Journal of Mathematical Research & Exposition
Vol.13, No.2, May 1993

A Nonparametric Least Square Estimation of the Slope Parameters in the Truncated Regression Model *

Lin Huazheng (Chengdu Science and Technology College)
Chai Genxiang (Sichuan University, China)

Abstract In this paper, we consider the truncated regression model. A new method of estimating the regression parameters based on truncted data is given. The residual distribution is allowed to be unspecified. Using the conclusions on nonparametric estimation of error distributions in our work early, we obtain consistency of the estimation under some regular conditions. An example is used to show that our results is an improvement of Heckman's work (1979) essentially.

§1. Introduction

Let the truncated regression model be defined as

$$y_i = x_i'\beta + e_i, \quad i = 1, 2, \cdots, \tag{1}$$

where $e_1, e_2, \cdots$ iid., $Ee_1 = 0, 0 < Ee_1^2 < \infty; x_i's(\in R)$ are known covariates vectors, $\beta(\in R)$ is an unknown vector of parameters. The datum (x_i, y_i) is observed only if $z_i \geq 0$ where $z_1, z_2, \cdots$ is a sequence of the iid. random variables. Let $\tilde{y}_1, \cdots, \tilde{y}_m$ be observed $(m = m(n) \leq n).\tilde{y}_1, \cdots, \tilde{y}_m$ is called as the truncated data of $y_1, \cdots, y_n$ in literature. Based on the observations $\tilde{y}_1, \cdots, \tilde{y}_m$, it is desired to estimate β. Truncated regression model is widely applied to ecnomics and social research field. Hence, a lot of scholars studied it and obtained many interesting results. In 1958, Tobin [1] proposed M. L.E of β under the assumption of normality. But he didn't make any theoretical analysis. In 1973, Amemiya [2] obtained the consistency and asympotically normality of M. L. E of β under the assumption of normality. Since then, no important theoretical work has been done about the model. For the case of unknown error distribution, Battacharya, et al (1983) [3] gave a nonparametric estimation of β that involved zero of a criterion function without representation formula for the estimators. Hence, it is inconvenient to the theoretical analysis and application. C. F. Wu, et al (1988) [1] considered nonparametric estimation of β, however, no theoretical results has been given.

*Received May 5 1990. This work is supported by Natural Science foundation Grant of China.

In this paper, we propose a nonparametric L.S. method to estimate β. Using the conclusions on nonparametric estimation of error distributions in our work early [5], we obtain the consistency of the estimation under some regular conditions.

Suppose

$$Z_i = t_i' r + u_i, \quad i = 1, 2, \cdots, \tag{2}$$

where $t_i's (\in R)$ are known and $r (\in R)$ is unknown. In what follows, we shall assume the following regular conditions

(i) $e_1, e_2, \cdots$ iid. $F(x/\sigma); u_1, u_2, \cdots$ iid. $F(x)$ and $Eu_1 = 0, 0 < E(u_1^2) < \infty$, where F and $\sigma > 0$ are both unknown;

(ii) F has a density function f, which is uniform continuous on R;

(iii)

$$E(e_1 | u_1 \geq y) = \sum_{j=1}^{k} \phi_j(f(y), F(y), y)\psi_j, \text{ for } y \in R \tag{3}$$

where k is a positive integer and $\phi_j(\cdot,\cdot,\cdot)'s$ are known functions, $\psi_j s$ unknown (the $\psi_j's$ may contain the model parmeters, but are independent on y).

(iv) Let $X_m' = \begin{pmatrix} x_1, & \cdots, & x_m \\ & \vdots & \\ \alpha_1, & \cdots, & \alpha_m \end{pmatrix}, S_m = X_m' X_m$, where

$$\alpha_i = \begin{pmatrix} \phi_1(f(-t_i'r), & F(-t_i'r), & -t_i'r) \\ & \vdots & \\ \phi_k(f(-t_i'r), & F(-t_i'r), & -t_i'r) \end{pmatrix}, \ i = 1, \cdots, m$$

and $v_{jj}^{(m)}$ is (j,j)-th entry of $S_m^{-1} (1 \leq j \leq p + k)$, then there exists a $\delta_j > 1$ such that

$$v_{jj}^{(m)} = O((\log m)^{-2} (\log \log m)^{-\delta_j}), \tag{4}$$

where the conventions is adopted that the first $m (< n)$ observations on Y_i are available $(1 \leq i \leq n)$, and $m = m(n) \to$, as $n \to \infty$.

