

Confidence Intervals based on Rank Statistics in Linear Models

K.H.Ng*, M.H.Lim and A.H.Pooi

Institute of Mathematical Sciences, Faculty of Science, University of Malaya, 50603 Kuala Lumpur, Malaysia.

*kokhaur@um.edu.my (Corresponding author)

Received on 13th December 2008, accepted in revised form 23rd October 2009.

ABSTRACT Consider the linear models of which the distributions of the errors are non-normal. We propose a method based on rank statistics for constructing confidence intervals for the parameters in the linear models. It is found that the proposed confidence intervals have coverage probabilities which are fairly close to the target value. Furthermore when the skewness of the distributions is large, the expected lengths of the proposed confidence intervals are found to be much shorter than those of the percentile bootstrap confidence intervals, and the classical confidence intervals which are derived by assuming that the errors are normally distributed.

ABSTRAK Pertimbangkan model linear dengan ralat yang bertaburan tak normal. Kita mencadangkan satu kaedah berasaskan statistik pangkat untuk membina selang keyakinan bagi parameter dalam model linear. Didapati kaedah yang dicadangkan mempunyai kebarangkalian liputan yang menghampiri nilai sasaran. Tambahan pula, apabila nilai kepincangan taburan adalah besar, kaedah yang dicadangkan menghasilkan selang keyakinan yang mempunyai panjang jangkaan yang jauh lebih pendek daripada selang keyakinan yang berdasarkan kaedah “bootstrap” persentil, dan kaedah klasikal yang mengandaikan bahawa taburan ralatnya adalah normal.

(Keywords: Bootstrap; Confidence intervals; Linear Regression; Quadratic-normal distribution.

INTRODUCTION

Consider the linear regression model which can be represented in the form

$$\mathbf{y} = \beta_0 \mathbf{x}_0 + \beta_1 \mathbf{x}_1 + \dots + \beta_{p-1} \mathbf{x}_{(p-1)} + \beta_p \mathbf{x}_p + \boldsymbol{\varepsilon} \quad (1)$$

where $\mathbf{y} = (y_1 \ y_2 \ \dots \ y_n)^T$ is the vector of observations, $\mathbf{X} = [\mathbf{x}_0 \ \mathbf{x}_1 \ \dots \ \mathbf{x}_p]$ is the matrix of explanatory variables, $\boldsymbol{\beta} = (\beta_0 \ \beta_1 \ \dots \ \beta_p)^T$ is the parameter vector, $\boldsymbol{\varepsilon} = (\varepsilon_1 \ \varepsilon_2 \ \dots \ \varepsilon_n)^T$ is the vector of random errors, and $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$ are independent and having the common cumulative distribution function(c.d.f) $G(\cdot)$.

When ε_i has a normal distribution with mean 0 and variance σ^2 , the usual $100(1-\alpha)\%$ classical confidence interval for the individual parameter β_i in Equation (1) is given by

$$\{\beta_i : \hat{\beta}_i - t_{\alpha/2, n-(p+1)} S_{\hat{\beta}_i} \leq \beta_i \leq \hat{\beta}_i + t_{\alpha/2, n-(p+1)} S_{\hat{\beta}_i}\} \quad (2)$$

where $\hat{\beta}_i$ is the least squares estimate of β_i , $t_{\alpha/2, n-(p+1)}$ is the $(1-\alpha/2)100\%$ point of the t -distribution with $(n-(p+1))$ degrees of freedom (df), $S_{\hat{\beta}_i} = \{a^{i+1, i+1}\}^{1/2} \hat{\sigma}$ is the standard error of $\hat{\beta}_i$,

$a^{i+1, i+1}$ is the $(i+1, i+1)$ entry of $(\mathbf{X}^T \mathbf{X})^{-1}$, and $\hat{\sigma}^2 = \mathbf{y}^T [\mathbf{I} - \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T] \mathbf{y} / (n - (p+1))$ is the residual mean square.

Apart from the classical confidence interval, there are many other confidence intervals which have been proposed in the literature for the individual parameter in the linear model with non-normal errors. A well-known method based on transformation of the response variables is given [4]. An alternative way of constructing confidence interval is by means of bootstrap [2], [3], [6], [7] and [12]. In performing the bootstrap for finding confidence interval, usually a large number N of estimates of the parameter β_i are calculated based on the N samples obtained through resampling. The estimates for β_i required in the bootstrap method may either be the ordinary least squares estimates or other types of estimates like the linear plus quadratic (LPQ) estimates [9] and [10] and the estimates based on EM algorithm [5].

Another alternative way of constructing confidence interval is by collecting the values $\beta_i^{(0)}$ of the parameter β_i for which the null hypothesis $H_0 : \beta_i = \beta_i^{(0)}$ is accepted. A way to test the null hypothesis that a particular parameter in $\boldsymbol{\beta}$ is equal to

zero is by using the statistics based on the ranks of residuals [1].

Presently, we first propose a rank test statistic for testing the null hypothesis $H_0 : \beta_i = \beta_i^{(0)}$ and construct the confidence interval given by $\{\beta_i : H_0 \text{ that } \beta_i = \beta_i^{(0)} \text{ is accepted}\}$.

In the case of normal errors, it is found that the confidence interval based on rank statistic is comparable to the classical confidence interval and percentile bootstrap confidence interval in terms of both coverage probability and expected length. When the errors have a skewed distribution, the coverage probabilities for the above three types of confidence intervals are all fairly close to the target value, but the expected length for the confidence interval based on rank statistic is much shorter than those of the classical confidence interval and percentile bootstrap confidence interval.

