Regression model with left truncated data

School of Mathematics  Sang Hui-yan  00001107

October 13, 2003

Abstract

For linear regression model with truncated data, Bhattacharya, Chernoff, and Yang(1983)(BCY) utilize Kendall’s $\tau$ and weighted medians to estimate $\beta_1$. Tsui, Jewell and Wu(1988)(TJW) proposed an alternative approach to estimating $\beta$ in the general multiple linear regression model. He and Yang proposed a non-iterative non parametric procedure for linear regression model with left-truncated data. This is important since both BCY and TJW use iterative procedures and requires a simultaneous estimation of the error distribution $F_\epsilon$ and $\beta$ at each step of the iteration. The main goal of this paper is to consider the regression method of He and Yang, and investigate the significance of the simulation experiments and explain in details of the simulation result.

1 Introduction

Consider the regression model

$$Y = \beta' Z + \epsilon$$

(1)

Where $Z$ is a random vector of covariates, $\beta$ is vetor of parameters interests and the error is uncorrelated with $Z$, mean 0 and cumulative distribution function $F_\epsilon$. Under random truncation the observation of $(Y, Z)$ is interfered by another independent random variable $T$ such that all three random quantities $Y, Z$ and $T$
are observable only if $Y \geq T$. Imaging in the UN-truncated situation, observable would have been a sequence of independent random vectors\{$(Y_k, Z_k, T_k, \ k = 1, 2, ...)$\}.

Under random truncation, however, some of these vectors would be missing and only a subsequence \{$(Y_{ki}, Z_{ki}, T_{ki}), i = 1, 2, ...$\} could be observed. This subsequence constitutes our data base for investigation. For convenience in exposition, the observable subsequence will be denoted by

\[(U_i, V_i, W_i), i = 1, 2, ... \quad \text{subject to} \quad U_i \geq V_i \] (2)
i.e. $U_i = Y_{ki}, V_i = T_{ki}, W_i = Z_{ki}$

our problem is to use data (2) to estimate the regression parameter $\beta$ and the error distribution function $F_\varepsilon$. Under model (1.1), the regression function is the conditional expectation $E(Y|Z = z) = \beta^T Z$, which is the basis for the ordinary least square estimation (OLS) with complete data. However, with truncated data. $E(Y|Z = z, Y \geq T)$ is no longer equal to $\beta^T Z$. Thus the OLS would not produce the right estimates for $\beta$ and $F_\varepsilon$. Truncation effects need would be appropriately treated.

Truncation occurs frequently in medical studies when one wants to study the length of survival after the start of the disease: If $Y$ denotes the time elapsed between the onset of the disease and death, and if the follow-up period starts $T$ units of time after the onset of the disease, then clearly $Y$ is left truncated by $T$. Truncated data are also common in actuarial, astronomic, demographic, epidemiologic, reliability testing, and other studies. (More examples and references of truncated data can be found in Andersen, Borgan, Gill, and Keiding 1993, sec. III.3; Efron and Petrosian 1994; Wang 1989; and Woodroofe 1985.)

Large data sets afford non parametric studies of the problem and free from the parametric constraints. The first non parametric estimator of the luminosity function was constructed by Lyden-Bell(1971) and studied systematically by Woodroofe(1985). This estimator will be one of the basic quantities in our study of the regression model. For simple regression, BCY utilize Kendall’s $\tau$ and weighted medians to estimate $\beta_1, F_\varepsilon$ and the error variance $\sigma_{\varepsilon^2}$ in model(1.1), in which the truncation variable $T$ need not be a constant.

The main goal of this paper is to introduce He’s method to evaluate the estimators in the regression model with lefted truncated data. He proves that under very weak conditions the estimates of $\beta, F_\varepsilon$ and $\sigma_{\varepsilon^2}$ are strongly consistent and
asymptotic normal. The derived asymptotic variance has a very simple form in contrast with the usual complicated expressions associated with random truncation and censoring models. Our results hold for arbitrary distributions of $Y$ and $T$ without the usual continuity restriction. It is also interesting to point out that our regression estimates reduce to the ordinary least square estimates if no truncation is present in the data. This is not the case in the iterative procedures of BCY and TJW. We consider the modelling of random data from groups subject to truncation. In this paper, we give the estimates of $\beta$, $F_\varepsilon$, and $\sigma_\varepsilon^2$. And show that under very weak conditions the estimates of $\beta$, $F_\varepsilon$, and $\sigma_\varepsilon^2$ are strongly consistent and asymptotic normal.

After the introduction, the chapter is organized as follows.

