b_i+\delta_i\right)\big/\left(b_i\delta_i\right)^2,\\
&e_{1[i,0,1,1]}=-\tfrac{1}{4}\bar{a}_i\left(1+a_i\right), e_{2[i,0,1,1]}=\tfrac{1}{2}\left(1+\hat{p}_i\right)\bar{a}_i, e_{3[i,0,1,1]}=\tfrac{1}{2}\hat{p}_i\ln a_i, e_{4[i,0,1,1]}=0.
\end{aligned}$$

Note that in all of the above calculations I first integrate with respect to ψ_i and then integrate with respect to φ_i by employing the Taylor's series expansion to expand the function about $\varphi_i=\bar{b}_i$ or 0.

Now we are ready to calculate the marginal probability density function of Equation A1 one by one

$$\int_{a_0}^{1}\int_{b_0}^{1}\int_{a_1}^{1}\int_{b_1}^{1}\frac{\sqrt{Var\left(\breve{\varepsilon}_{RD}\right)}}{\breve{\varepsilon}_{RD}}\prod_{i=0}^{1}d\psi_i d\varphi_i=-\left\{\begin{array}{l} I^{-1}\varepsilon_{RD}^{-1}\varsigma^{-1}+I^{-1}\varepsilon_{RD}^{-2}\begin{pmatrix}\bar{a}_0\overline{b}_0M_{[1,0,0,1]}\\ -\bar{a}_1\overline{b}_1M_{[0,0,0,1]}\end{pmatrix}\\ +\frac{1}{2}I\varepsilon_{RD}^{-1}\sum_{i=0}^{1}n_{[i]}^{-1}\begin{bmatrix}-\varsigma^{-1}q_i\\ +\bar{a}_{1-i}\overline{b}_{1-i}\begin{pmatrix}q_i\dot{M}_{[i,1,0,0]}\\ +M_{[i,0,1,0]}\end{pmatrix}\end{bmatrix}\\ +\frac{1}{2}I\varepsilon_{Rd}^{-2}n_{[1]}^{-1}\left[-q_1\,\bar{a}_0\overline{b}_0M_{[1,0,0,1]}\right.\\ +\bar{a}_0\overline{b}_0\begin{pmatrix}q_1\dot{M}_{[1,1,0,1]}\\ +M_{[1,0,1,1]}\end{pmatrix}-M_{[0,0,0,1]}\left(q_1\begin{pmatrix}\dot{M}_{[1,1,0,0]}\\ -\bar{a}_1\overline{b}_1\end{pmatrix}\right.\end{array}\right.$$

$$=-\left\{\begin{array}{l} I^{-1}\varepsilon_{RD}^{-1}\varsigma^{-1}+I^{-1}\varepsilon_{RD}^{-2}R_0\\ +\frac{1}{2}I\varepsilon_{Rd}^{-1}\begin{bmatrix}n_{[1]}^{-1}\left(q_1R_1+\bar{a}_0\overline{b}_0M_{[1,0,1,0]}\right.\\ +n_{[0]}^{-1}\left(q_0R_2+\bar{a}_1\overline{b}_1M_{[0,0,1,0]}\right.\end{bmatrix}\\ +\frac{1}{2}I\varepsilon_{RD}^{-2}\left[n_{[1]}^{-1}\left(q_1R_3+R_4\right)\right.\end{array}\right. \qquad \text{(A11)}$$

where

$$\begin{aligned}&R_0=\bar{a}_0\overline{b}_0M_{[1,0,0,1]}-\bar{a}_1\overline{b}_1M_{[0,0,0,1]},R_1\equiv\bar{a}_0\overline{b}_0\dot{M}_{[1,1,0,0]}-\varsigma^{-1},R_2\equiv\bar{a}_1\overline{b}_1M_{[1,0,1,0]}-\varsigma^{-1},\\ &R_3\equiv M_{[0,0,0,1]}\left(\bar{a}_1\overline{b}_1-\dot{M}_{[1,1,0,0]}\right)+\bar{a}_0\overline{b}_0\left(\dot{M}_{[1,1,0,1]}-M_{[1,0,0,1]}\right),\\ &R_4\equiv\bar{a}_0\overline{b}_0M_{[1,0,1,1]}-M_{[0,0,0,1]}M_{[1,0,1,0]}, \qquad \text{(A12)}\\ &R_5\equiv M_{[1,0,0,1]}\left(\dot{M}_{[[0,1,0,0]]}-\bar{a}_0\overline{b}_0\right)+\bar{a}_1\overline{b}_1\left(M_{[0,0,0,1]}-\dot{M}_{[0,1,0,1]}\right),\\ &R_6\equiv M_{[1,0,0,1]}M_{[0,0,1,0]}-\bar{a}_1\overline{b}_1M_{[0,0,1,1]}\end{aligned}$$

$\Rightarrow$

$$
\begin{aligned}
m_g^{(1)} &\equiv \omega\varsigma \iint\limits_{R\times\Omega} \frac{\sqrt{Var\left(\breve{\varepsilon}_{Rd}\right)}}{\breve{\varepsilon}_{Rd}} \prod_{i=0}^{1} p_i^{\eta-1} q_i^{\tau-1} dp_i dq_i d\psi_i d\varphi_i \\
&= -\varsigma \left\{
\begin{array}{l}
\dfrac{3N_0\varsigma^{-1}\tau}{\sqrt{\eta\tau}} + \left(9N_0R_0\tau + \tfrac{3}{2}N_1\left(n_{[1]}^{-1}R_1\tau + \bar{a}_0\bar{b}_0M_{[1,0,1,0]}\right)\dfrac{1}{\eta\left(\eta+\tau\right)\sqrt{\eta\tau}}\right. \\
+\tfrac{3}{2}N_1\left(n_{[0]}^{-1}R_2\tau + \bar{a}_1\bar{b}_1M_{[1,0,1,0]}\right)\dfrac{1}{\eta\left(\eta+\tau\right)^2\sqrt{\eta\tau}} \\
+\tfrac{9}{2}N_1\tau\left[\eta^2\left(\eta+\tau\right)^3\sqrt{\eta\tau}\right]^{-1}\left[n_{[1]}^{-1}\left(R_3\left(\eta+\tau\right)+R_4\right)+n_{[0]}^{-1}\left(R_5\left(\eta+\tau\right)+R_6\right)\right]
\end{array}
\right\} \\
&= \frac{-3\varsigma\left(\eta+\tau\right)\left\{\begin{array}{l}2N_0\eta\tau\left(\eta+\tau\right)\cdot \\ \left(\varsigma^{-1}\eta\left(\eta+\tau\right)+3R_0\right)\end{array}\right. + N_1\left[\begin{array}{l}\eta\tau\left(\eta+\tau\right)\left(n_{[1]}^{-1}R_1+n_{[0]}^{-1}R_2\right) \\ +\eta\left(n_{[1]}^{-1}\bar{a}_0\bar{b}_0+n_{[0]}^{-1}\bar{a}_1\bar{b}_1\right)M_{[1,0,1,0]}+3\cdot\end{array}\right.}{2\sqrt{\eta\tau}\eta^2\left(\eta+\tau\right)^3} \\
&\quad \left.\left.\tau\left(n_{[1]}^{-1}R_3+n_{[0]}^{-1}R_5\right)\right]\right\} - 9N_1\varsigma\tau\left(n_{[1]}^{-1}R_4+n_{[0]}^{-1}R_6\right)
\end{aligned}
\tag{A13}
$$

where for $i, j, k, l = 0, 1$ $M_{[i,j,k,\ell]}$ and $\dot{M}_{[i,j,k,\ell]}$ are given respectively by Equations A6-A10,

$$
\begin{aligned}
&\omega = \left\{\Gamma\left(\eta+\tau\right)/\left[\Gamma\left(\eta\right)\Gamma\left(\tau\right)\right]\right\}^2, \\
&\varsigma = \left(\bar{a}_0\bar{b}_0\bar{a}_1\bar{b}_1\right)^{-1}, \bar{a}_i = 1-a_i, \bar{b}_i = 1-b_i, \\
&N_0 = n_{[0]}^{-\frac{1}{2}}\left(1+\tfrac{1}{2}n_{[1]}^{-1}n_{[0]}-\tfrac{1}{8}n_{[1]}^{-2}n_{[0]}^{2}\right), \\
&N_1 = n_{[0]}^{\frac{1}{2}}\left(1-\tfrac{1}{2}n_{[0]}n_{[1]}^{-1}+\tfrac{3}{8}n_{[0]}^{2}n_{[1]}^{-2}\right)
\end{aligned}
\tag{A14}
$$

On the other hand, by integrating the following equation with respect to φ_i, ψ_i, $i = 0, 1$

$$\frac{\breve{\varepsilon}_{RD}}{\sqrt{Var\left(\breve{\varepsilon}_{RD}\right)}} = \frac{\breve{p}_1 - \breve{p}_0 - \varepsilon_{RD}}{\sqrt{A}}\left(1 - \frac{J}{2A}\right)$$
$$= I\left(\frac{u\left(\varphi_1\right)}{\Delta_1} - \frac{u\left(\varphi_0\right)}{\Delta_0} - \varepsilon_{RD}\right) - \tfrac{1}{2}I^3\left(\frac{u\left(\varphi_1\right)}{\Delta_1} - \frac{u\left(\varphi_0\right)}{\Delta_0} - \varepsilon_{RD}\right)J$$

This leads to

$$\int_{a_0}^{1}\int_{b_0}^{1}\int_{a_1}^{1}\int_{b_1}^{1} \frac{\breve{\varepsilon}_{RD}}{\sqrt{Var\left(\breve{\varepsilon}_{RD}\right)}} \prod_{i=0}^{1} d\psi_i d\varphi_i = I\left(R_0 - \varsigma^{-1}\varepsilon_{RD}\right)$$
$$-\tfrac{1}{2}I^3 \left\{ \begin{array}{l} n_{[1]}^{-1}\left[\begin{array}{l} q_1R_3 + R_4 \\ -\varepsilon_{RD}\left(q_1R_1 + \bar{a}_0\bar{b}_0 M_{[1,0,1,0]}\right) \end{array}\right] \\ +n_{[0]}^{-1}\left[\begin{array}{l} q_0R_5 + R_6 \\ -\varepsilon_{RD}\cdot\left(q_0R_2 + \bar{a}_1\bar{b}_1 M_{[0,0,1,0]}\right) \end{array}\right] \end{array} \right\} \quad \text{(A15)}$$

Further, we obtain by integrating Equation A15 with respect to *p*_{*i*}, *q*_{*i*}, *i* = 0, 1

$$m_g^{(2)} \equiv \omega\varsigma \iint_{\Omega\times R} \frac{\breve{\varepsilon}_{RD}}{\sqrt{Var\left(\breve{\varepsilon}_{RD}\right)}} \prod_{i=0}^{1} p_i^{\eta-1} q_i^{\tau-1} dp_i dq_i d\psi_i d\varphi_i$$
$$= \varsigma\left\{ \frac{N_1R_0}{(\eta+\tau)\sqrt{\eta\tau}} - \tfrac{1}{2}\left[\frac{N_2\left(n_{[1]}^{-1}R_3 + n_{[0]}^{-1}R_5\right)}{\eta(\eta+\tau)^2\sqrt{\eta\tau}} + \frac{N_2\left(n_{[1]}^{-1}R_4 + n_{[0]}^{-1}R_6\right)}{\eta\tau(\eta+\tau)^3\sqrt{\eta\tau}} \right] \right. \quad \text{(A16)}$$
$$= \frac{\varsigma\left\{ 2N_1R_0\eta\tau(\eta+\tau)^2 - N_2\left[\begin{array}{l} \tau(\eta+\tau)\left(n_{[1]}^{-1}R_3 + n_{[0]}^{-1}R_5\right) \\ +n_{[1]}^{-1}R_4 + n_{[0]}^{-1}R_6 \end{array}\right]\right\}}{2\left[\sqrt{\eta\tau}(\eta+\tau)\right]^3}$$

where ς, N_1, R_0, and R_j, $j = 3, 4, 5, 6$ are given respectively by Equations A12 and A14, and

$$N_2 \equiv \sqrt{n_{[0]}^3}\left(1 - \frac{3n_{[0]}}{2n_{[1]}} - \frac{3n_{[0]}^2}{8n_{[1]}^2} - \frac{n_{[0]}^3}{8n_{[1]}^3} + \frac{45n_{[0]}^4}{64n_{[1]}^4} - \frac{27n_{[0]}^5}{128n_{[1]}^5} + \frac{27n_{[0]}^6}{512n_{[1]}^6}\right)$$

Note that in calculating Equations A13 and A16 I used an approximation on the Gamma function: $\Gamma\left(z+a\right)/\Gamma\left(z+b\right) \approx z^{a-b}$ (Askey & Roy, 2010).

By integrating Equation 12 with respect to (φ_i, ψ_i) first and then (p_i, q_i) for $i = 0, 1$ we obtain $m_g\ (\eta, \tau)$ by substituting Equations A13 and A16 into Equation A17:

$$m_g\left(\eta,\tau\right) = \left(2\pi\right)^{-\frac{1}{2}} \left\{ m_g^{(1)}\left(\eta,\tau\right) - \tfrac{1}{2} m_g^{(2)}\left(\eta,\tau\right) + \tfrac{1}{8}\left[m_g^{(2)}\left(\eta,\tau\right)\right]^3 \right\} \qquad \text{(A17)}$$

Journal of Modern Applied Statistical Methods
November 2015, Vol. 14 No. 2, 219-235.

ISSN 1538 − 9472

Approaches for Detection of Unstable Processes: A Comparative Study

Yerriswamy Wooluru
JSS Academy of Tech. Ed.
Bangalore, India

Dr. D. R. Swamy
JSS Academy of Tech. Ed.
Bangalore, India

Dr. P. Nagesh
JSS Centre for Mgmt. Stud.
Mysore, India

A process is stable only when parameters of the distribution of a process or product characteristic remain same over time. Only a stable process has the ability to perform in a predictable manner over time. Statistical analysis of process data usually assume that data are obtained from stable process. In the absence of control charts, the hypothesis of process stability is usually assessed by visual examination of the pattern in the run chart. In this paper appropriate statistical approaches have been adopted to detect instability in the process and compared their performance with the run chart of considerably shorter length for assessing its patterns and ensuring the process stability.

Keywords: Process stability, run chart patterns, run test, unstable process

Introduction

The run chart is a most effective and widely used tool for monitoring the stability of a process by displaying the data to make process performance visible. As long as the series of points in time exhibit a random pattern, the process is assumed to have constant mean and standard deviation and no autocorrelation (i.e. stable). While run charts focus more on time pattern, a control chart focuses on acceptable limits of the process data. However, in many industrial situations, it becomes necessary to estimate process parameter whose stability cannot be monitored using control charts due to lack of data and time for establishing control limits. In the absence of properly established control charts, process stability can be evaluated with the help of run chart trend and its pattern, which can be detected by applying run rules and to conclude the assignable causes present in the process.

In run chart, each observation of a sample have a time variable representing the time of each data point is measured when data have time related behavior, run charts are familiar tools to visualize the process behavior. Also Deming (1986) pointed that when processes ought to behave randomly overtime, run charts can

Yerriswamy Wooluru is an Associate Professor. Email him at: jssateb.ac.in.

help to identify nonrandom behavior, which can unearth potential for improvement. Run charts can be used as one of the important tools for diagnosing and solving various industrial problems, nonrandom patterns are indicative of process instability. Depending on the causes of process instability the non-random patterns can be of different types. The SQC Handbook of Western Electric illustrated various types of unnatural or nonrandom patterns that may occur in the run chart (Western Electric, 1956). Among these, six types of non-random patterns of individual observations are upward shift, downward shift, increasing trend, decreasing trend, cyclic and systematic patterns.

Various statistical tools, such as Regression analysis, ANOVA method, SR test, INSR test, and Levene's test have been used to assess the process location and variation to detect statistical stability of the forging process. These tools have also been compared with run chart of considerably shorter length to assess the efficiency of the above statistical methods, and indicate the process stability.

Methodology

The methodology involves the following steps:

1. Understanding the basic concepts and tools to detect process stability of a manufacturing process.
2. Process data collection.
3. Approaches used for assessing the statistical stability of the process are
   a. Regression Analysis,
   b. SR method,
   c. INSR method,
   d. Run test
   e. ANOVA method
   f. Levene's test
4. Construction of Run chart using statistical software MINITAB
5. Compare the performance of the above approaches with Run chart.
6. Conclusion about the performance of the above methods.

Data collection and analysis

The data set pertaining to the critical quality characteristic i.e. inner diameter of piston rings for an automotive engine produced by forging process. The details of the operation and product specification are presented in Table 1. The required quality characteristic of 32 consecutive units are measured and presented in Table 2. The basic sample statistics are calculated and presented in Table 3.

Table 1. Product description

Part Name	Material	Operation	Specifications	Measuring Device
Piston ring	Cast steel	Forging	74.00 ± 0.05	Dial Gauge

*All dimensions are in mm.

Table 2. Measurements of Piston ring hole diameter in mm.

Sl. no.	Hole dia	Sl. no.	Hole dia	Sl. no.	Hole dia	Sl. no.	Hole dia
1	74.030	9	74.011	17	73.996	25	74.014
2	74.002	10	74.004	18	73.993	26	74.009
3	74.019	11	73.988	19	74.015	27	73.994
4	73.992	12	74.024	20	74.009	28	73.997
5	74.008	13	74.021	21	73.992	29	73.985
6	73.995	14	74.005	22	74.007	30	73.993
7	73.992	15	74.002	23	74.015	31	73.998
8	74.001	16	74.002	24	73.989	32	73.990

Table 3. Summary Statistics of the case study data.

Sample size	Mean	Median	Minimum	Maximum	Range	Std. Deviation
32	74.003	74.002	73.985	74.03	0.045	0.0115

Statistical Approaches to Detect Instability

Regression analysis

One way to quantify the change in location is to fit a straight line to the data using an index variable as the independent in the regression. In this case, the observed values are in the sequential run order and they are collected at equally spaced time intervals. In this study, index variable are $X = 1, 2, 3,\ldots N$ where N is the number of observations. If there is no significant drift in the location over time, the slope parameter would be zero. The scatter diagram of the data reveals a negative linear association. Therefore, it can be proceeded to find the equation of the regression line using MINITAB statistical software.

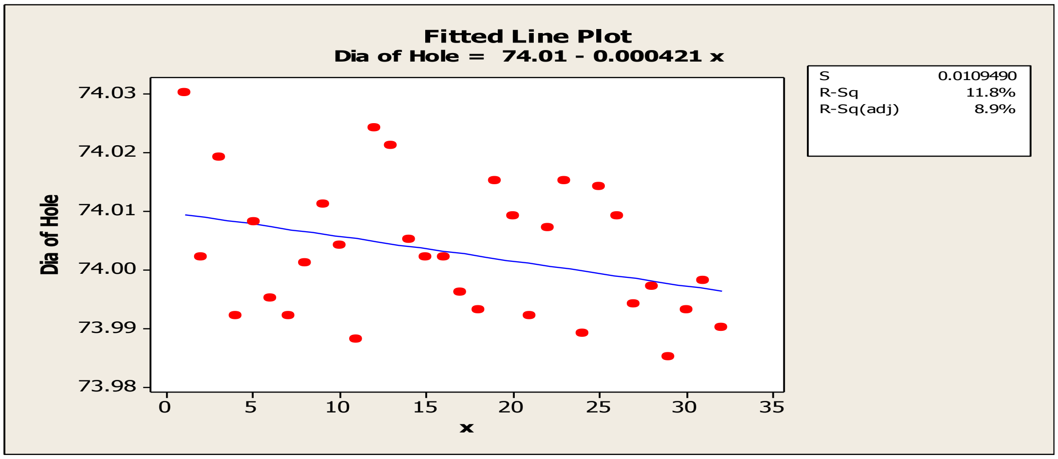


Figure 1. Output of regression analysis table for case study data.

The regression equation is Dia. of Hole = 74.0 - 0.000421 × (X)

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	0.0004831	0.0004831	4.03	0.054
Residual Error	30	0.0035964	0.0001199		
Total	31	0.0040795			

In the output of the regression analysis table for the case study data, the F-statistic is 4.03. The table value is 4.17 for F (0.05, 1, 30). Since $F_{calculated}$ is less than F_{table} value, and the p-value is greater than 0.05. It may be concluded that there is evidence that slope is almost equal to zero and ensure the process is stable over time.

SR method (standard deviation ratio method)

The SR test is derived from the square of the ratio of the standard deviation estimated using all the observations and the standard deviation estimated using sub group ranges/standard deviations/individual moving ranges. The basis of the SR test is that if the process is stable, all the approaches would yield similar estimates for the process standard deviation. In this case statistic, SR is computed as the ratio of the estimate of the long term variance and the estimate of the short term variance. The estimated sample variance based on the *N* observations will indicate the long term variance and the estimated variance based on the moving range (MR) method will reveal the short term variance.
Thus,

$$SR = \frac{\frac{1}{N-1}\sum_{i=1}^{N}\left(y-\bar{y}\right)^2}{\left(\overline{MR}/1.128\right)} \tag{1}$$

$$\bar{y} = \sum_{i=1}^{N} Y_i / N \tag{2}$$

$$\overline{MR} = \sum_{i=1}^{N-1} \left|y_{i+1} - y_i\right| / \left(N-1\right) \tag{3}$$

Ramirez and Runger (2006) assumed that an approximate *F*-distribution for SR, where the effective degree of freedom associated with the numerator and denominator are considered as (*N*-1) and 0.62 × (*N*-1) respectively and accordingly, it is recommended as an approximate *F*-test for SR.

Table 4. Calculation of Moving Range for the case study data.