CONFIDENCE INTERVAL BASED ON RANKS WHEN ERRORS ARE NON-NORMALLY DISTRIBUTED

In this section, we introduce a method based on ranks for finding confidence interval for the parameter β_j when the unknown cumulative distribution function (c.d.f) $G(\cdot)$ of the random errors is continuous. This method will make use of the quadratic-normal distribution. An introduction to the quadratic-normal distribution is as follows:

Suppose the random variable ε_i can be expressed as

$$\varepsilon_i = \begin{cases} \lambda_1 e_i + \lambda_2 (e_i^2 - \frac{(1+\lambda_3)}{2}) & , e_i \geq 0 \\ \lambda_1 e_i + \lambda_2 (\lambda_3 e_i^2 - \frac{(1+\lambda_3)}{2}) & , e_i < 0 \end{cases} , i = 1, 2, \dots, n. (3)$$

In the Equation (3), $e_i \sim N(0,1)$ and $\lambda = (\lambda_1, \lambda_2, \lambda_3)^T$ is a vector of constants [13]. The random variable ε_i is said to have a quadratic-normal distribution with parameters 0 and λ ($\varepsilon_i \sim QN(0, \lambda)$). The following are some basic results [13] regarding the quadratic-normal distribution.

The constants λ_i should be such that ε_i given by Equation (3) is a one-to-one function of e_i . If the function ε_i has one or more turning points, then the value of λ should be such that these points should

occur outside the range $|e_i| < Z_q$ ($q > 0$ is a small value and Z_q is the $(1-q)100\%$ point of the standard normal distribution). By examining the extreme values of ε_i , it can be shown that there are no turning points when $|e_i| < Z_q$ provided that $-\lambda_1 / (2\lambda_2\lambda_3) < -Z_q$ when $-\lambda_1 / (2\lambda_2\lambda_3) < 0$ and $-\lambda_1 / (2\lambda_2) > Z_q$ when $-\lambda_1 / (2\lambda_2) \geq 0$.

Let $m_k = E(\varepsilon_i^k)$, $1 \leq k \leq 4$. Next let $\bar{m}_3 = m_3 / \{m_2\}^{3/2}$ and $\bar{m}_4 = m_4 / \{m_2\}^2$ be respectively the measures of skewness and kurtosis of the quadratic-normal distribution. The set Ω of all possible values of (\bar{m}_3, \bar{m}_4) for which ε_i is either a one-to-one function or one of which $q \leq 0.001$ may be represented approximately by the shaded region in Figure 1.

A similar transformation introduced by Fleishman [8] takes the form $\varepsilon_i = a + be_i + ce_i^2 + de_i^3$ where a, b, c and d are constants. The corresponding ranges of \bar{m}_3 and \bar{m}_4 are approximately $[-1.82, 1.82]$ and $[2.32, 8.22]$ respectively. Compared to the transformation of Fleishman, Equation (3) gives a bigger set of possible values of (\bar{m}_3, \bar{m}_4) .

Without loss of generality, we may assume that β_p is the parameter of interest. To find a confidence interval for the parameter β_p , we may first let $\beta_p^{(0)}$ be a constant and solve the problem of testing the null hypothesis $H_0 : \beta_p = \beta_p^{(0)}$ against the alternative hypothesis $H_1 : \beta_p \neq \beta_p^{(0)}$.

The model in Equation (1) may first be written as $\mathbf{y}^{(m)} = \mathbf{y} - \beta_p^{(0)} \mathbf{x}_p = \mathbf{X}_1 \boldsymbol{\beta}^{(1)} + (\beta_p - \beta_p^{(0)}) \mathbf{x}_p + \boldsymbol{\varepsilon}$ (4)

where

$$\mathbf{X}_1 = [\mathbf{x}_0 \quad \mathbf{x}_1 \quad \dots \quad \mathbf{x}_{p-1}]$$

and

$$\boldsymbol{\beta}^{(1)} = (\beta_0 \quad \beta_1 \quad \dots \quad \beta_{p-1})^T .$$

Let $\hat{\boldsymbol{\beta}}^{(1)} = (\mathbf{X}_1^T \mathbf{X}_1)^{-1} \mathbf{X}_1^T \mathbf{y}^{(m)}$ be the least squares estimate of the parameter $\boldsymbol{\beta}^{(1)}$ in model $\mathbf{y}^{(m)} = \mathbf{X}_1 \boldsymbol{\beta}^{(1)} + \boldsymbol{\varepsilon}$ when H_0 is true. Then from Equation (4), we may write

$$\mathbf{y}^{(m)} - \mathbf{X}_1 \hat{\boldsymbol{\beta}}^{(1)} \equiv (\beta_p - \beta_p^{(0)}) \mathbf{x}_p + \boldsymbol{\varepsilon} . (5)$$

From Equation (5), we see that

$$\tilde{T} = (\mathbf{x}_p^T \mathbf{x}_p)^{-1} \mathbf{x}_p^T (\mathbf{y}^{(m)} - \mathbf{X}_1 \hat{\boldsymbol{\beta}}^{(1)}) \quad (6)$$

will tend to be near zero under H_0 .

$$\text{Let } T = (\mathbf{x}_p^T \mathbf{x}_p)^{-1} \mathbf{x}_p^T \mathbf{R} \quad (7)$$

where $\mathbf{R} = (R_1 \ R_2 \ \dots \ R_n)^T$ and R_i is the rank of the i th component in $\mathbf{y}^{(m)} - \mathbf{X}_1 \hat{\boldsymbol{\beta}}^{(1)}$ when the n components in $\mathbf{y}^{(m)} - \mathbf{X}_1 \hat{\boldsymbol{\beta}}^{(1)}$ are arranged in an ascending order.