In section 2, after the introduction of assumption, we show that is indeed a sequence of iid random vectors with a common distribution given in Lemma 1.1. The iid property is implicitly assumed in the literature, but it does not seem to be completely obvious. Then we give the nonparametric maximum likelihood estimators of $F(x)$ and $G(x)$, and we establish the asymptotic normality of the nonparametric estimates.

In section 3 on the basis of Lemma 1.1, we consider the weighted least square estimates for $\beta$, $F_\varepsilon$ and $\sigma_\varepsilon^2$. In section 2, we use He’s method to do the evaluation. We use a non-procedure to construct weighted least square estimates for $\beta$. The approach leads to relatively simple estimators. A consistent estimator for the error variance $\sigma_\varepsilon^2$ and the error distribution $F_\varepsilon$ are also provided. The conditions used in the section 2 are only the finiteness of $EY^2$ and $E(Z'Z)$ to ensure finite variances of the estimates. In this section, we give some illustrative examples to investigate the value of the estimators and their consistency.

In Section 4, we discuss the consistency and CLT of the estimators. He proved under very weak conditions a general theorem on the weak convergence to normality of the estimator of $\beta$ constructed in section 1 is derived. The asymptotic variance formula is surprisingly simple, and we did not encounter the usual complications in the computation of asymptotic variance in the truncation model.
2 Estimation of estimable functionals of distribution based on left truncated data

To model left-truncated data, we begin by modelling the complete data and then introduce the truncation mechanisms. We thus let $Y_1, Y_2, \ldots$ be independent random variables with a common distribution function $F$. These are only partially observed due to the presence of truncated variables ($T_j$), which are assumed to be independent random vectors independent of the $Y_j$. Imaging in the untruncated situation, observable would have been a sequence of independent random vectors $(Y_k, Z_k, T_k, \ k = 1, 2, \ldots)$. Under random truncation, however, some of these vectors would be missing and only a subsequence $(Y_{ki}, Z_{ki}, T_{ki}), \ i = 1, 2, \ldots$ could be observed. This subsequence constitutes our data base for investigation.

For convenience in exposition, the observable subsequence will be denoted by $(U_i, V_i, W_i), \ i = 1, 2, \ldots$ subject to $U_i \geq V_i$ i.e. $U_i = Y_{ki}, V_i = T_{ki}, W_i = Z_{ki}$ We generate For any real monotone function $g(x)$, let $g(x-)\) denote the left continuous version of $g(x)$ and let the curly brackets $gx$ denote the difference $g(x) - g(x-)$. For any two vectors $a = (a_1, a_2, \ldots, a_r)'$, $b = (b_1, b_2, \ldots, b_r)'$, the inequality $a \leq b$ means that $a_i \leq b_i, \ i = 1, 2, \ldots, r$. let

$$F(x) = P(Y \leq x), G(x) = P(T \leq x), \text{ and } F(x, z) = P(Y \leq x, Z \leq z). \quad (3)$$

For any cumulative distribution function $F$, let $\bar{F} = 1 - F$. Random truncation restricts the observation range of $Y$ and $T$ which we shall first specify. Let $(a_F, b_F)$ be the range of $Y$ or $F$ defined by

$$a_F = \inf \{x : F(x) > 0\} \quad b_F = \sup \{x : F_x < 1\}$$

. Let $a_G$ and $b_G$ be simply defined for the distribution $G$. Under random truncation, only the conditional distributions

$$F_0(x) = P[Y \leq x|Y \geq a_G] \quad \text{and} \quad G_0(x) = P[T \leq x|T \leq b_F]$$

can be estimated non parametrically.

Assuming that the $T_j$ are identically distributed, the probability

$$F(x) = P(Y \leq x), G(x) = P(T \leq x), \text{ and } F(x, z) = P(Y \leq x, Z \leq z).$$

plays a central role in the development of our estimates. Note that $F(Y_j)$ is the conditional probability of $Y_j$ being completely observed (i.e., untruncated and
uncensored) given its value; that is, \( \hat{F}(Y_j) = P(Y \leq Y_j | Y_j) \). The reciprocal of an estimate of this probability will serve as a weight from which we construct estimates of functionals of the distribution \( F(t) = P( Y \geq T ) \) in this section and of regression coefficients in Section 3.

Let \( \alpha = P(Y \geq T) \), so that \( 1 - \alpha \) denotes the truncation rate. In applications, \( T \) is typically not known in advance and some \( T \) is selected so that the size of the observed risk set at \( \alpha \) is not too small.

