Journal of Korean
Data & Information Science Society
2002, Vol. 13, No.2 pp 355 363

Bootstrap Confidence Intervals for Regression Coefficients under Censored Data

Kil Ho Cho[1] . Seong Hwa Jeong[2]

Abstract

Using the Buckley-James method, we construct bootstrap confidence intervals for the regression coefficients under the censored data. And we compare these confidence intervals in terms of the coverage probabilities and the expected confidence interval lengths through Monte Carlo simulation.

1. Introduction

In industrial life testing and medical follow-up studies, censoring is common because of time limits and other restrictions on data collection.

We consider the linear regression model

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \qquad (i = 1, \cdots, n)$$

where the responses Y_i are right censored, and the errors ε_i are independently and identically distributed(i.i.d.) random variables with mean zero and finite variance.

---

1. Professor, Department of Statistics, Kyungpook National University, Taegu, 702-701, Korea
E-mail : khcho@knu.ac.kr
2. Department of Statistics, Kyungpook National University, Taegu, 702-701, Korea

If Y_i is censored at time C_i, then the censored response Z_i can be written by

$$Z_i = Y_i\delta_i + C_i(1-\delta_i) = \min(Y_i, C_i) \quad (i = 1, \cdots, n),$$

where $C_i, \cdots, C_n$ are censoring values and δ_i are the indicator variables

$$\delta_i = \begin{cases} 1 & \text{if } Y_i \leq C_i \quad \text{(uncensored)} \\ 0 & \text{if } Y_i > C_i \quad \text{(censored)} \end{cases}$$

Miller(1976) considered the methods of estimating the regression coefficients of a linear regression model under type I censored data. Buckley and James(1979) suggested the method to accommodate censored observations to estimate regression coefficients.

Under uncensored data, the least squares estimators of β_0 and β_1 are the values a and b which satisfy the equations

$$\sum_{i=1}^{n}(Y_i - a - bX_i) = 0$$

and

$$\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - bX_i) = 0.$$

To accommodate censored observations, let

$$Y_i^* = Y_i\delta_i + E(Y_i \mid Y_i > C_i)(1-\delta_i) \quad i = 1, \cdots, n.$$

Then the estimator of β_1 by Buckley and James method is the value b which satisfies the equation $\sum_{i=1}^{n}(X_i - \overline{X})(Y_i^* - bX_i) = 0$. Since $E(Y_i \mid Y_i > C_i)$ is unknown, we adopt a self-consistency approach in Buckley and James(1979). Ritov(1990) and Currie(1996) investigated the asymptotic properties of the Buckley-James estimator.

The bootstrap method has been introduced by Efron(1979) who gave a series of examples to illustrate the validity of the approach for a wide class of statistics. The accuracy of the bootstrap approximation to the sampling distribution of various statistics has been investigated by Bickel and Freedman(1981). Efron(1981 and 1987) has developed methods for constructing approximate confidence intervals for a parameter. These intervals are constructed from the bootstrap distribution of an estimator.

In this paper, we consider the construction of bootstrap confidence intervals for the regression coefficients by the Buckley-James method. And we compare the coverage probabilities and the expected confidence interval lengths for these

confidence intervals through Monte Carlo simulation.

2. Bootstrap Confidence Intervals

The bootstrap method is a useful tool for constructing confidence interval or being applicable in situations where explicitly analytical solution cannot be found or the estimator is mathematically complicated. Also, the confidence intervals by bootstrap method do not require theoretical calculation.

Let $X = (X_1, X_2, \cdots, X_n)$ be the random sample from an unknown distribution F, and $\hat{\theta}(X_1, X_2, \cdots, X_n)$ be an estimator of θ. The bootstrap method is extremely simple in principle.

Step 1. Construct the sample probability distribution $\hat{F}$.

Step 2. With $\hat{F}$ fixed, draw a random sample of size n from $\hat{F}$, say

$$X_i^* \sim \hat{F}, \qquad i = 1, 2, \cdots, n.$$

Call this "the bootstrap sample",

$$X^* = (X_1^*, X_2^*, \cdots, X_n^*).$$

Step 3. Approximate the sampling distribution of $\hat{\theta}$ by the bootstrap distribution of

$$\hat{\theta}^* \equiv \hat{\theta}(X_1^*, X_2^*, \cdots, X_n^*).$$

We consider Buckley-James estimator for a linear regression model under censored data by bootstrap resampling method.