Then, a two-tail acceptance region $\{T : L \leq T \leq U\}$ may be used to test H_0 at the α level. As the c.d.f $G(\cdot)$ of the random errors is unknown, theoretically we cannot find the end points L and U . However, we may attempt to find L and U for each member in a wide class of distributions and show that there exist two constants L^* and U^* such that $L \cong L^*$, $U \cong U^*$ for most of the distributions in this class.

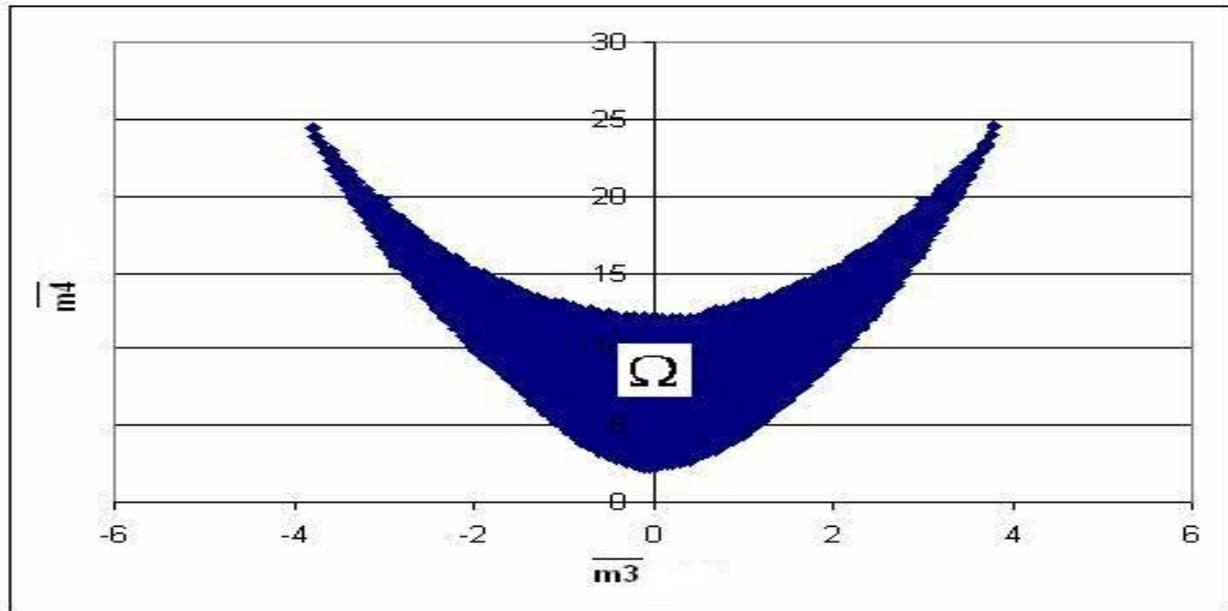


Figure 1. The shaded region represents all possible values of $(\bar{m}_3, \bar{m}_4) \in \Omega$

Let $QN(0, \bar{\lambda})$ denote a quadratic-normal distribution of which the variance is equal to 1. The parameter λ of a quadratic-normal distribution with variance σ^2 can be expressed in terms of $\bar{\lambda}$ as shown below:

$$\lambda_i = \sigma \bar{\lambda}_i, \quad i = 1, 2, \quad \lambda_3 = \bar{\lambda}_3.$$

Presently, we consider the class of quadratic-normal distribution with parameters 0 and $\lambda = (\lambda_1, \lambda_2, \lambda_3)^T$ such that the corresponding $(\bar{m}_3, \bar{m}_4) \in \Omega$. The process to find L^* and U^* is as follows.

Let the values of $\mathbf{X}_1, \boldsymbol{\beta}^{(1)}, \beta_p^{(0)}$ and σ be given. Firstly we select the values of (\bar{m}_3, \bar{m}_4) from the shaded region Ω such that for (\bar{m}_3, \bar{m}_4) not in the boundary of Ω , the values of \bar{m}_3 are in equal steps of size 0.1 while those of \bar{m}_4 are in equal steps of size 0.2. The total number of such chosen values of

(\bar{m}_3, \bar{m}_4) turns out to be $M = 2616$. Let $\bar{\lambda}^{(j)}$ be the value of parameter vector which corresponds to the j -th selected value. We generate N values of \mathbf{y} using the equation

$$\mathbf{y} = \mathbf{X}_1 \boldsymbol{\beta}^{(1)} + \boldsymbol{\varepsilon},$$

where $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$ are independent, $\varepsilon_i \sim QN(0, \lambda^{(j)})$, and $\lambda^{(j)}$ is given by

$$\lambda_i^{(j)} = \sigma \bar{\lambda}_i^{(j)}, \quad i = 1, 2, \quad \lambda_3^{(j)} = \bar{\lambda}_3^{(j)}.$$

We note that the value of $\mathbf{y}^{(m)} - \mathbf{X}_1 \hat{\boldsymbol{\beta}}^{(1)}$ in Equation (6) will be changed to $c(\mathbf{y}^{(m)} - \mathbf{X}_1 \hat{\boldsymbol{\beta}}^{(1)})$ if we change the value of σ to $c\sigma$. However, the value of \mathbf{R} in Equation (7) will remain the same when the value of σ is changed. Thus, the values of $L^{(j)}$ and $U^{(j)}$ will remain the same if we change the value of σ to other values.