Sl. no.	Hole dia (y_i)	$MR\left|\left(y_{i+1}-y_i\right)\right|$	Sl. no.	Hole dia (y_i)	$MR\left|\left(y_{i+1}-y_i\right)\right|$
1	74.030	-	17	73.996	0.006
2	74.002	0.028	18	73.993	0.003
3	74.019	0.017	19	74.015	0.022
4	73.992	0.027	20	74.009	0.006
5	74.008	0.016	21	73.992	0.017
6	73.995	0.013	22	74.007	0.015
7	73.992	0.003	23	74.015	0.008
8	74.001	0.009	24	73.989	0.016
9	74.011	0.010	25	74.014	0.025
10	74.004	0.007	26	74.009	0.005
11	73.988	0.016	27	73.994	0.015
12	74.024	0.036	28	73.997	0.003
13	74.021	0.003	29	73.985	0.007
14	74.005	0.016	30	73.993	0.008
15	74.002	0.003	31	73.998	0.005
16	74.002	0.000	32	73.990	0.008

$$\bar{y}=\frac{2368.09}{32}=74.0029,$$
$$\overline{MR}=\frac{0.373}{31}=0.0120, \tag{4}$$
$$\sigma'=\frac{\overline{MR}}{d_2}$$

d_2 = 1.128, Statistical constant for n = 2 (Montgomery, 2009, p.702)

$$\sigma'=\frac{0.012}{1.128}=0.0106\,.$$
$$\sum\left|\left(y_{i+1}-y_i\right)\right|=0.373$$
$$F\left(0.05,31,19.22\right)=1.93F\left(tab\right)$$

Because $SR = 0.012$, i.e., $(F\ calculated)$, $F\,(calculated) < F\,(table)$. Hence, it is concluded that the process is said to be stable.

Instability ratio test (INSR)

The instability ratio is defined as the ratio of the number of data points that have one or more violation of the Western Electric (1956) rules to the total number of data points plotted in the process behavior chart for the time period under assessment. The motivation for the INSR test is that if the process is stable, then it operates with common cause variation only and over time the observations move randomly about the central line and typically remain within the upper and lower control limits. The pattern exhibited in the run chart is called a random pattern.

Appearance of a nonrandom pattern, which can be detected by applying run rules, is indicative that there is either an assignable cause present in the process or the process output's variation has increased. Ramirez and Runger (2006) considered that the four most popular Western Electric (1956) rules for application of INSR method. Rules are as follows:

- 1 point out side of 3σ limits,
- 8 points in a row on one side of the central line,
- 2 of 3 points 2σ and beyond on the same side of the central line,
- 4 of 5 points 1σ and beyond on the same side of the central line.

Then the test statistic, INSR, is noted as follows

$$INSR = \frac{\text{Total number of violations with respect to the four rules in the chart}}{\text{Total number of observations plotted in the chart}} \times 100 \quad (5)$$

Table 5. Calculation of Moving Range for the case study data.

Sl. no.	(y_i)	$MR\left\|(y_{i+1}-y_i)\right\|$	Sl. no.	(y_i)	$MR\left\|(y_{i+1}-y_i)\right\|$
1	74.030	-	17	0.006	0.006
2	74.002	0.028	18	0.003	0.003
3	74.019	0.017	19	0.022	0.022
4	73.992	0.027	20	0.006	0.006
5	74.008	0.016	21	0.017	0.017
6	73.995	0.013	22	0.015	0.015
7	73.992	0.003	23	0.008	0.008
8	74.001	0.009	24	0.016	0.016
9	74.011	0.010	25	0.025	0.025
10	74.004	0.007	26	0.005	0.005
11	73.988	0.016	27	0.015	0.015
12	74.024	0.036	28	0.003	0.003
13	74.021	0.003	29	0.007	0.007
14	74.005	0.016	30	0.008	0.008
15	74.002	0.003	31	0.005	0.005
16	74.002	0.000	32	0.008	0.008

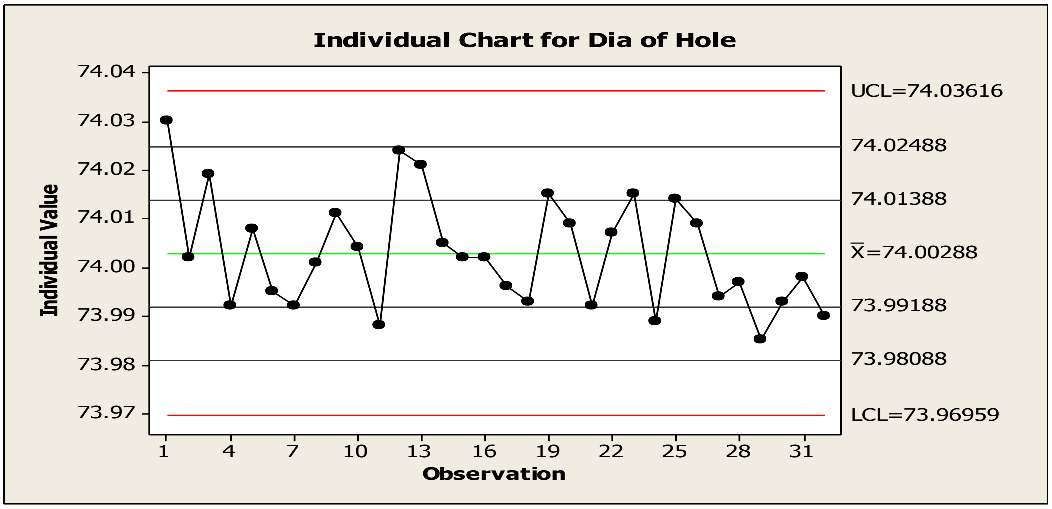


Figure 2. Run chart with 1σ, 2σ and 3σ control limits.

Process mean (μ) that represents the central line and the standard deviation (σ) that determines the distances of the control limits from the central line are usually unknown, and so these may be estimated from the N observations. The process means (μ) and standard deviation (σ) are estimated using arithmetic mean and moving ranges respectively.

Interpretation

a) 1 point out side of 3σ limits, (in Figure 2 no points violate this rule).

b) 8 points in a row on one side of the central line, (in Figure 2 no points violate this rule).

c) 2 of 3 points 2σ and beyond on the same side of the central line, (in Figure 2 no points violate this rule).

d) 4 of 5 points 1σ and beyond on the same side of the central line, (in Figure 2 no points violate this rule).

e) As no points violating the above 4 rules, INSR = 0.00, cutoff value for Run chart length ($N = 32$) is 3.125% [8], so the process is said to be stable.

Variation

To detect a change in variation in the process, Levene's test has been used it is based on the median rather than the mean. It assesses the assumptions that variance of the population from which different samples are drawn are equal. It tests the null hypothesis that the population variances are equal. If the resulting *p*-value of Levene's test is less than critical value (0.05), the obtained differences in the sample variances are unlikely to have occurred based on random sampling from a population with equal variances thus the null hypothesis of equal variances is rejected and it is concluded that there is a difference between the variances in the population. It also tests whether two sub samples in a given population have equal or different variances based on *p*-values.

Hypothesis Testing: Null hypothesis H_0 ; $\sigma_1 = \sigma_2 = \sigma_3 = \sigma_4$ (There is no change in variance)

Alternate hypothesis, H_0 ; $\sigma_1 \neq \sigma_2 \neq \sigma_3 \neq \sigma_4$ (There is change in variance)

Levine's Test has been carried out using the MINITAB software. Since the *p*-value is greater than 0.05, the null hypothesis is accepted and hence that there is no change in variance among the 4 sets in the sample data of 32 consecutive units.

ANOVA

This approach is to compare within subgroup variation to between subgroup variation to detect a difference in subgroup means and aimed at detecting changes in the process mean only. In this case study, N=32 individual observations are collected and the ANOVA method is applied by forming subgroups of size 2 using consecutive observations, i.e. there will be N/2 subgroups. Then the test statistic F is computed as the ratio of the mean sum of squares of subgroups (MS subgroup) and the mean sum of squares of errors (MS error).

Table 6. Analysis of Variance

Sl. no.	x_1	x_2	$\overline{x}_i$	$\overline{x}$	$(\overline{x}_i - \overline{x})^2$	$(x_{ji} - \overline{x}_i)^2$	$(x_{ji} - \overline{x})^2$
1	74.030	74.002	74.016	73.996	0.0004	0.000392	0.001192
2	74.019	73.992	74.0055	73.996	9.03E-05	0.000365	0.000545
3	74.008	73.995	74.0015	73.996	3.02E-05	8.45E-05	0.000145
4	73.992	74.001	73.9965	73.996	3.00E-07	4.05E-05	0.000041
5	74.011	74.004	74.0075	73.996	0.000132	2.45E-05	0.000289
6	73.988	74.024	74.006	73.996	0.0001	0.000648	0.000848
7	74.021	74.005	74.013	73.996	0.000289	0.000128	0.000706
8	74.002	74.002	74.002	73.996	0.000036	0.000000	0.000072
9	73.996	73.993	73.9945	73.996	2.30E-06	4.50E-06	0.000009
10	74.015	74.009	74.012	73.996	0.000256	0.000018	0.00053
11	73.992	74.007	73.9995	73.996	1.23E-05	0.000113	0.000137
12	74.015	73.989	74.002	73.996	0.000036	0.000338	0.00041
13	74.014	74.009	74.0115	73.996	0.00024	1.25E-05	0.000493
14	73.994	73.997	73.9955	73.996	3.00E-07	4.50E-06	0.000005
15	73.985	73.993	73.989	73.996	0.000049	0.000032	0.00013
16	73.901	73.87	73.8855	73.996	0.01221	0.000481	0.024901

Table 7. Resulted values from the ANOVA Analysis.

SS_{Factor} = 0.0277684	MS_{Factor} = 0.001
SS_E = 0.0026845	MS_E = 0.002
SS_T = 0.03045	F_o = 0.98

From F_{table}, $F_{critical} = 2.39$ and $F_{calculated} = 0.98$. Since $F_{cal.} < F_{0.05,15,16}$, the process position in time relating to a hole diameter data is not subjected to significant changes.

Run test for randomness in the sequence.

It tests the runs up and down or the runs above and below the mean by comparing the actual values to expect values. The statistic for comparison is the chi-square test [6]. All observations in the sample larger than the median value are given a positive sign and those below the median are given negative sign. A succession of values with the same sign is called a run and the number of runs '*a*' in the sequence of data points is found and it from the test statistic. For $n > 30$, this test statistic can be compared with a normal distribution with mean and the variance, the test is two-tailed. Data: Sample size: 32 observations, Median: 74.002

Table 8. Values above and below the median.

74.030	74.002	74.019	73.992	74.008	73.995	73.992	74.001
-	+	-	+	-	-	+	+
74.011	74.004	73.988	74.024	74.021	74.005	74.002	74.002
-	+	+	-	-	-	-	-
73.996	73.993	74.015	74.009	73.992	74.007	74.015	73.989
-	+	-	-	+	+	-	+
74.014	74.009	73.994	73.997	73.985	73.993	73.998	73.990
-	-	+	-	+	+	+	-

H_0: The sequence is produced in a random manner.
H_1: The sequence is not produced in a random manner.
Number of observations, $N = 32$, Number of runs, $a = 18$

$$\mu_a = \frac{2N - 1}{3} \tag{6}$$

$$\sigma_a^2 = \frac{16N - 29}{90} \tag{7}$$

$$\mu_a = \frac{2(32)-1}{3} = 21$$

$$\sigma_a^2 = \frac{16(32)-29}{90} = 5.37$$

For $N > 20$, the distribution of 'a' (number of runs) is reasonably approximated by a normal distribution, $N\left(\mu_a, \sigma_a^2\right)$. This approximation can be used to test the independence of the observations. In this case the standardized normal test statistic is developed by subtracting the mean from the observed number of runs 'a' and dividing by the standard deviation.

The test statistic is as follows.

$$Z_0 = \frac{a - \mu_a}{\sigma_a} \tag{8}$$

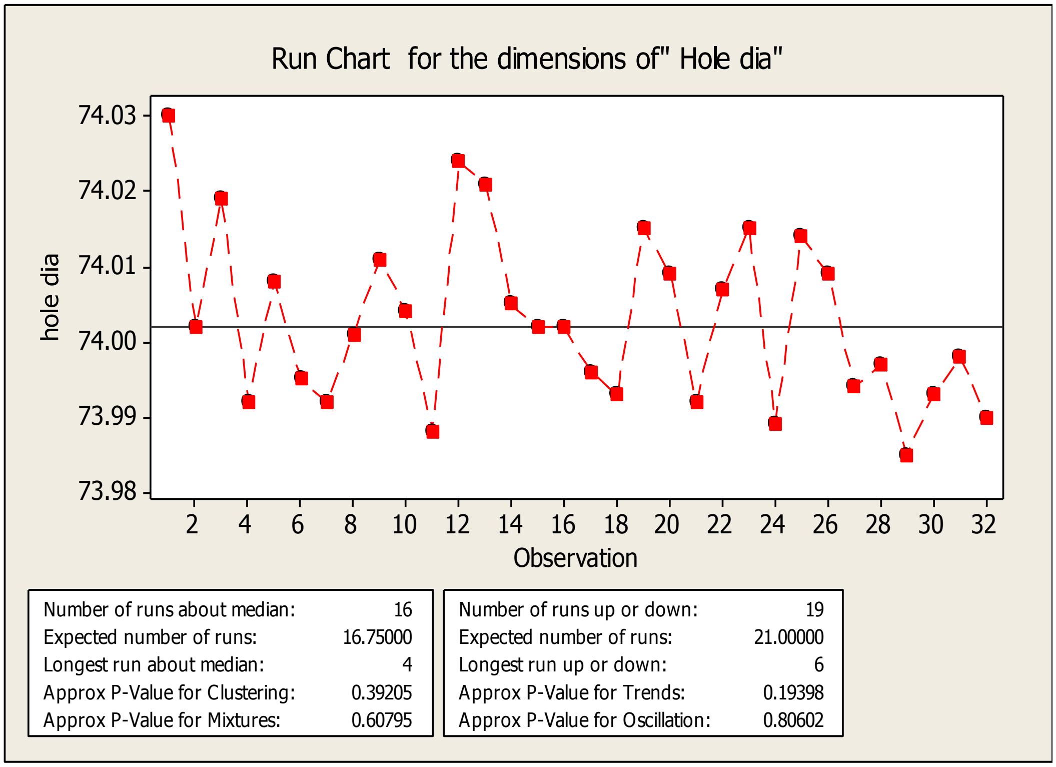
$$Z_0 = \frac{18-21}{2.32} = -1.30$$

Test statistic: Z_0 = -1.30, Significance level: α = 0.05
Critical value: $Z_{1-\alpha/2}$ = 1.96, Reject H_0, if $|Z| > 1.96$.

In this case, the test statistic (-1.30) is inside the critical region, the null hypothesis cannot be rejected and hence it is concluded that the data is random. The critical value $Z_{0.025} = 1.96$. Because $|Z_0| < Z_{0.025}$, the independence (randomness) of the sequence of the observations cannot be rejected.

Run chart analysis

A run chart is a line graph of data plotted over time. By collecting and charting data over time, trends or patterns in the process can be revealed. As run charts do not use control limits, they cannot exhibit if a process is stable. However, they can show that how the process is running. The run chart can be a valuable tool at the beginning of a manufacturing process, as it reveals important information about a process before collecting the enough data to create reliable control limits. Figure 3 shows the Run chart for the case study data constructed using statistical software MINITAB to assess the stability of the process.



Number of runs about median:	16
Expected number of runs:	16.75000
Longest run about median:	4
Approx P-Value for Clustering:	0.39205
Approx P-Value for Mixtures:	0.60795

Number of runs up or down:	19
Expected number of runs:	21.00000
Longest run up or down:	6
Approx P-Value for Trends:	0.19398
Approx P-Value for Oscillation:	0.80602

Figure 3. Construction of run chart using MINITAB-Statistical software.

The two tests (actual number of runs about median and number of runs up and down) have been conducted to check the randomness. In both the tests i.e., actual number of runs about median and number of runs up and down are close to the expected number of runs. It implies that the data come from random distribution. Clusters are groups of points in one area of the charts, cluster indicate variation due to special causes such as measurement problem. In this case, approximate *p*-value is 0.39205, it is greater than 0.05, hence it may be concluded that there is no clustering in the data. Process stability can be assured by observing the oscillation of data above and below the center line rapidly. In this case, Approximate *p*-value is 0.80602, it is greater than 0.05, so it may be conclude that there is no oscillating pattern in the data.

A mixture is characterized by an absence of points near the center line. It often indicates combined data from two populations or two processes operating at different levels. In this case, approximate *p*-value is 0.60795, it is greater than 0.05, hence it may be conclude that the data does not come from different process.

Trends are sustained and systematic sources of variation characterized by a group of points that drifts either up or down. Trends may warn that a process is about to go out of control and may be due to worn tools. In this case, approximate p-values is 0.19398, it is greater than 0.05, hence it is concluded that there is no trend in the data. The tests for non-random pattern are significant at the 0.05 level. All p-values for all the tests are greater than 0.05 (α) which suggests that the data come from a random distribution and process is stable.

Discussion

The data set pertaining to the quality characteristic i.e. inner diameter of piston rings for an automotive engine produced by forging process. Measurements for inner diameter of 32 consecutive units are measured and recorded. The various approaches have been used on the data in order to assess the stability of the forging process. Tests with respect to location, variation, randomness and sequence of data has been done through Regression analysis, ANOVA test, Run test, Levene's test, SR test, INSR test. The scatter plot reveals a least magnitude of negative linear association (almost zero).

In Regression analysis, R^2 value is 11.8%; it is can be stated that 11.8% of the total variation in the hole diameter occurs because of the variation in the observations sequence and remaining 88.2% is due to randomness and other causes of variation and also reveals that the relationship between the variables i.e. hole diameter and time is not significant. Also the F-test indicates that there is no considerable slope in the line.

In Levene's test, P-valve is greater than 0.05, so the null hypothesis cannot be rejected that there is no change in variance among the 4 sets in the sample data of 32 consecutive units.

In case of Instability ratio test, Calculated Instability Ratio (INSR) = 0.00, cutoff value for Run chart length ($N = 32$) is 3.125% [8], as instability ratio value is less than cutoff value, the process is said to be stable. In SR method, the test statistic SR is computed and compared with the F (table) value. F-Test for SR, conclude that the process is stable as SR = 0.012 i.e. (F calculated) is less than F (0.05, 31, 19.22) = 1.93 i.e., (F table). In case of ANOVA method, $N = 32$ individual observations, it is applied by forming subgroups of size 2 using consecutive observations, i.e. there will be $N/2$ subgroups.

Then the test statistic F is computed as the ratio of the mean sum of squares of subgroups (MS subgroup) and the mean sum of squares of errors (MS error). From F_{table}, $F_{critical} = 2.39$ and $F_{calculated} = 0.98$. Since $F_{calculated} < F_{0.05,15,16}$, the process position in time relating to a hole diameter is not subjected to significant changes. Run Test for randomness of the sequence is concluded that the data is random. The Table 9 presents the summary of results of the various statistical methods.

Table 9. Summary results of the statistical method.

Sl. no.	Statistical method	Result	Stable/Unstable
1	Regression	F(calculated) < F(table), p > 0.05	Stable
2	SR-method	F(calculated) < F(table)	Stable
3	Instability Ratio method	Instability ratio < cutoff value,	Stable
4	Levene's Test	p > 0.05	Stable
5	ANOVA method	F(calculated) < F(table),	Stable
6	Run Test	Z_0(calculated) < $Z_{1-\alpha/2}$(table),	Stable
7	Run Chart	p > 0.05,All cases	Stable

Alternative approaches were presented to assess the stability of the process and compared with the run chart. Process stability has been detected using the approaches such as Regression analysis, SR method, INSR method, Levene's test, ANOVA method. Even though all the approaches yield the same result (i.e., process is stable), above mentioned approaches have their own advantages and limitations. As the exact distribution of SR is not known and assumed an approximate F-distribution for SR, it can be applied only when the number of observations is larger than or equal to 32. The advantage of ANOVA approach is that the F-test conducted using the 'between' and 'within' sums of squares is well defined and it is applicable even when the available number of observations is small but it requires practitioner's to have background in statistics. Run test indicated that the data points are independent and random, hence it is concluded that there is no shift in location. INSR Test is more effective test as it uses rules similar to run chart and it works well for large number of samples. For small number of samples like 32-100 subgroups it leads to a Type-I error (i.e. probability of declaring a stable process as unstable) as high as 0.35. Ramirez and Runger recommended taking the 95th percentile point of the distribution of INSR as the cutoff value. With aim to increase the effectiveness, it has been recommended using the ANOV and the INSR tests. All the statistical methods indicates the presence of statistical stability in the case study data but run chart using the statistical software MINITAB gives more effective and accurate result compared to the other methods for assessing stability of the process.

References

Banks, J. (1989). *Principles of quality control*. New York, NY: John Wiley & Sons.

Banks, J., Carson II, J. S., Nelson, B. L., Nicol, D. M. (2001). *Discrete-event simulation* (3rd ed.). Prentice - Hall of India.

Champ, C. W. & Woodall, W. H. (1987). Exact results for Shewhart control chart for supplementary runs rules. *Technometrics*, *29*(4), 393-399. doi:10.1080/00401706.1987.10488266

Czarski, A. (2009). Assessment of a long-term and short-term process capability in the approach of analysis of variance (ANOVA), *Metallurgy and Foundry Engineering*, *35*(2), 111-119. doi:10.7494/mafe.2009.35.2.111

Deming, W. E. (1986). *Out of the crisis*. Cambridge, MA: Massachusetts Institute of Technology, Center for Advanced Engineering Study.

Gouri, S. K. (2010). A quantitative approach for detection of unstable processes using a run chart. *Quality Technology and Quantitative Management*, *7*(3), 231-247.