The residual, $\hat{\varepsilon}$, is

$$\hat{\varepsilon} = Y - X\hat{\beta} \equiv (\hat{\varepsilon}_1, \hat{\varepsilon}_2, \cdots, \hat{\varepsilon}_n).$$

Then the centered residuals, $\tilde{\varepsilon}_i$, are given by

$$\tilde{\varepsilon}_i = \hat{\varepsilon}_i - \frac{1}{n}\sum_{j=1}^{n} \hat{\varepsilon}_j, \qquad i = 1, 2, \cdots, n.$$

Let $\hat{F}_n$ be the empirical distribution of $\tilde{\varepsilon}_i$. That is, $\hat{F}_n$ puts mass $1/n$ at $\tilde{\varepsilon}_i$ and $\int x \, d\hat{F}_n = 0$. Given Y, let $\varepsilon_1^*, \cdots, \varepsilon_n^*$ be conditionally independent random variables with common distribution $\hat{F}_n$. And let ε^* be the n-dimensional vector whose i-th component is ε_i^*. We put

$$Y^* = X\hat{\beta} + \varepsilon^*.$$

Informally, ε^* is obtained by resampling the centered residuals. Now, we construct bootstrap sample (Z_i^*, δ_i^*, X_i), where

$$Z_i^* = \min(Y_i^*, C_i^*)$$

and

$$\delta_i^* = \begin{cases} 1, & \text{if } Y_i^* \le C_i^* \\ 0, & \text{if } Y_i^* > C_i^* \end{cases} .$$

Then the estimators of regression coefficients by Buckley-James method are given by

$$\hat{\beta}_1^* = \{\sum^u Y_i^* (X_i - \bar{X}) + \sum^c \bar{Y}_i^* (\hat{\beta}_1^*)(X_i - \bar{X})\} / \left\{\sum_{i=1}^n (X_i - \bar{X})^2\right\}$$

and

$$\hat{\beta}_0^* = n^{-1}\{\sum^u Y_i^* + \sum^c \bar{Y}_i^* (\hat{\beta}_1^*)\} - \hat{\beta}_1^* \bar{X},$$

where $\sum^u$ and $\sum^c$ denote summations over the uncensored and the censored values only, respectively.

We construct the bootstrap confidence intervals of the regression coefficients under censored data as the followings;

Step 1. Generate bootstrap samples $(Z_i^{*b}, \delta_i^{*b}, X_i^b)$, $i = 1, \cdots, n$, $b = 1, \cdots, B$.

Step 2. Calculate the bootstrap estimators $\hat{\beta}^{*b} \equiv (\hat{\beta}_0^{*b}, \hat{\beta}_1^{*b})$ corresponding to each bootstrap samples, $(Z_i^{*b}, \delta_i^{*b}, X_i^b)$, $i = 1, \cdots, n$, $b = 1, \cdots, B$.

Step 3. Construct the confidence intervals of regression coefficients by three bootstrap methods that are percentile, bias corrected(BC) percentile, and bias corrected and accelerated(BCa) percentile.

Percentile confidence interval is constructed by using the empirical distribution function of bootstrapping estimations. Let $\hat{G}$ be the empirical distribution function of the bootstrap distribution of $\hat{\beta}^*$. If the bootstrap distribution can be obtained by Monte Carlo method, then $\hat{G}$ is estimated from the bootstrap replications $\hat{\beta}^{*1}, \cdots, \hat{\beta}^{*B}$ as

$$\hat{G}(t) = \frac{1}{B}\sum_{b=1}^{B} I(\hat{\beta}^{*b} \le t)$$

where t is a real value and B is a replication number of bootstrap.