For each generated value of \mathbf{y} , we compute the value of T (see Equation (7)) and estimate the values $L^{(j)}$ and $U^{(j)}$ such that

$$P(T < L^{(j)}) \cong P(T > U^{(j)}) \approx \alpha/2.$$

Consider the case when $p = 1$ and $n = 30$. When the histogram for the M values of $L^{(j)}$ (or $U^{(j)}$) are plotted (see Figures 2 and 3), we see that the spread of the lower limit $L^{(j)}$ of the acceptance region (or the upper limit $U^{(j)}$ of the acceptance region) is very small.

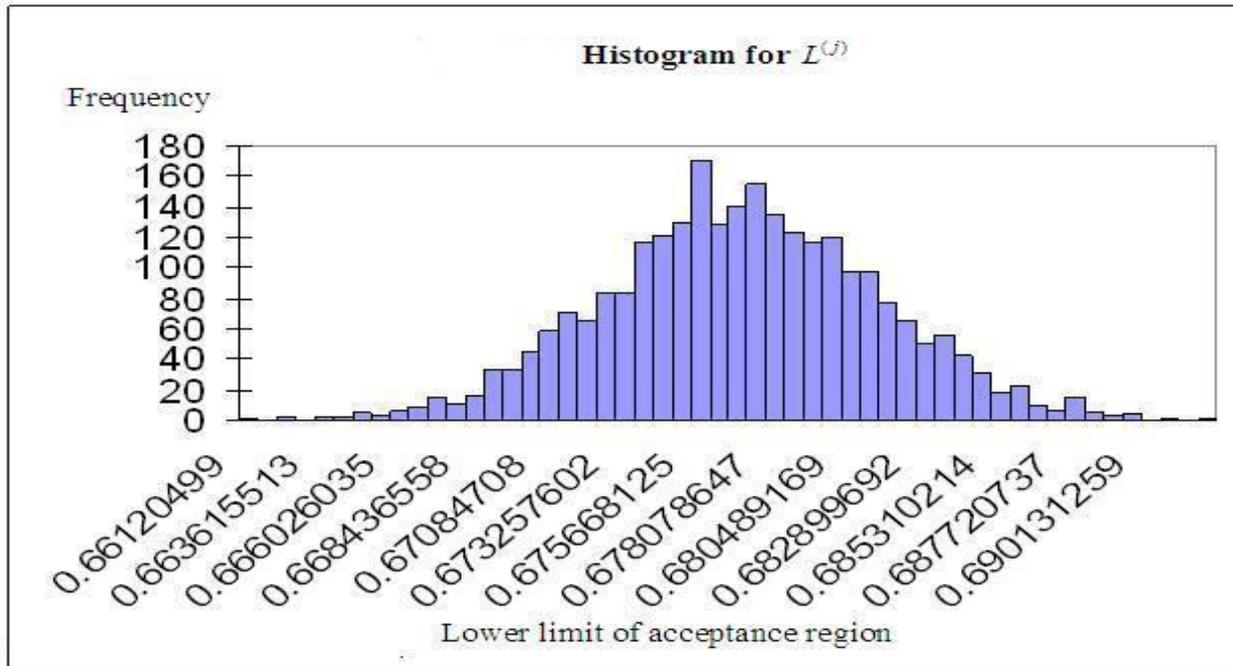


Figure 2. Histogram for $M = 2616$ values of $L^{(j)}$

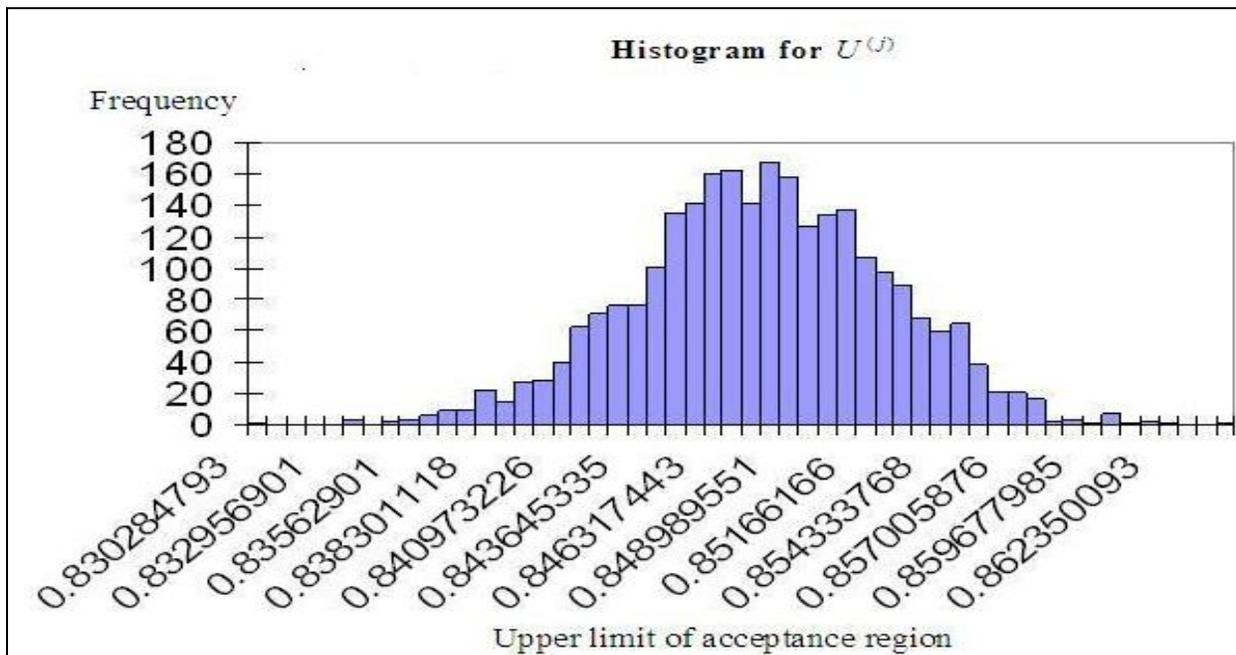


Figure 3. Histogram for $M = 2616$ values of $U^{(j)}$

Let L^* (or U^*) be the average value of the M values $L^{(1)}, L^{(2)}, \dots, L^{(M)}$ (or $U^{(1)}, U^{(2)}, \dots, U^{(M)}$). Furthermore, let $A^{(j)} = P(L^* \leq T \leq U^* | \bar{\lambda}^{(j)})$ be the probability evaluated at $\bar{\lambda}^{(j)}$ of the acceptance region $[L^*, U^*]$.

When $p = 1$ and $n = 30$, the histogram for the M values of $A^{(j)}$ is shown in Figure 4. The histogram shows that the spread of $A^{(j)}$ is fairly small. The acceptance region $\{T : L^* \leq T \leq U^*\}$ may thus be