In §2, we shall derive the estimation method and some Lemmas. The main results will be given in §3 and two examples are given in §4.

§2. The estimation of β and some lemmas

Let the least square estimation of r based on $\{t_1, z_1\}, \cdots, \{t_n, z_n\}$ be $\hat{r}_n$ and the residul $\hat{u}_i = z_i - t_i' \hat{r}_n, i = 1, \cdots, n$. Denote the emprical distribution functions of $\hat{u}_1, \cdots, \hat{u}_n$ by $\hat{F}_n$. Then the estimation of $f(x)$ is defined as follows:

$$\hat{f}_n(x) = (2a_n)^{-1} [\hat{F}_n(x + a_n) - \hat{F}_n(x - a_n)], x \in R, \tag{5}$$

where $a_n > 0$ is a window width with $a_n \to 0$ as $n \to \infty$. Under the regular condition (iii), we have

$$E(Y_i|Z_i \geq 0) = x_i'\beta + E(e_i|u_i \geq -t_i'r) = x_i'\beta + \sum_{j=1}^{k} \phi_j(f(-t_i'r), F(-t_i'r), -t_i'r)\psi_j. \qquad (6)$$

A nature estimation of α, hence, can be difined as

$$\hat{\alpha}_{ni} = \begin{pmatrix} \phi_1(\hat{f}_n(-t_i'\hat{r}_n), & \hat{F}_n(-t_i'\hat{r}_n), & -t_i'\hat{r}_n \\ & \vdots & \\ \phi_k(\hat{f}_n(-t_i'\hat{r}_n), & \hat{F}_n(-t_i'\hat{r}_n), & -t_i'\hat{r}_n) \end{pmatrix}, \quad i = 1, 2, \ldots, m. \qquad (7)$$

Denote

$$\hat{X}_m = \begin{pmatrix} x_1', & \hat{\alpha}_{n1}' \\ \vdots & \vdots \\ z_m', & \hat{\alpha}_{nm}' \end{pmatrix}_{m\times(p+k)} \qquad (8)$$

and

$$\tilde{Y} = (\tilde{y}_1, \cdots, \tilde{y}_m)', \ \psi = (\psi_1, \cdots, \psi_k)'. \qquad (9)$$

Laslty, we defined L.S. E of (β, ψ) as follows

$$\begin{pmatrix} \hat{\beta}_n \\ \hat{\psi}_n \end{pmatrix} = (\hat{X}_m'\hat{X}_m)^{-1}\hat{X}_m'\tilde{Y}. \qquad (10)$$

Next, we shall give two-fold lemmas used in the sequal.

Lemma 1 *Under the regular conditions (i)-(iii), suppose the sequence of trial points $\{t_i\}$ in model (2) satisfies:*

a) *There exists a $M > 0$ such that $\|t_i\| \leq M, i = 1, 2, \cdots$;*

b) $(1/n)Q_n \to \Sigma(> 0)$ *as $n \to \infty$, where $Q_n = \sum_{i=1}^{n} t_i t'$ and*

$$0 < a_n \to 0, \sqrt{n}a_n/(\log n) \to \infty, \ \textit{as } n \to \infty. \qquad (11)$$

If v_n and v are random variables with $v_n \to v$, a.s., then

$$\hat{f}_n(v_n) \to f(v), \ \text{a.s..} \qquad (12)$$

Proof The conditions of lemma can be used to establish (see [5])

$$\sup_x |\hat{f}_n(x) - f(x)| \to 0, \ \text{a.s.}$$

and by the uniform continuity of f, it follows (12).

Lemma 2 *Under the assumpation of Lemma 1, if the random variables v_n and v satisfy that $v_n \to v$, a.s. then*

$$\hat{F}_n(v_n) \to F(v), a.s. \qquad (13)$$

Proof Since

$$(1/n)Q_n \to \Sigma \Longrightarrow nQ_n^{-1} \to \Sigma^{-1} \Longrightarrow \|Q_n^{-1}\|^{-1/2}\log\|Q_n^{-1}\| = O(n^{-1/2}\log n),$$

it is well known (see Theorem 2.9 in [6]) that $\hat{r}_n \to r$, a.s. Denote the empirical distribution function of $u_1, \cdots, u_n$ by F_n, then $\sup_x |F_n(x) - F(x)| \to 0$, a.s.