Grant, E. L. & Leavenworth, R. S. (1996). *Statistical quality control* (7th ed.). New York. NY: McGraw-Hill.

Montgomery, D. C. (2001). *Introduction to statistical quality control* (4th ed.). New York, NY: John Wiley & Sons.

Montgomery, D. C. (2009). *Introduction to statistical quality control* (pp. 702) (6th ed.). New York, NY: John Wiley & Sons.

Nelson, L. S. (1984). The Shewhart control chart -- test for special causes. *Journal of Quality Technology*, *16*(4), 237-239.

Nelson, L. S. (1985). Interpreting Shewhart X-bar control charts. *Journal of Quality Technology*, *17*(2), 114-116.

Pham, D. T. & Wani, M. A. (1997). Feature based control chart pattern recognition. *International Journal of Production Research*, *35*(7), 1875-1890. doi:10.1080/002075497194967

Prabhuswamy. M. S. & Nagesh. P. (2007). Process capability analysis made simple through graphical approach. *Kathmandu University Journal of Science, Engineering and Technology*, *1*(3).

Prabhuswamy. M. S. & Nagesh, P. (2010-2011). Process capability validation and short - Long term process capability analysis with case study. *Proceedings of ETIMES-2006*.

Ramirez, B. & Runger, G. (2006). Quantitative techniques to evaluate process stability. *Quality Technology*, *18*(1), 53-68. doi:10.1080/08982110500403581

Western Electric (1956). *Statistical quality control handbook*. Indianapolis, IN: Western Electric Company.

Journal of Modern Applied Statistical Methods
November 2015, Vol. 14, No. 2, 236-256.

ISSN 1538 − 9472

Contrails: Causal Interference Using Propensity Scores

Dean S. Barron
twobluecats.com
Long Beach, CA

Joe H. Brown
Irvine, CA

Contrails are clouds caused by airplane exhausts, which geologists contend decrease daily temperature ranges on Earth. Following the 2001 World Trade Center attack, cancelled domestic flights triggered the first absence of contrails in decades. Resultant exceptional data capacitated causal inference analysis by propensity score matching. Estimated contrail effect was 6.8981°F.

Keywords: Contrails, contrails effect, airplane exhaust, causal inference, propensity score, resampling, logistic regression, regression, MCMC

Introduction

Contrails are the clouds formed as a result of the introduction of relatively warm water vapor from airplane engine exhausts into surrounding cold, moist, atmospheric air (the word "contrails" is a contraction of two words, "condensation trails.") Under salient conditions, such mixing within the airplane engine exhaust plume saturates the atmospheric air, causing condensation of water droplets upon the exhaust particles. In turn, these newly formed droplets freeze into ice particles that constitute contrails (Schumann, 2005; EPA, 2000). The process also depends on non-atmospheric factors, such as engine and fuel characteristics (Wendler & Stuefer, 2002). The contrail formation process typically occurs at altitudes over 25,000 ft. and temperatures below −40°C.

Geologists asserted that contrails (1) decrease the daily high temperature by blocking incoming sunlight, (2) increase nightly low temperatures by preventing escape of greenhouse gases, and, therefore (3) decrease the daily temperature range on the Earth's surface below (e.g., Meerkotter, et al., 1999). This contrails effect was estimated to be 1.98°F or 3.24°F (Travis, et al., 2002); the greater of these is hereafter referred to as the Travis estimate.

Dean Barron is a statistical consultant at twobluecats.com. Email him at dean@twobluecats.com.

A contrail may dissipate quickly or linger for hours. Persistent contrails may grow expansively and then frequently morph or incorporate into cloud cover (EPA, 2000). Over individual geographic areas, the presence of contrails depends on existent conditions.

Therefore, observing the actual temperature range in the absence of contrails was impossible in areas where contrails had always been present. After the World Trade Center attacks of 11 September 2001, however, all flights in the United States were suspended for several days. Thus, a complete absence of contrails prevailed, including those locations where contrails had been present continuously for decades.

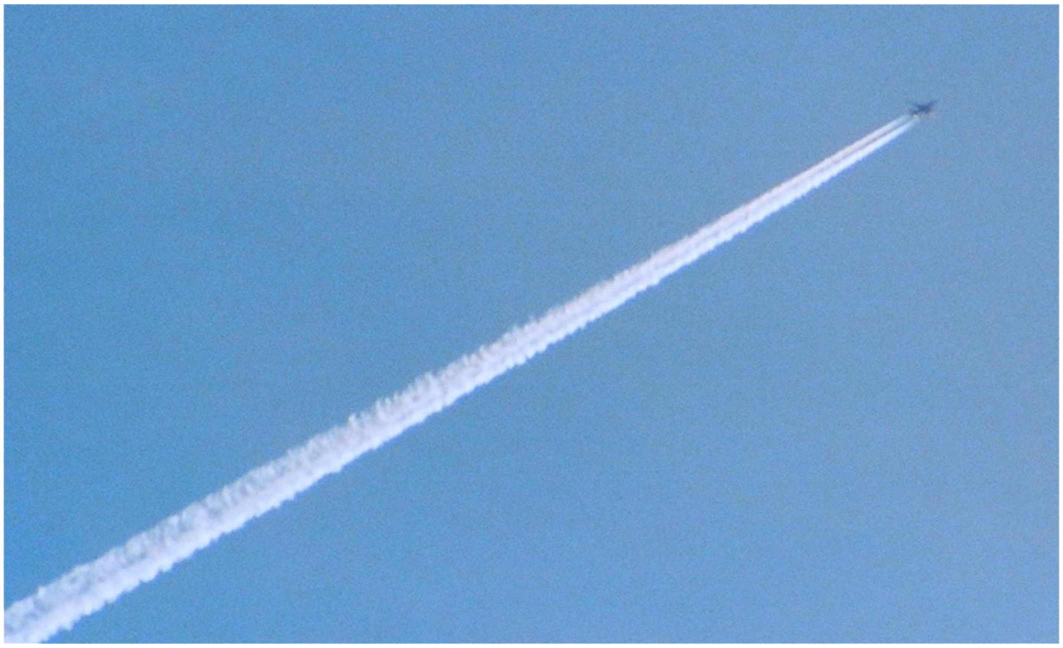


Figure 1. Contrail (Barron, 2013)

Purpose of the Study

Data situations with such counterfactuals are precisely the forte of analysis using causal inference. A propensity score (PS) was modeled and then used to match from the control group without replacement for the treatment group. Additionally, regression analysis and Bayesian Markoff Chain Monte Carlo (MCMC) were performed.

Data were obtained from The National Climatic Data Center (NCDC), which had daily historical data since 1929 from approximately 300 countries and 30,000 cities. The treatment group was defined as United States (hereafter, referred to as, "domestic") stations data from September 12-13, 2001, taking advantage of the absence of contrails. The control group was defined as all non-treatment station readings, both domestic and international. The data was subjected to random sampling and quality control.

The contrails effect, which, in causal inference terminology is the Average Treatment Effect on the Treated (ATT), was estimated to be 6.8981°F ($p < 0.0001$), compared with 6.5513°F ($p < 0.0001$) from the naive regression, and 6.5195°F ($\alpha = 0.05$ HPD Interval 5.7795, 7.2552) from MCMC simulation. All were more than twice the Travis estimate. The propensity score matching approach was determined to be preferable due to its superior covariate characteristics.

Methodology

Data

The NCDC weather-related database stores daily data as collected by the National Weather Service (NWS) Automated Surface Observing System (ASOS) in downloadable .txt format inside triple-compressed op-op.gz-tar formatted files (NCDC, 2010). The study data were restricted to measurements from stations that were operational in 2001.

These observations were further limited to 0-4 weeks before and after each September 12-13 for each of the three superimposable calendar years 1990, 2001, and 2007. Treatment variable, *CONTRAILS*0, was defined:

$$CONTRAILS0 = \begin{cases} 1, & \text{treatment, domestic 12 SEP2001 and 13 SEP2001} \\ 0, & \text{control, otherwise} \end{cases} \quad (1)$$

Contrail formations above airports have different characteristics than above non-airport locations. Because contrails generally do not form until aircraft reach 25,000 foot altitudes, contrails above airports typically derive from aircraft flights which had originated from other airports. Hence, airports might or might not have contrails (Mims, Chambers & Oostra, n.d.). Therefore, for this analysis, all airports were excluded from the control dataset only.

A two-stage stratified random sampling scheme was then imposed. Domestic data formed the first group. The United States was the only nation that stopped flights, therefore, neighboring Mexico and Canada formed the second group. Belgium and France were chosen as European counterparts for the third group. All other countries constituted the fourth group.

The first sampling stage selected 1,607 stations as treatment and 8,805 as control; from this, the second random sampling stage selected 278 and 440, respectively. The latter corresponded to a possible 3,214 and 478,250 observations, respectively. This data sampling procedure was designed to facilitate the required manual identification and subsequent elimination of airport locations.

Resultant samples sizes contained 556 treatment and 22,810 control observations, of which only 503 treatment and 4,737 control actually contained data. Further quality control on missing critical variables (*dewp*, *slp*, *wdsp*, *visib*, and *temperature-related*), dropped the final analysis dataset to 322 treatment and 2,557 control observations.

In addition to the variables contained in the NCDC database, the adjusted latitude was calculated using the formula (2) to correct for gravity (Bauer, et al., 2000). Normal gravity is defined as the gravity which would be observed were planet Earth to be a perfect ellipsoid with associated perfect rotation. The corrected latitude reflects deviations from ideal conditions, and is a function of only the latitude.

$$latitudecorr = 9.78\times10^{5} * \begin{pmatrix} 1+5.28\times10^{-3} * \sin^{2}\left(latitude\right) \\ +2.35\times10^{-5} * \sin^{4}\left(latitude\right) \end{pmatrix} \qquad (2)$$

Variables that were included in the propensity logistic regression model are described in Table 1.

Table 1. Variables

ID	Variable	Required non-missing	Description
1	*CONTRAILS0*		1 = TREATMENT, Absence of contrails 0 = CONTROL, contrails present / contrails effect
2	*temp*	YES	mean temperature for the day in degrees Fahrenheit
3	*dewp*	YES	mean dew point for the day in degrees Fahrenheit
4	*slp*	YES	mean sea level pressure for the day
5	*visib*	YES	mean visibility for the day in miles
6	*wdsp*	YES	mean wind speed for the day in knots
7	*MXSPD*		maximum sustained wind speed
8	*PRCP*		total precipitation
9	*p133fog*		fog / FRSHTT character 1
10	*p134rain*		rain or drizzle / FRSHTT character 2
11	*p135snow*		snow or ice pellets / FRSHTT character 3
12	*p137thun*		thunder / FRSHTT character 5
13	*elev*		elevation in meters
14	*latitudecorr*		absolute value latitude in degrees
15	*latitudeabs*		latitude correction for gravity in milligalileos
16	*temprange*	YES	temperature range in degrees Fahrenheit

Analysis

Causal inference, regression analysis, and Bayesian MCMC were used. The several shades of each resulted in a total of 10 different methods, hereafter referred to as *METHOD*1 through *METHOD*10.

Causal Inference

The Propensity Score (PS) was the predicted value from the linear first order logistic regression model of *CONTRAILS*0 as a function of the covariates. All variables were retained to maximize R^2.

For *METHOD*1, the PS of a treatment observation was compared with the PS of any remaining unmatched control observation. Matching by the absolute smallest PS difference, a greedy strategy was implemented in descending PS order of treatment observations. The ATT estimate for *CONTRAILS*0 was equal to the *temprange* difference of treatment and control groups from the matched observation pairs, and evaluated by *t* test.

In *METHOD*2, resampling was performed to examine if the dataset perhaps had yielded a coincidentally favorable match. Nine treatment group sample sizes, n*trt*, (288, 216, 162, 136, 108, 96, 81, 72, 68) were resampled ($n = 180$) at a corresponding specified control to treatment observational ratio (2, 3, 4, 5, 6, 7, 8, 9, 10, respectively). Because there was also potential for relative abundance of a control subregion to impact results, each of the four control subregions were equally represented, as calculated in (3).

$$\mathrm{n}cntl_{\text{subregion}} = \left(1/4\right) * \left(\mathrm{n}trt\right) * \left(\text{control-to-treatment observational ratio}\right) \tag{3}$$

Care was taken to select whole numbers and ensure that $\mathrm{n}cntl_{\text{subregion}} < 174$, because that was the sample size of the smallest subregion.

For each individual resample, the ATT was calculated identically as in *METHOD*1. For each n*trt* level, the ATT was calculated as the mean of its 180 samples; the overall ATT was the mean of the 1,620 runs.

In *METHOD*3, the tails of the dataset were trimmed to only the region of overlapping PS ranges of the treatment and control observations. The PS minima and maxima were determined for treatment ($\mathrm{PS}min_{trt}$, $\mathrm{PS}max_{trt}$) and control ($\mathrm{PS}min_{cntl}$, $\mathrm{PS}max_{cntl}$). A new PS range was set from the maximum minimum ($\max(\mathrm{PS}min_{trt}, \mathrm{PS}min_{cntl})$) to the minimum maximum ($\min(\mathrm{PS}max_{trt}, \mathrm{PS}max_{cntl})$) by dropping external values. In *METHOD*4, resampling was also performed.

For the best among the four methods, the resultant matched pairs and frequency distributions of the selected countries were analyzed. Patterns of the matched pairs were noted.

Regression

Three regressions were conducted to provide baseline comparisons for the propensity matching results, and to provide parameter estimates for other variables (4, 5, 6).

$$temprange = f\,(CONTRAILS0, \text{full model with all variables}) \tag{4}$$

$$temprange = f\,(CONTRAILS0, \text{best stepwise/backward elimination result}) \tag{5}$$

$$temprange = f\,(CONTRAILS0) \tag{6}$$

Hereafter these are referred to as *METHOD*5, *METHOD*6, and *METHOD*7, respectively. Resampling was performed on the best of the three, hereafter, referred to as *METHOD*8.

Bayesian

Two MCMC regression simulations were run, based upon (5) and (6), referred to as *METHOD*9 and *METHOD*10, respectively. Blocking strategy was determined by a correlations and resultant convergence characteristics. Non-informative priors were implemented first. When not feasible, the parameter estimates from the corresponding regression were to be used as informative priors.

The *CONTRAILS* estimates from all methods and Travis were compared. The MCMC simulation *METHOD*9 posterior estimates for *CONTRAILS*0 were analyzed to determine the percentage that were greater than each *CONTRAILS*0 estimate. The probability that a particular *CONTRAILS*0 estimate was an underestimation corresponds to this percentage.

Covariate and contrail effect estimate comparisons

Covariate differences between the matched treatment and control groups were calculated to reveal differences between the groups, which were compared with differences from the analysis dataset. Transition from significant to not significant was used as evidence of amelioration of covariate mean differences.

Omnibus distribution tests

Distributional differences between treatment and control groups were subjected to omnibus tests. These were Kolmogoroff-Smirnoff (KS), Cramér-von Mises (CM), and “oando” (see the Master’s thesis of the first author, Barron, 2007).

Results

Causal Inference

Propensity Score Logistic regression for PS was performed including all covariates with intercept using the final dataset (n*ttl* = 2879). The resultant model of *CONTRAILS*0 was statistically significant ($X^2 = 289.0694$, df = 14, p-value < 0.0001). The area under the ROC curve c-value = 0.785, Somers' $D = 0.570$, Kendall's Tau-$a = 0.113$, and standard definition of percentage behavior explained by model, $R^2 = 0.1127$. All correlations with

*CONTRAILS*0 and maximum likelihood parameter estimates are detailed in Table 2.

Table 2. Correlations and Propensity Score (PS) Logistic Regression Results

ID	Variable	Correlation	Parameter Estimate	Wald X^2	*p*-value
0	*Intercept*	N/A	-68.1469	3.7370	0.0532
1	*CONTRAILS*0	1	N/A	N/A	N/A
2	*temp*	0.1267	0.1246	109.8874	<0.0001
3	*dewp*	0.0225	-0.1089	89.6451	<0.0001
4	*slp*	0.0989	0.0801	36.5800	<0.0001
5	*visib*	-0.0942	-0.0795	35.8712	<0.0001
6	*wdsp*	-0.0581	-6.6053E-03	0.1791	0.6721
7	*MXSPD*	-0.0439	-3.4946E-03	2.4711	0.1160
8	*PRCP*	-0.0217	-0.1344	0.3487	0.5549
9	p133*fog*	0.0914	1.3716	49.4748	<0.0001
10	p134*rain*	-0.0449	0.2519	2.0147	0.1558
11	p135*snow*	-0.0404	-1.0575	1.0007	0.3171
12	p137*thun*	0.0871	1.3485	21.3320	<0.0001
13	*elev*	0.0063	-4.8630E-04	10.7487	0.0010
14	*latitudecorr*	0.0003	-1.6200E-05	0.2089	0.6476
15	*latitudeabs*	-0.0895	-2.0420E-02	4.5209	0.0335
16	*temprange*	0.3119	N/A	N/A	N/A

To determine if there would be sufficient PS coverage to enable matching of treatment and control, the PS range was divided into four bins with equal *n*-treatment counts. Spread was adequate (Table 3).

Table 3. Propensity Score (PS) Frequency Distributions by Bin

	PS RANGE	n*trt*	n*cntl*	RATIO
BIN1	0.0002, 0.1160	80	1790	22.38
BIN2	0.1160, 0.1724	81	357	4.41
BIN3	0.1724, 0.2725	81	267	3.30
BIN4	0.2726, 0.8380	80	143	1.79
TOTAL	0.0002, 0.8380	322	2557	7.94

***METHOD*1 / Matched Pairs, No Resampling, No Overlap** Mean difference of *temprange* between matched pairs, the ATT estimate, was 6.8981 (*t* = 9.91, *p* < 0.0001, 95%CI 5.5293, 8.2670). The mean absolute distance between matched propensity scores was 0.0035 (median < 0.0001, range < 0.0001, 0.1033).

***METHOD*2 / Matched Pairs, Resampling, No Overlap** The results consistently approximate the ATT estimate obtained with the non sampled data. The 1620 runs from the 9 different combinations had *temprange* mean = 6.7871 (median = 6.7847, range 2.5779, 10.8118). The mean of PS matched mean absolute distances was 0.0194 (median = 0.0133, range 0.0005, 0.1040). The results of the runs of the n*trt* and control to treatment observational ratios appear in Table 6.

***METHOD*3 / Matched Pairs, No Resampling, Overlap** Trimming down to the overlap region reduced the dataset to n*trt* = 321 n*cntl* = 2525 n*ttl* = 2846. Compared with the analysis dataset, this was a reduction of only one treatment and 32 control observations. The mean difference of *temprange* was 6.8931 (t = 9.88, p < 0.0001), with a mean absolute distance between matched propensity scores of 0.0032 (median < 0.0001, range < 0.0001, 0.0978).

***METHOD*4 / Matched Pairs, Resampling, Overlap** The contrail effect estimates were slightly higher than those without the overlap strategy. The 9 different combinations averaged *temprange* = 6.9654 (median = 6.9352, range 3.2071, 10.7119). The mean of PS matched mean absolute distances was 0.0141 (median = 0.0072, range 0.0005, 0.0960). The results of the runs of the various n*trt* and control to treatment observational ratios are also summarized in Table 6.

Analysis of the matches

The majority of treatment - control pairs appeared either once or twice. There were 200 distinct ordered pairs within the 322 matches, of which 171 (85.50%) had fewer than three occurrences. Only four appeared five or more times, California-France (11), Texas-France (8), Texas-United States (6), and California-Mexico (5).

All four strata of control country groups were represented in the matches. Despite the boost in percent observations secondary to designation as separate subgroups, the Relative Risk (RR) of selection for CANADA/MEXICO and BELGIUM/FRANCE were only somewhat lower than OTHER

INTERNATIONAL. Not surprisingly, the UNITED STATES group had the highest RR, as in Table 4.

Table 4. Subgroup Counts in Control Data

STRATUM	COUNTRIES	n*match*	n*cntl*	row % *match*	RR	column %*match*	column %*cntl*
1	UNITED STATES	41	174	23.56%	2.00	12.73%	6.80%
2	CANADA/MEXICO	65	631	10.30%	0.77	20.19%	24.68%
3	BELGIUM/FRANCE	102	851	11.99%	0.93	31.68%	33.28%
4	OTHER INTL	114	901	12.65%	1.01	35.40%	35.24%
	TOTAL	**322**	**2557**	**12.59%**			

Twenty-five countries were included in the control population. The highest percentage of matched control observations was 50% selected for Australia (n*match* = 9 n*cntl* = 18); the lowest was the lone 0% for Georgia (0/19). Over two-thirds had RR for selection between 1/3 and 3 (17/25).

Regression

***METHOD*5, *METHOD*6, AND *METHOD*7 / Naive Regression, No Resampling** The full regression model (*METHOD*5) with all covariates was statistically significant ($R^2 = 0.5713$, $F = 254.32$, $p < 0.0001$) with a parameter estimate for *CONTRAILS*0 of 6.5513. Both backward elimination and stepwise arrived at identical models with eight independent variables (METHOD6, forced *CONTRAILS*0 inclusion, $R^2 = 0.5698$, $F = 475.26$, $p < 0.0001$). The parameter estimate for *CONTRAILS*0 was 6.5173 (standard error = 0.3769, $t = 17.29$, $p < 0.0001$).

The minimal *CONTRAILS*0 only model, *METHOD*7, was also statistically significant but exhibited a much lower $R^2 = 0.0973$ ($F = 310.01$, $p < 0.0001$) with a much higher parameter estimate of 9.3605 (standard error = 0.5316, $t = 17.61$, $p < 0.0001$). Among the three, *METHOD*6 was selected as preferred. Parameter estimates appear in Table 5.