used to test $H_0 : \beta_p = \beta_p^{(0)}$ at the α level. The above method for finding the acceptance region based on the rank test statistics is similar to that described in [11] which uses rank test statistic for testing the interaction in a two way layout.

A confidence interval for β_p is now given by

$$\delta(\mathbf{y}) = \left\{ \beta_p^{(0)} : H_0 \text{ that } \beta_p = \beta_p^{(0)} \text{ is accepted} \right\},$$

and the coverage probability of the confidence interval is

$$P^{(j)} = P(\beta_p \in \delta(\mathbf{y}) | \lambda = \bar{\lambda}^{(j)}).$$

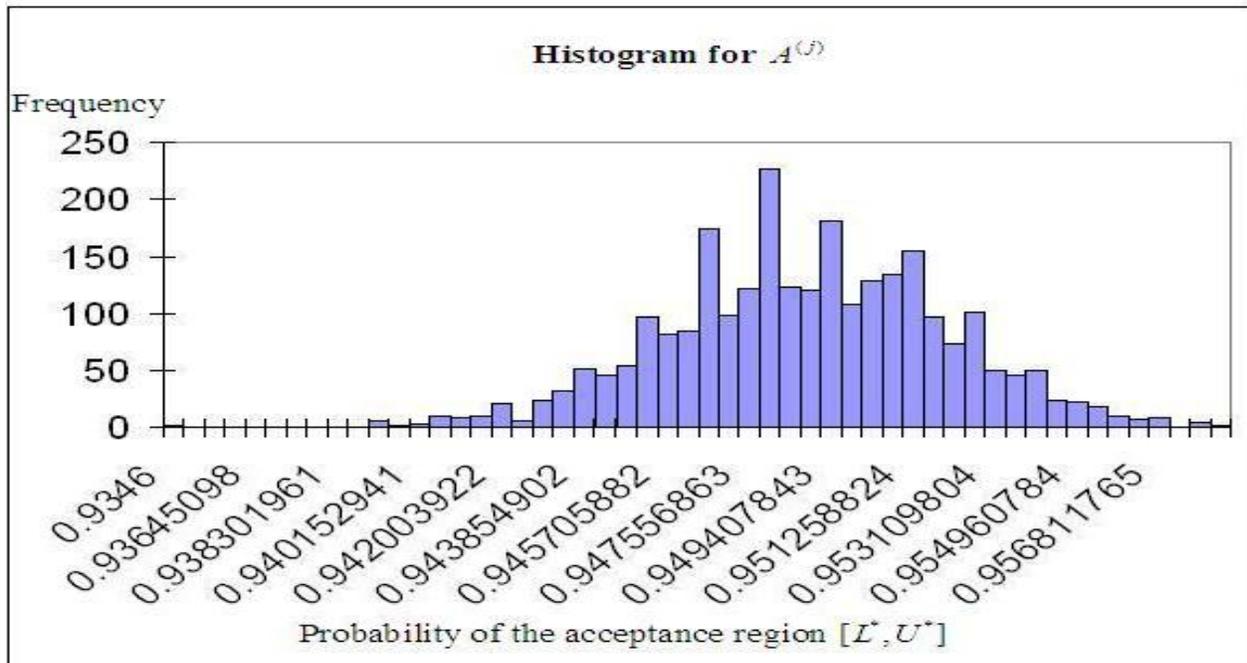


Figure 4. Histogram for $M = 2616$ values of $A^{(j)}$

NUMERICAL RESULTS

To estimate the coverage probabilities of the confidence intervals in Sections 1 and 2, we first generate N values of \mathbf{y} . For each generated value of \mathbf{y} , we find the classical confidence interval, percentile bootstrap confidence interval and confidence interval based on ranks. We next use the proportion \hat{p} of \mathbf{y} (out of N values of \mathbf{y}) of which the corresponding confidence interval covers the true value of β_p . The value of \hat{p} is then an estimate of the coverage probability. The results for \hat{p} are shown in Tables 1 and 2.

