Confidence Intervals Based on Rank Statistics in Simple Linear Models

By:

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

(Paper presented at the ***International Conference on Applied Probability and Statistics*** held on 1-3 December 2008 in Hanoi, Vietnam)

Confidence Intervals based on Rank Statistics in Simple Linear Models

K.H.Ng[1], M.H.Lim[2] and A.H.Pooi[3]
Institute of Mathematical Sciences,
University of Malaya,
Malaysia

Abstract

Consider the simple linear models with non-normal errors. A method based on rank statistics has been proposed in our earlier work for constructing confidence interval for the slope parameter. When the skewness of the distributions of the errors is large, and the values of the explanatory variable are in equal steps, the method based on rank statistics has been shown to produce confidence interval of which the expected length is shorter than those of the bootstrap confidence interval, and the classical confidence interval which is derived by assuming that the errors are normally distributed. An important step in the method based on rank statistics is the determination of the acceptance region for testing the null hypothesis that the slope parameter is zero. Presently we show that the acceptance region depends basically on the skewness and kurtosis of the values of the explanatory variable. This finding suggests that when the values of the explanatory variable are given, we may construct a confidence interval for the slope parameter by using an acceptance region which has been determined for other set of values of the explanatory variable having the same measures of skewness and kurtosis. Thus in implementing the method in practice, we may use pre-determined acceptance regions. We also investigate the performance of the confidence interval based on rank statistics when the values of the explanatory variable are not in equal steps. Our simulation studies show that the method based on rank statistics continues to give better performance.

Keyword: Rank Statistics;; Confidence intervals; Linear Model; Quadratic-normal distribution; Bootstrap.

[1]Corresponding author: E-mail: kokhaur@um.edu.my

1. Introduction

Consider the linear model which can be represented in the form

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

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

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

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

where $\hat{\beta}_i$ is the least squares estimate of β_i, $t_{\alpha/2, n-2}$ is the $(1-\alpha/2)100\%$ point of the t-distribution with $(n-(p+1))$ degrees of freedom(df), $S_{\hat{\beta}_i} = \left\{a^{i+1,i+1}\right\}^{1/2} \hat{\sigma}$ 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 in [1]. An alternative way of constructing confidence interval is by means of bootstrap

(see for example, [2],[3],[4],[5] and [6]). 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 (see [7] and [8]) and the estimates based on EM algorithm ([9]).

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 a fixed constant is by using the statistics based on the ranks of residuals (see for example [10]).

In the previous simulation study (see [11]), we consider the case when $p = 1$ and the x_i are in equal steps. The simulated results show that in the case of normal errors, the confidence interval based on hypothesis testing using rank statistics is comparable to the classical confidence interval and 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 statistics is much shorter than those of the classical confidence interval and bootstrap confidence interval.

Presently, we show that when $p = 1$, the acceptance region of the test based on rank statistics for the slope parameter basically depends on the measures of skewness and kurtosis of the values of the explanatory variable. This finding implies that when the values of $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$ are given, we may first determine the measures of skewness and kurtosis of the values $x_1 . x_2, \cdots, x_n$ and use an acceptance region which has been previously

determined for another set of values of the explanatory variable having the same measures of skewness and kurtosis. The use of pre-determined acceptance regions would reduce significantly the computing time required for implementing the method for finding confidence interval.

We also perform more simulation studies for the case when the values of the explanatory variable are not in equal steps. Our simulation results show that when the distribution of the errors is skewed, rank test continues to yield shorter confidence intervals for the slope parameter.

2. Confidence Interval when Errors are Non-normally Distributed

In this section, we give an outline of the method (see[11]) based on ranks for finding a confidence interval for the slope parameter β_1 in the simple linear model.

Let $\beta_1^{(0)}$ be a constant and consider the problem of testing the null hypothesis $H_0 : \beta_1 = \beta_1^{(0)}$ against the alternative hypothesis $H_1 : \beta_1 \neq \beta_1^{(0)}$.

The simple linear model

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad , i = 1,2,\cdots,n \tag{3}$$

may first be written as

$$y_i^{(m)} = y_i - \beta_1^{(0)} x_i = \beta_0 + (\beta_1 - \beta_1^{(0)}) x_i + \varepsilon_i \, , \; i = 1,2,\cdots,n \tag{4}$$

Let $E_i = y_i^{(m)} - \frac{1}{n}\sum_{i=1}^{n} y_i^{(m)}$. When H_0 is true,

$$\widetilde{T} = \frac{\sum_{i=1}^{n} x_i E_i}{\sum_{i=1}^{n} x_i^2}$$

will tend to be near zero.

