A Comparison of Confidence Intervals from *R*-estimators in Regression

by

Karen J. George[1], Joseph W. McKean[2], William R. Schucany[3], & Simon J. Sheather[4]

[1] Department of Information Science and Engineering, University of Canberra
[2] Department of Mathematics & Statistics, Western Michigan University
[3] Department of Statistical Science, Southern Methodist University
[4] Australian Graduate School of Management, University of New South Wales

Technical Report No. SMU/DS/TR-277

A Comparison of Confidence Intervals from R-estimators in Regression

Karen J. George
Department of Information Science and Engineering
University of Canberra

Joseph W. McKean
Department of Mathematics and Statistics
Western Michigan University

William R. Schucany
Department of Statistical Science
Southern Methodist University

Simon J. Sheather
Australian Graduate School of Management
University of New South Wales

March 26, 1995

Abstract

The small sample properties of a number of approaches to finding a confidence interval for a slope parameter in rank regression are investigated. An approach based on the bootstrap percentile-t procedure is shown to have excellent overall performance.

Keywords: Bootstrap; Jackknife; Monte Carlo; Standard error.

1 Introduction

Rank-based methods are a popular robust alternative to least squares for analysing data from a linear model. Advantages of these procedures include bounds on the influence of outliers in the y space and the ability to handle both symmetric and asymmetric error distributions; see, for example, Hettmansperger (1984).

There is a large body of literature on rank-based methods beginning with the landmark papers of Jureckova (1971) and Jaeckel (1972). Much of this literature focusses on asymptotic properties. For these methods to be of widespread practical use, their finite sample properties need to investigated and understood. One such crucial property concerns the validity of hypothesis tests and confidence intervals, that is, whether the levels and confidence coefficients are close to their nominal values. McKean and Sheather (1991) provide a survey of the small sample properties of R-estimates and their acompanying analyses of linear models. They review the results of previous studies as well as discuss the results of their own simulations. They conclude that approaches based on the bootstrap and the jackknife have produced encouraging results for R-estimates based on sign and Wilcoxon scores, respectively, but that further refinement of these procedures is necessary to achieve widespread small sample validity.

In this paper we compare the small sample properties of a number of approaches to finding a confidence interval for a slope parameter based on an R-estimate. Section 2 of this paper provides a short summary of the properties of R-estimates. The focus here is upon Wilcoxon scores, since these are the most widely used scores. Asymptotic confidence intervals for regression parameters depend on a scale parameter that must be estimated in practice. Different approaches to estimating this parameter are discussed in Section 3. Finally, in Section 4 the results of a Monte Carlo study are reported which compare the performance of various finite sample methods for obtaining a confidence interval, based on an R-estimator with Wilcoxon scores, for the slope parameter in straight line regression. An approach based on the bootstrap percentile-t procedure is shown to have excellent overall performance.

2 R-estimates of a Linear Model

Consider a linear model of the form

$$\mathbf{Y} = \beta_0 \mathbf{1} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}\,, \tag{2.1}$$

where $\mathbf{1}$ is a vector of 1's, $\mathbf{X}$ is an $n \times p$ matrix of known constants, β_0 is an intercept, $\boldsymbol{\beta}$ is a $p \times 1$ vector of parameters, and $\boldsymbol{\epsilon}$ is an $n \times 1$ vector of random errors which are independent and identically distributed with distribution function F and density function f. We shall assume that $\mathbf{X}$ has full column rank p and that med $\epsilon_i = 0$. Let σ^2 denote the variance of ϵ_i. It will be convenient to write the ith row of $\mathbf{X}$ as $\mathbf{x}_i'$.

The R-estimate of $\boldsymbol{\beta}$ proposed by Jaeckel (1972) is a value of $\boldsymbol{\beta}$ which minimizes the dispersion function,

$$D(\boldsymbol{\beta}) = \sum_{i=1}^{n} a\left(R\left(Y_i - \mathbf{x}_i'\boldsymbol{\beta}\right)\right)\left(Y_i - \mathbf{x}_i'\boldsymbol{\beta}\right), \tag{2.2}$$

where $R(u_i)$ denotes the rank of u_i among $u_1, \ldots, u_n$, and $a(1) \leq \cdots \leq a(n)$ is a given set of rank scores. The scores are generated as $a(i) = \phi\left(i/(n+1)\right)$ where ϕ is an increasing and bounded function defined on (0,1) which is standardized so that $\int \phi = 0$ and $\int \phi^2 = 1$. In this paper we shall only consider Wilcoxon scores generated by $\phi(u) = \sqrt{12}\,(u - \frac{1}{2})$ and we denote the resulting R-estimate of $\boldsymbol{\beta}$ as $\widehat{\boldsymbol{\beta}}$. The intercept β_0 is estimated by the median of the residuals $e_i = Y_i - \mathbf{x}_i'\widehat{\boldsymbol{\beta}}$.

