View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Universiti Putra Malaysia Institutional Repository

Pertanika J. Sci. & Technol. **18 (1)**: 209–221 (2010)
ISSN: 0128-7680
© Universiti Putra Malaysia Press

Bootstrapping the Confidence Intervals of R^2_{MAD} for Samples from Contaminated Standard Logistic Distribution

Fauziah Maarof[1], Lim Fong Peng[1*] and Noor Akma Ibrahim[2]
[1]*Department of Mathematics, Faculty of Science,*
[2]*Institute for Mathematical Research, Universiti Putra Malaysia,*
43400 UPM, Serdang, Selangor, Malaysia
[*]*E-mail: fongpeng2805@yahoo.com*

ABSTRACT

This paper investigates the confidence intervals of R^2_{MAD}, the coefficient of determination based on median absolute deviation in the presence of outliers. Bootstrap *bias-corrected accelerated* (BC_a) confidence intervals, known to have higher degree of correctness, are constructed for the mean and standard deviation of R^2_{MAD} for samples generated from contaminated standard logistic distribution. The results indicate that by increasing the sample size and percentage of contaminants in the samples, and perturbing the location and scale of the distribution affect the lengths of the confidence intervals. The results obtained can also be used to verify the bound of R^2_{MAD}.

Keywords: Bootstrap, R^2_{MAD}, confidence interval, contaminated standard logistic distribution

INTRODUCTION

In statistics, the median absolute deviation (MAD) is a resistant measure of the variability of a univariate sample. It can be used to estimate the scale parameter of a distribution, for which variance and standard deviation do not exist, such as the Cauchy distribution. Even when working with distributions for which the variance exists, the MAD has advantages over the standard deviation. For instance, the MAD is more resilient to outliers in a data set. For standard deviation, the distances from the mean are squared. Then, on the average, large deviations are weighted more heavily. In the MAD, the magnitude of the distances of a small number of outliers is irrelevant. Recently, problem related with robust statistic is often discussed and investigated by many researchers. Among other, Crouxa *et al.* (2002) discussed the breakdown behaviour of the maximum likelihood (ML) estimator in the logistic regression model, while Bondell (2008) constructed a new family of procedures to estimate the parameters in the general semi-parametric biased sampling model and then compared it to the existing robust logistic regression procedures via a simulation study and real data example. For the R^2_{MAD}, a review of the statistical literature indicates that the coefficient of determination based on the median absolute deviation is seldom discussed, especially in relation to its application. Many researchers are more interested in investigating the sampling properties, as well as the confidence intervals of the coefficient of determination (R^2) and the adjusted coefficient of determination ($\overline{R}^2$). According to Ohtani (2000), the sampling properties of the R^2 and $\overline{R}^2$ have been examined by Koerts and Abrahamse (1970), Rencher and Pun (1980), Cramer (1987), Carrodus and Giles (1992), Meepagala (1992), Ohtani and Hasegawa (1993), Ohtani (1994), and Srivastava and Ullah (1995), whereas their confidence intervals were examined by Helland (1987) who proposed a simple approximate confidence interval. Ohtani (2000) considered estimating the standard errors of R^2 and $\overline{R}^2$, and constructing their confidence intervals using bootstrap method.

Received: 26 December 2008
Accepted: 26 June 2009
[*]Corresponding Author

In general, bootstrapping is a statistical method used to estimate the sampling distribution of an estimator by sampling with replacement from the original sample, most often with the purpose of deriving robust estimates of standard errors and confidence intervals of a population parameter. It has become an important tool in statistical inference since the pioneering article by Efron (1979). In 2004, Hossain and Khan discussed the non-parametric bootstrapping procedure for the multiple logistic regression model associated with Davidson and Hinkley's (1997) "boot" library in R statistical software. They estimated the sampling distribution of a statistic empirically without making any assumptions about the form of the population, and without deriving the sampling distribution explicitly using the non-parametric bootstrapping. As stated in Press and Zellner (1978), R^2 is rarely accompanied by a measure of precision. It is difficult to calculate since the distribution of R^2 is rather complex and dependent on unknown parameters. Then, Ohtani (2000) stated that the bootstrap method proposed by Efron (1979) is often useful and easier to execute than the Bayesian method when the calculation of precision and the construction of a confidence interval are difficult. Thus, the bootstrap method is considered in this study as a tool for the construction of the confidence intervals of R^2_{MAD}.