Table 5. Regression and MCMC results

		REGRESSION		**MCMC**	
ID	**Variable**	***METHOD*6**	***p*-value**	***METHOD*9**	**95% HPD Interval**
		Estimate		**Posterior Mean**	
0	*Intercept*	17.5910	<0.0001	17.5903	16.1170 , 19.0984
1	*CONTRAILS*0	6.5173	<0.0001	6.5195	5.7795 , 7.2552
2	*temp*	0.5114	<0.0001	0.5112	0.4728 , 0.5493
3	*dewp*	-0.5473	<0.0001	-0.5471	-0.5857 , -0.5091
4	*slp*	ELIMINATED	>0.05	NA	NA
5	*visib*	ELIMINATED	>0.05	NA	NA
6	*wdsp*	-0.5771	<0.0001	-0.5770	-0.6207 , -0.5318
7	*MXSPD*	ELIMINATED	>0.05	NA	NA
8	*PRCP*	ELIMINATED	>0.05	NA	NA
9	p133*fog*	ELIMINATED	>0.05	NA	NA
10	p134*rain*	-1.8677	<0.0001	-1.8689	-2.4019 , -1.3160
11	p135*snow*	-3.9595	<0.0001	-3.9653	-5.7258 , -2.1473
12	p137*thun*	1.5057	0.0278	1.5053	0.1708 , 2.8489
13	*elev*	0.0026	<0.0001	0.0026	0.0023 , 0.0029
14	*latitudecorr*	ELIMINATED	>0.05	NA	NA
15	*latitudeabs*	ELIMINATED	>0.05	NA	NA
16	*temprange*	DEP VAR	>0.05	DEP VAR	NA

**Note*. Regression parameter estimates then served as MCMC priors

***METHOD*8 / Naive Regression, Resampling** Because *METHOD*6 was preferred over the reduced model, only the former was subjected to resampling. For the 1620 runs, the *CONTRAILS*0 estimate had mean = 6.6889 (median = 6.6336, range 4.4486, 8.8783). The mean *F* was 150.6304, mean $R^2 = 0.5752$. Each of the 1620 individual runs were statistically significant ($p < 0.0001$). The results of the runs of the various n*trt* and control to treatment observational ratios are also summarized in Table 6.

Table 6. Resampling results

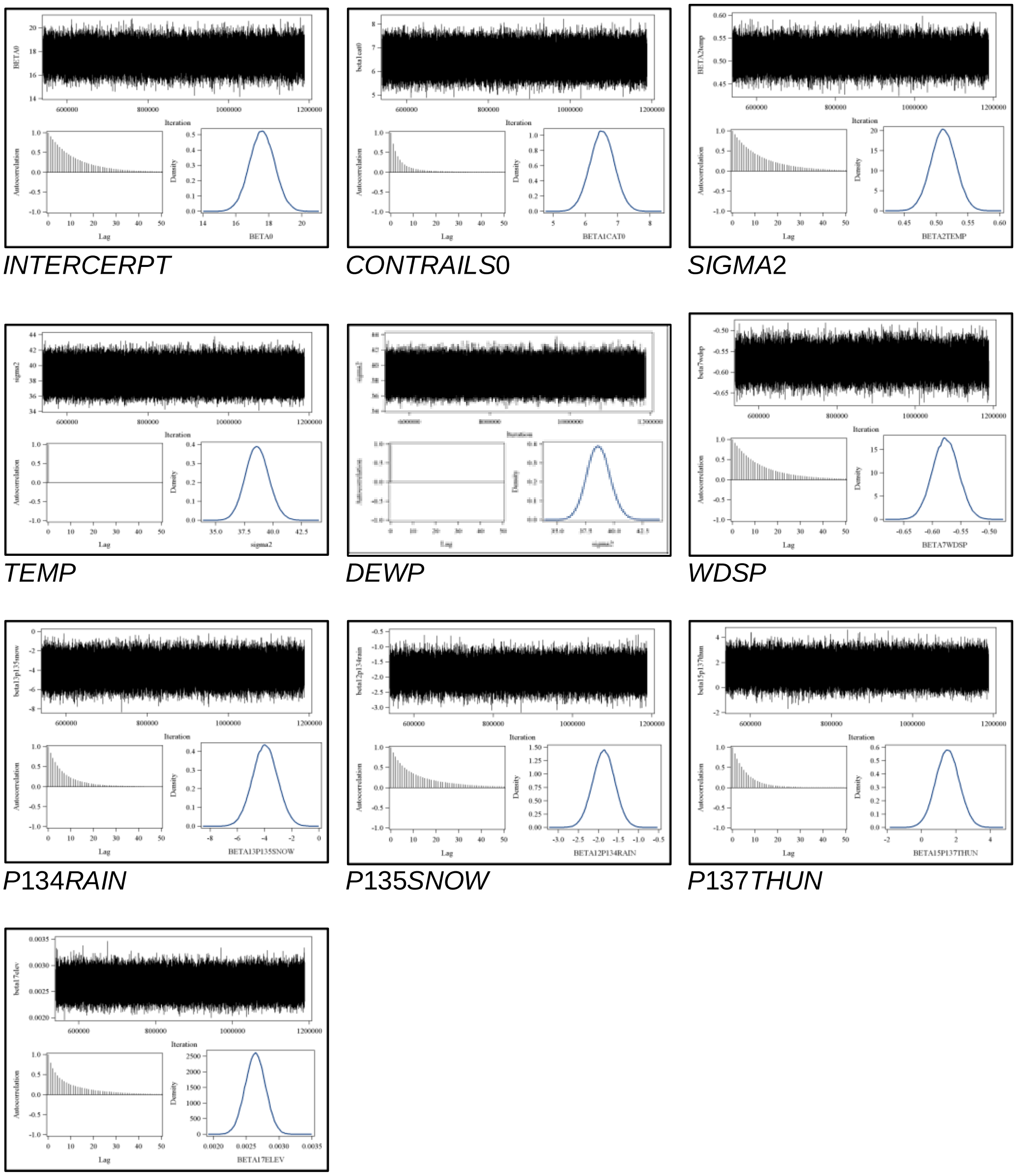
			***METHOD*2**		***METHOD*4**		***METHOD*8**	
			PROP / NO OVERLAP		**PROP / OVERLAP**		**REGRESSION**	
n*trt*	**Ratio**	**Runs**	***CONTRAILS*0 ESTIMATE**	**|ΔPS|**	***CONTRAILS*0 ESTIMATE**	**|ΔPS|**	***CONTRAILS*0 ESTIMATE**	**R^2**
136.3	4.8	1620	6.7871	0.0194	6.9654	0.0141	6.6889	0.5752
288	2	180	7.0458	0.0615	7.1010	0.0545	6.5149	0.5750
216	3	180	6.6921	0.0275	6.8047	0.0218	6.5790	0.5743
162	4	180	6.6644	0.0181	6.8320	0.0125	6.6631	0.5750
136	5	180	6.6833	0.0136	6.8779	0.0086	6.7120	0.5746
108	6	180	6.6957	0.0123	6.9054	0.0074	6.7504	0.5741
96	7	180	6.8237	0.0107	7.0324	0.0061	6.7797	0.5747
81	8	180	6.8053	0.0108	7.0089	0.0058	6.7693	0.5753
72	9	180	6.7258	0.0107	6.9667	0.0053	6.7130	0.5768
68	10	180	6.9478	0.0098	7.1593	0.0046	6.7184	0.5771

Note: Top row is mean for all runs; other rows are means for that resample level

Bayesian

***METHOD*9 / MCMC, Best Model** The best model estimated *CONTRAILS*0 as 6.5195 (α = 0.05 HPD Interval 5.7795, 7.2552). The model was a normal posterior predictive distribution with normal priors for effects and inverse gamma for variance. Non-informative priors failed to generate a reasonable model, based upon diagnostic plots or Geweke. Therefore, informative priors were set as the estimates from the best model regression, *METHOD*6 (Table 5). Variances were set at 100, except elev which was 5×10^{-8}. The MCMC was performed with five blocks: (1) *CONTRAILS*0, (2) *Intercept*, *temp*, *dewp*, *wdsp*, (3) *p*134*rain*, *p*135*snow*, *p*137*thun*, (4) *elev*, and (5) σ^2. The groups were based on correlations and commonality of data collection.

Acceptance rates ranged from 0.2200 to 0.3040 at the end of the tuning period, 540k burn-in, and 648k sampling. Visually, the diagnostic plots revealed convergence of parameter means, increasingly diminished autocorrelations, and normal posterior density distributions (Table 5, Figure 2). Geweke diagnostic was 0.6480 for *CONTRAILS*0, and ≥ 0.1181 for all others. All were ≥ 0.05, indicative that the final 50% of runs featured posterior parameter estimates that were not statistically different than of the initial 10%.



INTERCERPT *CONTRAILS*0 *SIGMA*2

TEMP *DEWP* *WDSP*

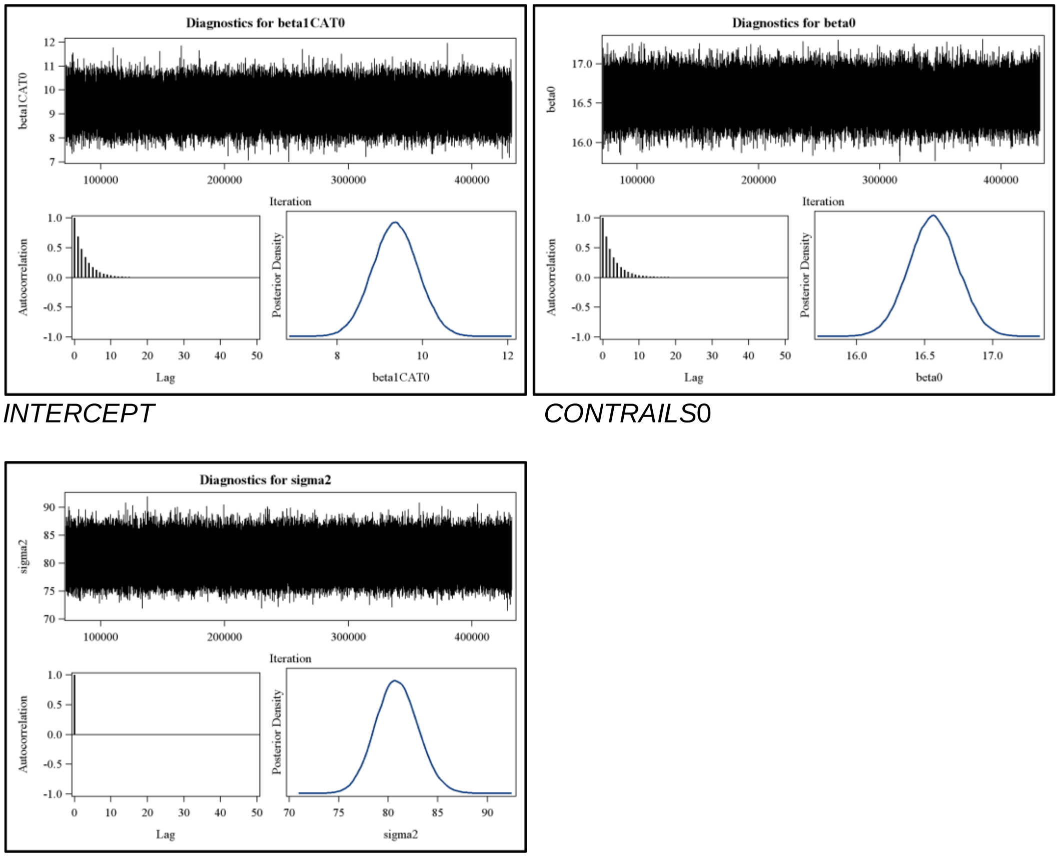
*P*134*RAIN* *P*135*SNOW* *P*137*THUN*

ELEV

Figure 2. Diagnostic Plots, *METHOD*9

METHOD 10 / MCMC, Minimal Model The minimal model estimated *CONTRAILS*0 as 9.3609 ($\alpha = 0.05$ HPD Interval 8.3327, 10.4145). The model was a normal posterior predictive distribution with normal priors for effects and inverse gamma for variance. The MCMC was performed with three blocks, one for each of *beta*0 (intercept), *beta*1*CAT*0 (*CONTRAILS*0), and σ^2. Non-informative priors were used because they proved sufficient.

Acceptance rates were from 0.3528 to 0.3720 for end-tuning period, 72k burn-in, and 360k sampling. Visually, the diagnostic plots also revealed convergence of means, increasingly diminished autocorrelations, and normal posterior density distributions (Figure 3). Geweke diagnostics were all > 0.1700.



INTERCEPT

*CONTRAILS*0

*SIGMA*2

Figure 3. Diagnostic Plots, *METHOD*10

***CONTRAILS*0 Estimate Testing** The *CONTRAILS*0 estimates from MCMC simulation *METHOD*9 analysis revealed that 15.65% were greater than the estimate from *METHOD*1. For any run, the minimum MCMC posterior estimate was 4.8125; the maximum was 8.2761. Thus, all 180k runs were greater than the Travis estimate (Table 7).

Table 7. Comparison of *CONTRAILS*0 Estimates

METHOD	TYPE	DESCRIPTION	*CONTRAILS*0 ESTIMATE	*METHOD*9 PERCENT >*CONTRAILS*0
1	PROPENSITY		6.8981	15.65%
2	PROPENSITY	RESAMPLING	6.7871	23.79%
3	PROPENSITY	OVERLAP	6.8931	15.96%
4	PROPENSITY	RESAMPLING & OVERLAP	6.9654	11.79%
5	REGRESSION	FULL MODEL	6.5513	46.77%
6	REGRESSION	BEST MODEL	6.5173	50.41%
7	REGRESSION	MINIMAL MODEL	9.3605	0.00%
8	REGRESSION	BEST MODEL & RESAMPLING	6.6889	32.71%
9	MCMC	BEST MODEL	6.5195	50.15%
10	MCMC	MINIMAL MODEL	9.3609	0.00%
		Travis estimate	3.24	100.00%

Covariate and contrail effect estimate comparisons Of the 14 covariates, the original data had 10 with statistically significant mean differences between the treatment and control groups, as indicated by the bold figures in Table 8. With *METHOD*1, for all covariates, one fails to reject the H_0 that the means in the treatment and control groups are equal.

Table 8. Comparison of Covariates of Original and Propensity Score Matched Data

		ORIGINAL (n*ttl*=2879)				**MATCHED (n*ttl*=644)**			
ID	**Variable**	**variance**	**df**	***t*-value**	***p*-value**	**variance**	**df**	***t*-value**	***p*-value**
1	*CONTRAILS*0				NA				NA
2	*temp*	Unequal	505	-9.02	**<0.0001**	Unequal	595	-0.18	0.8573
3	*dewp*	Unequal	439	-1.36	0.1746	Unequal	621	0.13	0.8983
4	*slp*	Unequal	453	-6.26	**<0.0001**	Equal	642	-0.16	0.8738
5	*visib*	Unequal	696	8.37	**<0.0001**	Unequal	518	-0.37	0.7148
6	*wdsp*	Unequal	750	5.34	**<0.0001**	Unequal	517	0.01	0.9933
7	*MXSPD*	Unequal	2614	6.60	**<0.0001**	Unequal	326	0.74	0.4586
8	*PRCP*	Unequal	455	1.38	0.1697	Unequal	582	-0.40	0.6861
9	p133*fog*	Unequal	366	-3.89	**<0.0001**	Equal	642	0.31	0.7530
10	p134*rain*	Equal	2877	2.41	**0.0159**	Equal	642	-0.77	0.4431
11	p135*snow*	Unequal	970	4.13	**<0.0001**	Unequal	578	0.58	0.5635
12	p137*thun*	Unequal	351	-3.22	**0.0014**	Equal	642	-0.79	0.4307
13	*elev*	Unequal	448	-0.39	0.6951	Unequal	634	-1.60	0.1106
14	*latitudecorr*	Equal	2877	-0.02	0.9859	Equal	642	-1.33	0.1845
15	*latitudeabs*	Unequal	518	6.50	**<0.0001**	Unequal	584	0.37	0.7111
16	*TEMPRANGE*				NA				NA

**Note.* Differences that are statistically significant at $\alpha = 0.05$ are in bold

The *CONTRAILS*0 estimate from PS matching using all observations without overlap was 6.8981. This was a statistically significantly difference from the Travis estimate ($t = 5.26$, $p < 0.0001$).

Except for the minimal models (*METHOD*7, *METHOD*10), the contrails effect estimate within the 95% confidence interval of *METHOD*1, and therefore did not represent statistical difference. Due to its simplicity, *METHOD*1 was preferred over the other causal inference methods; due to covariate egalities, it was preferred over the regression and MCMC methods.

Omnibus distribution tests

Distributional differences were tested by three omnibus tests, Kolmogoroff-Smirnoff (KS), Cramér-von Mises (CM), and oando (Barron, 2007). The control and treatment group distributional differences were statistically significant for KS for the analysis dataset (n*trt* = 322, n*cntl* = 2557, D = 0.4412, $p < 0.001$) and propensity matched data subset (n*trt* = n*cntl* = 322, D = 0.3571,

$p < 0.001$). CM also indicated statistical significance for both datasets (CM = 25.2482, $p < 0.001$ and CM = 8.6263, $p < 0.001$).

Oando performed with resampling yielded inconclusive results. For 4000 resampling runs, the mean p-value = 0.1988, 32.73% p-value ≤ 0.05, and 11.20% p-value > 0.50. For 180 resampling runs, the matched dataset had mean p-value = 0.2871, 22.22% p-value ≤ 0.05, and 20.56% p-value > 0.50.

Due to the definition of oando which weights by the rank of the gap from the prior observation, the result is possibly reflective of a non-homogenous range. In response, the analysis dataset was partitioned according to *temprange* rank. The low and high ends were curves; the middle was linear. *REGION*1 was defined as the union of the low (*REGION*1A) and high (*REGION*1B); *REGION*2 was defined as the middle (Figure 4).

The *temprange* difference between control and treatment represented the *CONTRAILS*0 estimate. For the entire analysis dataset, the union of *REGION*1 and *REGION*2, the *temprange* difference was 9.3605. For mid-*tempranges* of 12.20 to 20.22, the *CONTRAILS*0 estimate for *REGION*2 was not statistically different from zero. However, for *REGION*1 the contrails effect estimate was 11.4521 (p-value < 0.0001).

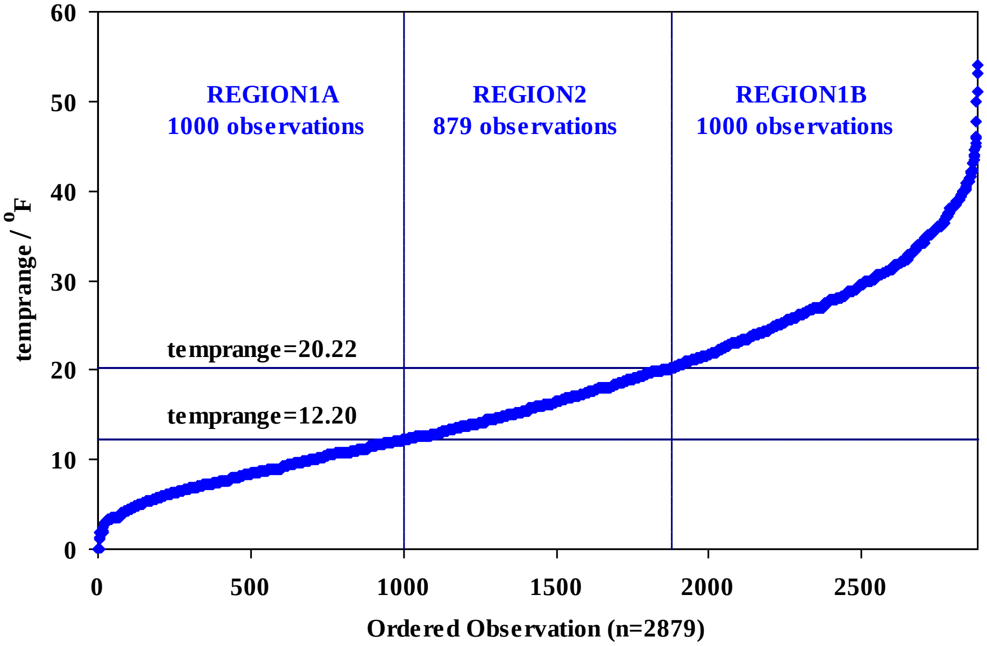


Figure 4. Partition of Ordered Observations of Analysis Dataset

Conclusion

The contrail effect was estimated to equal a 6.8981°F decrease in the daily temperature range at ground level on planet Earth using propensity matching, *METHOD*1. This result was statistically different from the Travis estimate of 3.24°F.

Although rarely studied, daily temperature range does impact animal populations and population dynamics (Viterbi, et al., 2012). Smaller daily temperature ranges have been shown to decrease the black grouse bird population in Italy (Viterbi, et al., 2012) as well as to influence Moluccan Woodcock population density in Indonesia (Eden, et al., 2013). The impact upon other species may also be significant (Eden, et al., 2013).

In pursuit of fuel economy, modern engines sport a greater efficiency of propulsion. However, aircraft equipped with such engines generate contrails starting at lower altitudes (Schumann, et al., 2000), and up to higher altitudes (Schumann, 2000). More persistent contrails could shadow even more of the Earth's surface than the 16% EPA estimation (EPA, 2000).

Analysis variables were solely based upon the NCDC datasets. Other data might have been useful, for example, temperature and other atmospheric measurements taken at altitudes at 25,000 feet; NCDC measures only at ground level. The restriction to data from a single source obviated the need to judge relative reliability of different databases, measurement devices, and data collection procedures.