**Assumption:** We shall assume through the paper that \( a_G \leq a_F \). Under this condition \( F_0 = F \). Otherwise \( \beta \) will not be identifiable by the model (1.1) A counter example is provided at the end of Section 3. We also assume that

\[
\alpha = P[Y \geq T] > 0.
\]

Otherwise, no data can be observed and the problem becomes trivial.

The iid property is implicitly assumed in the literature, but it does not seem to be completely obvious. He proves that the i.i.d property of the observations is preserved under random truncation.

**Lemma 1.1** the random vectors \((U_i, V_i, W_i)\) in (2) are i.i.d. with common distribution

\[
F^*(u, v, w) = P[U_1 \leq u, V_1 \leq v, W_1 \leq w] = P(Y \leq u, T \leq v, Z \leq w | Y \geq T) = \frac{1}{\alpha} \int_{a_G \leq x \leq u} \int_{z \leq w} G(x \wedge v) F(dx, dz)
\]

where \( x \wedge v = \min(x, v) \).

**Proof.** Let the symbol \( \Sigma^* \) stand for

\[
\sum^* = \sum_{m_1=1}^{\infty} \sum_{m_2=m_1+1}^{\infty} \cdots \sum_{m_n=m_{n-1}+1}^{\infty}
\]
Then the joint distribution of (1.2) can be derived as follows.

\[
P[(U_i, V_i, W_i) \leq (u_i, v_i, w_i); \ 1 \leq i \leq n]
\]

\[
= \sum_{i=1}^{n} P\{(Y_{m_i}, T_{m_i}, Z_{m_i}) \leq (u_i, v_i, w_i), Y_{m_i} \geq T_{m_i}; 1 \leq i \leq n\} \cap \{Y_j < T_j; j \neq m_i, 1 \leq j < m_n\}
\]

\[
= \sum_{i=1}^{n} (1 - \alpha)^{m_1-1}(1 - \alpha)^{m_2-m_1-1} \ldots (1 - \alpha)^{m_n-m_{n-1}-1} \prod_{i=1}^{n} \alpha P(Y_i \leq u_i, T_i \leq v_i, Z_i \leq w_i | Y_i \geq T_i)
\]

\[
= \prod_{i=1}^{n} P(Y_i \leq u_i, T_i \leq v_i, Z_i \leq w_i | Y_i \geq T_i)
\]

Hence, the random vectors in (2) are iid with the common distribution stated in the lemma.

It follows immediately from Lemma 1.1 that the marginal distributions of \(U\) and \(V\) are

\[
F^*(x) = P[U \leq x] = \frac{1}{\alpha} \int_{-\infty}^{x} G(s) dF(s)
\]

\[
G^*(x) = P[V \leq x] = \frac{1}{\alpha} \int_{-\infty}^{x} \bar{F}(s-) dG(s) \ (3)
\]

As a convention, we denote by a superscript * any distribution function that is associate with the truncated random variables. Here and after, the integral sign \(\int_{a}^{b}\) stands for \(\int_{(a,b)}\). The integral sign without limits refers to the integral from \(-\infty\) to \(+\infty\).

To construct estimates for \(\beta\) and \(F_{\epsilon}\), we shall need the following empirical distribution function.

\[
F^*_n(u) = \frac{1}{n} \sum_{j=1}^{n} I[U_j \leq u]
\]

\[
G^*_n(v) = \frac{1}{n} \sum_{j=1}^{n} I[v_j \leq v]
\]

\[
F^*_n(u, w) = \frac{1}{n} \sum_{j=1}^{n} I[U_j \leq u, W_j \leq w_j]
\]

\[
R^*_n(s) = \frac{1}{n} \sum_{j=1}^{n} I[V_j \leq s \leq U_j] = G^*_n(s) - F^*_n(s-)
\]
Where \( I[A] \) denotes the indicator function of the event \( A \).

Our estimation problem requires first the estimation of \( \alpha, F(x), G(x) \) and the joint distribution function \( F(y, z) \) given by (2.1).

Optional non parametric estimators for \( F(x) \) and \( G(x) \) are the well known product-limit estimates given respectively by

\[
F_n(x) = 1 - \prod_{u_i \leq x} [1 - \frac{r(u_i)}{R_n(u_i)}] \quad \text{and} \quad G_n(u_i) = 1 - \prod_{v_i > x} [1 - \frac{s(v_i)}{R_n(v_i)}] \quad (2.5)
\]

where \( r(u_i) = \sharp\{k \leq n : u_k = u_i\} \) for \( 1 \leq i \leq n \). And \( s(v_i) = \sharp\{k \leq n : x_k = v_i\} \) for \( 1 \leq i \leq n \). where an empty product is set equal to one.