In this paper, the confidence intervals of the mean and standard deviation of the sampling distribution of R^2_{MAD}, based on the samples from the contaminated standard logistic distributions, are constructed using the bootstrap method.

In the next section, a short background of the standard logistic distribution and the R^2_{MAD} is given, and this is followed by a simulation study (Section 2), the bootstrap procedure (Section 3) as well as discussions of the results (Section 4) and the conclusion (Section 5).

R^2_{MAD} FROM CONTAMINATED STANDARD LOGISTIC DISTRIBUTION

The Standard Logistic Distribution

The logistic distribution has been used for various growth models and for a certain type of regression which is appropriately known as logistic regression.

In the probability theory and statistics, the logistic distribution is a continuous probability distribution with no shape parameter. Hence, the probability density function (pdf) of the logistic distribution has only one shape (i.e. the bell shape) which resembles the normal distribution but with heavier tails.

In general, let x be a random variable, which is generated from logistic distribution, then the probability density function (pdf) of the logistic distribution is given by:

$$f(x) = \frac{e^{w}}{\sigma(1+e^{w})^{2}};\ -\infty < x < \infty,\ -\infty < \mu < \infty, \sigma < 0$$

where, $w = \dfrac{x-\mu}{\sigma}$, μ = location parameter and σ = scale parameter, and its cumulative distribution function (cdf) is:

$$F(x) = \frac{e^{w}}{1+e^{w}}.$$

Therefore, for the standard logistic distribution $(\mu = 0, \sigma = 1)$, the pdf is

$$f(x) = \frac{e^{x}}{(1 + e^{x})^{2}}$$

and its cdf is:

$$F(x) = \frac{e^{x}}{1 + e^{x}},$$

for $-\infty < x < \infty$.

The Median Absolute Deviation Correlation Coefficient, r_{MAD}

This section presents a correlation analogue of the MAD, r_{MAD}. Let (X, H) be a bivariate distribution with parameters, μ_x, μ_h, σ_h, ρ.

Define

$$x_i' = x_i - med(x) \quad h_i' = h_i - med(h)$$

and $MAD_x = med|\, x_i - med(x_i)\,|$ and $MAD_h = med|\, h_i - med(h_i)\,|$;

$i = 1, 2, \ldots, n$, where *med(.)* is the median function.

According to Gideon (2007), the definition of r_{MAD} is as follows:

$$r_{MAD}(x,h) = \frac{1}{2}\left[med\left(\left| \frac{x_i'}{MAD_x} + \frac{h_i'}{MAD_h} \right| \right) - med\left(\left| \frac{x_i'}{MAD_x} - \frac{h_i'}{MAD_h} \right| \right) \right] \tag{1}$$

Then, $R^2{}_{MAD} = (r_{MAD})^2$. Gideon verified that it is evidently not true that $|\, r_{MAD}\,| \leq 1$. Thus, r_{MAD} may sometimes go beyond + 1 or – 1.

Generation of R^2_{MAD} for the Standard Logistic Distribution

Let $X_{(i)}$ be the i^{th} order statistics and T_i is a set of constants ($i = 1, \ldots, n$). A regression test is a graphical method which is related to probability plots, in which the order statistics $X_{(i)}$ (on the vertical axis) are plotted against T_i (on the horizontal axis). After that, a straight line is fitted to the points, and R^2 will indicate the association between $X_{(i)}$ and T_i.