Let

$$T = \frac{\sum_{i=1}^{n} x_i R_i}{\sum_{i=1}^{n} x_i^2}$$

where R_i is the rank of E_i when the values $E_1, E_2, \cdots, E_n$ are arranged in an ascending order. When H_0 is true, we would expect that T will also tend to be near zero.

Let $QN(0,\lambda)$ be the quadratic-normal distribution with parameters 0 and λ (see [12]) and $\overline{m}_3$ together with $\overline{m}_4$ be respectively the measures of skewness and kurtosis of the quadratic-normal distribution.

Let the constants $L(\lambda)$ and $U(\lambda)$ be such that

$$P\{L(\lambda) \le T \le U(\lambda) \mid \lambda\} = 1 - \alpha$$

Next let L^* and U^* be respectively the average values of $L(\lambda)$ and $U(\lambda)$ over the values of λ which are feasible. Then, a confidence interval for β_1 is given by

$$\delta(\mathbf{y}) = \{\beta_1^{(0)} : L^* \le T \le U^*\}$$

and the coverage probability of the confidence interval is

$$P(\lambda) = P(\beta_1 \in \delta(\mathbf{y}) \mid \lambda)$$

The values of L^* and U^* can be found by using simulation. When $n = 30$ or 35, Table I gives the values of L^* and U^* for a variety of values for the measures of skewness $(\overline{m}_3^{(x)})$ and kurtosis $(\overline{m}_4^{(x)})$ of the explanatory variable. When $n = 30$ or 35 but the values of $\overline{m}_3^{(x)}$ and $\overline{m}_4^{(x)}$ computed from the data are not in Table I, interpolation may be used.

3. Numerical Results

To estimate the coverage probability $P(\boldsymbol{\lambda})$ of the confidence interval in Section 2, we first generate N values of $\mathbf{y}$. For each generated value of $\mathbf{y}$, we find the confidence intervals using the classical method, bootstrap, and the procedure in Section 2. We next compute 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_1. The value of $\hat{p}$ is then an estimate of the corresponding coverage probability.

Tables II to IV show the results of the coverage probabilities and expected lengths of confidence intervals for β_1 when the values of the explanatory are negatively skewed, symmetrical or positively skewed.

Tables II to IV show that irrespective of the sizes of the skewness and kurtosis of the values of the explanatory variables and the distribution of the errors, the coverage probabilities for the three types of confidence intervals are all fairly close to the target value 0.95. But, when the skewness of the distribution of the errors is large, the expected length of the confidence interval based on rank statistics is much shorter than those of the classical confidence interval and bootstrap confidence interval.

4. Concluding remarks

Suppose $x = x^*$ is given and we are interested in finding a prediction interval for the future observation when $x = x^*$. The method based on ranks may first be applied to find a confidence interval for the parameter $\eta = \beta_0 + \beta_1 x^*$. A prediction interval may then be obtained by enlarging the confidence interval for η. It would then be interesting to investigate the performance of the resulting prediction interval. Future research may also be carried out to investigate the possibility of using rank statistics to construct simultaneous confidence intervals.

References

1. Box, G.E.P, Cox, D.R.(1964). An analysis of transformations. *Journal of Royal Statistical Society Series B*, **26**:211-252.
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. Efron,B.(1982). The Jacknife, the bootstrap and other resampling plans. SIAM, Philadelphia.
5. Efron, B.(1987). Better bootstrap confidence intervals (with discussion). *Journal of the American Statistical Association*, **82**:171-200.
6. Loh,W-Y.(1987). Calibrating confidence coefficients. *Journal of the American Statistical Association*, **82**:155-162.
7. Knautz, H.(1993). Nichtlineare Schatzung des Parametervektors im linearen Regressionsmodell. *Mathematical Systems in Economics*, **33**, Anton Hain, Frankfurt a.M.
8. Knautz, H.(1999). Nonlinear unbiased estimation in the linear regression model with nonnormal disturbances. *Journal of Statistical Planning and Inference,* **81**:293-309.
9. Bartolucci, F. and Scaccia, L.(2005). The use of mixtures for dealing with non-normal regression errors. *Computational Statistics & Data Analysis*, **48**:821-834.
10. Adichie, J.N.(1978). Rank tests of sub-hypotheses in the general linear regression. *The Annals of Statistics*, **6**:1015-1025.
11. K.H.Ng, M.H.Lim and A.H.Pooi (2008) Confidence Intervals based on Rank Statistics in Linear Models. Proceeding The 9th Islamic Countries Coference on Statistical Sciences 2007 (ICCS-IX), 1646-1660.

