Discussion Paper

CHUNG-HUA INSTITUTION FOR ECONOMIC RESEARCH
75 Chang Hsing St., Taipei, Taiwan
Republic of China

Discussion Paper

CHUNG-HUA INSTITUTION FOR ECONOMIC RESEARCH
75 Chang Hsing St., Taipei, Taiwan
Republic of China

Discussion Paper

No: 8405

The Bootstrap and the Censored Regression

CHIHWA KAO

Chung-Hua Institution for Economic Research

This paper proposes a bootstrap estimate for the standard error of the Buckley-James estimator in a censored regression and when the error distribution is unknown.

The Bootstrap and the Censored Regression

CHIHWA KAO*

Abstract

This paper proposes a bootstrap estimate for the standard error of the Buckley-James estimator in a censored regression and when the error distribution is unknown.

Key Words: EM algorithm; Buckley and James estimator; Kaplan and Meier Estimator.

* The author is Associate Research Fellow,
Chung-Hua Institution for Economic Research,
P. O. Box: 36-306,
Taipei, Taiwan, R.O.C.

I) Introduction

This paper is concerned with the estimation of the standard error for the Buckley & James (BJ) estimator (1979) in a linear model when the data is randomly right censored and the error distribution is unknown. The BJ estimator is a non-parametric version of the EM algorithm introduced by Dempster, Laird and Rubin (1977). It is known that there is no firm theoretical fundation for the estimated standard error of the BJ estimator BJ proposed, although BJ did provide some idea of the large sample behavior through Monte Carlo simulations. In this paper, I propose a bootstrap (Efron 1979a, 1979b, 1981a, 1981b) estimate for the standard error of the BJ estimator. The bootstrap is a technique for estimating standard errors. The first step is to estimate the residuals. The second step is to resample the residuals from the empirical distribution which is obtained by the Kaplan-Meier estimate (1958) to generate pseudo-data, and the model can be refitted to the pseudo-data to obtain the new estimates of the standard error of the BJ estimator.

Consider the linear regression

$$T_i = \beta' X_i + \varepsilon_i,\ i = 1, \ldots , n,$$

where β and X_i are k by 1 and ε_i are independent with common distribution F. Assume that F has mean zero and variance σ^2. One cannot observe $T_i, \ldots , T_n$ but

$$y_i = \min[T_i, C_i]$$

and

$$\delta_i = I[T_i \leq C_i],$$

Where I[·] denotes the indicator function and C_i are the censored variables. The problem considered above is to estimate β based on $(y_1, \delta_1), \ldots, (y_n, \delta_n)$.

The description of the BJ estimator below is adapted from Miller and Halpern (1982). BJ (1979) proposed an estimate of β, herein called the BJ estimator, based on an expectation identity:

$$E\{y_i\delta_i + E(T_i|T_i>y_i)(1-\delta_i)\} = \beta' X_i.$$

BJ substitute an estimate for the conditional expectation $E(T_i|T_i>y_i)$ based on the Kaplan & Meier estimator (1958) into the variable

$$y_i = y_i\delta_i + E(T_i|T_i>y_i)(1-\delta_i)$$

and solve the least squares normal equations iteratively.

If $\delta_i = 1$, let $\hat{y}_i = y_i$, but if $\delta_i = 0$, let

$$\hat{y}_i = \hat{\beta}'X_i + \frac{\sum_{\hat{\varepsilon}_k>\hat{\varepsilon}_i} \hat{w}_k\hat{\varepsilon}_k}{1- \hat{F}(\hat{\varepsilon}_i)},$$

where $\hat{\varepsilon}_i = y_i - \hat{\beta}'X_i$, $\hat{F}$ is the Kaplan & Meier estimator based on the δ_i, i.e.,

$$\hat{F}(\varepsilon) = 1- \prod_{\hat{\varepsilon}(i)\leqq\varepsilon} \left(\frac{n-i}{n-i+1}\right)^{\delta_i},$$

and the weights $\hat{w}_k$ are the size of the jump assigned to ε_i by the $\hat{F}$.

Then the BJ estimator of β at the (k+1)st step, $\hat{\beta}_{k+1}$, is the least squares estimator

$$\hat{\beta}_{k+1} = (\sum_i X_i\hat{y}_i)(\sum_i X_iX_i')^{-1}.$$

The iteration is continued until $\hat{\beta}_k$ converges to a limiting value $\tilde{\beta}$.