The dispersion function D is a nonnegative and convex function of $\boldsymbol{\beta}$. The finite algorithm developed by George and Osborne (1990) is based on the reduced gradient algorithm (Osborne, 1985) and provides the exact R-estimate $\widehat{\boldsymbol{\beta}}$. The k-step Newton algorithm developed by McKean and Hettmansperger (1978) provides a consistent estimate which is an approximation to $\widehat{\boldsymbol{\beta}}$. The statistical package MINITAB now includes R-estimates based on the k-step algorithm.

Under mild regularity conditions, Jureckova (1971) showed that $\widehat{\boldsymbol{\beta}}$ is asymptotically normal $N\left(\boldsymbol{\beta}, \tau^2\left(X_c'X_c\right)^{-1}\right)$, where $\tau = \left(\sqrt{12}\int f^2(x)dx\right)^{-1}$ and X_c is X centered by subtracting the column means. A nominal 100(1 - α) % confidence interval for $\boldsymbol{\beta}$ based on the R-estimate

and suggested by this asymptotic distribution is given by

$$\widehat{\beta}_i \pm z_{(\frac{\alpha}{2})}\tau\sqrt{(X_c'X_c)_{ii}^{-1}} \tag{2.3}$$

where $z_{(\frac{\alpha}{2})}$ is the $(1-\frac{\alpha}{2})$ percentile of the standard normal distribution and $(X_c'X_c)_{ii}^{-1}$ denotes the ith diagonal entry of $X_c'X_c$. Estimates of τ are discussed in the next section.

3 Different Approaches to Estimating τ

Koul, Sievers and McKean (1987) proposed an estimator of $\gamma = \int f^2(x)dx$ and hence τ. It is based on H_n, the empirical distribution function of $|e_1 - e_2|$, where e_i denotes the ith residual $(i = 1, \ldots, n)$. Let F_n denote the empirical distribution function of $e_1, \ldots, e_n$ and $\hat{f}_K$ denote the kernel estimate of f, that is,

$$\hat{f}_K(x) = \frac{1}{nh}\sum_{j=1}^{n} K\left(\frac{x - e_j}{h}\right),$$

where the kernel K is a density function symmetric about 0 and the bandwidth $h \to 0$ as $n \to \infty$. Koul, Sievers and McKean (1987) showed that their estimator, $\widehat{\gamma}_{KSM}$ can be written as

$$\widehat{\gamma}_{KSM} = \int_{-\infty}^{\infty} \hat{f}_K(x)dF_n(x) = \frac{1}{n^2h}\sum_{i=1}^{n}\sum_{j=1}^{n} K\left(\frac{e_i - e_j}{h}\right)$$

with K equal to the rectangular kernel. They recommended one use $h = q_{n,\alpha}n^{-1/2}$, where $q_{n,\alpha}$ is the α_{th} quantile of H_n. Under mild regularity conditions, Koul, Sievers and McKean (1987) showed that their estimate is uniformly consistent under either symmetric or skewed errors. For a similar estimate of γ see Aubuchon and Hettmansperger (1989).

George and Osborne (George, 1993) propose an estimator of τ derived from the asymptotic linearity result for rank statistics of Jureckova (1969). First, note that the partial derivatives (gradient) of $D(\beta)$ exist almost everywhere and where they exist are equivalent to the negative of the regression rank statistic $S(\beta)$ given by Jureckova. Note, whereever the gradient of $D(\beta)$ doesn't exist, $S(\beta)$ does exist but is multi-valued. Next, this linearity

result suggests a quadratic approximation to (2.2)

$$D(\beta) \approx D(\beta^0) - (\beta - \beta^0)'S(\beta^0) + \frac{1}{2\tau}(\beta - \beta^0)'X_c'X_c(\beta - \beta^0), \tag{3.1}$$

where β^0 denotes the true parameter value. A thorough discussion of the above can be found in Hettmansperger (1984, p. 232-239).