12. 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.

Table I: The values of L^* and U^* when $n=30$ or $n=35$

		$n=30$		$n=35$	
$\overline{m}_3^{(x)}$	$\overline{m}_4^{(x)}$	L^*	U^*	L^*	U^*
-3.5	21.3	-6.3277	-1.1420	-6.7501	-0.0346
-3.1	18.6	-6.9907	-1.9008	-7.0246	-1.0189
-2.8	17.2	-7.4557	-2.4250	-7.6517	-1.6869
-2.5	14	-7.0963	-1.9805	-6.9646	-0.9410
-2.3	13.6	-7.5328	-2.4701	-7.5870	-1.5938
-2.1	12.8	-7.7380	-2.6901	-7.8449	-1.8581
-2.0	12.6	-7.8895	-2.8567	-8.0563	-2.0784
-1.9	10.4	-7.3115	-2.1695	-7.0674	-1.0217
-1.8	14	-8.5872	-3.6674	-9.1633	-3.2713
-1.6	10.8	-8.1158	-3.0729	-8.2680	-2.2718
-1.5	13.4	-8.8767	-3.9772	-9.5450	-3.6690
-1.2	11.4	-8.8582	-3.9062	-9.3769	-3.4445
-1	5.2	-7.1944	-1.8195	-6.4420	-0.2097
-0.9	7	-8.0794	-2.9133	-7.8850	-1.7868
-0.8	8	-8.5514	-3.4597	-8.6454	-2.5893
-0.7	8	-8.6840	-3.6003	-8.8250	-2.7706
-0.6	7.6	-8.7009	-3.5993	-8.8004	-2.7294
-0.5	6.8	-8.5904	-3.4424	-8.5532	-2.4436
-0.4	4.4	-7.7509	-2.4147	-7.0822	-0.8689
-0.34817	4.33657	-7.7999	-2.4583	-7.1327	-0.9128
-0.3	10.4	-9.6455	-4.7149	-10.3509	-4.4162
-0.2	4.8	-8.2401	-2.9404	-7.7787	-1.5643
-0.15999	3.659049	-7.6839	-2.2668	-6.8326	-0.5684
-0.1	6.8	-9.0499	-3.9089	-9.1370	-3.0102
-0.07277	2.602047	-6.8889	-1.3095	-5.5354	0.7674
-0.02312	2.947682	-7.3358	-1.8096	-6.1899	0.1213
0	6.6	-9.0970	-3.9381	-9.1564	-3.0132
0.025628	3.408262	-7.7682	-2.3034	-6.8507	-0.5469
0.1	2.6	-7.1356	-1.5245	-5.8009	0.5473
0.10437	3.064088	-7.6017	-2.0735	-6.5261	-0.1912
0.2	3.6	-8.0897	-2.6231	-7.2582	-0.9273
0.3	3.6	-8.1820	-2.6933	-7.3398	-0.9856
0.4	4	-8.4968	-3.0405	-7.8028	-1.4491
0.5	4.6	-8.8662	-3.4634	-8.3731	-2.0382
0.6	5.2	-9.1839	-3.8271	-8.8684	-2.5516
0.7	5.6	-9.3984	-4.0611	-9.1813	-2.8699
0.8	6.2	-9.6625	-4.3644	-9.5935	3.3015
0.9	7.4	-10.0590	-4.8551	-10.2802	-4.5056
1	8.2	-10.0313	-5.1588	-10.6967	-4.5096
1.1	9.4	-10.6226	-5.5481	-11.2373	-5.1195
1.2	11.2	-10.9931	-6.0361	-11.9191	-5.9161
1.3	13.4	-11.3656	-6.5349	-12.6144	-6.7522
1.5	7.8	-10.5359	-5.1524	-10.5658	-4.1477
1.6	11.2	-11.2856	-6.2405	-12.1131	-6.0090
1.7	14	-11.7340	-6.8812	-13.0083	-7.1120
1.9	10.8	-11.4391	-6.2341	-12.0175	-5.7389