Suppose $F_0(x) \equiv F(w)$ with $w = (x - \mu)/\sigma$, where μ is the location parameter and σ is the scale parameter. When a sample of values W_i are taken from *F(w),* with $\mu = 0$ and $\sigma = 1$, the sample X_i is then constructed from $F_0(x)$ by calculating:

$$X_{(i)} = \mu + \sigma W_{(i)}, \quad i = 1, \ldots, n.$$

Now, let

$$m_i = E(W_{(i)}).$$

Then

$$E(X_{(i)}) = \mu + \sigma m_i \tag{2}$$

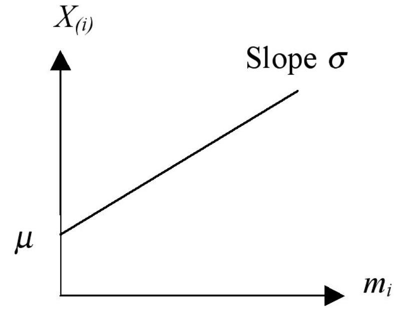

Fig. 1

and a plot of $X_{(i)}$ against m_i is approximately a straight line with an intercept μ on the vertical axis and slope σ, as shown in *Fig. 1*.

Nevertheless, for logistic distribution, m_i is difficult to calculate. Therefore, the expression (2) can be replaced by:

$$X_{(i)} = \mu + \sigma T_i + \varepsilon_i \tag{3}$$

where ε_i is the error term. When $T = m_i$, ε_i will have mean zero.

However, an approximate value of m_i is used for the samples from the standard logistic distribution,

$$H_i = \log\{i/(n+1)\}, \tag{4}$$

which is given by D'Agostino and Stephens (1986).

Then, each pair (X_i, H_i) is substituted into formula (1) for the standard logistic distribution in the previous section.

The study proceeded with the simulation of data to obtain the sampling distribution of R^2_{MAD} from the contaminated standard logistic distribution. The sampling distribution of R^2_{MAD} is simulated for the logistic contaminants (logistic (2, 0.2), logistic (0, 0.2) and logistic (2, 1)) and normal contaminants (normal (3, 0.2)). The contaminants percentages used are 5%, 15%, and 25% with n equals to 20, 40, and 100. For n = 20 with 5% logistic (2, 0.2) contaminants, the contaminated sample contained 95% of the sample from the standard logistic distribution and 5% sample from the logistic (2, 0.2) distribution. The same procedure was used for the other contaminations.

In this study, the contamination in the generation of samples is important as it can be used to observe the effect of the sample size, percentage of contamination and distribution of contaminant on the confidence intervals for the mean and standard deviation of R^2_{MAD}. The logistic (2, 0.2), logistic (0, 0.2), and logistic (2, 1) contaminants were therefore chosen to determine the behaviour of the confidence intervals for the mean and standard deviation of R^2_{MAD}, when the values of location and scale of logistic distribution increased, whereas any values could be chosen as the location and scale of the normal contaminants, provided they were not far from the parameter values of logistic contaminants.

For the cases without outliers, the following steps are used in calculating R^2_{MAD} for the standard logistic distribution.

Step 1. Using S-plus, a random sample of order statistics $\{X_{(1)}, X_{(2)}, ..., X_{(n)}\}$ is generated from the standard logistic distribution with $\mu = 0$ and $\sigma = 1$ for the sample of size *n*.

Step 2. Calculate the approximate value of m_i, i.e., H_i, using formula (4).

Step 3. Regress X_i 's on H_i 's and calculate r_{MAD} using (1). This is followed by the calculation of R^2_{MAD}.

Step 4. Steps 1 – 3 are repeated 1000 times to generate (1000) R^2_{MAD} for the standard logistic distribution.

Step 5. Steps 1 – 4 are repeated for the different sample sizes (n = 20, 40, 100).

However, for cases with outliers, the samples of order statistics generated for use in the calculation of R^2_{MAD} are obtained from the contaminated standard logistic distribution for logistic (2, 0.2), logistic (0, 0.2), logistic (2, 1), and normal (3, 0.2). Similar procedures in the simulation for the cases without outlier (Steps 2 – 5) are repeated. Then, (1000) R^2_{MAD} for the standard logistic distribution with 5%, 15% and 25% contaminants are obtained for n = 20, 40, 100, respectively.