Let $\hat{G}^{-1}(\alpha)$ be the 100α percentile of $\hat{\beta}^{*}$,

$$\hat{G}^{-1}(\alpha) = \{ t : \hat{G}(t) \ge \alpha \}.$$

Therefore, $\hat{G}^{-1}(\alpha)$ is the $B\alpha$th value in the ordered list of $\hat{\beta}^{*b}$ and $\hat{G}^{-1}(1-\alpha)$ is the $B(1-\alpha)$th value in the ordered list of $\hat{\beta}^{*b}$. Then, $100(1-\alpha)\%$ percentile confidence interval of regression coefficient is

$$\left(\hat{G}^{-1}\left(\frac{\alpha}{2}\right), \hat{G}^{-1}\left(1-\frac{\alpha}{2}\right) \right)$$

Bias corrected confidence interval is the method to construct interval by correcting the bias of the bootstrap distribution. Bias correction is obtained by centering the standard normal cumulative distribution function(cdf). The bias correction $\hat{z}_0$ is given by

$$\hat{z}_0 = \Phi^{-1}(\hat{G}(\hat{\beta})) = \Phi^{-1}\left[\frac{1}{B}\sum_{b=1}^{B} I(\hat{\beta}^{*b} \le \hat{\beta})\right]$$

where Φ^{-1} is the inverse function of the standard normal cdf. Then $100(1-\alpha)\%$ bias corrected percentile confidence interval is

$$\left(\hat{G}^{-1}(\alpha_1), \hat{G}^{-1}(\alpha_2) \right)$$

where $\alpha_1 = \Phi(2\hat{z}_0 + z_{\alpha/2})$ and $\alpha_2 = \Phi(2\hat{z}_0 + z_{1-\alpha/2})$.

If $\hat{G}(\hat{\beta}) = 0.5$, that is, if half of the bootstrap distribution of $\hat{\beta}^{*}$ is less than the observed value $\hat{\beta}$, then $z_0 = 0$ and the percentile method and the BC method agree.

Bias corrected and accelerated(BCa) method has more satisfactory coverage probability than percentile method. We compute the bias correction $\hat{z}_0$ in BC method and acceleration $\hat{a}$ given by

$$\hat{a} = \frac{\sum_{i=1}^{n} (\hat{\beta}(\cdot) - \hat{\beta}(i))^3}{6 \left\{ \sum_{i=1}^{n} (\hat{\beta}(\cdot) - \hat{\beta}(i))^2 \right\}^{2/3}}$$

where

$$\hat{\beta}(\cdot) = \frac{1}{n} \sum_{i=1}^{n} \hat{\beta}(i),$$

and $\hat{\beta}(i)$ is the estimator $\hat{\beta}$ definded on the subsample which arises when the i th observation has been deleted. Then $100(1-\alpha)\%$ BCa confidence interval of regression coefficient is

$$\left(\hat{G}^{-1}(\alpha_3), \quad \hat{G}^{-1}(\alpha_4) \right)$$

where

$$\alpha_3 = \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z_{\alpha/2}}{1 - \hat{a}(\hat{z}_0 + z_{\alpha/2})} \right)$$

and

$$\alpha_4 = \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z_{1-\alpha/2}}{1 - \hat{a}(\hat{z}_0 + z_{1-\alpha/2})} \right)$$

If $\hat{z}_0$ and $\hat{a}$ are zero then $\alpha_3 = z_{\alpha/2}$ and $\alpha_4 = z_{1-\alpha/2}$. Thus the percentile method and the BCa method agree.

3. Example and Simulation Study

As the data for the example, we use Stanford heart transplant data which was given by Miller(1976). The date for the start of the transplantation program at Stanford was on October 1 in 1967, and the cut-off date for this analysis was on April 1 in 1974. During this time interval, 69 patients received heart transplants, and the lengths of their survival, in days, after transplantation are considered as response variables. And ages of patients at transplant are considered as an explanatory variable. In accordance with Buckley-James method we take the response variables to log survival times to the base 10.

Now, we compute the bootstrap regression coefficients for heart transplant data

through the 1,000 bootstrap replications, and construct confidence intervals of them by the percentile, BC and BCa methods. The results are listed in Table 3.1.