Without any loss of generality, we assume that $u_1, u_2, \cdots$ and $v_1, v_2, \cdots$ are defined on same probability space, and denote any sample point by ω. Define

$$A = \{\hat{r}_n \to r, v_n \to v, \sup_x |F_n(x) - F(x)| \to 0, \text{ as } n \to \infty\},$$

then fixed $\omega \in A$. For simplity, we still use the notations $\hat{F}_n(x), F_n(x)$ and $\hat{r}_n$, omitted the dependence on ω. By the definition of A, for any $\varepsilon > 0$ there exists a $N = N(\omega, \varepsilon) > 1$ such that $\|\hat{r}_n - r\| < \varepsilon$, for $n \geq N$. Hence for $n \geq N$

$$\begin{aligned} |\hat{F}_n(x) - F(x)| &\leq (1/n)\sum_{i=1}^n I_{(x<u_i\leq x+t_i'(\hat{r}_n - r))} \\ &\quad + (1/n)\sum_{i=1}^n I_{(x+t_i'(\hat{r}_n - r)<u_i\leq x)} \\ &\leq (1/n)\sum_{i=1}^n I_{(x<u_i\leq x+M\varepsilon)} + (1/n)\sum_{i=1}^n I_{(x-M\varepsilon<u_i\leq x)} \\ &\triangleq I_{n1} + I_{n2}. \end{aligned}$$

Denote $f_0 = \sup_x f(x)$, by uniform continuity of f, we have $f_0 < \infty$. It is evident that

$$\sup_x |F(x+M\varepsilon) - F(x)| = \sup_x \int_x^{x+M\varepsilon} f(s)ds \leq f_0 M\varepsilon$$

and

$$\begin{aligned} \sup|I_{n1}| &= \sup|F_n(x+M\varepsilon) - F_n(x)| \\ &\leq \sup|F_n(x+M\varepsilon) - F(x+M\varepsilon)| \\ &\quad + \sup|F(x+M\varepsilon) - F(x)| + \sup|F_n(x) - F(x)| \\ &\leq 2\sup|F_n(y) - F(y)| + f_0 M\varepsilon. \end{aligned}$$

From the definition of A, it follows that $\limsup_n \sup_x |I_{n1}| \leq f_0 M\varepsilon$. By arbitrarity of $\varepsilon > 0$, it gets $\sup_x |I_{n1}| \to 0$ as $n \to \infty$, similarly $\sup_x |I_{n2}| \to 0$ as $n \to \infty$. So we have that

$$\sup_x |\hat{F}_n(x) - F_n(x)| \to 0 \text{ as } n \to \infty. \tag{14}$$

Since

$$\begin{aligned} |\hat{F}_n(v_n) - F(v)| &\leq |\hat{F}_n(v_n) - F_n(v_n)| + |F_n(v_n) - F(v_n)| + |F(v_n) - F(v)| \\ &\leq \sup_x |\hat{F}_n(x) - F_n(x)| + \sup_x |F_n(x) - F(x)| + |F(v_n) - F(v)|. \end{aligned}$$

By the continuity of F and (14) it follows that $\hat{F}_n(v_n) \to F(v)$. as $n \to \infty$. Note $P(A) = 1$, so (13) is obtained.

§3. Main theorem

For simplity, we discuss the case of $k = 1$, it is easy to extend to general case.

Theorem 1 *Under the regular conditions (i), (ii), (iii), the trial points* $\{t_i\}$ *in model (2) satisfy that*

a) *There exists a* $M > 0$, *such that* $\|t_i\| \leq M, i = 1, 2 \cdots$;

b) $(1/n)Q_n \to \Sigma > 0, n \to \infty$, *where* $Q_n = \sum_{i=1}^{n} t_i t_i'$.

If $0 < a_n \to 0, \sqrt{n}a_n/(\log n) \to \infty$ *and*

$$\hat{S}_m^{-1}(i,i) \to 0, i = 1, \cdots, p+1, \tag{15}$$

where $\hat{S}_m = \hat{X}_m'\hat{X}_m, \hat{S}_m^{-1}(i,i)$ *is* (i,i)*-th entry of* $\hat{S}_m^{-1}$. *Then*

$$\begin{pmatrix} \hat{\beta}_n \\ \hat{\psi}_n \end{pmatrix} \to \begin{pmatrix} \beta \\ \psi \end{pmatrix} \text{ in pr.} \tag{16}$$