BOOTSTRAP CONFIDENCE INTERVALS FOR THE MEAN AND STANDARD DEVIATION OF R^2_{MAD}

Bootstrap Procedure

In this study, the *bias-corrected accelerated* (BC_a) confidence intervals were computed by simulation with the sample sizes of 20, 40, and 100, since it possessed a higher degree of correctness as compared to other types of bootstrap confidence intervals. According to Efron (1987), the BC_a bootstrap confidence interval is known to lead to second-order accuracy in correctness, while the *percentile* (*P*), the *bias-corrected* (*BC*), and the asymptotic normal confidence intervals have the first-order accuracy only.

The implementation of the bootstrap procedure can be summarized as follows: Let $\hat{\theta}^{*(\alpha)}$ indicates the 100•α th percentile of B bootstrap replications $\hat{\theta}^{*}(1), \hat{\theta}^{*}(2), ..., \hat{\theta}^{*}(B)$. The percentile interval $(\hat{\theta}_{lo}, \hat{\theta}_{up})$ of the intended coverage 1 - 2α is directly obtained from these percentiles:

$$\left(\hat{\theta}_{lo}, \hat{\theta}_{up}\right) = \left(\hat{\theta}^{*(\alpha)}, \hat{\theta}^{*(1-\alpha)}\right)$$

In this study, the step above was repeated 1000 times (B = 1000), and then the averages and standard errors of the mean and standard deviation of the R^2_{MAD} were obtained. Furthermore, 95% confidence intervals for the mean and standard deviation of the R^2_{MAD} could be calculated.

The R^2_{MAD}, which was used in the bootstrap procedure above, was generated using the method presented in the previous section. The bootstrap estimates of the mean and standard deviation of R^2_{MAD} and their confidence intervals for the various cases considered in this investigation were obtained using the S-plus software.

RESULTS AND DISCUSSION

This section presents a discussion on the results of the bootstrap estimates for the mean and standard deviation of the sampling distribution of R^2_{MAD} simulated for various sample sizes, percentage of contamination, as well as distribution of the contaminants, location and scale of the distribution, from which the samples were drawn. Tables 1, 2, 3, 4, and 5 show 95% confidence intervals of R^2_{MAD} for the cases with and without outliers. Tables 6 and 7 show the range of confidence intervals for the mean and standard deviation of R^2_{MAD}, respectively.

As expected for the cases without outliers, Table 1 shows that the range of the confidence intervals for the mean and standard deviation of R^2_{MAD} has become smaller when the sample size increases. For the cases with outliers, on the other hand, Tables 2, 3, 4 and 5 show that there are some similar results just like for the cases without outlier. The range of the confidence intervals for the mean and standard deviation of R^2_{MAD} become smaller when the sample size increases. The same pattern prevails for each type of the contaminants (see Tables 6 and 7). Meanwhile, for each sample size, the results showed that increasing the percentage of contaminants would widen the confidence intervals for the mean and standard deviation of R^2_{MAD} for each type of contaminants, except for the confidence intervals for the standard deviation of R^2_{MAD} with logistic (2, 1) contaminants (see Tables 6 and 7).

In the cases with normal contaminants, the results show that the confidence intervals for the mean of R^2_{MAD} are wider as compared to the cases with logistic contaminants. The trend can be seen more clearly when it is compared to logistic (2, 1) contaminants, as indicated in Tables 6. For the standard deviation, however, the range of the confidence intervals for the normal contaminants does not show any specific trend as compared to the cases with logistic contaminants, as given in Table 7.