It follows by (2.5) that the cumulative hazard function of \( F \), defined by

\[
\Lambda(x) = \int_{-\infty}^{x} \frac{dF(u)}{1 - F(u)}
\]

can be estimate by

\[
\Lambda_n(x) = \int_{-\infty}^{x} \frac{dF_n(u)}{1 - F_n(u)} = \frac{dF^*_n(u)}{R_n(u)} \quad (2.6)
\]

For \( \alpha \) we use estimate

\[
\alpha_n = \frac{G_n(x)\bar{F}_n(x-)}{R_n(x)} \quad (2.7)
\]

for any \( x \) such that \( R_n(x) > 0 \)

It is shown in He and Yang(1998b) that \( \alpha_n \) is equivalent to the more familiar estimate \( \int G_n(x)dF_n(x) \) studied in the literature. Clearly \( \alpha_n \) has a much simpler form. The fact that \( \alpha_n \) is independent of \( x \) is instrumental for an easy construction of estimates of \( F(y, z) \) and \( \beta \) as to be carried out below.

By corollary of He and Yang(1998b), we have

\[
\alpha_n \to \alpha_0 \equiv \int G_0(x)dF(x) = \alpha/G(b_F) \quad (2.8)
\]

a.s., as \( n \to \infty \)

where \( G_0(x) \) is given by (1.2). Obviously, we have

\[
\alpha/G(x) = \alpha/G_0(x), \quad \text{for all} \quad x \in (a_G, b_F) \quad (2.9)
\]
It is differential form is

\[ F(dy, dz) = \frac{1}{G(y)F^*(dx, dz)} , \quad y \geq a_G \]  \hspace{1cm} (2.10)

This gives an estimation equation for \( F(y, z) \). Replacing \( F^*, G, \alpha \), by the corresponding estimates given in (2.4),(2.5) and (2.7), yields an estimator for \( F(y, z) \) as

\[ F_n(y, z) = \alpha \int_{u \leq y} \int_{w \leq z} \frac{1}{G_n(u)} dF^*_n(u, w) \]  \hspace{1cm} (2.11)

This estimator is in agreement with the product-limit estimate \( F_n(y) \) given by (2.5) in the sense that its marginal distribution \( F_n(y, \infty) = F_n(y) \). To see this, we evaluate

\[
F_n(y, \infty) = \alpha_n \int_{u \leq y} \frac{1}{G_n(u)} dF^*_n(u) \\
= \int_{u \leq y} G_n(u) \frac{F(u)}{G_n(u)} dF^*_n(u) \\
= \int_{u \leq y} \bar{F}_n(u^-) d\wedge_n(u) = F_n(y)
\]

The second equality follows from (2.7) and the last one from (2.6)

3 Weighted Estimators of the Regression Model

BCY uses the simple linear regression model

\[ Y = \beta_0 + \beta_1 Z + \varepsilon \]

to estimate the slope \( \beta_1 \) where the truncating variable \( T \) is assumed to be a known constant, say \( t_0 \). The slope \( \beta_1 \) corresponds to the Hubble constant in the study of luminosity function of galaxies. The variables \( Y \) and \( Z \) represents respectively measurements of the apparent luminosity and the red shift of a galaxy. The selection of bias known Malmquist(1920) or Scott(1957) due to their pioneer work on the subject, see Peeble(1993). Both Malmquist and Betty Scott studied the problem with about two dozens of observations of \((Y, Z)\) and under normality assumption. The sparse data back then is in sharp contrast with the rich survey data in the order of tens of thousands that is now available. The new evidence shows that parametric models like normal distribution and some others do not fit the data adequately in many situations and are commonly open to question, see, e.g., Nicoll and Segal (1980)
In this section, we will construct the estimates for $\beta$ by utilizing a new method. Notice certain moment relations. Let 
\[
Z_0 = Y, Z = (Z_1, Z_2, \ldots, Z_r)' \quad \text{and} \quad \mu_{ij} = E(Z_i Z_j), \quad 0 \leq i, j \leq r
\]
Let $\tau$ denote the matrix $(\mu_{ij})_{i,j=1}^n$. Put $\gamma = (\mu_{01}, \ldots, \mu_{0r})'$. Multiplying the regression model (1) by $Z'$ and taking expectation yields the following equation.
\[
\gamma = \tau \beta \quad (12)
\]
Components of $\gamma$ and $\tau$ can be estimated by the sample moments $\hat{\mu}_{ij}$ given by
\[
\hat{\mu}_{ij} = \int z_i z_j dF_n(z_0, z) = \frac{\alpha_n}{n} \sum_{k=1}^n \frac{W_{kj} W_{kj}}{G_n(U_k)} , a \leq i, j \leq r.
\]
where $W_{k0} \equiv U_k, k = 1, 2, \ldots, n$, and $[W_{k1}, \ldots, W_{kr}] = W_k$ is the truncated observation of the vector $Z_k$ and $G_n$ is given in section 1. Then $\gamma$ and $\tau$ can be estimated by
\[
\hat{\gamma} = (\tilde{\mu}_{01}, \ldots, \tilde{\mu}_{0r})', \quad \hat{\tau} = (\tilde{\mu}_{ij})_{i,j=1}^r \quad (14)
\]
Equation (12) thus provides an estimating equation for $\beta$, An estimate for $\beta$ is any value $\hat{\beta}$ that satisfies
\[
\hat{\gamma} = \hat{\tau} \hat{\beta}
\]
It can be checked that the estimator $\tilde{\beta}$ is a vector of weighted least square estimates (LSE) which minimizes the quadratic form
\[
Q(\beta) = \sum_{k=1}^n \frac{1}{G_n(U_k)} (U_k - \beta_1 W_{k1} - \cdots - \beta_r W_{kr})^2.
\]
When $\alpha = 1$, no truncation presents. Then
\[
(U_k, W_k) = (Y_k, Z_k) \quad \text{and} \quad G_n(U_k) = 1
\]
for all $k$. Consequently $\tilde{\beta}$ reduces to the ordinary LSE.