Writing $\beta - \beta^0 = \beta - \widehat{\beta} + \widehat{\beta} - \beta^0$ and substituting this into (3.1), George and Osborne obtain the following approximation

$$D(\beta) \approx D(\beta^0) + \frac{1}{2\tau}(\beta - \widehat{\beta})'X_c'X_c(\beta - \widehat{\beta}) + \frac{1}{\tau}(\beta - \widehat{\beta})'X_c'X_c(\widehat{\beta} - \beta^0), \tag{3.2}$$

ignoring small values from $S(\beta^0) \approx 0$ and terms involving $(\widehat{\beta} - \beta^0)^2$. Choosing a mesh of values centered and symmetric about $\widehat{\beta}$, (3.2) defines a set of linear equations with the unknowns $D(\beta^0), 1/\tau, 1/\tau(\widehat{\beta} - \beta^0)$. Due to symmetry, the normal equations simplify so that $1/\tau$ can be estimated independently of the other unknowns. This estimate is given by

$$\widehat{\tau}_{GO} = \frac{bv - aw}{b^2 - ac}, \tag{3.3}$$

where

$$\begin{aligned}
a &= 4p + 1, \\
b &= 5\sum_{i=1}^{p} \delta_i^2 \left(X_c'X_c\right)_{ii}, \\
c &= \frac{17}{2}\sum_{i=1}^{p} \left(\delta_i^2 (X_c'X_c)_{ii}\right)^2, \\
v &= D(\widehat{\beta}) + \sum_{i=1}^{p} \left\{D(\widehat{\beta} + \delta_i I_i) + D(\widehat{\beta} - \delta_i I_i) + D(\widehat{\beta} + 2\delta_i I_i) + D(\widehat{\beta} - 2\delta_i I_i)\right\}, \\
w &= \frac{1}{2}\sum_{i=1}^{p} \delta_i^2 X_{ci}'X_{ci} \left\{D(\widehat{\beta} + \delta_i I_i) + D(\widehat{\beta} - \delta_i I_i) + 4D(\widehat{\beta} + 2\delta_i I_i) + 4D(\widehat{\beta} - 2\delta_i I_i)\right\},
\end{aligned}$$

X_{ci} denotes the ith column of X_c, I_i denotes the ith unit vector, and p is the rank of X. The set of values $\{\delta_i\}, i = 1 \ldots p$, are chosen to so that the resulting mesh is within a region where the linear trend of the equivalent rank statistics are expected to be strongest. The full details of this algorithm are given in George (1993).

4 Monte Carlo Study

A Monte Carlo study was carried out to compare different approaches to finding confidence intervals, based on an R-estimator with Wilcoxon scores, for the slope parameter in straight line regression. The straight line regression model used was

$$Y_i = \beta_0 + \beta_1 x_i + 0.1\epsilon_i \quad i = 1, \ldots, n. \tag{4.1}$$

where $\beta_0 = 2$ and $\beta_1 = 0.5$. The random errors ϵ_i were fixed to have standard deviation one and mean zero and were chosen from three distributions, normal, Laplace and lognormal. Two types of x were used, evenly spaced between 5 and 6 and uniformly distributed between 5 and 6. Three sample sizes (n = 10, 20 and 50) and two confidence coefficients (90% and 95%) were included in the study. These choices result in 36 combinations of x, n, error distribution and confidence coefficient. For each configuration, 1000 pseudo-random samples were drawn. Each of these samples was generated using the 48-bit linear congruential random number generator ERAND48, which is available in most Unix implementations. Care was taken to ensure that each configuration was started with a different randomly drawn seed. The two results that are summarised here are the coverage probabilities and the lengths of the confidence intervals for β_1.