For the cases with logistic contaminants, on the other hand, the range of the confidence intervals for the mean and standard deviation of R^2_{MAD} increases with an increase in the location of the logistic distribution as shown in Tables 6 and 7. Meanwhile, increasing the scale of the logistic distribution was found to widen the confidence intervals for the mean and standard deviation of R^2_{MAD} for the samples with 5% contaminations. For the samples with 15% and 25% contamination, however, the range of the confidence intervals for the mean and standard deviation of R^2_{MAD} was found to decrease with the increase in the scale of the logistic distribution, as indicated in Tables 6 and 7. In all the constructed BCa 's, R^2_{MAD} is bounded between -1 and 1.

TABLE 1

95% Confidence intervals of R^2_{MAD} for cases without outliers

N		Observed	BCa (95%)
20	mean	0.8099	(0.8044, 0.8157)
	sd	0.09474	(0.08967, 0.09993)
40	mean	0.8468	(0.8416, 0.8514)
	sd	0.07802	(0.07352, 0.08291)
100	mean	0.8966	(0.8934, 0.8998)
	sd	0.04993	(0.04669, 0.05359)

TABLE 2
95% Confidence intervals of R^2_{MAD} for cases with logistic (2, 0.2) contaminants

n		Contaminant	Observed	BCa (95%)
20	mean	5%	0.7999	(0.7935, 0.806)
		15%	0.7973	(0.79, 0.8034)
		25%	0.7798	(0.7718, 0.7871)
	sd	5%	0.1001	(0.09538, 0.1064)
		15%	0.1042	(0.09896, 0.1105)
		25%	0.1253	(0.1184, 0.1324)
40	mean	5%	0.8486	(0.8437, 0.8532)
		15%	0.8375	(0.8323, 0.8428)
		25%	0.8043	(0.797, 0.8107)
	sd	5%	0.08016	(0.07583, 0.08518)
		15%	0.08438	(0.07979, 0.0894)
		25%	0.1094	(0.104, 0.1164)
100	mean	5%	0.8976	(0.8946, 0.9006)
		15%	0.8809	(0.8773, 0.8844)
		25%	0.8426	(0.8376, 0.8476)
	sd	5%	0.0494	(0.04668, 0.05245)
		15%	0.0581	(0.0548, 0.06213)
		25%	0.07979	(0.07524, 0.0846)

TABLE 3
95% Confidence intervals of R^2_{MAD} for cases with logistic (0, 0.2) contaminants

n		Contaminant	Observed	BCa (95%)
20	mean	5%	0.8005	(0.7945, 0.8063)
		15%	0.7921	(0.7845, 0.7973)
		25%	0.7809	(0.7738, 0.7874)
	sd	5%	0.09811	(0.09291, 0.1032)
		15%	0.1053	(0.09955, 0.1107)
		25%	0.1069	(0.1014, 0.1127)
40	mean	5%	0.8501	(0.8453, 0.8546)
		15%	0.8346	(0.829, 0.8393)
		25%	0.8234	(0.8182, 0.829)

Table 3 continued

	sd	5%	0.073	(0.06845, 0.078)
		15%	0.08091	(0.07599, 0.085)
		25%	0.0864	(0.08186, 0.09121)
100	mean	5%	0.8972	(0.8942, 0.9005)
		15%	0.8806	(0.8773, 0.8836)
		25%	0.8593	(0.8557, 0.863)
	sd	5%	0.04961	(0.04706, 0.05267)
		15%	0.05056	(0.04789, 0.0536)
		25%	0.05913	(0.05602, 0.06256)

TABLE 4
95% Confidence intervals of R^2_{MAD} for cases with logistic (2, 1) contaminants

n		Contaminant	Observed	BCa (95%)
20	mean	5%	0.7983	(0.7914, 0.8042)
		15%	0.8033	(0.7972, 0.8093)
		25%	0.8017	(0.796, 0.8079)
	sd	5%	0.1014	(0.09669, 0.108)
		15%	0.09906	(0.0934, 0.1048)
		25%	0.1001	(0.09489, 0.1051)
40	mean	5%	0.8464	(0.8414, 0.8513)
		15%	0.8446	(0.839, 0.849)
		25%	0.8468	(0.8412, 0.8514)
	sd	5%	0.07908	(0.07467, 0.08404)
		15%	0.07882	(0.07462, 0.08395)
		25%	0.07769	(0.07359, 0.082)
100	mean	5%	0.8975	(0.8944, 0.9005)
		15%	0.8963	(0.893, 0.8991)
		25%	0.8926	(0.8893, 0.8958)
	sd	5%	0.05049	(0.0473, 0.05402)
		15%	0.05162	(0.04829, 0.05629)
		25%	0.05237	(0.04934, 0.05593)