The error variance of the multiple regression model (1) is
\[
\hat{\sigma}^2 = E\varepsilon^2 = \mu_{00} + \beta' \tau \beta - 2\beta' \gamma = \mu_{00} - \beta' \gamma.
\]
Substituting the moment estimates for the $\mu_{ij}$ yields an estimate for $\sigma^2$ given by
\[
\hat{\sigma}^2 = \hat{\mu}_{00} - \hat{\beta}' \hat{\gamma}.
\]
The error distribution $F_\epsilon$ can be estimated by

$$F_\epsilon(y) = \int I[s - \hat{\beta}' z \leq y] dF_n(s, z)$$

Some illustrative examples

The following numerical examples illustrate some of the issues discussed above. We use S-plus and SAS to do the regression including computing and graphics

Example 1

considers the simple model $y_i = \beta z_i + \epsilon_i$, in which the $\epsilon_i$ are i.i.d. $N(0, 0.25)$ and $z_i$ are i.i.d. uniformly distributed on $[0,2]$ and independent of the $\epsilon_i$. A sample of 100 data $(Z_i, Y_i, T_i)$ was generated from this model with $\beta = 1$. The $y_i$ are subject to left truncation by i.i.d. $N(0, 1)$ random variables $t_j$ that are independent of the $(x_j, \epsilon_j)$. $Y_i$ are only observed when $Y_i \geq T_i$, then some of the vectors would be missing and only a subsequence

$$(U_i, V_i, W_i), i = 1, 2, ..., N \quad \text{subject to} \quad U_i \geq V_i$$

The data are plotted in Figure 1. a

plot of raw data: ○, uncensored data; +, censored data. The solid and the dotted lines represents the true and fitted (Using He’s method) regression lines, respectively.

For complete data, we use classical regression method to fit the linear model. The result are given by R in form 1.a. since $p < .001$, so we adopt the regression model, the estimator was found to be $\hat{\beta} = 0.9532$.

Under the truncation, Only 79 sequences of $(U_i, V_i, W_i), i = 1, 2, ..., 79 \quad \text{subject to} \quad U_i \geq V_i$ can be observed. we use $79/100 = 0.79$ to estimate the missing rate, $\alpha = P[Y \geq T] = 0.83$. In this situation, we use Use He’s method, the preliminary estimator was found to be $\hat{\alpha} = 0.850188$.

In the computation process, we also notice that for all $x$ such that $(R_n(x) > 0)$,

$$\alpha_n = \frac{G_n(x)F_n(x^-)}{R_n(x)}$$

are equal to 0.7876.