The value of $(\bar{m}_3, \bar{m}_4) = (0.0, 3.0)$ in Tables 1 and 2 corresponds to the case when $\lambda^T = (\lambda_1, 0, -1)$. Equation (3) shows that ε_i then has a normal distribution. The remaining five values of (\bar{m}_3, \bar{m}_4) appearing in Tables 1 to 2 correspond respectively to the cases when the values of $\bar{\lambda}^T = \{\lambda_1 / (m_2)^{1/2}, \lambda_2 / (m_2)^{1/2}, \lambda_3\}$ for $\varepsilon_i / (m_2)^{1/2}$ are equal to (1.11551, -0.07269, -1.0), (0.792009, 0.12745, -1.0), (0.305646, 0.503374, -0.60584), (0.144808, 0.76962, -0.19564) and (0.005644, 0.875607, -0.07203). The third value (0.0, 4.0) for (\bar{m}_3, \bar{m}_4) corresponds to the case

when ε_i is symmetrical but having narrow waist. The last three values for (\bar{m}_3, \bar{m}_4) correspond to the cases when ε_i is skewed to the right and having large kurtosis.

Tables 1 and 2 show that when ε_i has a normal distribution, the coverage probabilities and expected lengths of the confidence intervals found by using the methods in Sections 1 and 2 are comparable. When the distributions of the errors are non-normal, the coverage probabilities of the three types of confidence intervals considered are still comparable and close to the target value 0.95. However, the expected lengths of the confidence intervals based on rank statistics are much shorter than those of the bootstrap and classical confidence intervals when the distributions of the errors are skewed.

To have an insight into the three types of confidence intervals, we plot 100 simulated confidence intervals for β_1 in Figures 5 to 7. In the figures, the upper limits of the 100 confidence intervals based on rank statistics have been arranged in an ascending order. We see from Figure 5 that the end points of the confidence intervals based on rank statistics are fairly similar to those of the classical confidence intervals. Figures 6 and 7 show that in the case when the distribution of the errors is skewed, the three types of confidence intervals cover the true value 3 of β_1 for 97-98% of the time. These figures also show how much shorter could the confidence interval based on rank statistics be in comparison with the bootstrap and classical confidence intervals.

Table 1. Coverage probabilities and expected lengths of confidence intervals for β_1 when $p = 1, n = 30; x_i = i, 1 \leq i \leq 30; \beta_0 = 2, \beta_1 = 3, \sigma = 1$ and $\alpha = 0.05$.
($N = 5000$, standard error of coverage probability ≈ 0.003)

No	\bar{m}_3	\bar{m}_4	CP.RK	CP.Bt	CP.CL	EL.RK	EL.Bt	EL.CL
1	0	2.6	0.9500	0.9394	0.9526	0.088	0.082	0.086
2	0	3.0	0.9492	0.9380	0.9528	0.087	0.082	0.086
3	0	4.0	0.9460	0.9398	0.9528	0.083	0.082	0.085
4	1.0	9.0	0.9511	0.9416	0.9561	0.067	0.083	0.083
5	3.0	20.0	0.9494	0.9284	0.9560	0.051	0.084	0.082
6	3.8	24.5	0.9381	0.9328	0.9547	0.032	0.084	0.075

Table 2. Coverage probabilities and expected lengths of confidence intervals for β_1 when $p = 1, n = 40; x_i = i, 1 \leq i \leq 40; \beta_0 = 2, \beta_1 = 3, \sigma = 1$ and $\alpha = 0.05$.
($N = 5000$, standard error of coverage probability ≈ 0.003)

No	\bar{m}_3	\bar{m}_4	CP.RK	CP.Bt	CP.CL	EL.RK	EL.Bt	EL.CL
1	0	2.6	0.9424	0.9396	0.9458	0.058	0.053	0.055
2	0	3.0	0.9422	0.9398	0.9460	0.056	0.053	0.055
3	0	4.0	0.9424	0.9404	0.9460	0.053	0.053	0.055
4	1.0	9.0	0.9423	0.9436	0.9504	0.043	0.053	0.054
5	3.0	20.0	0.9407	0.9340	0.9542	0.031	0.054	0.053
6	3.8	24.5	0.9419	0.9364	0.9590	0.018	0.054	0.048

In Tables 1 and 2,

CP.RK = Coverage probability of confidence interval based on rank statistic;

CP.Bt = Coverage probability of bootstrap confidence interval;

CP.CL = Coverage probability of classical confidence interval;

EL.RK = Expected length of confidence interval based on rank statistic;

EL.Bt = Expected length of bootstrap confidence interval;

EL.CL = Expected length of classical confidence interval.