TABLE 5
95% Confidence intervals of R^2_{MAD} for cases with normal (3, 0.2) contaminants

n		Contaminant	Observed	BCa (95%)
20	mean	5%	0.7982	(0.7915, 0.8045)
		15%	0.8055	(0.7989, 0.8117)
		25%	0.7941	(0.7869, 0.8026)
	sd	5%	0.1011	(0.09564, 0.1071)
		15%	0.107	(0.1008, 0.1137)
		25%	0.1216	(0.1147, 0.1289)
40	mean	5%	0.8473	(0.8426, 0.8523)
		15%	0.8391	(0.8338, 0.8444)
		25%	0.8219	(0.8156, 0.8274)
	sd	5%	0.07848	(0.074, 0.08299)
		15%	0.08671	(0.08248, 0.09124)
		25%	0.099	(0.09409, 0.1051)
100	mean	5%	0.8955	(0.8925, 0.8985)
		15%	0.8805	(0.8765, 0.8843)
		25%	0.8471	(0.8425, 0.851)
	sd	5%	0.05051	(0.0479, 0.05347)
		15%	0.06278	(0.05962, 0.06738)
		25%	0.07173	(0.06882, 0.07524)

TABLE 6
The range of the confidence intervals for the mean of R^2_{MAD}

Contaminant	Contamination	n	Range
Logistic (2, 0.2)	5%	20	0.0125
		40	0.0095
		100	0.006
	15%	20	0.0134
		40	0.0105
		100	0.0071
	25%	20	0.0153
		40	0.0137
		100	0.01

Table 6 continued

Logistic (0, 0.2)	5%	20	0.0118
		40	0.0093
		100	0.0063
	15%	20	0.0128
		40	0.0103
		100	0.0063
	25%	20	0.0136
		40	0.0108
		100	0.0073
Logistic (2, 1)	5%	20	0.0128
		40	0.0099
		100	0.0061
	15%	20	0.0121
		40	0.01
		100	0.0061
	25%	20	0.0119
		40	0.0102
		100	0.0065
Normal (3, 0.2)	5%	20	0.013
		40	0.0097
		100	0.006
	15%	20	0.0128
		40	0.0106
		100	0.0078
	25%	20	0.0157
		40	0.0118
		100	0.0085

TABLE 7
The range of the confidence intervals for the standard deviation of R^2_{MAD}

Contaminant	Contamination	*n*	Range
Logistic (2, 0.2)	5%	20	0.01102
		40	0.00935
		100	0.00577
	15%	20	0.01154
		40	0.00961
		100	0.00733
	25%	20	0.014
		40	0.0124
		100	0.00936
Logistic (0, 0.2)	5%	20	0.01029
		40	0.00955
		100	0.00561
	15%	20	0.01115
		40	0.00901
		100	0.00571
	25%	20	0.0113
		40	0.00935
		100	0.00654
Logistic (2, 1)	5%	20	0.01131
		40	0.00937
		100	0.00672
	15%	20	0.0114
		40	0.00933
		100	0.008
	25%	20	0.01021
		40	0.00841
		100	0.00659
Normal (3, 0.2)	5%	20	0.01146
		40	0.00899
		100	0.00557
	15%	20	0.0129
		40	0.00876
		100	0.00776
	25%	20	0.0142
		40	0.01101
		100	0.00642