Figure 1.b and 1.c plots of distributions of $U$ and $V$ respectively. The solid lines represent the product-limit estimates of $F(x)$ and $G(x)$ with truncated data,
The dotted lines represent the empirical estimates of $F(x)$ and $G(x)$ with complete data. From figure 1.b and 1.c, we derive that the two distribution $\hat{F}_n$ and $F$ are highly consistent. They also can illustrate the asymptotic properties of $F(x)$ and $G(x)$, as $N \to +\infty$.

we have got the estimates of $\alpha$, $F_x$ and $G_x$, As mentioned in the section 1, equation (12) provides an estimating equation for $\beta$, $\hat{\beta} = 0.9068$.

$$Q(\beta) = \sum_{k=1}^{n} \frac{1}{G_n(U_k)} (U_k - \beta W)^2 = 26.09488$$

To see the consistency of the estimates, we can evaluate further from samples of 50, 100, 150 data $(Z_i, Y_i, T_i)$, generated from the model in Example 1(with $\beta = 1$).

legend of the form
Figure 2: 1.a

$n$: the number of sample of complete data.
$N$: the number of sample under truncation.
$\alpha$: the empirical estimate of the truncated rate
$\hat{\alpha}$: the estimate of the truncation rate.
$\hat{\sigma}^2$: the error variance of the multiple regression model with truncated data.
$\tilde{\sigma}^2$: the error variance of residuals of the regression model with complete data.
$\hat{Q}(\beta)$: the weighted least square estimate with truncated data.
$\tilde{Q}(\beta)$: the ordinary least square estimates with complete data.
$\hat{\beta}$: the estimate of beta with complete data.
$\tilde{\beta}$: the estimate of beta with truncated data.

Form 1.a
Example 2 considers the simple model \( Y = \beta Z + \varepsilon \), in which the \( \varepsilon \) are i.i.d. Generate \( \varepsilon \) from normal distribution with mean=15, standard deviance = 1. \( Z = (Z_1, Z_2, Z_3) \) are i.i.d. \( Z_1 \) are subject to normal distribution with mean 17 and standard deviance 2, \( Z_2 \) are subject to normal distribution with mean 5 and standard deviance 2, \( Z_3 \) uniformly distributed on [5,15] and independent of the \( \varepsilon \). Three samples of 50, 100, 150 data \( (Z_i, Y_i, T_i) \) were generated from this model respectively with \( \beta = (\beta_1, \beta_2, \beta_3) = (0.3, 0.45, 0.2) \), the intercept \( \beta_0 = 10 \). \( Y \) are subject to left truncation by i.i.d. Normal random variables \( T \) that are independent of the \( (Y, \varepsilon) \). Only a subsequence

\[ (U_i, V_i, W_i), i = 1, 2, ..., N \quad \text{subject to} \quad U_i \geq V_i \]

We generate a sequence i.i.d variables \( T = (T_1, T_2, ..., T_n) \) subject to normal distribution with standard deviance 1. We can change the mean of variable \( T \) to

<table>
<thead>
<tr>
<th>( n )</th>
<th>( N )</th>
<th>( \tilde{\alpha} )</th>
<th>( \hat{\alpha} )</th>
<th>( \tilde{\sigma}^2 )</th>
<th>( \hat{\sigma}^2 )</th>
<th>( \hat{Q}(\beta) )</th>
<th>( \check{Q}(\beta) )</th>
<th>( \hat{\beta} )</th>
<th>( \check{\beta} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>32</td>
<td>0.84</td>
<td>0.8610</td>
<td>0.2148</td>
<td>0.2163</td>
<td>10.52633</td>
<td>9.494798</td>
<td>0.9335616</td>
<td>0.9453221</td>
</tr>
<tr>
<td>100</td>
<td>79</td>
<td>0.79</td>
<td>0.7876</td>
<td>0.2657</td>
<td>0.2413</td>
<td>26.30978</td>
<td>26.09488</td>
<td>0.9068152</td>
<td>0.9532342</td>
</tr>
<tr>
<td>150</td>
<td>126</td>
<td>0.84</td>
<td>0.8639</td>
<td>0.2488</td>
<td>0.2661</td>
<td>37.08403</td>
<td>33.39995</td>
<td>0.8716478</td>
<td>0.9473892</td>
</tr>
</tbody>
</table>

Figure 3: 1.a
estimate the model under different missing rate. The Form 2.a, 2.b, 2.c below are given the result of simulation. Note that the symbols represent the same variance defined in the legend of Form 1.a.