Two omnibus tests, Kolmogoroff-Smirnoff and Cramér-von Mises, confirmed distributional differences between treatment and control groups, supportive of the propensity score matching results. The third, oando, revealed that the data might be an amalgam of two regions, center and extremes. Future explorations could introduce an indicator variable reflecting such a partition, or, fractionate into individual analyses.

The correlation between daily mean temperature and *CONTRAILS*0 of 0.1267 was consistent with an association of higher mean temperatures at ground level with absence of contrails. This was in agreement with the minority; most prior studies have indicated a net warming effect, but inconclusively (Mims, Chambers & Oostra, n.d.). The NCDC data calculates its reported daily mean temperatures based upon the actual operating hours for that specific station (Lott, 2010). Mean temperature theoretically might also be defined as the mean of 24 hourly readings, or many other possible variants. Alternatively, the median measurement might be a reasonable reflection of central tendency. These considerations could cloud conclusions. Because the contrails effect upon daily mean temperature was not the focus of this analysis, techniques employed were not aimed at obtaining such an estimate. Therefore, although interesting, any inferences regarding daily mean temperature are merely ancillary.

Author Contributions

The seminal concept of subjecting contrail-related data to causal inference was by J.H.B., and subsequently developed and discussed by J.H.B. and D.S.B. The logistic regression was coded in SAS by J.H.B. The remainder of the research, SAS coding, and all writing was performed by D.S.B.

References

Barron, D. S. (2007). *Kolmogoroff-Smirnoff Enhancement* (Master's thesis). Retrieved from ProQuest Dissertations and Thesis Full Text. (1451232).

Barron, D. S. (Photographer). (2013). *Contrail 33* [photograph], Personal collection.

Bauer, P., Read, A, & Johnson, P. (2000). *The gravity geophysical method*. Retrieved from http://geoinfo.nmt.edu/geoscience/projects/astronauts/gravity_method.html

Bristow, K. L & Campbell, G. S. (1984). On the relationship between incoming solar radiation and daily maximum and minimum temperature. *Agricultural and Forest Meteorology*, *31*(2), 159-166. doi:10.1016/0168-1923(84)90017-0

Cottee-Jones, H. E. W., Mittermeier, J. C., & Redding, D. W. (2013). The Moluccan Woodcock Scolopax rochussenii on Obi Island, North Moluccas, Indonesia: a ‘lost’ species is less endangered than expected. *Forktail* (29), 88-93.

Gelman, A. & Hill, J. (2006) *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge, UK: Cambridge University Press.

Gray, D. M. & Male, D. H. (1981). *Handbook of snow: Principles, processes, management and use.* Toronto, Canada: Pergamon Press.

Lott, N. (2010). The National Climatic Data Center. *Federal climate complex: Global surface summary of day data, version 7* [Data file]. Retrieved from ftp://ftp.ncdc.noaa.gov/pub/data/gsod/readme.txt

Meerkotter, R., Schumann, U., Doelling, D. R., Minnis, P., Nakajima, T., & Tsushima, Y. (1999). Radiative forcing by contrails. *Annales Geophysicae, 17*(8), 1080-1094. doi:10.1007/s00585-999-1080-7

Mims, F. M., Chambers, L. & Oostra, D. H. (n.d.). C*ontrail studies*. Retrieved from http://mynasadata.larc.nasa.gov/science_projects/contrail-studies/

Moore, M. (Director & Writer). (2004a). *Fahrenheit 9/11* [Motion picture]. United States: Fellowship Adventure Group (presented by), Dog Eat Dog Films, Miramax Films (uncredited).

Moore, M. (2004b). *What Fahrenheit 9/11 says about the Saudi flights out of the country after September 11*. Retrieved from http://www.michaelmoore.com/words/fahrenheit-911-facts/what-fahrenheit-911-says-about-the-saudi-flights-out-of-the-country-after-september-11

The National Climatic Data Center. (2010). *Integrated Surface Data, version 7* [Data files]. Available from ftp://ftp.ncdc.noaa.gov/pub/data/gsod/

National Retail Federation (2013). *FAQs 4-5-4 calendar*. Retrieved from http://www.nrf.com/modules.php?name=Pages&sp_id=392

Schumann, U. (2000). Influence of propulsion efficiency on contrail formation. *Aerospace Science and Technology 4*(6), 391–401. doi:10.1016/S1270-9638(00)01062-2

Schumann, U., Busen, R., & Plohr, M. (2000). Experimental test of the influence of propulsion efficiency on contrail formation. *Journal of Aircraft, 37*(6), 1083-1087. doi:10.2514/2.2715

Schumann, U. (2005). Formation, properties and climatic effects of contrails. *Comptes Rendus Physique, 6*(4-5), 549-565. doi:10.1016/j.crhy.2005.05.002

Travis, D. J., Carleton, A. M., & Lauritsen, R. G. (2002). Climatology: Contrails reduce daily temperature range, brief communications. *Nature, 418*, 601. doi:10.1038/418601a

United States Environmental Protection Agency. (2000). *Aircraft Contrails Factsheet*. EPA430-F-00-005, September 2000 Air and Radiation (6205J).

Viterbi, R., Imperio, S., Alpe, D., Peverelli, V.B., & Provenzale, A. (2012). Climatic control and population dynamic of black grouse in the Western Italian Alps. *Istituto di Scienze dell'Atmosfera e del Clima, CNR, Relazioni finale*. Retrieved from http://www.regione.piemonte.it/agri/area_tecnico_scientifica/osserv_faun/progetti/dwd/alcotra/relFinali/1T.pdf

Wendler, G. & Stuefer, M. (2002). *Improved contrail forecasting techniques for the subarctic setting of Fairbanks, Alaska*. Geophysical Institute University of Alaska, University of Alaska Fairbanks Special Report UAG R-329.

Journal of Modern Applied Statistical Methods
November 2015, Vol. 14, No. 2, 257-274.

ISSN 1538 − 9472

Statistical Modeling of Migration Attractiveness of the EU Member States

Tatiana Tikhomirova
Plekhanov Russian University of Economics
Москва, Russia

Yulia Lebedeva
Plekhanov Russian University of Economics
Москва, Russia

Identifying the relationship between the migration attractiveness of the European Union countries and their level of socio-economic development is investigated. An approach is proposed identify influences on migration socio-economic characteristics, by aggregating and reducing their diversity, and substantiating the cause-and-effect relationships of the studied phenomenon. A stable classification of countries scheme is developed according to the attractiveness of migration on aggregate factors, and then an econometric model of a binary choice using panel data for 2008-2010 was applying, quantifying the impact of aggregate designed factors on immigration and emigration.

Keywords: Immigration attractiveness, immigration, emigration, applied statistics, multivariate statistical techniques, multi-dimensional space, the panel data.

Introduction

Migration is “one of the most important challenges of the 21st century” (Albertinelli et al., 2011; “Migrants in Europe”, 2012). This phenomenon, caused by rising unemployment, increasing crime, the destruction of the traditional indigenous way of life, increasing the burden on the budget, and many other negative consequences, particularly when unregulated or illegal urges governments of developed countries to take certain measures to regulate migration flows within the appropriate migration policies.

The problems of developing an effective migration policy are also becoming more pressing in the European Union (EU), particularly because of the open borders within the framework of this community. In such a situation, the regulation of migration within the EU is usually associated with exposure to the factors generating the process and the living conditions of the population, of which, according to experts, the most important is the difference in the levels of living of the population and socio-economic development of the community.

Prof. Dr. Tikhomirova is Professor of Economics. Email Tatiana Tikhomirova at t_tikhomirova@mail.ru. Email Yulia Lebedeva at live_wire@mail.ru.

An important stage of the development and validation of measures of migration policy in the EU is the clusterization of countries within homogeneous groups in terms of socio-economic development and identification of the main reasons - factors that determine the patterns of intra-group and between group processes. In this regard, the construction of a sustainable clusterization of EU countries in terms of the attractiveness of migration is considered, as well as the identification of factors that have an impact on migration, and cause the differences in the development of EU countries, is relevant.

In this study, the 29 countries of the European Union: Austria, England, Belgium, Bulgaria, Hungary, Germany, Denmark, Greece, Ireland, Spain, Italy, Cyprus, Latvia, Lithuania, Luxembourg, Malta, Netherlands, Norway, Poland, Portugal, Romania, Slovakia, Slovenia, Finland, France, Croatia, Czech Republic, Sweden, Estonia were examined for the period of 2008 to 2010, on 84 social, economic and political indicators: compensation of employees, GDP per capita in PPS, life expectancy at birth by sex, the number of pupils and students, self-reported unmet need for medical examination or treatment by income quintile, etc. (Sartori, 2012; "Migration and migrant population statistics", 2015; Institul National de Statistica, n.d.; National Statistics Office, Malta, n.d.). All these characteristics are given in comparable units of measurement.

In general, the source data set is a parallelepiped (see Figure 1), where the axis P_i belongs to EU member states $i = \overline{1.29}$, the axis X_j belongs the previously mentioned socio-economic and demographic characteristics of EU countries $j = \overline{1.84}$, and the axis t_k is time interval, $t = \overline{1.3}$.

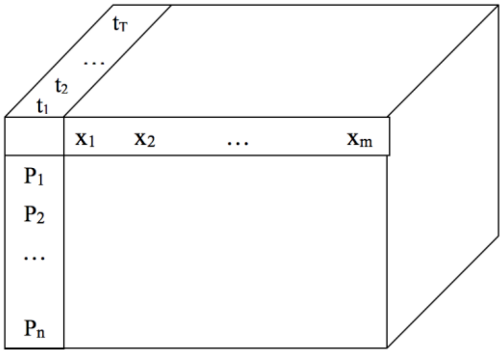


Figure 1. Parallelepiped of initial data on indicators of the attractiveness of the EU member states migration

When considering the information set numerous problems appear: 1) selecting the informative features that have a significant statistical effect on the migration, 2) reducing the dimensions of the array of information and the transition to the matrix representation of the data, 3) selecting the correct mathematical tools for analyzing small samples in which the number of signs exceeds the number of objects (29 × 84), making it impossible to construct the set econometric models, 4) recovering the gaps in the baseline data, 5) leveling the effect of multicollinearity between variables without significant loss of information content of the feature space (Tikhomirov, Tikhomirova, Oushmaev, 2011).

The first problem (the assessment of the relationship between factors and migration attractiveness of countries) was solved in several stages. With the help of multiple correlation analysis those features that have the greatest impact on statistical indicators of officially registered immigrants and emigrants were selected from the total number of socio-economic and demographic indicators. It was found that 32 of the 84 characteristic have a significant impact on immigration and 9 characteristic have a significant impact on emigration.

In the next step the combined influence of selected characteristics on migration attractiveness of countries was investigated, using the approach proposed by the authors: scaling of countries by aggregated, randomized indicators. This approach lies in the fact that the selected indicators are assigned levels according to the following principle: if the data has a direct correlation to the corresponding endogenous variable, the number of officially registered emigrants, or the number of registered immigrants, i.e., the correlation coefficient between the factor variable and efficient variable is significant and positive, then the ranks assigned to each variable are as follows: the observation with the highest value is assigned the maximum rank and levels are in descending order.

If the variable has an inverse relationship with the endogenous variable, i.e. the correlation coefficient is significant and negative, then the ranks are assigned to each variable in the reverse order: the observation with the largest value has a rank corresponding to one and then ranks are arranged in ascending order. Then the sum of the ranks corresponding to all variables influencing the emigration and immigration for each country is calculated:

$$R_i = \sum_j R_{ij} \tag{1}$$

where i is the serial number of the country $i=\overline{1.29}$; j is the serial number of variable X_j, $j=\overline{1.9}$ (for emigration) and $j=\overline{1.32}$ (for immigration), R_i is the sum of the ranks in the country with the number of i, R_{ij} is the rank assigned to the i^{th} observation of the j^{th} variable.

In the next step, the percentage deviation of the sum of the rank of each country from the median level of emigration and immigration in the EU respectively is calculated:

$$\Pi_i = \frac{(R_i - M)}{M} * 100\% \qquad (2)$$

where M is the median for all R_i.

The scaling was produced with respect to values of percentage: from the largest percentage to the lowest value of percentage. The results of the calculations by the variables of the attractiveness of emigration are presented in Figure 2. Figure 3 shows the ranking of EU countries, which was built for the number of officially registered emigrants per thousand inhabitants.

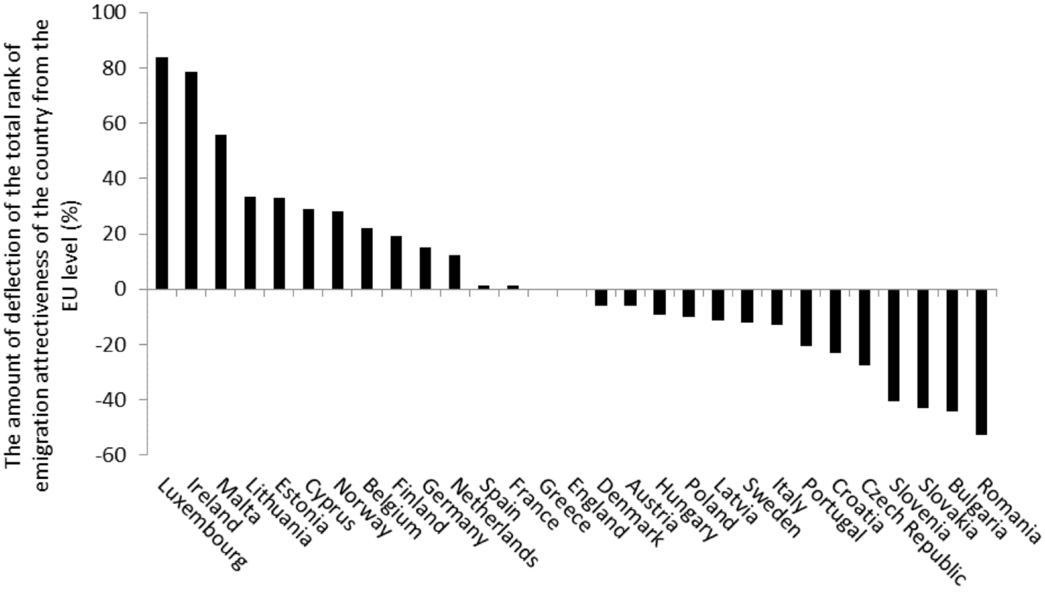


Figure 2. Distribution of the EU countries with respect of emigration attractiveness

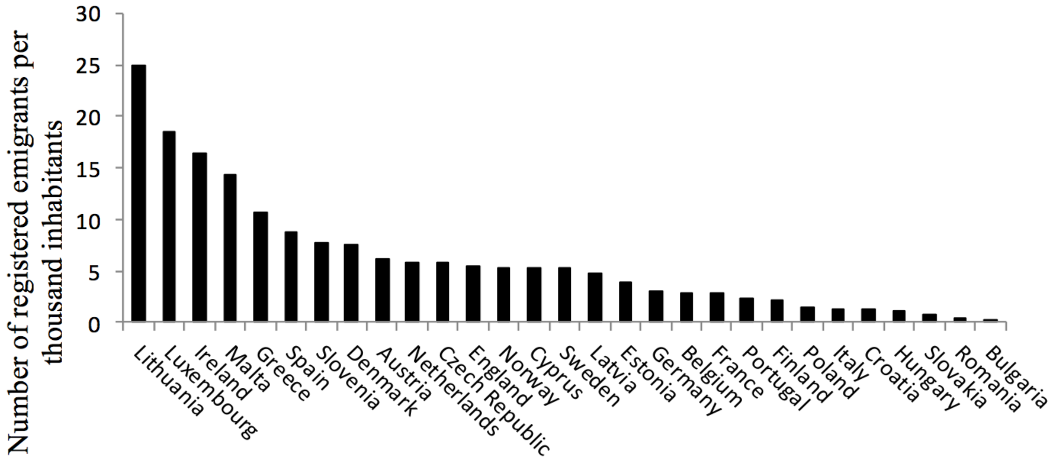


Figure 3. The distribution of EU countries by number of registered emigrants per thousand inhabitants

Comparing the histograms placed in Figure 2 and Figure 3 indicates that these extreme values are the same. This indicates that the selected explanatory variables are really informative and their joint effect on the attractiveness of emigration is significant. Moreover, in Figure 2 and Figure 3, heterogeneity of the EU countries by selected characteristics is observed (their scatter relative to the EU median level is greater than 70%), which leads to the need for clustering of countries by studied characteristics. Similar results were obtained during the distribution of countries by number of immigrants and the characteristics that affect the level of immigration.

The next problem which we solved in this paper was caused by multicollinearity selected features. Statistics of Pearson has confirmed the presence of multicollinearity in features of emigration attractiveness $\left(\chi^2_{est} = 60.8 \succ \chi^2_{tab.} = 50.9\right)$ on 99% confidence level. It should be noted that the multicollinearity of the features of immigration attractiveness was not statistically established.

Cluster analysis (see Figure 4), which was built for the 9 variables of emigration attractiveness, also shows a relationship between them.

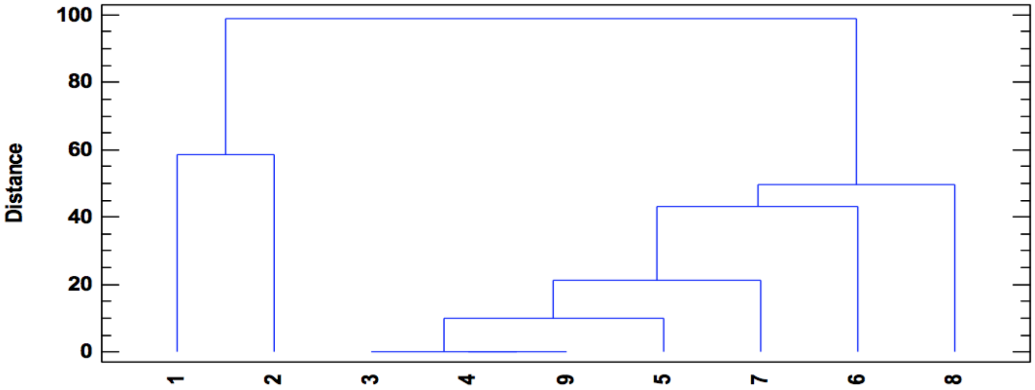


Figure 4. Cluster analysis dendrogram constructed for the 9 variables of emigration attractiveness using method farthest neighbor (the square of the Euclidean metric)

The dendrogram (Figure 4) shows that the variables numbered 3, 4, 9 are collinear. Step by step, we removed one variable, which had the least variation, from consideration. This meant that the space variables of emigration attractiveness of EU countries were reduced with no loss of informativity and the problem of multicollinearity was solved. Statistics of Pearson after the removal of collinear variables (no. 4 and 9) was: $\left(\chi^2_{est.} = 32.8 \prec \chi^2_{tab.} = 38.9\right)$ for $\alpha = 0.01$.

A similar approach was applied to reduce the feature space on the immigration attractiveness of EU countries.

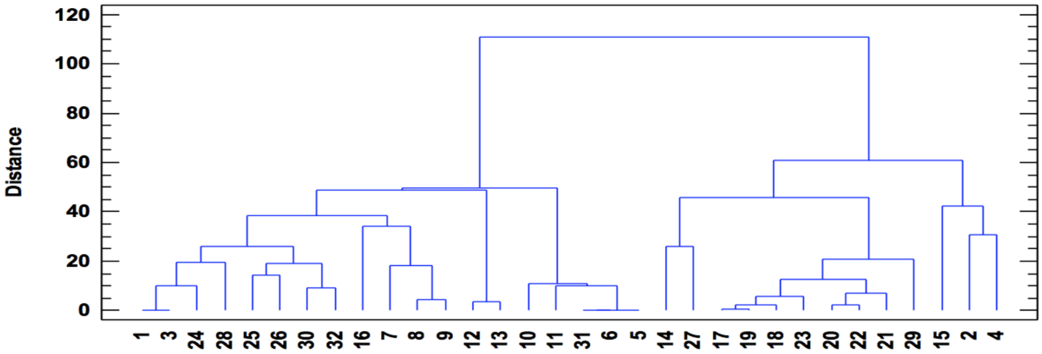


Figure 5. The dendrogram of the cluster analysis for the 32 features of immigration attractiveness of the EU (the square of the Euclidean metric)

As follows from the dendrogram (Figure 5), many of the characteristics are closely related, although the statistic of Pearson did not confirm the presence of multicollinearity for them, apparently, due to the excessive multi-dimensionality (the number of variables, in this case exceeds the number of observations). For example, variables number 1 and 3, as well as 5 and 6, are collinear so for further research it is advisable to leave only one of each pair, based on the principle that the most preferred variable is the one with higher variability.

Reduction of the feature space of immigration attractiveness was conducted in several iterations. Moreover, in each iteration of the classification we built on a selected set of variables until the requirements of sustainability were met (Tikhomirov et al. 2011). In order to obtain a stable classification 9 iterations took place. The final dendrogram of cluster analysis of EU countries by immigration attractiveness based on many of its defining characteristics is represented in Figure 6.