4.1 Three confidence interval methods

The first type of confidence interval considered is motivated by the asymptotic theory reported in Section 2. Following the recommendations in McKean and Sheather (1991) t critical values with $n - 2$ degrees of freedom (denoted by $t_{n-2,(\frac{\alpha}{2})}$) were used in place of standard normal values in (2.3) giving

$$\hat{\beta}_1 \pm t_{n-2,(\frac{\alpha}{2})}\hat{\tau}\sqrt{(X_c'X_c)^{-1}} \tag{4.2}$$

as a nominal 100(1 - α)% confidence interval for β_1. Two different estimates, $\hat{\tau}$ were considered, namely, those of Koul, Sievers and McKean (1987) and George and Osborne (1992). We shall refer to these two estimates as the KSM and the GO estimates, respectively.

The second type of confidence interval considered is based on the jackknife. Originally introduced by Quenouille (1949) as a bias reduction technique, the jackknife provides a general tool for estimating variances. Let $\widehat{\beta}_{(i)}$ denote the estimate of β obtained when the i-observation is removed ($i = 1, \ldots, n$). Then the i-th pseudo-value is $n\widehat{\beta} - (n-1)\widehat{\beta}_{(i)}$. The jackknife variance estimate (Tukey, 1958) is just $1/n$ times the sample variance of the n pseudo-values. Schucany and Sheather (1989) showed that the jackknife variance estimator for R-estimators based on Wilcoxon scores is strongly consistent in the one- and two-sample location problems. They conjectured that the same is true in regression. A confidence interval based on the jackknife variance estimate is given by

$$\hat{\beta}_1 \pm t_{m-1,(\frac{\alpha}{2})}\sqrt{\widehat{var}_J}, \tag{4.3}$$

where $\widehat{var}_J$ is the jackknife variance estimate of $\hat{\beta}_1$ and m equals the number of distinct pseudo values. This formula for the degrees of freedom was first suggested by Mosteller and Tukey (1977).

The bootstrap percentile-t confidence intervals are the third type included in the study. This method has produced encouraging finite sample results for R-estimators based on sign scores (Schrader and McKean, 1987). Confidence intervals for β_1 involve bootstrapping a studentized statistic, $T = \frac{\hat{\beta}_1 - \beta_1}{\widehat{SD}}$ where $\widehat{SD}$ is a consistent estimate of the standard deviation of $\hat{\beta}_1$. The algorithm used to produce the confidence interval is:

1. Compute $\hat{\beta}_0$ and $\hat{\beta}_1$, the R-estimates of β_0 and β_1, the resulting residuals $e_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$ $i = 1, ..., n$ and $\widehat{SD}$, a consistent estimate of the standard deviation of $\hat{\beta}_1$.

2. Inflate the residuals by multiplying each of them by $\sqrt{n/(n-2)}$, as recommended in Stine (1989, p. 256).

3. Draw a bootstrap sample $e_1^*, ..., e_n^*$ with replacement from these inflated residuals.

4. Calculate a bootstrap sample of y's via $y_i^* = \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i^* \quad i = 1, ..., n$.

5. Based on the bootstrap sample compute $\hat{\beta}_1^*$ and $\widehat{SD}^*$, the R-estimate of β_1 and its estimated standard deviation.

6. Calculate the bootstrap version of the studentized statistic, $t^* = \frac{\hat{\beta}_1^* - \hat{\beta}_1}{\widehat{SD}^*}$.

7. Repeat steps 3-6, 1000 times.

Let t_α^* be the αth sample quantile of the 1000 t^*, then a $100(1-\alpha)\%$ bootstrap percentile-t confidence interval for β_1 is given by

$$(\hat{\beta}_1 - t^*_{1-\alpha/2}\widehat{SD}, \hat{\beta}_1 - t^*_{\alpha/2}\widehat{SD}). \tag{4.4}$$

Two different estimates of $\widehat{SD}$, the standard deviation of $\hat{\beta}_1$ were used in (4.4), namely, the KSM and the GO estimates. For more details about both the bootstrap and the jackknife see Efron and Tibshirani (1993).

4.2 Implementation of the Koul-Sievers-McKean and the George-Osborne estimators of τ

In this implementation of the KSM estimate, we used the recommendations in McKean and Sheather (1991, p. 10). Thus, δ was set to be 0.8, the sample standard deviation was chosen as the intial scale estimate and the resulting estimate of τ was multiplied by the bias correction $\sqrt{n/(n-2)}$.

For the GO estimator, the quadratic in (3.2) was fitted to $D(\beta)$ on 5 points. These 5 points form an equally spaced grid within a region over which **S** is approximately linear. See George (1993) for specific details.