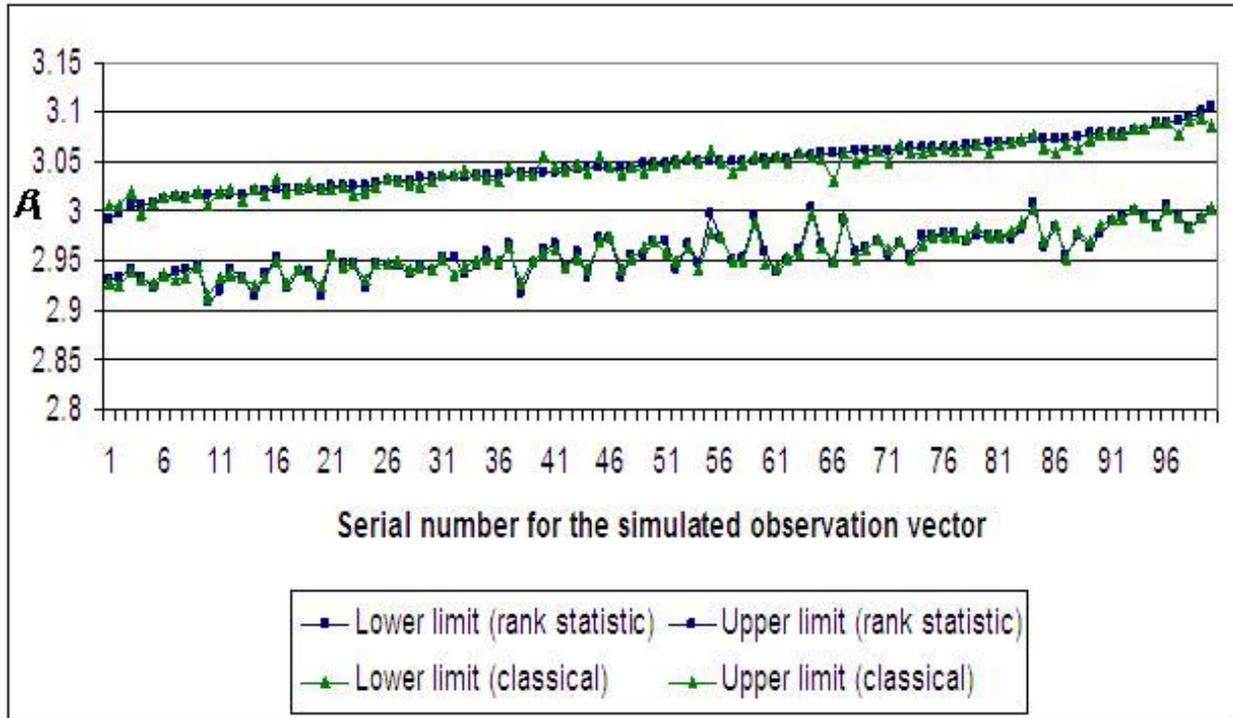


Figure 5. 100 simulated confidence intervals for β_1 when $\bar{m}_3 = 0, \bar{m}_4 = 3.0; p = 1, n = 30; x_i = i, 1 \leq i \leq 30; \beta_0 = 2, \beta_1 = 3, \sigma = 1$ and $\alpha = 0.05$.

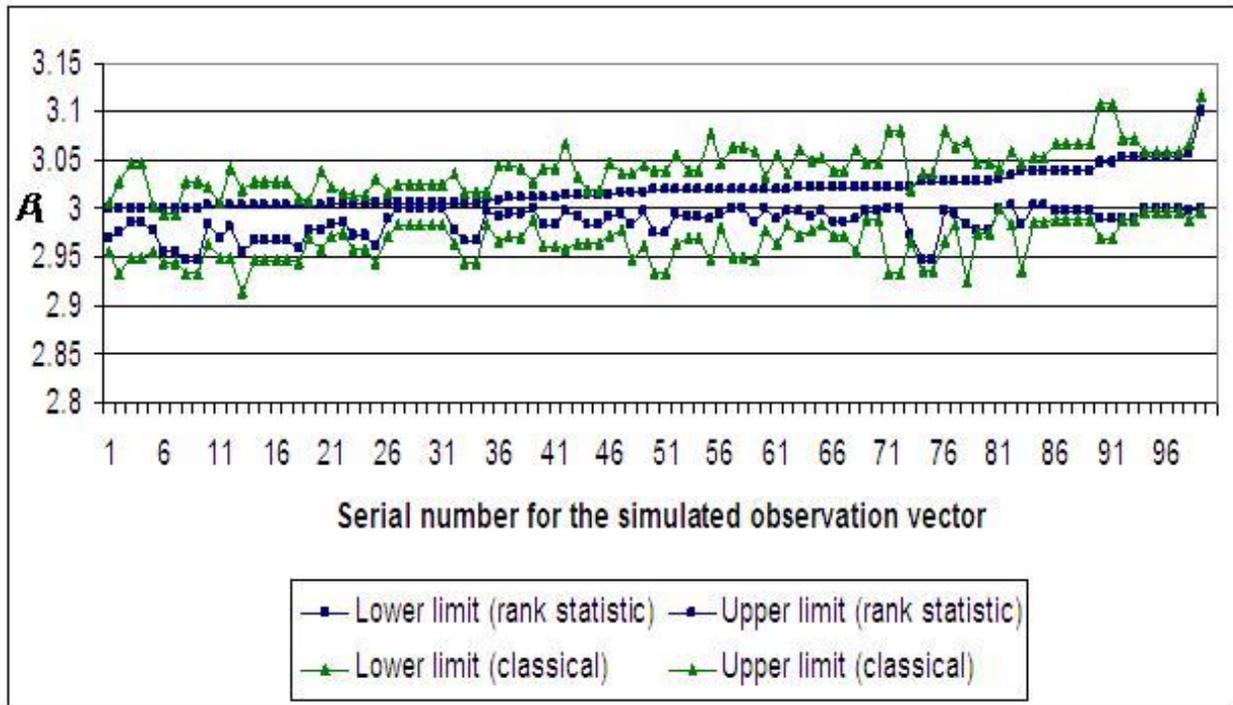


Figure 6. 100 simulated confidence intervals for β_1 when $\bar{m}_3 = 3.8, \bar{m}_4 = 24.5; p = 1, n = 30; x_i = i, 1 \leq i \leq 30; \beta_0 = 2, \beta_1 = 3, \sigma = 1$ and $\alpha = 0.05$.