**Form 2.a**

<table>
<thead>
<tr>
<th>$n$</th>
<th>$N$</th>
<th>$\hat{\alpha}$</th>
<th>$\hat{\alpha}$</th>
<th>$\hat{\sigma}^2$</th>
<th>$\hat{\sigma}^2$</th>
<th>$\hat{Q}(\beta)$</th>
<th>$\hat{Q}(\beta)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>50</td>
<td>1</td>
<td>1</td>
<td>0.9666534</td>
<td>0.9473203</td>
<td>47.36602</td>
<td>47.36602</td>
</tr>
<tr>
<td>50</td>
<td>47</td>
<td>0.94</td>
<td>0.9489414</td>
<td>0.9666534</td>
<td>0.9348812</td>
<td>47.36602</td>
<td>46.30362</td>
</tr>
<tr>
<td>50</td>
<td>46</td>
<td>0.92</td>
<td>0.9253977</td>
<td>0.9666534</td>
<td>0.9813385</td>
<td>47.36602</td>
<td>48.78073</td>
</tr>
<tr>
<td>50</td>
<td>42</td>
<td>0.84</td>
<td>0.8424177</td>
<td>0.9666534</td>
<td>0.8127674</td>
<td>47.36602</td>
<td>40.52174</td>
</tr>
<tr>
<td>50</td>
<td>25</td>
<td>0.5</td>
<td>0.5022321</td>
<td>0.9666534</td>
<td>0.9194556</td>
<td>47.36602</td>
<td>45.76845</td>
</tr>
<tr>
<td>100</td>
<td>100</td>
<td>1</td>
<td>1</td>
<td>1.066348</td>
<td>1.055685</td>
<td>105.5685</td>
<td>105.5685</td>
</tr>
<tr>
<td>100</td>
<td>92</td>
<td>0.92</td>
<td>0.9190125</td>
<td>1.066348</td>
<td>1.045550</td>
<td>105.5685</td>
<td>104.6673</td>
</tr>
<tr>
<td>100</td>
<td>79</td>
<td>0.79</td>
<td>0.7980488</td>
<td>1.066348</td>
<td>0.9782337</td>
<td>105.5685</td>
<td>96.83676</td>
</tr>
<tr>
<td>100</td>
<td>59</td>
<td>0.59</td>
<td>0.5916286</td>
<td>1.066348</td>
<td>1.228288</td>
<td>105.5685</td>
<td>122.4907</td>
</tr>
</tbody>
</table>

Under complete data, set $n = 100$, we use classical regression method to estimate $\beta$. The result give by R is $(10.53880, 0.27236, 0.45054, 0.19188)$. And $p$ value $\leq 0.001$. We accept the linear regression model.

If $n = 50$, the estimate for $\beta$ is $(8.93548, 0.34962, 0.49634, 0.21399)$.

Note: Form 2.b give the estimation of the $\beta$.

**Form 2.b**
Based on the form, compare the estimates of the error variance of the multiple regression model defined in the beginning of this section under different missing rate. Under fixed number of the complete random sequence i.i.d variables, We find that the estimators of the regression model are growing more accurately when the truncated rate gets smaller.

On the other hand, If relatively large data are generated, the simulation result is closer to the true values of the variables in the regression model. To evaluate the consistency of the estimates of variables such as parameter $\beta$, weighted least square estimates $Q(\beta)$, the error variance of the regression model $\hat{\sigma}^2$

We also noted that when $\alpha = 1$, that is to say no truncation presents. Then $G_n(U_k = 1)$ for all k. Consequently $\beta$ reduces to the ordinary least square estimates(OLE).

4 strong consistency and CLT of the estimates

In He’s paper, he proved the consistency of the estimates derived in the section 2 and section 3.

The consistency of any estimate will be established if we show that for any measurement $\phi(x, z)$ such that $\int |\phi(x, z)|dF_n(x, z) < \infty$. the following convergence takes place.

$$\int \phi(x, z)dF_n(x, z) \rightarrow \int |\phi(x, z)|dF(x, z)a.s.$$
integrals with respect to the variables $x, z$

**Lemma 4.1** Assume that $F(x)$ and $G(x)$ are continuous, and $a_F \geq a_G$, Then for any measurable $\phi(x, z) \geq 0, S_n \leq S$, a.s

**Theorem 4.2** Under Lemma 1.1 and the condition of Lemma 4.1, we have

$$\int \phi(x, z) dF_n(x, z) \to \int \phi(x, z) F(dx, dz), \quad \text{a.s.}$$

Applying Theorem 4.2 and the Cramer-world theorem, we obtain immediately the strong consistent property of the weighted LSE $\hat{\beta}$ as well as of $\hat{\sigma}^2$ and $\hat{F}_z$.

**Theorem 4.3** suppose the condition of Lemma 1.1 hold. Let the moments $\mu_{jj} < \infty$ for $j = 0, 1, 2, \ldots, r$. Then as $n \to \infty$, with probability 1, $\hat{\mu}_{ij} \to \mu_{ij}$, a.s.

Hence, if $\tau = (\mu_{ij})_{i,j=1,\ldots,r}$ is nonsingular, with probability 1 we have $\hat{\beta} \to \beta, \hat{\sigma}^2 \to \sigma^2$. Further more, at each continuous point $y$ of $F_e$, $\hat{F}_e(y) \to F_e(y)$.