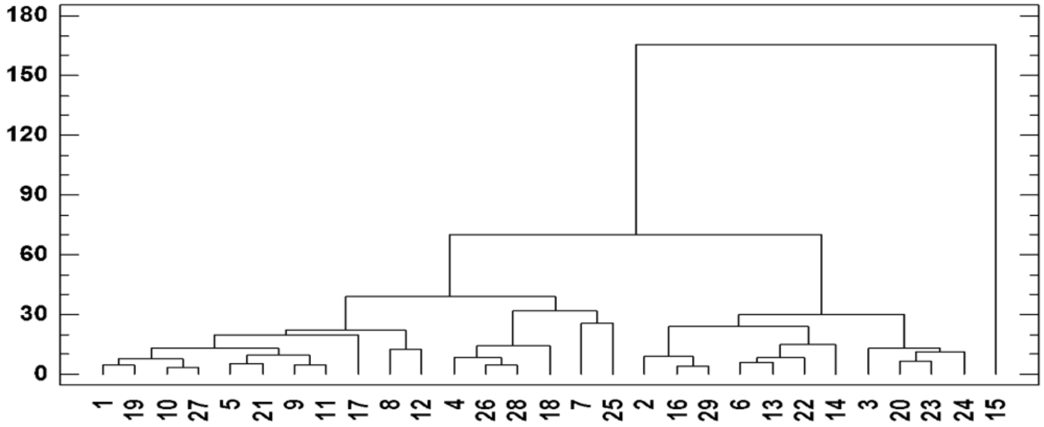


Figure 6. The dendrogram of cluster analysis of immigration attractiveness of the EU countries by using far neighbor method (the square of the Euclidean metric)

From the obtained clustering of countries it follows that they can be divided into two groups of immigration attractiveness (see Table 1 and Figure 7). The first group includes countries of the former capitalist camp, and the second group has the countries of the former socialist camp. Luxembourg (the object no.15) is located out of the general mass of EU countries and cannot be added to either of the groups.

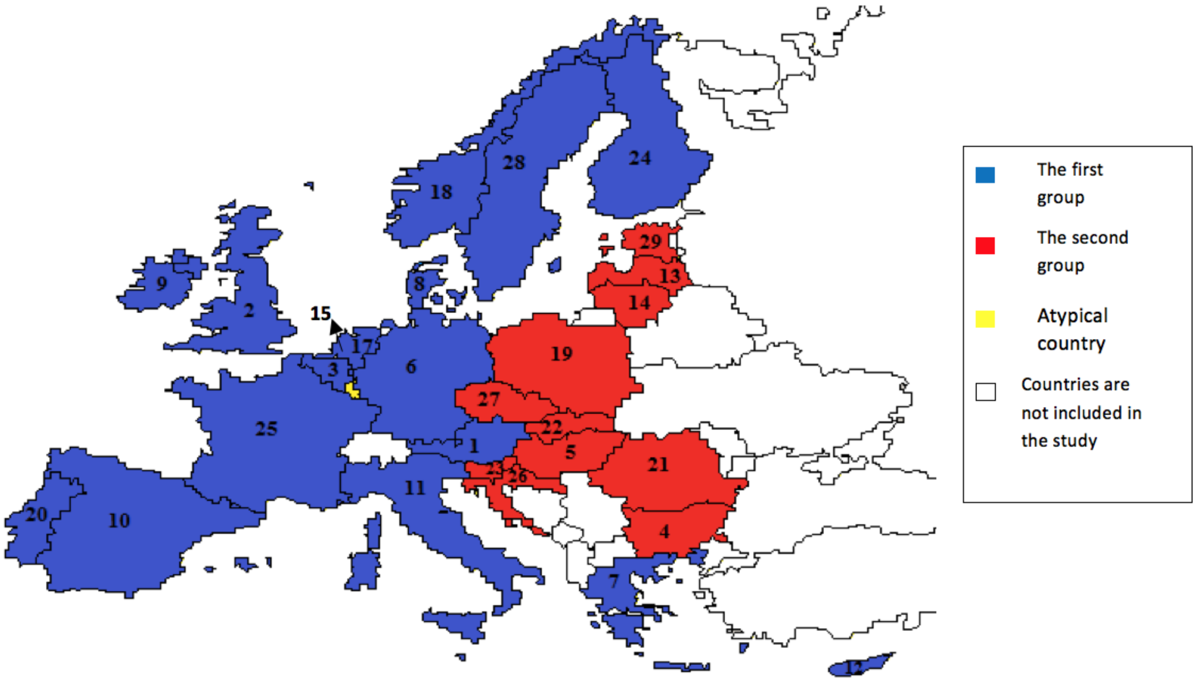
Table 1. The distribution of EU countries by immigration attractiveness by homogeneous groups

1 group	Belgium, Denmark, Germany, Ireland, Greece, Spain, France, Italy, Cyprus, Malta, the Netherlands, Austria, Portugal, Finland, Sweden, England, Norway.
2 group	Bulgaria, The Czech Republic, Estonia, Latvia, Lithuania, Hungary, Poland, Romania, Slovenia, Slovakia, Croatia.
Unclassified country	Luxembourg.

On the map (Figure 7) these three groups are displayed. The blue color indicates the countries included in the first group, and the red indicates those in the second. Countries which are not included in the review are white.

The quality of the classification was confirmed by discriminant analysis. The percentage of correctly classified cases (in the application of discriminant analysis) was 100%. From the results of the discriminant analysis, shown in Figure 8, it follows that the groups of countries are located far enough away from each other to indicate their significant differences in immigration attractiveness.

As a result of the statistical analysis 15 variables which have an impact on immigration were selected from 84 variables, such as: final consumption expenditure of households and non-profit organizations serving households as a percentage of GDP, net national income as a percentage of GDP, direct investment flows abroad as a percentage of GDP, natural decline in population per thousand residents, the number of students in higher education per one thousand inhabitants (the number of graduates between the ages of 20-29 years in mathematics, science and technology per thousand population), employment rate by highest level of education attained (the percentage of age group 20-64 years), overcrowding rate by tenure status (the percentage of owner, with mortgage or loan), the percentage of individuals aged 16 to 74 using the Internet for ordering goods or services from other EU countries, the percentage of individuals in aged 16 to 74 using a mobile phone via UMTS (3G) to access the Internet, number of deaths due to accidents, selected from standardized death rate by 100000 inhabitants, individuals seeking information on the Internet with the purpose of learning, life expectancy at birth (healthy life years) and 7 variables which have an impact on emigration: the gross fixed capital formation, defined as investment’s percentage of GDP; the gross fixed capital formation, defined as investment’s percentage of GDP; population of foreigners by citizenship; the



1	Austria	6	Germany	11	Italy	16	Malta	21	Romania	26	Croatia
2	England	7	Greece	12	Cyprus	17	Netherlands	22	Slovakia	27	Czech Republic
3	Belgium	8	Denmark	13	Latvia	18	Norway	23	Slovenia	28	Sweden
4	Bulgaria	9	Ireland	14	Lithuania	19	Poland	24	Finland	29	Estonia
5	Hungary	10	Spain	15	Luxembourg	20	Portugal	25	France		

Figure 7. The dendrogram of cluster analysis of immigration attractiveness of the EU countries by using far neighbor method (the square of the Euclidean metric)

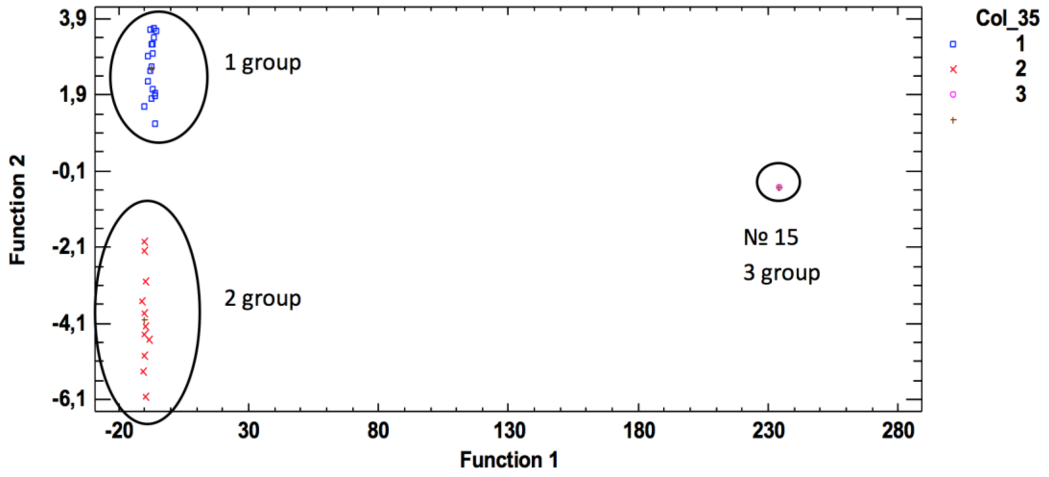


Figure 8. The distribution of EU countries by immigration attractiveness in projections of the discriminant functions

gender differences in the risk of poverty, the percentage from the group 65 years or over; the percentage of individuals aged 16 to 74 using the Internet for ordering goods or services from other EU countries; the volume of passenger transport relative to GDP.

In order to construct models of immigration and emigration attractiveness in the EU countries, the method of principal components was applied to selected variables. At this stage, aggregate variables were built. They affect the attractiveness of immigration and emigration, and are used in econometric modeling as regressors.

It should be noted that the classification of EU countries, held on principal components, retained their membership in the group (Figure 10).

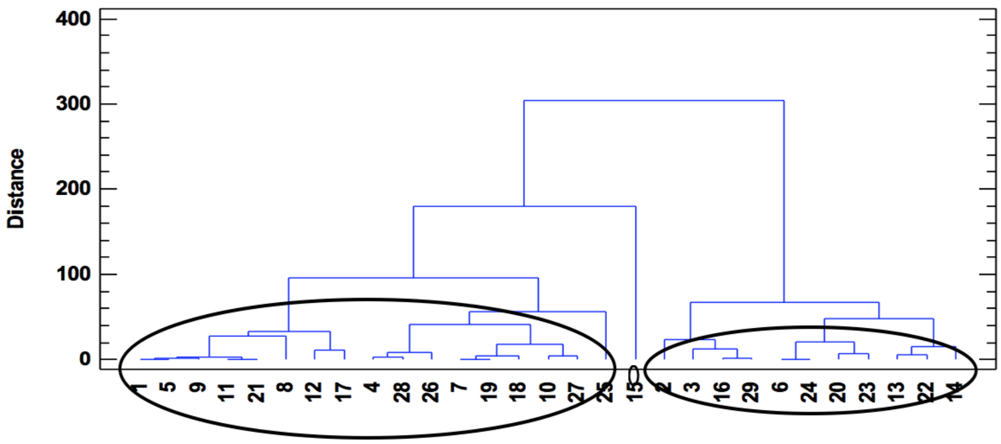


Figure 9. The dendrogram of the cluster analysis of the EU immigration appeal, based on principal components

This classification of countries is stable, which is confirmed by the results of the discriminant analysis (Figure 10).

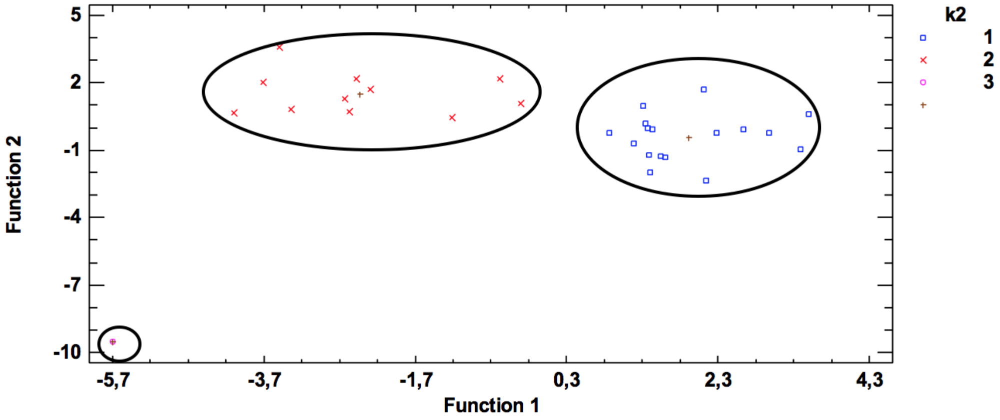


Figure 10. The distribution of the EU countries by main components of immigration attractiveness in projections of the discriminant functions

Similar calculations were carried out by emigration in EU countries. Classification of the EU countries of emigration attractiveness is robust and is presented in Figure 11 and Table 2.

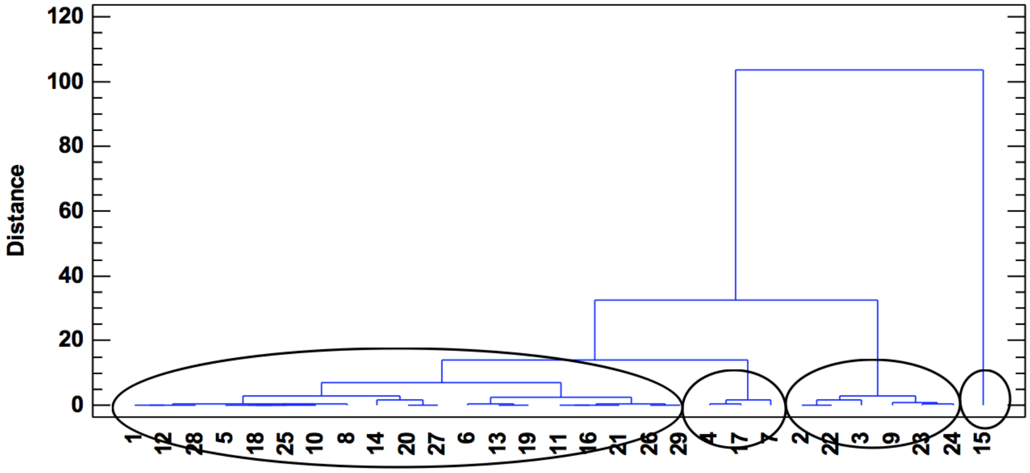


Figure 11. The distribution of the EU countries by main components of immigration attractiveness in projections of the discriminant functions

Table 2. Distribution of the EU countries by emigration attractiveness by homogeneous groups

1 group	Austria, England, Belgium, Hungary, Germany, Greece, France, Italy, Cyprus, Latvia, Lithuania, Netherlands, Norway, Poland, Portugal, Finland, Croatia, Sweden, Estonia.
2 group	Denmark, Ireland, Malta.
3 group	Bulgaria, Spain, Romania, Slovakia, Slovenia, Czech Republic
Unclassified country	Luxembourg.

Consider the results of principal component analysis for the characteristics of the immigration and emigration attractiveness of EU countries. The system of equations of principal components of immigration is as follows:

$$\begin{cases} F_1 = -0.22x_2 + 0.14x_4 - 0.25x_5 - 0.3x_7 - 0.1x_{14} + 0.03x_{15} - 0.28x_{16} - 0.36x_{18} \\ \quad +0.3x_{24} - 0.30x_{25} - 0.3x_{26} - 0.25x_{27} + 0.24x_{28} + 0.35x_{33} + 0.22x_{34} \\ F_2 = 0.27x_2 - 0.09x_4 - 0.12x_5 - 0.02x_7 - 0.46x_{14} - 0.39x_{15} + 0.19x_{16} - 0.07x_{18} \\ \quad -0.06x_{24} - 0.19x_{25} - 0.29x_{26} - 0.39x_{27} - 0.32x_{28} + 0.19x_{33} + 0.28x_{34} \\ F_3 = 0.34x_2 + 0.62x_4 - 0.42x_5 - 0.27x_7 + 0.15x_{14} - 0.04x_{15} + 0.34x_{16} - 0.01x_{18} \\ \quad +0.01x_{24} + 0.16x_{25} - 0.08x_{26} + 0.04x_{27} + 0.25x_{28} + 0.07x_{33} - 0.04x_{24} \\ F_4 = 0.47x_2 - 0.03x_4 + 0.13x_5 + 0.39x_7 + 0.32x_{14} + 0.49x_{15} + 0.00x_{16} + 0.17x_{18} \\ \quad -0.15x_{24} - 0.26x_{25} - 0.28x_{26} + 0.00x_{27} + 0.01x_{28} + 0.22x_{33} + 0.16x_{24} \end{cases} \tag{3}$$

In the component F_1 the following variables have the greatest weight: 7 (the natural population change per 1000 inhabitants), 18 (the overcrowding rate by tenure status), 24 (the percentage of individuals aged 16 to 74 using the Internet for ordering goods or services from other EU countries), 25 (the percentage of individuals in aged 16 to 74 using a mobile phone via UMTS (3G) to access the Internet), and 26 (the percentage of the inhabitance from age 16 to 74 who use a laptop with wireless connection to access the Internet). Moreover, all variables except 18 have positive weights. It was decided that the F_1 describes the technical equipment of the country.

In the component F_2 the following variables have the greatest weight: 14 (the students in the tertiary education system per 1000 inhabitants), 15(science and technology graduates, defined as tertiary graduates in science and technology per 1000 of population aged 20-29 years and graduates in mathematics, science and technology per 1000 of population aged 20-29), 27 (number of deaths due to accidents, selected from standardized death rate by 100000 inhabitants), and 28 (individuals seeking information on the Internet with the purpose of learning, from individuals aged 16 to 74, who used the Internet within the last three months before the survey). All variables included in the component have negative weights. In this situation, it was assumed that F_2 is responsible for the low level of skills of the economically active population.

In the component F_3 the following variables have the greatest weight: 2 (the final consumption expenditure of households and non-profit institutions serving households measured as percentage of GDP), 4 (the net national income), 16 (the employment rate by highest level of education attained, from the age group 20-64 years. All variables included in this component have positive weights, so we considered it appropriate to characterize the F_3 as the level of production.

In the component F_4 the following variables have the greatest weight: 2 (the final consumption expenditure of households and non-profit institutions serving households as a percentage of GDP), 7 (the natural population change per 1000 inhabitants), 14 (the students in the tertiary education system per 1000 inhabitants), and 15 (science and technology graduates, defined as tertiary graduates in science and technology per 1000 of population aged 20-29 years and graduates in mathematics, science and technology per 1000 of population aged 20-29). F_4 can be interpreted as a country with a production oriented economy.

Principal component analysis applied to the variables of emigration attractiveness identified the following factors:

$$\begin{cases} F_1 = 0.55x_1 - 0.16x_2 - 0.39x_3 - 0.26x_5 - 0.43x_6 - 0.42x_7 - 0.31x_8 \\ F_2 = 0.15x_1 + 0.76x_2 - 0.19x_3 + 0.04x_5 - 0.39x_6 + 0.15x_7 + 0.43x_8 \\ F_3 = 0.16x_1 + 0.10x_2 + 0.24x_3 + 0.70x_5 - 0.30x_6 + 0.22x_7 - 0.53x_8 \end{cases} \tag{4}$$

In the component F_1 the following variables have the greatest weight: 1 (the gross fixed capital formation, defined as investment's percentage of GDP), 6 (the gender differences in the risk of poverty, the percentage from the group 65 years or over), 7 (percentage of individuals aged 16 to 74 using the Internet for ordering goods or services from other EU countries). The first variable has a positive weight, and the other two have negative. This suggests that F_1 is responsible for the underdevelopment of the domestic market of a country.

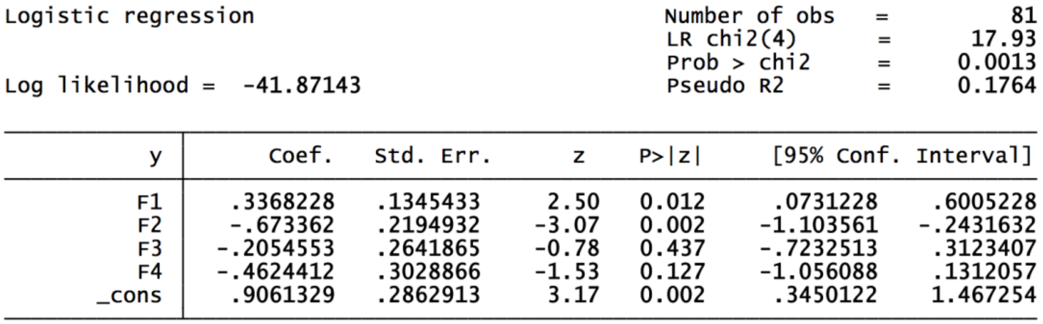
In the component F_2 the following variables have the greatest weight: 2 (the net national income) and 8 (the volume of passenger transport relative to GDP). All variables included in this component have positive weights. In this regard, F_2 can be interpreted as the skill level of the economically active population in a country.

In the component F_3 the following variables have the greatest weight: 5 (population of foreigners by citizenship) and 8 (the volume of passenger transport relative to GDP). It was decided that the component F_3 is responsible for the shortage of labor in a country.

For the studied countries, binary choice econometric models were built by panel data using principal components, which allowed us to quantify the degree of influence of identified factors in the migration attractiveness of the EU countries. The logit model for immigration in the EU countries, which is based on principal components (see Equation 3) is presented in Table 3.

As can be seen from the results of Table 3 coefficients of regressors F_1 and F_2 are statistically different from zero. In our case, confidence intervals for the parameter estimates $\hat{\beta}_1$ and $\hat{\beta}_2$ do not cover the zero on 95% confidence level. Factors F_3 and F_4 are not statistically significant, so on the second iteration of the modeling process they were removed from consideration (see Table 4).