Table 3.1 Regression coefficients and confidence intervals

Regression coefficients	$\hat{\beta}_0 = 3.5962$	$\hat{\beta}_1 = -0.0278$
Percentile interval	(1.1575 , 4.9726)	(-0.0761 , -0.0009)
BC interval	(2.1690 , 5.6167)	(-0.0635 , -0.0025)
BCa interval	(2.1032 , 5.2167)	(-0.0635 , -0.0025)

Also, we investigate the bootstrap estimates and the bootstrap confidence intervals through the Monte Carlo simulation. We assume that as a lifetime distribution, Weibull distribution(Weib) with parameters $\alpha = 1.0$ and $\beta = 0.8$ (decreasing failure rate), $\beta = 1.0$ (constant failure rate) and $\beta = 1.5$(increasing failure rate) are considered. And as censoring distribution, exponential(Exp) distribution with parameters which the censoring rates are approximately 10% and 30%, respectively, are considered. The Monte Carlo simulations are performed for all combinations of lifetime distributions and censoring distributions with sample sizes 40 and 60. The simulation experiment is performed 1,000 replications and 1,000 bootstrap replications for each case. The criteria used to compare the confidence intervals are coverage probability and the expected length of confidence interval. We use 0.90 as a nominal probability. The results of these simulations are given in Table 3.2. We do not list the results for sample size 60 because of similarity as sample size 40. From this table, we can observe the following facts;

(1) The coverage probabilities of bootstrap confidence intervals are nearly nominal coverage 0.90.

(2) The coverage probabilities of all the confidence intervals decrease as censoring rate increases.

(3) The coverage probabilities of all the confidence intervals increase as n increases.

(4) BCa method among the bootstrap confidence intervals has the shortest expected confidence length.

Therefore, we can know that the bootstrap method is very effective for confidence intervals of regression coefficients under the censored data.

Table 3.2 Bootstrap confidence intervals

	Methods	Censoring rate	10%		30%	
			β_0	β_1	β_0	β_1
Weib(1.0, 0.8)	Percentile	Length	16.941	0.358	9.414	0.393
		Coverage	0.883	0.887	0.869	0.883
	BC	Length	13.081	0.355	10.180	0.362
		Coverage	0.891	0.897	0.880	0.894
	BCa	Length	11.197	0.337	9.341	0.351
		Coverage	0.900	0.901	0.900	0.906

	Methods	Censoring rate	10%		30%	
			β_0	β_1	β_0	β_1
Weib(1.0, 1.0)	Percentile	Length	15.974	0.351	9.638	0.217
		Coverage	0.879	0.888	0.873	0.887
	BC	Length	15.537	0.395	12.417	0.228
		Coverage	0.899	0.897	0.897	0.895
	BCa	Length	10.192	0.365	9.340	0.161
		Coverage	0.901	0.901	0.899	0.902

	Methods	Censoring rate	10%		30%	
			β_0	β_1	β_0	β_1
Weib(1.0, 1.5)	Percentile	Length	15.968	0.351	11.794	0.335
		Coverage	0.891	0.894	0.880	0.890
	BC	Length	13.614	0.338	14.301	0.341
		Coverage	0.894	0.895	0.892	0.897
	BCa	Length	13.964	0.337	10.162	0.306
		Coverage	0.895	0.901	0.896	0.902

References

1. Bickel, P. and Freedman, D. (1981). Some asymptotic theory for the bootstrap. *Annals of Statistics*, 9, 1196-1217
2. Buckley, J. and James, I. R. (1979). Linear regression with censored data. *Biometrica*, 66, 429-436.
3. Currie, I. D. (1996). A note on Buckley-James estimators for censored data. *Biometricka*, 83, 912-915
4. Efron, B. (1979). Bootstrap methods; Another look at the jackknife. *Annals of Statistics*, 7, 1-26.
5. Efron, B. (1981). Nonparametric standard errors and confidence interval. *Canadian Journal of Statistics*, 9, 139-172.
6. Efron, B. (1987). Better bootstrap confidence interval, *Journal of the American Statistical Association*, 82, 171-185.
7. Freedman, D. (1981). Bootstrapping regression model. *Annals of Statistics*, 9, 1218-1228.
8. Miller, R. G. (1976), Least squares regression with censored data, *Biometrika*, 63, 449-464.
9. Ritov, Y. (1990), Estimation in a linear regression model with censored data. *Annals of Statistics*, 18, 303-328.

[2002년 9월 접수, 2002년 10월 채택]