4.3 Comparison of results

Figures 1 and 2 contain plots of the empirical confidence coefficients versus sample size for each of the following 5 confidence intervals:

a. The jackknife interval based on (4.3)

b. The asymptotic interval (4.2) based on the GO estimate of τ

c. The asymptotic interval (4.2) based on the KSM estimate of τ

d. The bootstrap-t interval (4.4) based on the GO estimate of τ

e. The bootstrap-t interval (4.4) based on the KSM estimate of τ

The nominal value of the confidence coefficient is marked on both plots with an unbroken line. Also marked on both plots with broken lines are values of the empirical confidence coefficient which are 1.96 standard errors above and below the nominal value. The performance of the bootstrap percentile-t intervals based on the Koul, Sievers and McKean standard error estimate is the most impressive with almost all the empirical confidence coefficients lying within 1.96 standard errors from the nominal value. Confidence intervals a, b and c, that is, those based on (4.2) and (4.3) typically over cover, while e, the bootstrap percentile-t intervals based on the George and Osborne standard error estimate often significantly under cover. The under coverage of the bootstrap-t interval based on the George and Osborne estimator may be due to the fact that there are ties in the bootstrap residuals.

Figures 3 and 4 contain plots of the ratios of the average length of confidence intervals a, b and c to the average length of the bootstrap percentile-t intervals based on the Koul, Sievers and McKean standard error estimate, that is, interval e. The bootstrap percentile-t intervals based on the George and Osborne standard error estimate were not included in the comparsion of interval lengths, since the empirical confidence coefficient of these intervals is consistently more than 1.96 standard errors below the nominal value in small samples. Again the performance of the Koul, Sievers and McKean bootstrap percentile-t intervals is most impressive, since it provided the shortest average interval length in almost every situation. On the other hand, the jackknife confidence intervals generally have the longest average lengths of all the intervals considered.

5 References

Aubuchon, J.C. and Hettmansperger, T.P. (1989). Rank-based inference for linear models: asymmetric errors. *Statist. Probab. Lett.*, **8**, 97-107.

Efron, B. and Tibshirani, R.J. (1993). *An Introduction ot the Bootstrap.* Chapman and Hall, New York.

George, K. (1993). An asymptotic method for estimating the variance of R-estimators in regression. Manuscript in preparation.

George, K. and Osborne, M. (1990). The efficient computation of linear rank statistics. *J. Stat. Comp. Simul.,* **35**, 227-237.

Hettmansperger, T.P. (1984). *Statistical Inference Based on Ranks.* Wiley, New York.

Jaeckel, L.A. (1972). Estimating regression coefficients by minimizing the dispersion of the residuals. *Ann. Math. Statist.,* **43**, 1449-1458.

Jureckova, J. (1969). Asymptotic linearity of a rank statistic in regression. *Ann. Math. Statist.,* **40**, 1889-1900.

Jureckova, J. (1971). Nonparametric estimate of regression coefficients. *Ann. Math. Statist.,* **42**, 1328-1338.

Koul, H., Sievers, G.L. and McKean, J.W. (1987). An estimator of the scale parameter for the rank analyses of linear models under general score functions. *Scand. J. Statist.,* **14**, 131-141.

McKean, J.W. and Hettmansperger, T.P. (1978). A robust analysis of the general linear model based on one-step R-estimates. *Biometrika*, **65**, 571-579.

McKean, J.W. and Sheather, S.J. (1991). Small sample properties of robust analyses of linear models based on R-estimates: a survey. *Directions in Robust Statistics and Diagnostics,*

Part II, Stahel, W. and Weisberg, S. (eds.), Springer-Verlag, New York, 1-20.

Mosteller, F. and Tukey, J.W. (1977). *Data Analysis and Regression.* Addison-Wesley, Reading, MA.

Quenouille, M.H. (1949). Approximate tests of correlation in time series. *J. Royal. Statist. Soc. B*, **11**, 68-84.

Schrader, R.M. and McKean, J.W. (1987). Small sample properties of least absolute values analysis of variance. *Statistical Analysis Based on the L_1-Norm and Related Methods,* Y. Dodge, North-Holland, Amsterdam, 307-321.

Schucany, W.R. and Sheather, S.J. (1989). Jackknifing R-estimates. *Biometrika*, **76**, 393-398.