Continued from Table I

		$n = 30$		$n = 35$	
$\overline{m}_3^{(x)}$	$\overline{m}_4^{(x)}$	L^*	U^*	L^*	U^*
2.2	13.2	-12.0326	-6.9501	-12.9642	-6.8007
2.4	13.4	-12.2417	-7.0517	-13.0515	-6.7697
2.6	13.8	-12.5173	-7.7834	-13.1839	-6.7375
2.7	18	-12.9775	-8.1452	-14.5062	-8.6176
3	18	-13.3415	-8.3116	-14.6459	-8.5228
3.4	20.2776	-14.1895	-9.0059	-15.5075	-9.2155
3.8	24.5	-14.9789	-10.0354	-17.1199	-10.9746

Table II Coverage probabilities and expected lengths of confidence intervals for β_1 when $n=30; \overline{m}_3^{(x)}=-3.5, \overline{m}_4^{(x)}=21.3; \beta_0=2, \beta_1=3, \sigma=1$ **and** $\alpha=0.05$.
($N=1000$, standard error of coverage probability ≈ 0.0069)

No	$\overline{m}_3$	$\overline{m}_4$	CP.RK	CP.Bt	CP.CL	EL.RK	EL.Bt	EL.CL
1	-3.8	24.4	0.958	0.936	0.939	0.272	0.673	0.632
2	-1.0	9.0	0.961	0.952	0.949	0.566	0.667	0.661
3	-0.1	2.8	0.962	0.960	0.967	0.721	0.651	0.675
4	0.0	3.0	0.962	0.958	0.968	0.714	0.652	0.675
5	1.0	9.0	0.961	0.949	0.956	0.567	0.665	0.659
6	3.8	24.5	0.962	0.949	0.945	0.278	0.668	0.626

Table III: Coverage probabilities and expected lengths of confidence intervals for β_1 when $n=30; \overline{m}_3^{(x)}=0, \overline{m}_4^{(x)}=3.0; \beta_0=2, \beta_1=3, \sigma=1$ **and** $\alpha=0.05$.
($N=1000$, standard error of coverage probability ≈ 0.0069)

No	$\overline{m}_3$	$\overline{m}_4$	CP.RK	CP.Bt	CP.CL	EL.RK	EL.Bt	EL.CL
1	-3.8	24.4	0.952	0.945	0.953	0.248	0.732	0.688
2	-1.0	9.0	0.960	0.951	0.962	0.580	0.720	0.719
3	-0.1	2.8	0.960	0.956	0.956	0.760	0.708	0.734
4	0.0	3.0	0.960	0.955	0.956	0.752	0.709	0.733
5	1.0	9.0	0.960	0.954	0.959	0.577	0.718	0.717
6	3.8	24.5	0.958	0.945	0.957	0.256	0.728	0.680

Table IV: Coverage probabilities and expected lengths of confidence intervals for β_1 when $n = 30; \overline{m}_3^{(x)} = 3, \overline{m}_4^{(x)} = 20.0; \beta_0 = 2, \beta_1 = 3, \sigma = 1$ and $\alpha = 0.05$.
($N = 1000$, standard error of coverage probability ≈ 0.0069)

No	$\overline{m}_3$	$\overline{m}_4$	CP.RK	CP.Bt	CP.CL	EL.RK	EL.Bt	EL.CL
1	-3.8	24.4	0.954	0.948	0.954	0.396	1.007	0.948
2	-1.0	9.0	0.956	0.958	0.959	0.829	0.995	0.991
3	-0.1	2.8	0.956	0.955	0.961	1.057	0.976	1.011
4	0.0	3.0	0.956	0.956	0.961	1.048	0.976	1.010
5	1.0	9.0	0.956	0.955	0.958	0.828	0.993	0.988
6	3.8	24.5	0.956	0.943	0.953	0.408	1.005	0.938

In Tables II to IV,
CP.RK = Coverage probability of confidence interval based on rank statistics;
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 statistics;
EL.Bt = Expected length of bootstrap confidence interval;
EL.CL = Expected length of classical confidence interval.