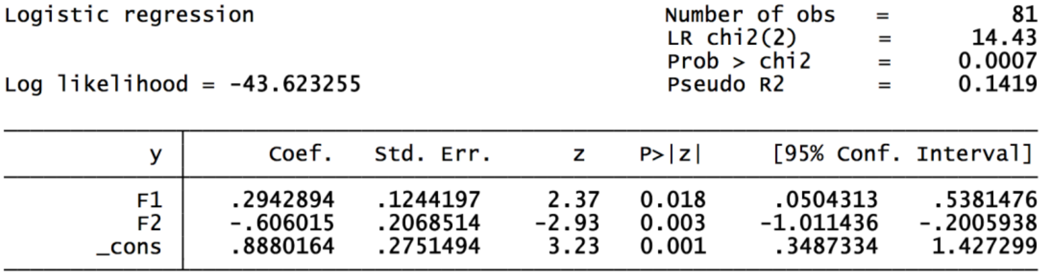
Table 3. Distribution of the EU countries by emigration attractiveness by homogeneous groups



```
Logistic regression                               Number of obs   =         81
                                                  LR chi2(4)      =      17.93
                                                  Prob > chi2     =     0.0013
Log likelihood =  -41.87143                       Pseudo R2       =     0.1764

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          F1 |   .3368228   .1345433     2.50   0.012     .0731228    .6005228
          F2 |   -.673362   .2194932    -3.07   0.002    -1.103561   -.2431632
          F3 |  -.2054553   .2641865    -0.78   0.437    -.7232513    .3123407
          F4 |  -.4624412   .3028866    -1.53   0.127    -1.056088    .1312057
       _cons |   .9061329   .2862913     3.17   0.002     .3450122    1.467254
------------------------------------------------------------------------------
```

Table 4. Statistical characteristics of the quality of the logit model of immigration



```
Logistic regression                               Number of obs   =         81
                                                  LR chi2(2)      =      14.43
                                                  Prob > chi2     =     0.0007
Log likelihood = -43.623255                       Pseudo R2       =     0.1419

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          F1 |   .2942894   .1244197     2.37   0.018     .0504313    .5381476
          F2 |   -.606015   .2068514    -2.93   0.003    -1.011436   -.2005938
       _cons |   .8880164   .2751494     3.23   0.001     .3487334    1.427299
------------------------------------------------------------------------------
```

Logit model of immigration in EU countries has the form:

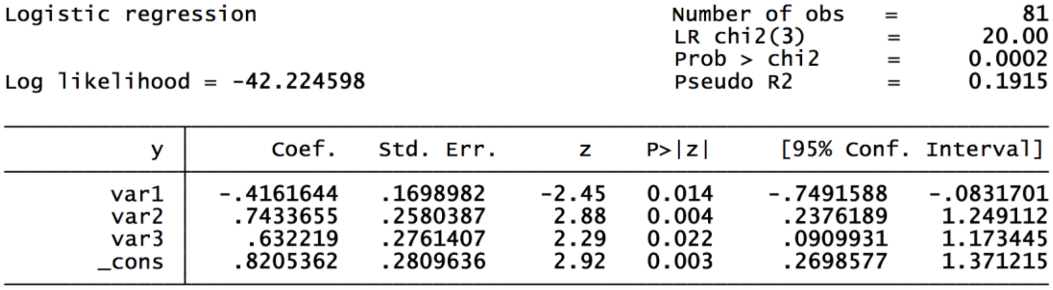
$$P\left(y_i=1\middle|x_i\right)=\frac{e^{0.9+0.3F_1-0.6F_2}}{1+e^{0.9+0.3F_1-0.6F_2}}\, ИP\left(y_i=1\middle|x_i\right)=\frac{1}{1+e^{0.9+0.3F_1-0.6F_2}} \qquad (5)$$

Factors F_1 and F_2 have an impact on immigration in the EU countries. The first factor F_1 has a positive impact, but factor F_2 has a negative one. It can be reasonably argued that an increase in technical equipment (development of IT technologies) and decrease of low-skilled economically active population increases the probability of a favorable immigration situation in the country. The level of well-being and the production orientation do not have a significant impact on the immigration attractiveness of the country. The logit model of emigration attractiveness, built on the principal components (see Equation 4) has the form:

$$P\left(y_i = 1 \middle| x_i\right) = \frac{e^{0.8-0.4F_1+0.7F_2+0.6F_3}}{1+e^{0.8-0.4F_1+0.7F_2+0.6F_3}} \text{И} P\left(y_i = 1 \middle| x_i\right) = \frac{1}{1+e^{0.8-0.4F_1+0.7F_2+0.6F_3}} \qquad (6)$$

The results of the calculation of the migration logit model for our binary data are shown in Table 5.

Table 5. Statistical characteristics of the quality of the logit model of emigration



```
Logistic regression                             Number of obs   =         81
                                                LR chi2(3)      =      20.00
                                                Prob > chi2     =     0.0002
Log likelihood = -42.224598                     Pseudo R2       =     0.1915

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        var1 |  -.4161644   .1698982    -2.45   0.014    -.7491588   -.0831701
        var2 |   .7433655   .2580387     2.88   0.004     .2376189    1.249112
        var3 |    .632219   .2761407     2.29   0.022     .0909931    1.173445
       _cons |   .8205362   .2809636     2.92   0.003     .2698577    1.371215
------------------------------------------------------------------------------
```

Components F_1, F_2, F_3 and the constant have a significant impact on the amount of emigration. Econometric modeling of the attractiveness of emigration has revealed that with an increase in the production component of the economy and the labor shortage in the country, as well as a reduction of the development of the internal market, the probability of the country's emigration attractiveness grows.

Conclusions

From 15 variables that influence the number of immigrants in the EU, we identified four latent factors of immigration attractiveness: F_1 describes the technical equipment of the country; F_2 – the low level of skills of the economically active population; F_3 – level of production; F_4 – as a production oriented economy of the country. From 9 variables that influence the number of immigrants in the EU, we identified three latent factors of immigration attractiveness: F_1 – underdevelopment of the domestic market of the country; F_2 – a scientific backwardness of the country; F_3 – the shortage of labor in the country.

Using an iterative approach of cluster analysis, discriminant analysis, and factor analysis we have received the stable classification of countries by the level of immigration and emigration. The countries were divided into two groups according to the immigration attractiveness. The first group included former capitalist countries (Belgium, Denmark, Germany, Ireland, Greece, Spain, France, Italy, Cyprus, Malta, the Netherlands, Austria, Portugal, Finland, Sweden, England, Norway) and the second included ex-socialist countries (Bulgaria, Czech Republic, Estonia, Latvia, Lithuania, Hungary, Poland, Romania, Slovenia, Slovakia, Croatia). This result has great importance, since during the study it was revealed that the EU, which has long sought to achieve economic and social equality, has not been able to overcome the historically formed significant differences in the levels of development. Luxembourg was not identified in any group, which confirms that Luxembourg has the economic status of a free economic zone.

According to the emigration attractiveness, EU countries were divided into three stable groups. The first group included Austria, Belgium, Great Britain, Hungary, Germany, Greece, Italy, Cyprus, Latvia, Netherlands, Norway, Poland, Portugal, Finland, France, Croatia, Sweden, Estonia, the second included Denmark, Ireland, Malta, and the third included Bulgaria, Spain, Romania, Slovakia, Slovenia, Czech Republic. Again Luxembourg was not identified in any group.

Econometric modeling of the immigration attractiveness allowed us to explain that increasing in technical equipment (development of IT technologies) and increasing skills of the economically active population increases the likelihood of a successful immigration situation in the country. Immigration situation does not change with the growth of the welfare of a country and the industrial economy orientation.

Econometric modeling of emigration attractiveness revealed that it is determined by an increase in the production component of the country's economy and labor shortages. With increasing underdevelopment of the domestic market the likelihood of a favorable emigration environment is decreased. The results of this study may be of practical interest for a variety of community and government organizations in making effective decisions in the field of migration policy by influencing the work of selected factors, as well as to predict the level of migration attractiveness in different countries.

References

Albertinelli, A., Knauth, B., Kraszewski, K., & Thorogood, D. (Eds.). (2011). *Migrants in Europe: A statistical portrait of the first and second generation* (2011 Edition). Luxembourg: Publications Office of the European Union. doi:10.2785/5318

Institutul National de Statistica. (n.d.). *Institutul National de Statistica.* Retrieved from http://www.insse.ro/cms/

Migrants in Europe: A statistical portrait of the first and second generation, 2011 edition. (2012, March 19 – April 1). *Demoscope Weekly*, 503-504. Retrieved from http://demoscope.ru/weekly/2012/0503/biblio03.php.

Migration and migrant population statistics. (2015, May). In *Eurostat: Statistics Explained.* Retrieved from http://ec.europa.eu/eurostat/statistics-explained/index.php/Migration_and_migrant_population_statistics

National Statistics Office, Malta. (n.d.). *NSO Online.* Retrieved from http://nso.gov.mt/en/Pages/NSO-Home.aspx

Sartori, F. (2012). EU Member states granted citizenship to more than 800 000 persons in 2010. *Statistics in focus*, 45. http://ec.europa.eu/eurostat/documents/3433488/5585336/KS-SF-12-045-EN.PDF/241b75f4-2dfb-4aca-b0ff-1b6c34a78dc8

Tikhomirov, N. P., Tikhomirova, T. M., Oushmaev O. S. (2011) *Econometric methods and multivariate statistical analysis: A Textbook*. Moscow: Economics.

Journal of Modern Applied Statistical Methods
November 2015, Vol. 14, No. 2, 275-281.

ISSN 1538 − 9472

Statistical Software Applications & Review
Caution for Software Use of New Statistical Methods (R)

Akiva J. Lorenz
Dallas Independent School Dist.
Dallas, TX

Barry S. Markman
Wayne State University
Detroit, MI

Shlomo S. Sawilowsky
Wayne State University
Detroit, MI

Open source programming languages such as R allow statisticians to develop and rapidly disseminate advanced procedures, but sometimes at the expense of a proper vetting process. A new example is the least trimmed squares regression available in R's lqs() in the MASS library. It produces pretty regression lines, particular in the presence of outliers. However, this procedure lacks a defined standard error, and thus it should be avoided.

Keywords: R, lqs(), least trimmed squares regression

Introduction

As new methods appear software vendors race to disseminate them, providing a competitive edge in increasing sales. In the past half century there were numerous examples where this led to the inclusion of procedures that were inappropriate or destructive. For example, consider the mainframe version of SPSS's general linear model command in the 1980's. Option 9, a contrast coding least squares regression approach, due to Overall and Spiegel (1969), was subsequently shown to test no known statistical hypothesis (see, e.g., Blair & Higgins, 1978a; Blair, 1978; Blair & Higgins, 1978b). Another example is the R aov () function "when conducting analysis of covariance" which "does not work correctly" (Schumacker, 2015, p. 288).

Dr. Lorenz is an Evaluation Analyst with the Dallas Independent School District. Email him at akiva@wayne.edu. Dr. Markman is a Professor in the College of Education. Email him at barry.markman@wayne.edu. Dr. Sawilowsky is a Professor in the College of Education. Email him at shlomo@wayne.edu.

One of many modern approaches to regression is the least squared trimmed means, where the sum of squared residuals are replaced with the "sum of the *q* smallest squared residuals, where *q* is roughly *n*/2" (Verzani, 2004, p. 100). Hence, this is essentially an M (maximum likelihood) estimator. It is invoked in R via the lqs() function located in the MASS package.

Rousseuw and Leroy (1987) indicated least trimmed means regression is resistant to outliers (see also Verzani, 2004, p. 100). Ripley (2004) noted that least trimmed *squares* is based on minimizing "the sum of squares for the smallest *q* of the residuals," where q takes on various values (e.g., S+ and R sets *q* to 90% as the default). The result is a regression model that "maximizes accuracy to the *q*% of data. The quantile squared residual... [with] floor((*n* + *p* + 1)/2)" (Ripley, n.d.), where *n* are data points and *p* are the regressors. lqs() is exact with one regressor. (For further details, see Fox, 2002. Note that this method is ill equipped to recover if there are no outliers, when ordinary regression should have been used. Once data are trimmed, they are removed from further calculations whether they should have been eliminated or not.)

Unfortunately, the lqs () function is not associated with a defined standard error. (This is a common problem with maximum likelihood applications. For example, see Holford, (2002, p. 45) regarding a 2×2 table with zero frequencies in a cell). Hence, the purpose of this study is demonstrate this concern with respect to lqs().

Methodology

The number of repetitions per experiment was 100,000, conducted on an Intel Sandy Bridge i7-2600K 3.4GHz CPU-based computer, with ultra-high speed Corsair Vengeance Low Profile 4x4GB RAM, Crucial M4 256GB solid state hard drive, and the Windows 7 Ultimate 64 bit operating system. This equipment was necessary due to the well known lack of speed of the R platform, and even so the results compiled in each table took more than 45 minutes to complete. Data were produced using R rnorm(). To determine the veracity of the coding, the normal theory ordinary least squares method was used for comparison using R's lm().

Standard error of beta and the lqs() method.

The t test is defined as beta divided by the standard error of beta (Brase & Brase, 2013, p. 536; Mann, 1995, p. 667), which is then associated with $df = N - 2$ for the t (or *Z* for large samples) distribution. It is generally not optimal to use the normal theory formula for the standard error (i.e., the standard deviation divided by the sample size), because it is not robust to non-normally distributed. (There are potential alternatives, such as the Winsorized sample standard deviation, or a jackknife or bootstrap approximation. See, e.g., Sawilowsky & Fahoome, 2003, p. 22, 376 - 382. However, there are limitations to those alternatives.)

Wilcox (1996) provided alternatives in computing the standard error for other hypothesis tests (e.g., the sample median), but that was only after a test was presented using the robust estimator in the numerator combined with the normal curve theory standard error in the denominator (see, e.g., p. 120). The same approach could be used here, with the *p*-value associated with beta obtained from lqs() determined via the normal curve theory standard error (i.e., which is produced by the lm() routine).

Results

Using the standard error under lm(y ~ x) (i.e., beta associated with the ordinary least squares regression) as the denominator for the test of beta obtained from lqs() was found to be unsatisfactory, with inflated Type I errors from between 7.3 and 104 times nominal alpha, as noted in Table 1 below.

Table 1. Type I error rates for $n_1 = n_2 = 30$; $r = 0.0$; 100,000 repetitions

α	**Test**	
	lm()	**lqs()**
0.050	0.04972	**0.36455**
0.010	0.01041	**0.21966**
0.001	0.00102	**0.10248**

Note: Values in bold exceed Bradley's (1978) liberal definition of robustness.

An attempt was made to improve the standard error used in lqs() by replacing the original *y* values with the fitted values of *y* obtained from lqs(). The standard error of the estimate (SE_E, or residual standard deviation) was based on

$$SE_E = \sqrt{\sum_i^n \frac{\left(y - y'\right)^2}{n-2}} \tag{1}$$

where *y*' was obtained as fitted values from lqs() instead of the fitted values from lm(). The standard error of beta (SE_b) is determined by

$$SE_b = \frac{SE_E}{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2} \tag{2}$$

Assembling the t test on beta as a ratio of beta divided by (2),

$$t = \frac{b}{SE_b} \tag{3}$$

the obtained t is significant if

$$\left|t_{obt}\right| \geq t_{\frac{\alpha}{2},n-2}$$

Although as noted in Table 2 there was improvement in the Type I error rates, the inflation was nevertheless from between 5.8 and 39.4 times nominal alpha, which is not acceptable. (Note the values for lm() differed slightly from those in Table 1 above due to the change in the seed number).

Table 2. Type I error rates for $n_1 = n_2 = 30$; $r = 0.0$; 100,000 repetitions

α	**Test**	
	lm()	**lqs()**
0.050	0.05029	**0.29371**
0.010	0.01061	**0.14499**
0.001	0.00109	**0.04151**

Note: Values in bold exceed Bradley's (1978) liberal definition of robustness.

Regarding the least median squares (lms) option (i.e., “method = lms” option in lqs (), which can be used to invoke a variety of robust methods), subsequent to a Monte Carlo simulation Paranagama (2010) concluded, "In practice, the use of LMS is limited by the absence of formulas for standard errors” (p. 35). This difficulty applies to the default method (least trimmed squares), and hence, lqs() must be abandoned if the purpose of conducting the linear model is to compute a t test on beta until an adequate standard error for the least squares regression algorithm can be found.

Conclusion

An appropriate standard error has not been derived for the lqs() method. Because the t test on β requires the standard error, various options were considered: (1) the *p*-value associated with β obtained from lqs () was determined via the normal curve theory standard error via the lm() procedure, which failed because it produced Type I errors as large as 104 times nominal α, and (2) the standard error was obtained by replacing the original *y* values with the fitted values of *y* obtained from lqs(), which was an improvement, but also failed because it produced Type I errors as large as 39.4 times nominal α.

The lqs() procedure produces pretty regression equations, and visually fits data in situations with outliers better than the normal theory lm(). However, the absence of a defined standard error precludes its usage in practice. Moreover, the method is not even being close to maintaining nominal alpha. The matter will become increasingly serious as applied researchers continue to be attracted to its highly publicized robustness regression lines, ease of availability in R, and implement it in applied work. For example, lqs() was used by Fan, Lu, Madnick, and Cheung (2001) in a study on data integration in information systems, Abo-Khalil and Abo-Zied (2012) in a study of sensorless control of wind turbines, and Gidnaa and Domínguez-Rodrigo (2013) in a study of human femoral length from fragmented specimens.

In conclusion, new methods should be avoided until such time that they are fully vetted. If this caution was true in the past with expensive, major commercial software such as SPSS, then how much more so caution should be invoked when using free, open source software such as R.

References

Abo-Khalil, A. G., & Abo-Zied, H. (2012). Sensorless control for DFIG wind turbines based on support vector regression, *IECON 2012 - 38th Annual Conference on IEEE Industrial Electronics Society*, 3475 - 3480. doi:10.1109/IECON.2012.6389341

Blair, R. C. (1978). I’ve been testing some statistical hypotheses: Can you guess what they are?, *The Journal of Educational Research*, *72*(2), 116-118.

Blair, R. C., & Higgins, J. J. (1978a). Comments on "Contrast coding in least squares regression analysis," *American Educational Research Journal*, *15*(1), 149-151.

Blair, R. C., & Higgins, J. J. (1978b). Tests of hypotheses for unbalanced factorial designs under various regression/coding method combinations. *Educational and Psychological Measurement*, *38*, 621-631.

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

Brase, C. H., & Brase, C. P. (2013). *Understanding basic statistics*. 6th ed. Boston, MA: Brooks/Cole, Cengage Learning.

Fan, W, Lu, H., Madnick, S. E. & Cheung, D. W.-L. (2001). Discovering and reconciling value conflicts for numerical data integration. *Information Systems*, *26*(8), 635–656. doi:10.1016/S0306-4379(01)00043-6

Fox, J. (2002). Robust Regression. *Appendix to An R and S-Plus Companion to Applied Regression*. *R-Project.org*. Retrieved from http://cran.r-project.org/doc/contrib/Fox-Companion/appendix-robust-regression.pdf.

Gidnaa, A. O., & Domínguez-Rodrigo, M. (2013). A method for reconstructing human femoral length from fragmented shaft specimens. *HOMO - Journal of Comparative Human Biology*, *64*(1), 29–41.

Holford, T. R. (2002). *Multivariate methods in epidemiology*. Oxford: Oxford University Press.

Mann, P. S. (1995). *Introductory statistics*. 2nd ed. NY: Wiley.

Overall, J. E., & Spiegel, D. K. (1969). Concerning least squares analysis of experimental data. *Psychological Bulletin*, *72*(5), 311-322. doi:10.1037/h0028109

Paranagamap, T. D. (2010). *A simulation study of the robustness of the least median of squares estimator of slope in a regression through the origin model* (Unpublished Master Thesis). Manhattan, KS: Kansas State University. Retrieved from http://krex.k-state.edu/dspace/bitstream/handle/2097/7045/ThilankaParanagama2010.pdf

Ripley, B. D. (n.d.). *Resistant Regression*. Retrieved from http://astrostatistics.psu.edu/su07/R/html/MASS/html/lqs.html

Ripley, B. D. (2004). *Robust statistics*. University of Oxford. Retrieved from http://www.stats.ox.ac.uk/pub/StatMeth/Robust.pdf

Rousseeuw, P. J., & Leroy, A. M. (1987). *Robust regression and outlier detection*. NY: Wiley.

Sawilowsky, S. S., & Fahoome, G. (2003). *Statistics through Monte Carlo simulation with Fortran*. Rochester Hills, MI: JMASM.

Schumacker, R. E. (2015). *Learning statistics Using R*. Los Angeles: Sage.

Verzani, J. (2004). *Using R for introductory statistics*. Boca Raton: Chapman & Hall/ CRC.

Wilcox, R. R. (1996). *Statistics for the social sciences*. San Diego, CA: Academic Press.

Journal of Modern Applied Statistical Methods
November 2015, Vol. 14, No. 2, 282-292.

ISSN 1538 − 9472

JMASM Algorithms and Code: Two Group Program for Cohen's d, Hedges' g, η^2, R_{adj}^2, ω^2, ε^2, Confidence Intervals, and Power

David A. Walker
Northern Illinois University
DeKalb, IL

The purpose of this research is to provide an application for users interested in a SPSS syntax program to determine an array of commonly-employed effect sizes and confidence intervals not readily available in SPSS functionality, such as the standardized mean difference and r-related squared indices, for a between-group design.

Keywords: Effect size, confidence intervals, SPSS, syntax

Introduction

The purpose of this research is to provide an application for researchers and practitioners interested in a SPSS syntax program (Walker, 2015) to determine an array of commonly-employed effect sizes and confidence intervals not readily available in SPSS functionality, such as the standardized mean difference and *r*-related squared indices, for a between-group design using descriptive statistics: means, standard deviations, and sample sizes.

As a brief *précis*, in the social sciences, there has been a sustained effort by researchers, editorial boards, and professional organizations for mandatory reporting of effect sizes with statistical significance testing (American Educational Research Association, 2006; American Psychological Association [APA], 2010; Cohen, 1992; Ferguson, 2009; Levine & Hullett, 2002; Thompson, 1998; Wilkinson & The APA Task Force on Statistical Inference, 1999). Cohen (1988, p. 10) noted that an effect size, "…serves as an index of degree of departure from the null hypothesis." When reported with statistically significant results, effect sizes can provide information, for example, pertaining to the extent

Dr. Walker is a Professor of Educational Research, Technology, and Assessment in the College of Education. Email him at dawalker@niu.edu.

of the difference between means or the magnitude of a relationship in terms of the proportion of the total variance accounted for in an outcome (Cohen, 1988). Effect sizes can also be employed to indicate the functional, applied effect of an outcome (Nickerson, 2000).