Stine, R. (1989). An introduction to bootstrap methods. *Soc. Meth. R.*, **18**, 243-291.

Tukey, J.W., (1958). Bias and confidence in not-quite so large samples (abstract). *Ann. Math. Statist.,* **29**, 614.

Figure Legends

Figures 1 and 2: Plot of the empirical confidence coefficients of each of the five methods versus sample size for nominal 90% and 95% confidence intervals; intervals based on (4.3) are denoted by 'a', those based on (4.2) and the GO estimate are denoted by 'b', those based on (4.2) and the KSM estimate are denoted by 'c', those based on (4.4) and the GO estimate are denoted by 'd', and those based on (4.4) and the KSM estimate are denoted by 'e'.

Figures 3 and 4: Plot of the ratios of the average length of confidence intervals based on (4.2) and (4.3) to the average length of (4.4) based on the KSM estimate; intervals based on (4.3) are denoted by 'a', those based on (4.2) and the GO estimate are denoted by 'b', those based on (4.2) and the KSM estimate are denoted by 'c'. Ratios greater than 1 imply that the average length of the coded interval exceeds that for the bootstrap-t with the KSM estimate.

FIGURE 1.

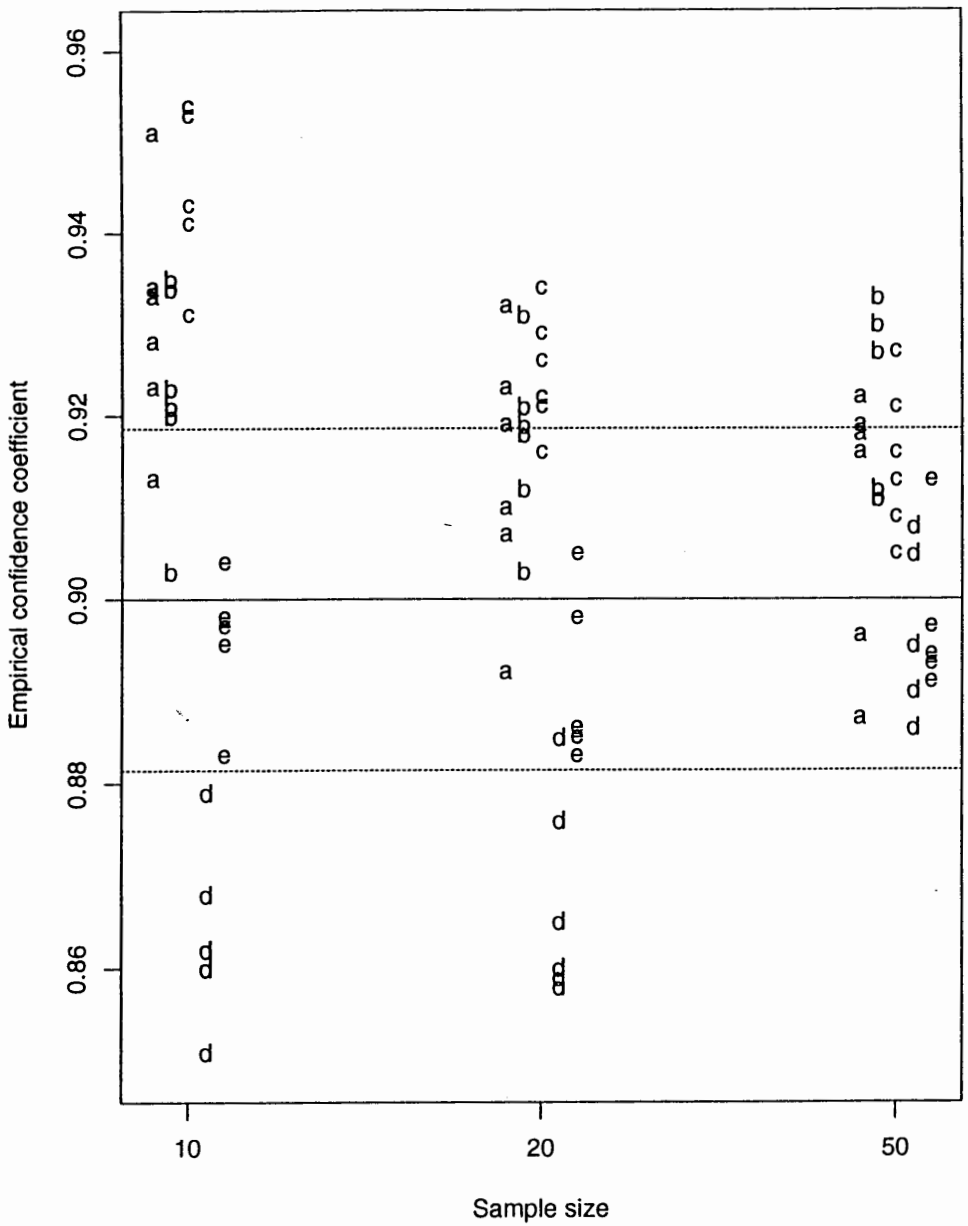

FIGURE 2.

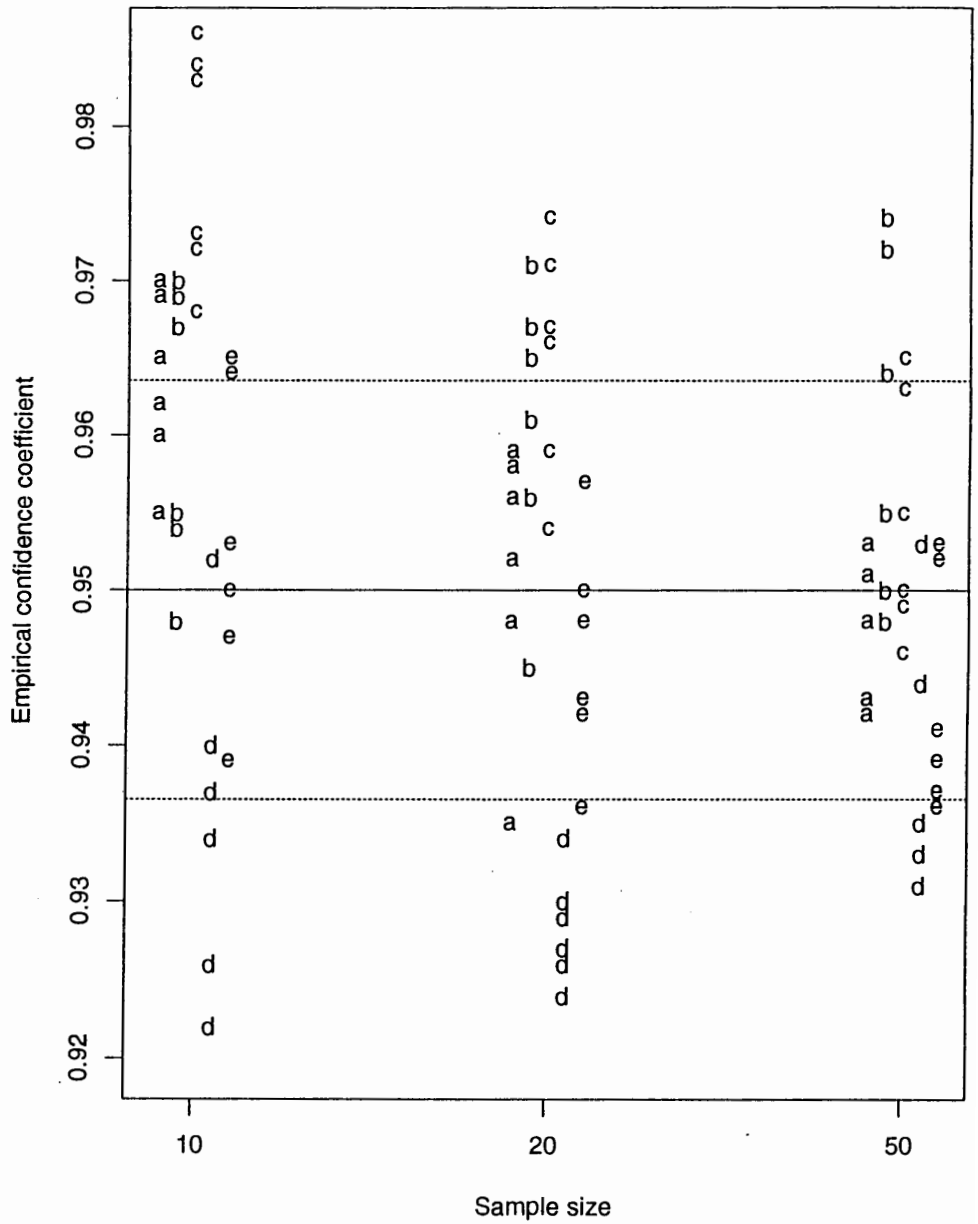

FIGURE 3.