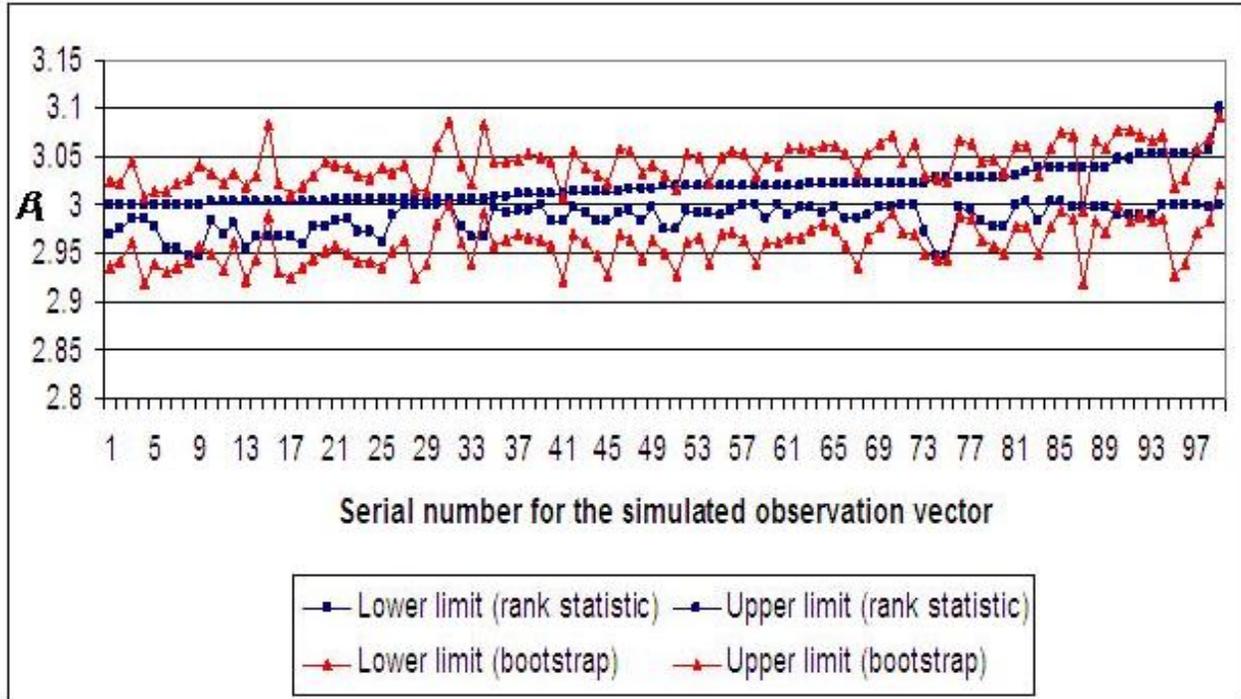


Figure 7. 100 simulated confidence intervals for β_1 when $\bar{m}_3 = 3.8, \bar{m}_4 = 24.5; p = 1, n = 30; x_i = i, 1 \leq i \leq 30; \beta_0 = 2, \beta_1 = 3, \sigma = 1$ and $\alpha = 0.05$.

CONCLUSIONS

The proposed method based on rank statistics for finding confidence intervals for the individual parameters can be applied for a wide range of distributions of the errors. An obvious advantage of the proposed confidence interval is its shorter expected length when the errors have a skewed distribution.

It is expected that the end points L^* and U^* of the acceptance region in Section 2 would depend on the matrix \mathbf{X}_1 when $p \geq 2$. Thus, future research may be carried out to devise efficient methods for finding the end points when the value of \mathbf{X}_1 is given.

ACKNOWLEDGEMENTS

The authors are very grateful to the referee for making various suggestions which have considerably improved the article.

REFERENCES

1. Adichie, J.N.(1978). Rank tests of sub-hypotheses in the general linear regression. *The Annals of Statistics*, **6**:1015-1025.

2. Beran,R.(1987). Prepivoting to reduce level error of confidence sets. *Biometrika*, **74**:457-468.
3. Beran,R.(1988). Balanced simultaneous confidence sets. *Journal of the American Statistical Association*, **83**:679-686.
4. Box, G.E.P and Cox, D.R.(1964). An analysis of transformations. *Journal of Royal Statistical Society Series B*, **26**:211-252.
5. Bartolucci, F. and Scaccia, L. (2005). The use of mixtures for dealing with non-normal regression errors. *Computational Statistics & Data Analysis*, **48**:821-834.
6. Efron,B.(1982). The Jackknife, the bootstrap and other resampling plans. SIAM, Philadelphia.
7. Efron, B.(1987). Better bootstrap confidence intervals (with discussion). *Journal of the American Statistical Association*, **82**:171-200.
8. Fleishman, A.I (1978). A method for simulating non-normal distributions. *Psychometrika*, **43(4)**:521-532.
9. Knautz, H.(1993). Nichtlineare Schatzung des Parametervektors im linearen Regressionsmodell. *Mathematical Systems in Economics*, **33**, Anton Hain, Frankfurt a.M.
10. Knautz, H. (1999). Nonlinear unbiased estimation in the linear regression model with nonnormal disturbances. *Journal of Statistical Planning and Inference*, **81**:293-309.

11. Koh, Y.B. and Pooi, A.H. (2007). Test for Two-factor Interaction based on Ranks. Technical Report No. 13/2005. Institute of Mathematical Sciences, University of Malaya.
12. Loh, W-Y. (1987). Calibrating confidence coefficients. *Journal of the American Statistical Association*, **82**:155-162.
13. Pooi, A.H. (2003). Effects of non-normality on confidence intervals in linear models. Technical Report No. 6/2003. Institute of Mathematical Sciences, University of Malaya.