Ferguson (2009) and Thompson (2009) proposed that effect sizes differentiate generally into the subsequent categories: 1) variance accounted for measures such as squared indices of *r*; 2) corrected estimates, typically employed to reduce estimation bias, such as R_{adj}^2; and 3) standardized mean differences, for example, Cohen's *d*. The current study's program will extrapolate effect sizes from all of these categories.

Cohen (1988) suggested that for *r*-related squared indices, which indicate the proportion of variance in the dependent variable accounted for by the effect of the independent variable, values of .01, .06, and .14 should serve as markers of small, medium, and large effects, respectively. Further, Cohen (1988) defined the values of effect sizes for the standardized difference between means as small = .20, medium = .50, and large = .80. However, it should be appropriately noted that it is at the discretion of the researcher to determine the context in which qualifying labels such as "small," "medium," and "large" effects are being defined when using any effect size index. This caution has been stated by Glass, McGaw, and Smith (1981) with reiteration from Cohen (1988) and Thompson (2009).

Lastly, there has been an emphasis in the literature (APA, 2010; Cohen, 1994; Sapp, 2004; Vacha-Haase & Thompson, 2004; Wilkinson & The APA Task Force on Statistical Inference, 1999) that not only should effect sizes be reported with statistically significant results, but confidence intervals ought to complement said point estimate indices for more comprehensive analysis and interpretation of outcomes. As noted by Levin and Robinson (2003, p. 235), "Reporting and interpreting effect sizes (with corresponding confidence intervals) in multiple experiment studies where the effect of interest is replicated (i.e., its direction is confirmed) may provide readers with more useful information concerning the believability and magnitude of the effect…"

Two group program

The SPSS syntax program will create an internal matrix table to assist users in determining the effects pertaining to the standardized mean difference and/or the proportion of variance in the dependent variable accounted for by the effect of the independent variable for two groups. The preponderance of the ensuing formulas are derived from Aaron, Ferron, and Kromrey (1998), Cohen (1988), Cohen and Cohen (1983), Cooper and Hedges (1994), and Richardson (1996).

The variance accounted for effect size measures include eta squared (η^2: Note equal to R^2 (Beasley & Schumacker, 1995)), which is known to be a positively-biased index, particularly with small sample sizes, and is defined as:

$$\eta^2 = \frac{d^2}{\left(d^2 + 4\right)} \tag{1}$$

where d = Cohen's d value.

Additionally, correction indices for η^2, such as omega squared (ω^2), epsilon squared (ε^2), and $R_{adj}{}^2$, all algebraically and theoretically-related measures (Cohen, 1988), are part of the program and formulated as:

$$\omega^2 = \frac{\left(t^2 - 1\right)}{\left(t^2 + N_1 + N_2 - 1\right)} \tag{2}$$

$$\varepsilon^2 = 1 - \left(1 - \eta^2\right) \times \frac{\left(N_1 + N_2 - 1\right)}{\left(N_1 + N_2 - 2\right)} \tag{3}$$

$$R_{adj}^2 = \eta^2 - \left(\left(1 - \eta^2\right) \times \left(\frac{2}{\left(N_1 + N_2 - 3\right)}\right)\right) \tag{4}$$

where t is the t value derived from the model as

$$d \times SQRT\left[\frac{\left(N_1 \times N_2\right)}{\left(N_1 + N_2\right)}\right] \tag{5}$$

M_1, $SD_{1,}$ N_1 and M_2, $SD_{2,}$ N_2 are the means, standard deviations, and sample sizes for Group 1 and Group 2, respectively.

Finally, Cohen's d is a measure of standardized mean difference and is defined as:

$$d = \frac{(M_1 - M_2)}{SQRT\left[\frac{(SD_1^2 + SD_2^2)}{2}\right]} \tag{6}$$

Note that Kraemer (1983) indicated the formula for *d* is optimal when both sample sizes are relatively equal and also large. Further, Cohen's *d* is recognized as a biased estimate (Hedges, 1981) and; thus, Hedges' g is a correction measure for this concern. It should be mentioned; however, that *d* and g are approximately equivalent when *n* = 30 (Hedges & Olkin, 1985). Hedges' g is defined as:

$$g = d \times \left(1 - \left(\frac{3}{(4 \times (N_1 + N_2) - 9)}\right)\right) \tag{7}$$

For the syntax program, the squared indices' estimated confidence intervals (CI) are set at 90% and based on the work of Cohen, Cohen, West, and Aiken (2003). For these estimated CIs, it is agreed that the sample size should be > 60, which, comparatively, assumes negligible error and; therefore, the absence of an adjustment for noncentrality. The error term for the approximated CI is defined as:

$$\begin{aligned} R^2E = SQRT\left(4 \times \eta^2\right) \times \left(1 - \eta^2\right)^2 \times \frac{(N - 1 - 1)^2}{(N^2 - 1)} \\ \times (N + 3) \ldots R^2 \pm R^2E \times 1.645 \end{aligned} \tag{8}$$

For the standardized mean difference CIs, these are set at 95%. The program's estimated CI formula is based on previous research by Grissom and Kim (2005), Hedges and Olkin (1985), and Steiger (2004). Bird (2002) found that if *d* is < 2.00, which in social science research frequently can be the circumstance with middling-sized effects (Richard, Bond, & Stokes-Zoota, 2003; Rosnow & Rosenthal, 2003), adjustment for noncentrality is not compulsory. The error term for this approximated CI is defined as:

$$d1; g1 = \frac{N}{(N_1 \times N_2)} + \frac{d^2}{(2 \times N)} \ldots d \pm d1 \times 1.96 \tag{9}$$

Note: For any CI within the program, the user can alter it by changing the *Z* value within the syntax, for example, to values such as 1.28 (80% CI), 1.645 (90% CI), 1.96 (95% CI), or 2.58 (99% CI), where Cohen (1990, p. 1310) observed "I don't think that we should routinely use 95% intervals: Our interests are often better served by more tolerant 80% intervals."

Results

As seen in Appendix A, the user would put the two-group descriptive data (M_1, SD_1, N_1 for Group 1 and M_2, SD_2, N_2 for Group 2) in the space between BEGIN DATA and END DATA along with the total sample size (*N*). Thus, these descriptive data in the example from the program are, in group order, 16.45 2.23 30 11.77 4.66 34 64 and represent continuous data for the dependent variable (Depression Score) and categories for the independent variable Group (i.e., Group 1 [Treatment] and Group 2 [Control]).

Once the program is run, the results show that the matrix produced will cluster the effect sizes by the categories noted previously: standardized means difference, squared index, and corrected squared indices. Additionally, the matrix generates an overall model *post-hoc* power value, which is predicated on alpha established at .05 and the particular sample sizes for Group 1 and Group 2.

As can be seen in the results from Table 1, the standardized mean difference effect size for Cohen's *d* was 1.256 or a "large" effect of over one standard deviation difference in Depression Score between Group 1 and Group 2 with 95% CI at (1.109, 1.403) and overall model power = .999, where power ≥ .80 is desired in social science research (Nunnally, 1978). The correction for Cohen's *d*, Hedges' *g*, was very comparable in value at 1.241 (1.094, 1.387).

Table 1. Standardized Mean Difference, Confidence Intervals, and Model Post-Hoc Power.

Cohen's *d*	$95\%_{CIL}$	$95\%_{CIU}$	Hedges' g	$95\%_{CIL}$	$95\%_{CIU}$	Power
1.256	1.109	1.403	1.241	1.094	1.387	0.999

Note. CI = Confidence Interval; L = Lower; U = Upper.

For the squared and corrected squared indices, the results in Table 2 indicated that the proportion of variance in the dependent variable accounted for by the effect of the independent variable was "large" overall for all of the various indices. As would be expected, these effect size measures ranged from a low of

25.9% for the correction R_{adj}^2 (90% CI .089, .430) to a high of 28.3% for the non-corrected η^2 (90% CI .107, .458).

Table 2. Proportion of Variance in the DV Accounted for by the Effect of the IV and Confidence Intervals.

η^2	90%$_{CIL}$	90%$_{CIU}$	R_{adj}^2	90%$_{CIL}$	90%$_{CIU}$	ω^2	90%$_{CIL}$	90%$_{CIU}$	ε^2	90%$_{CIL}$	90%$_{CIU}$
.283	.107	.458	.259	.089	.430	.274	.100	.448	.271	.098	.444

References

Aaron, B., Kromrey, J. D., & Ferron, J. M. (1998, November). *Equating r-based and d-based effect size indices: Problems with a commonly recommended formula*. Paper presented at the annual meeting of the Florida Educational Research Association, Orlando, FL.

American Educational Research Association. (2006). Standards for reporting on empirical social science research in AERA publications. *Educational Researcher, 35*, 33-40.

American Psychological Association. (2010). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: Author.

Beasley, T. M., & Schumacker, R. E. (1995). Multiple regression approach to analyzing contingency tables: Post hoc and planned comparison procedures. *Journal of Experimental Education, 64*, 79-93.

Bird, K. D. (2002). Confidence intervals for effect sizes in analysis of variance. *Educational and Psychological Measurement, 62*, 197–226. doi:10.1177/0013164402062002001

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist, 45*, 1304-1312.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*, 155-159.

Cohen, J. (1994). The earth is round (*p* < .05). *American Psychologist, 49*, 997-1003.

Cohen, J., & Cohen, P. (1983). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah: NJ: Erlbaum.

Cooper, H., & Hedges, L. V. (Eds.). (1994). *The handbook of research synthesis*. New York: Russell Sage Foundation.

Ferguson, C. J. (2009). An effect size primer: A guide for clinicians and researchers, *Professional Psychology: Research and Practice*, *40*(5), 532-538. doi:10.1037/a0015808

Glass, G. V., McGaw, B., & Smith, M. L. (1981). *Meta-analysis in social research*. Beverly Hills, CA: Sage.

Grissom, R. J., & Kim, J. J. (2005). *Effect sizes for research: A broad practical approach*. New York: Psychology Press.

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, *6*, 107-128.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. San Diego: Academic Press.

Kraemer, H. C. (1983). Theory of estimation and testing of effect sizes: Use in meta-analysis. *Journal of Educational Statistics*, *8*, 93-101.

Levin, J. R., & Robinson, D. H. (2003). The trouble with interpreting statistically nonsignificant effect sizes in single-study investigations. *Journal of Modern Applied Statistical Methods*, *2*(1), 231-236. Available at http://digitalcommons.wayne.edu/jmasm/vol2/iss1/23/

Levine, T. R., & Hullett, C. R. (2002). Eta squared, partial eta squared, and misreporting of effect size in communication research. *Human Communication Research, 28*(4), 612-625. doi:10.1111/j.1468-2958.2002.tb00828.x

Nickerson, R. S. (2000). Null hypothesis significance testing: A Review of an old and continuing controversy. *Psychological Methods*, *5*(2), 241-301. doi:10.1037/1082-989X.5.2.241

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.

Richard, F. D., Bond, C. F., Jr., & Stokes-Zoota, J. J. (2003). One hundred years of social psychology quantitatively described. *Review of General Psychology, 7*(4), 331-363. doi:10.1037/1089-2680.7.4.331

Richardson, J. T. E. (1996). Measures of effect size. *Behavior Research Methods, Instruments, & Computers, 28*, 12-22.

Rosnow, R., & Rosenthal, R. (2003). Effect sizes for experimenting psychologists. *Canadian Journal of Experimental Psychology, 57*(3), 221–237. doi:10.1037/h0087427

Sapp, M. (2004). Confidence intervals within hypnosis research. *Sleep and Hypnosis, 6*(4), 169–176.

Steiger, J. H. (2004). Beyond the F test: Effect size confidence intervals and tests of close fit in the analysis of variance and contrast analysis. *Psychological Methods, 9*(2), 164–182. doi:10.1037/1082-989X.9.2.164

Thompson, B. (1998). Statistical significance and effect size reporting: Portrait of a possible future. *Research in the Schools, 5*(2), 33-38.

Thompson, B. (2009). A brief primer on effect sizes. *Journal of Teaching in Physical Education, 28*, 251-254.

Vacha-Haase, T., & Thompson, B. (2004). How to estimate and interpret various effect sizes. *Journal of Counseling Psychology, 51*(4), 473-481. doi:10.1037/0022-0167.51.4.473

Walker, D. A. (2015). Two group program for Cohen's *d*, Hedges' g, η^2, R_{adj}^2, ω^2, ε^2, confidence intervals, and power. [Computer program]. DeKalb, IL: Author.

Wilkinson, L., & The APA Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54*(8), 594-604.

Appendix A: SPSS syntax two group program for Cohen's *d*, Hedges’ g, η^2, R_{adj}^2, ω^2, ε^2, confidence intervals, and power.

```
DATA LIST LIST /M1 SD1 (2F9.3) N1 (F8.0) M2 SD2 (2F9.3) N2 N (2F8.0).
***********************************************************************
Put your two-group data (M1, SD1, N1 for Group 1 and M2, SD2, N2 for Group 2) in
the space between BEGIN DATA and END DATA along with the total sample size (N)
***********************************************************************.
BEGIN DATA
16.45 2.23 30 11.77 4.66 34 64
END DATA.
COMPUTE POOLD = ((N1-1)*(SD1**2)+(N2-1)*(SD2**2))/((N1+N2)-2).
COMPUTE COHEND = ABS((M1-M2)/SQRT(POOLD)).
COMPUTE D1 = N/(N1*N2) + COHEND**2/(2*N).
COMPUTE HEDGESG = COHEND*(1-(3/(4*(N1 + N2)-9))).
COMPUTE G1 = N/(N1*N2) + HEDGESG**2/(2*N).
COMPUTE CRITICAL = 0.05.
COMPUTE K = 1.
COMPUTE H = (2*N1*N2)/(N1+N2).
COMPUTE NCP = ABS((COHEND*SQRT(H))/SQRT(2)).
COMPUTE ALPHA = IDF.T(1-CRITICAL/2,N1+N2-2).
COMPUTE POWER1 = 1-NCDF.T(ALPHA,N1+N2-2,NCP).
COMPUTE POWER2 = 1-NCDF.T(ALPHA,N1+N2-2,-NCP).
COMPUTE B = POWER1 + POWER2.
COMPUTE ETA2 = COHEND**2/(COHEND ** 2 + 4).
COMPUTE EPSILON = 1-(1-ETA2) * (N1  +  N2-1) / (N1  +  N2-2).
COMPUTE TTEST = COHEND  * SQRT((N1  *  N2) /(N1  +  N2)).
COMPUTE OMEGA = (TTEST**2-1)/(TTEST**2 + N1 + N2 -1).
COMPUTE SEETA1 = (1-ETA2)/SQRT(N1 + N2-1).
COMPUTE SEETA2 = 2/(N1 + N2 - 2).
COMPUTE SEETA3 = SQRT(SEETA2 + 4*ETA2).
COMPUTE SEETA = SEETA1 * SEETA3.
COMPUTE TTEST = COHEND  * SQRT((N1  *  N2) /(N1  +  N2)).
COMPUTE ADJR2 = ETA2 - ((1-ETA2)*(2/(N1 + N2 -3))).
COMPUTE ADJR2A = (((4*ADJR2)*(1-ADJR2)*(N-K-1)**2)).
COMPUTE ADJR2B = (N**2-1)*(N+3).
COMPUTE ADJR2C = ADJR2A/ADJR2B.
COMPUTE ADJR21 = SQRT(ADJR2C).
**********************************************************************
NOTE: Confidence Intervals can be altered below by changing the Z = value to
    either 1.96 = (95%) or 2.58 = (99%) For the squared indices, they are at 90%
**********************************************************************.
COMPUTE Z = 1.645.
COMPUTE ADJR2L = (ADJR2-(Z*ADJR21)).
COMPUTE ADJR2H = (ADJR2+(Z*ADJR21)).
COMPUTE OMEGA = (TTEST**2-1)/(TTEST**2 + N1 + N2 -1).
COMPUTE SEE1 = (1-EPSILON)/SQRT(N1  +  N2-1).
COMPUTE SEE2 = 2/(N1 + N2 - 2).
COMPUTE SEE3 = SQRT(SEE2 + 4*EPSILON).
COMPUTE SEEPSILON = SEE1 * SEE3.
COMPUTE SEO1 = (1-OMEGA)/SQRT(N1  +  N2-1).
COMPUTE SEO2 = 2/(N1 + N2 - 2).
COMPUTE SEO3 = SQRT(SEO2 + 4*OMEGA).
COMPUTE SEOMEGA = SEO1 * SEO3.
COMPUTE ETAA = (((4*ETA2)*(1-ETA2)*(N-K-1)**2)).
COMPUTE ETAB = (N**2-1)*(N+3).
COMPUTE ETAC = ETAA/ETAB.
COMPUTE ETA1 = SQRT(ETAC).
COMPUTE ETAL = (ETA2-(Z*ETA1)).
COMPUTE ETAH = (ETA2+(Z*ETA1)).
COMPUTE OMEGAA = (((4*OMEGA)*(1-OMEGA)*(N-K-1)**2)).
COMPUTE OMEGAB = (N**2-1)*(N+3).
COMPUTE OMEGAC = OMEGAA/OMEGAB.
COMPUTE OMEGA1 = SQRT(OMEGAC).
COMPUTE OMEGAL = (OMEGA-(Z*OMEGA1)).
COMPUTE OMEGAH = (OMEGA+(Z*OMEGA1)).
COMPUTE EPSILONA = (((4*EPSILON)*(1-EPSILON)*(N-K-1)**2)).
COMPUTE EPSILONB = (N**2-1)*(N+3).
COMPUTE EPSILONC = EPSILONA/EPSILONB.
COMPUTE EPSILON1 = SQRT(EPSILONC).
COMPUTE EPSILONL = (EPSILON-(Z*EPSILON1)).
COMPUTE EPSILONH = (EPSILON+(Z*EPSILON1)).
**********************************************************************
NOTE: Confidence Intervals for Cohen's d are at 95%
**********************************************************************.
```

```
COMPUTE Z = 1.96.
COMPUTE GH = (HEDGESG+(G1*Z)).
COMPUTE GL = (HEDGESG-(G1*Z)).
COMPUTE DH = (COHEND+(D1*Z)).
COMPUTE DL = (COHEND-(D1*Z)).
EXECUTE.
FORMAT POOLD to DL (F9.3).
VARIABLE  LABELS  COHEND  'Cohens  d'/B  'Power'/ETA2  'Eta  Squared'/OMEGA  'Omega
    Squared'/EPSILONL '90% CI Lower'/
EPSILONH '90% CI Upper'/OMEGAL '90% CI Lower'/ OMEGAH '90% CI Upper'/ETAL '90%
    CI Lower'/ADJR2L '90% CI Lower'/
GL  '95%  CI  Lower'/  GH  '95%  CI  Upper'/HEDGESG  'Hedges  g'/ADJR2H  '90%  CI
    Upper'/ADJR2 'Adjusted R2'/DL '95% CI Lower'/
DH '95% CI Upper'/ETAH '90% CI Upper'/EPSILON 'Epsilon Squared'/.
REPORT FORMAT=LIST AUTOMATIC ALIGN(CENTER)
  /VARIABLES= COHEND DL DH HEDGESG GL GH B
  /TITLE "Standardized Mean Difference, Confidence Intervals, and Model Post-Hoc
    Power".
REPORT FORMAT=LIST AUTOMATIC ALIGN(LEFT)
MARGINS (*,150)
  /VARIABLES=  ETA2  ETAL  ETAH  ADJR2  ADJR2L  ADJR2H  OMEGA  OMEGAL  OMEGAH  EPSILON
    EPSILONL EPSILONH
  /TITLE "Proportion of Variance in the DV Accounted for by the Effect of the IV
    and Confidence Intervals".
```

Journal of Modern Applied Statistical Methods
November 2015, Vol. 14, No. 2, 293-294

ISSN 1538 − 9472

Instructions for Authors

Authors wishing to submit to *JMASM* may do so using the submission form at the journal's website, http://digitalcommons.wayne.edu/jmasm. Three areas are appropriate for *JMASM*:

1. Development or study of 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. Development or 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.

Elegant derivations, as well as articles with no take-home message to practitioners, have low priority. Articles based on Monte Carlo (and other computer-intensive) methods designed to evaluate new or existing techniques or practices, particularly as they relate to novel applications of modern methods to everyday data analysis problems, have high priority.

Work appearing in *Regular Articles*, *Brief Reports*, and *Emerging Scholars* is externally peer reviewed, with input from the Editorial Board; work appearing in *Statistical Software Applications and Review* and *JMASM Algorithms and Code* is internally reviewed by the Editorial Board. *JMASM* charges neither article processing fees nor submission fees.

Please observe the following guidelines when preparing manuscripts:

1. *JMASM* uses a modified American Psychological Association style guideline.
2. Articles should be submitted without a title page or abstract. There should be no material identifying authorship except in the fields of the submission form. Include a statement in the cover letter indicating that proper human subjects protocols were followed where applicable, including informed consent.
3. Manuscripts should be prepared in Microsoft Word (.doc or .docx) 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.
4. The text maximum is 20 pages double spaced, not including tables, figures, graphs, and references. Use 11 point Times Roman font.
5. 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.
6. The submission form requires an Abstract with a 50 word maximum, and 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.

7. 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.

8. 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.

9. 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 the spacebar twice after a period.

Journal of
Modern Applied
Statistical Methods

ISSN: 1538−9472 http://digitalcommons.wayne.edu/jmasm

PUBLISHED biannually (May, November) in partnership by:

JMASM, Inc.
PO Box 48023
Oak Park, MI 48237
ea@jmasm.com

Wayne State University Library System
Purdy Library
Detroit, MI 48202
digitalcommons@wayne.edu

WAYNE STATE
UNIVERSITY
LIBRARY SYSTEM