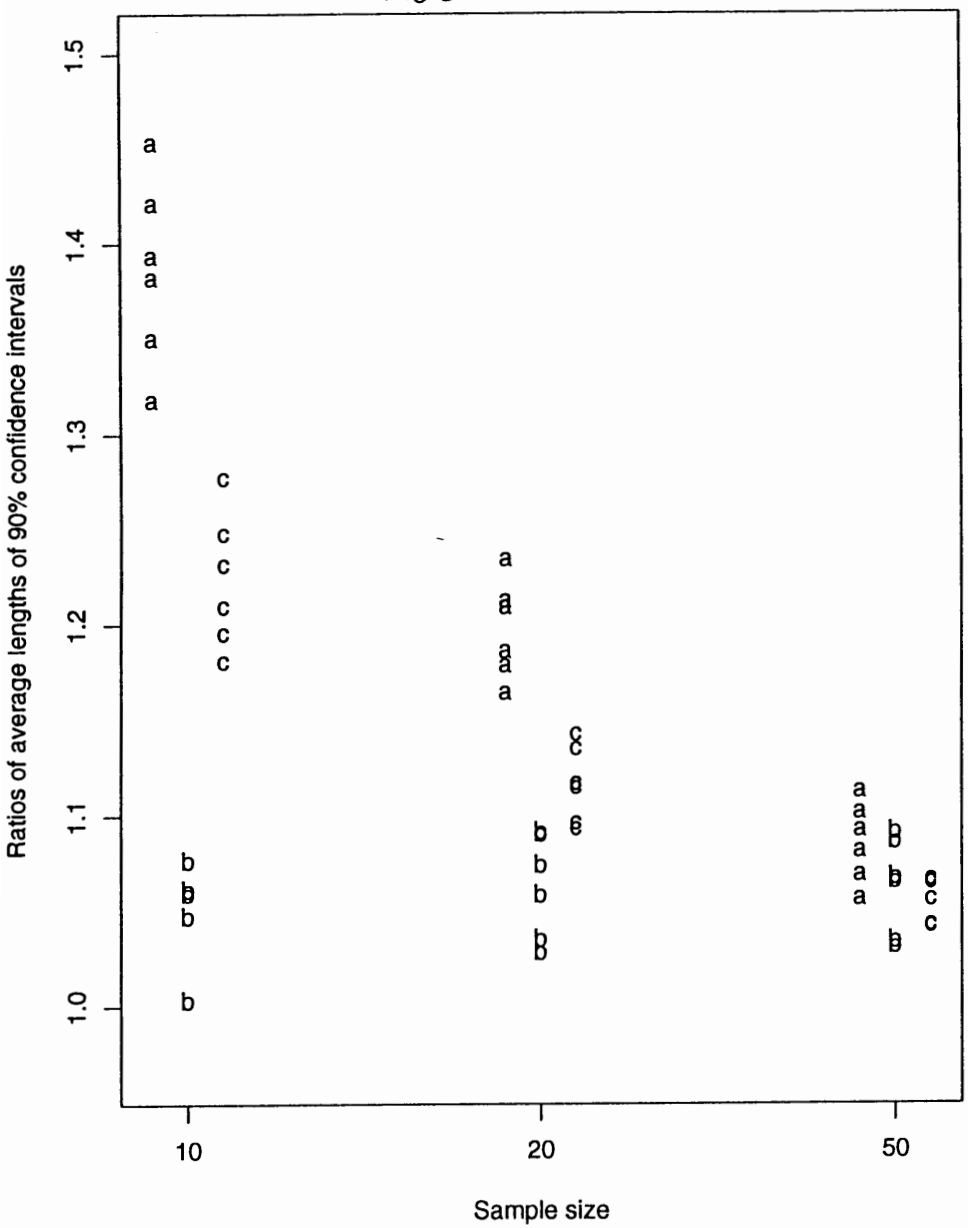

FIGURE 4.