The normal convergence of the errors $\sqrt{n}(\hat{\beta} - \beta)$ is established in Theorem 4.5 which follows from a more general central limit theorem to be given in the Theorem 4.4 below. We need several preliminary results as stated below in Lemmas 4.1 and 4.2.

**Theorem 4.4**

$$\mu = \int g(x, z) F(dx, dz), \quad \bar{g}(s) = \int_{x \leq s} (g(x, z) - \mu) F(dx, dz).$$

Let $F$ and $G$ be continuous, $g(x, z)$ be any measurable function. Suppose the following finite conditions hold,

$$\int_{-\infty}^{b_F} \frac{dG}{1 - F} < \infty, \int \frac{dF}{G^2} < \infty$$

and $\int \frac{\bar{g}^2(x, z)}{G(x)} F(dx, dz) < \infty$

Then as $n \to \infty$,

$$\sqrt{n} \int g(x, z) F_n(dx, dz) - \int g(x, z) F(dx, dz) \to N(0, \sigma^2)$$

where

$$\sigma^2 = \alpha \int \frac{(g(x, z) - \mu)^2}{G(x)} F(dx, dz) + \alpha \int \frac{\bar{g}^2}{FG^2} dG < \infty$$
\[ \mu = \int g(x, z) F(dx, dz), \quad \bar{g}(s) = \int_{x \leq s} (g(x, z) - \mu) F(dx, dz). \]

We now apply Theorem 4.1 to obtain the normal convergence of \( \sqrt{n}(\hat{\beta} - \beta) \).

For this purpose let
\[ g_j(y, z) = z_j(y - \beta' z), \quad h(y, z) = (y - \beta' z^2) \]
and
\[ \tilde{g}_j(s) = \int_{y \leq s} g_j(y, z) F(dy, dz), \quad \tilde{h}(s) = \int_{y \leq s} (h(y, z) - \sigma^2 z^2) F(dy, dz). \]

Let \( Z = (Z_1, Z_2, \ldots, Z_r)' \) as given in model and let \( \tau = E(Z'Z) \)

**Theorem 4.5** Let \( F, G \) be continuous and
\[ \int_{-\infty}^{b_F} \frac{dG}{1 - F} < \infty, \quad \int dFG^2 < \infty \]

. Suppose that tau is nonsingular and
\[ E \left[ \frac{|Z_i Z_j| \varepsilon^2 + \varepsilon^4}{G(Y)} \right] < \infty, \quad i, j = 1, 2, \ldots, r. \]

Then, as \( n \to \infty \),

(i) \( \sqrt{n}(\hat{\beta} - \beta) \) has asymptotic multivariate-normal distribution with zero mean and covariance matrix \( \tau^{-1} \Sigma \tau^{-1} \), where the matrix \( \Sigma = (\sigma_{i,j}) \) with components
\[ \sigma_{i,j} = \alpha \int \frac{g_i(y, z) g_j(y, z)}{G(y)} F(dy, dz) + \int \frac{\tilde{g}_i \tilde{g}_j}{FG^2} dG \]

(ii) \( \sqrt{n}(\hat{\sigma}^2 - \sigma^2) \) has asymptotic normal distribution with zero mean and variance
\[ \tau^2 = \alpha \int \frac{(h(y, z) - \sigma^2 z^2)}{G(y)} F(dy, dz) + \alpha \int \frac{\tilde{h}^2}{FG^2} dG \]

5 Conclusion

In this article we have introduced a versatile and easily implementable estimation method for left-truncated data. The method requires a single preprocessing...
of the data to produce a product-limit estimate of the distribution that is now available in most statistical computing packages and then applying one of the regression routines, to the data in conjunction with the weights produced in the preprocessing stage. Asymptotic theory and simulation studies in Sections 2 and show that this easily implementable method has good statistical properties.

In the context of regression, He’s estimates are natural competitors of the iterative least squares estimate of BCY and TCW and other estimates for left-truncated data. Our comparison of the performance of He’s estimate with that of the BCY estimate in a simulation study has shown the advantage of the former. Both BCY and TJW use iterative procedures and requires a simultaneous estimation of the error distribution \( F_\varepsilon \) and \( \beta \) at each step of the iteration. In the simulation experiments and in the process of analyzing in details of the simulation result, we did not encounter the usual complications in the computation of asymptotic variance in the truncation model. Moreover, we prove under very weak conditions a general theorem on the weak convergence to normality of the estimator of \( \beta \) constructed in section 1 is derived. The asymptotic variance formula is simple.

References