BJ suggested the estimate of the covariance matrix of $\tilde{\beta}$ to be

$$\hat{\sigma}_\mu^2 (\sum_i \delta_i X_i X_i')^{-1}, \tag{1}$$

where

$$\hat{\sigma}_\mu^2 = \frac{1}{n_u - 2} \sum_i \delta_i (\hat{\varepsilon}_i - \frac{1}{n_u} \sum_j \delta_j \hat{\varepsilon}_j)^2,$$

and n_u is the number of uncensored observations.

BJ provided some idea of the large sample behavior of (1) through Monte Carlo simulations, but as we mentioned early, there is no theoretical justification for (1).

II) The Bootstrap

We consider a bootstrap method of estimating the standard errors of the BJ estimator. The procedures of the bootstrap can be stated as follows:

1) Let $\hat{F}$ be the Kaplan and Meier estimator based on the δ_i, i.e.,

$$\hat{F}(\varepsilon) = 1 - \prod_{e_i \leqq \varepsilon} (\frac{n-i}{n-i+1})^{\delta_i},$$

where $e_i = y_i - \beta_{BJ}' X_i$, and β_{BJ} is the BJ estimator. That is, the probability distribution which puts mass $1/n$ at each observed point (e_i, δ_i).

2) Use a random number generator to draw n new points (e_i^*, δ_i^*) independently and with replacement from $\hat{F}$, so that each new point is an independent random selection of one of the n original

data points. These new points, which we call the bootstrap sample are a subset of the original points (e_i, δ_i). Some of the original points will have been selected zero times, some once, some twice, etc.

3) Compute β^*_{BJ}, the BJ estimator for the bootstrap sample.

4) Repeat steps (2) and (3) a large number of times, each time using an independent set of new random numbers to generate the new bootstrap sample. Call the resulting sequence of the bootstrap BJ estimators $\beta^{*1}_{BJ}, \beta^{*2}_{BJ}, \cdots, \beta^{*n}_{BJ}$.

5) Let the bootstrap standard error of the BJ estimator be $\hat{\sigma}_{Boot}$. The value of $\hat{\sigma}_{Boot}$ is approximated, by the sample standard error of the β^*_{BJ} values,

$$\hat{\sigma}_{Boot} = \sqrt{\frac{\sum_j (\beta^{*j}_{BJ} - \beta^{*\cdot}_{BJ})^2}{n-1}},$$

where $\beta^{*\cdot}_{BJ} = (\sum_j \beta^{*j}_{BJ})/n$.

We expect the bootstrap estimate, $\hat{\sigma}_{Boot}$, performs better than (1), since the bootstrap estimate is the nonparametric maximum likelihood estimate of the standard error. That is, the bootstrap estimate, $\hat{\sigma}_{Boot}$ is simply the standard deviation of the quantity of interest β_{BJ}, if the unknown distribution F is taken equal to the Kaplan and Meier estimate $\hat{F}$. (Efron 1981b) We know that the Kaplan and Meier $\hat{F}$ is the nonparametric maximum likelihood estimate of the unknown distribution F (Kaplan and Meier 1958), i.e., $\hat{F}$ has most of properties of the maximum likelihood estimate: Consistency, asymptotic normality and asymptotic efficiency.

III) Conclusion

This paper proposes an alternative method, called the bootstrap, of estimating the standard error of the BJ estimator. I believe that this paper contributes to the studies of the estimation of censored regression models.

References

Buckley, J. and James, I. (1979), "Linear Regression with Censored Data," Biometrika, 66, 429-436.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum Likelihood from Incomplete Data Via the EM Algorithm," Journal of the Royal Statistical Society, Ser. B, 39, 1-38.

Efron, B. (1979a), "Bootstrap Methods: Another Looks at the Jackknife," Annals of Statistics, 7, 1-26.

__________ (1979b), "Computers and the Theory of Statistics: Thinking the Unthinkable," SIAM Review, 21, 460-480.

__________ (1981a), "Censored Data and the Bootstrap," Journal of the American Statistical Association, 76, 312-319.

__________ (1981b), "Nonparametric Estimates of Standard Error: The Jackknife, the Bootstrap and Other Methods," Biometrika, 68, 589-599.

Kaplan, E. L. and Meier, P. (1958), "Nonparametric Estimation from Incomplete Observations," Journal of the American Statistical Association, 53, 457-481.

Miller, R. and Halpern, J. (1982), "Regression with Censored Data," Biometrika, 69, 521-531.