CONCLUSIONS

As expected, the range of the confidence intervals for the mean and standard deviation of R^2_{MAD} becomes smaller when the sample size increases for both the cases with and without outliers, based on the results presented in the previous section. The same trend can be seen for each distribution of the contaminants. Meanwhile, increasing the percentage of contaminants widens the confidence intervals for each sample size, except for the confidence intervals for the standard deviation of R^2_{MAD} with logistic (2, 1) contaminants. Increasing the location and scale of the distribution of logistic contaminants has also been found to affect the range of confidence intervals for the mean and standard deviation of R^2_{MAD}. The results obtained are used to verify the bound of R^2_{MAD}, i.e. whether or not $\left|r_{MAD}\right|$ is always less than 1.

ACKNOWLEDGEMENT

The authors are grateful to Prof. Rahmatullah Imon for useful comments. A special thank you is due to the Graduate Research Fellowship, supported by Universiti Putra Malaysia, Malaysia.

REFERENCES

Ahmed Hossain and Abdullah Khan H. T. (2004). Nonparametric bootstrapping for multiple logistic regression model using r. *BRAC University Journal*, *I* (2), 109 – 113.

Carrodus, M.L. and Giles, D.E.A. (1992). The exact distribution of R^2 when regression disturbances are autocorrelated. *Economics Letters, 38*, 375-380.

Crouxa, C., Flandre, C. and Haesbroeck, G (2002). The breakdown behavior of the maximum likelihood estimator in the logistic regression model. *Statistics & Probability Letters, 60*, 377–386.

Cramer, J.S. (1987). Mean and variance of R^2 in small and moderate samples. *Journal of Econometrics, 35*, 253-266.

D'Agostino, R. B. and Stephens, M. A. (1986). *Goodness-of-fit Techniques*. Marcel Dekker, Inc.

Davidson, A. C. and Hinkley, D. V. (1997). *Bootstrap Methods and Their Application*. Cambridge: Cambridge University Press.

Efron, B. (1979). Bootstrap methods: Another look to the jackknife. *The Annals of Statistics, 7*, 1-26.

Efron, B. (1987). Better bootstrap confidence intervals. *Journal of the American Statistical Association, 82,* 171 – 200.

Gideon, R A. (2007). The correlation coefficients. *Journal of Modern Applied Statistical Methods*, *6*(2), 517 – 529.

Hall, P. (1988). Theoretical comparisons of bootstrap confidence intervals (with discussion). *The Annals of Statistics, 16*, 927-985.

Helland, I.S. (1987). On the interpretation and use of r^2 in regression analysis. *Biometrics, 43*, 61-69.

Howard, D.B. (2008). A characteristic function approach to the biased sampling model, with application to robust logistic regression. *Journal of Statistical Planning and Inference*, *138*, 742 – 755.

Koerts, J. and Abrahamse, A.P.J. (1970). The correlation coefficient in the general linear model. *European Economics Review, 1,* 401-427.

Meepagala, G. (1992). The small sample properties of r^2 in a misspecified regression model with stochastic regressors. *Economics Letters, 40*, 1-6.

Ohtani, K. and Hasegawa, H. (1993). On small sample properties of r^2 in a linear regression model with multivariate t errors and proxy variables. *Econometric Theory, 9*, 504-515.

Ohtani, K. (1994). The density function of r^2 and $\overline{R}^2$, and their risk performance under asymmetric loss in misspecified linear regression models. *Economic Modelling, 11,* 463-471.

Ohtani, K. (2000). Bootstrapping R^2 and adjusted R^2 in regression analysis. *Economic Modelling*, *17*, 473-483.

Press, S.J. and Zellner, A. (1978). Posterior distribution for the multiple correlation coefficient with fixed regressors. *Journal of Econometrics, 8*, 307-321.

Rencher, A.C. and Pun, F.C. (1980). Inflation of R^2 in best subset regression. *Technometrics, 22*, 49-53.

Srivastava, A.K. and Ullah, A. (1995). The coefficient of determination and its adjusted version in linear regression models. *Econometric Review, 14*, 229-240.