Proof Since $E(e_1|u_1 \geq y)$ is dependent on y and $rk(x_1, \cdots, x_m) = p$, hence $\hat{X}_m'\hat{X}_m$ is nonsingular. By [5], $\sqrt{n}(\hat{r}_n - r) \to N_p(0, \sigma^2\Sigma^{-1})$ in dis., thus $\hat{r}_n \to r$ a.s., and $t_i'\hat{r}_n \to t_i'r$ a.s.. For any integer $q \geq 1$, let $\hat{X}_q$ is a $q \times (p+1)$ matrix with i-th row equal to $(x_i', \hat{\alpha}_{ni})$. By Lemmas 1 and 2, it follows that $\hat{f}_n(-t_i'\hat{r}_n) \to f(-t_i'r)$ a.s., and $\hat{F}_n(-t_i'\hat{r}_n) \to F(-t_i'r)$ a.s.. Since $\phi_1(x_1, x_2, x_3)$ is continuous,

$$\phi_1(\hat{f}_n(-t_i'\hat{r}_n), \hat{F}_n(-t_i'\hat{r}_n), -t_i'\hat{r}_n) \to \phi_1(f(-t_i'r), F(-t_i'r), -t_i'r), \text{ a.s.}$$

Hence $\hat{\alpha}_{ni} \to \alpha_i$ a.s., $i = 1, \cdots, q$. Since $(\hat{X}'q\hat{X}_q)^{-1}\hat{X}_q'\tilde{Y}_q$ is continuous in $\hat{\alpha}_{n1}, \cdots, \hat{\alpha}_{nq}$ where $\hat{Y}$ is a $q \times 1$ matrix with i-th row equal to $\tilde{Y}_i$, thus

$$(\hat{X}_q'\hat{X}_q)^{-1}\hat{X}_q'\tilde{Y}_q \to (\tilde{X}_q'\tilde{X}_q)^{-1}\tilde{X}_q'\tilde{Y}_q \text{ a.s.,} \tag{17}$$

where $\tilde{X}_q = \begin{pmatrix} x_1', & \alpha_1 \\ \vdots & \vdots \\ x_q', & \alpha_q \end{pmatrix}$. By arbitrarity of q, it follows that

$$(\hat{X}_m'\hat{X}_m)^{-1}\hat{X}_m'\tilde{Y} - (X_m'X_m)^{-1}X_m'\tilde{Y} \to 0, \text{ a.s.} \tag{18}$$

Similarly $(\hat{X}_m'\hat{X}_m)^{-1} - (X_m'X_m)^{-1} \to 0$, a.s. In terms of (15), we have that $(X_m'X_m)^{-1} \to 0$, and hence

$$(X_m'X_m)^{-1}X_m'\tilde{Y} \to \begin{pmatrix} \beta \\ \psi \end{pmatrix}, \text{ in pr,}$$

From (18), we get

$$(\hat{X}_m'\hat{X}_m)^{-1}\hat{X}_m'\tilde{Y} \to \begin{pmatrix} \beta \\ \psi \end{pmatrix}, \text{ in pr,}$$

that is

$$\begin{pmatrix} \hat{\beta}_n \\ \hat{\psi}_n \end{pmatrix} \to \begin{pmatrix} \beta \\ \psi \end{pmatrix}, \text{ in pr.}$$

thus, we complete the proof.

Theorem 2 *Under the regular conditions (i)-(iv). The trial points* $\{t_i\}$ *in model (2) satisfy that*

a) *There exists a* $M > 0$ *such* $||t_i|| \le M, i = 1, 2 \cdots$;

b) $(1/n)Q_n \to \Sigma > 0$ *as* $n \to \infty$.

If $0 < a_n \to 0, \sqrt{n}a_n/(logn) \to \infty$. *Then*

$$\begin{pmatrix} \hat{\beta}_n \\ \hat{\psi}_n \end{pmatrix} \to \begin{pmatrix} \beta \\ \psi \end{pmatrix}, \text{ a.s.}$$

Proof By regular condition (iv)

$$(X_m'X_m)^{-1}X_m'\tilde{Y} \to \begin{pmatrix} \beta \\ \psi \end{pmatrix}, \text{ a.s.}$$

Using the same argument in Theorem 1, we have that

$$(\hat{X}_m'\hat{X}_m)^{-1}\hat{X}_m'\tilde{Y} - (X_m'X_m)^{-1}X_m'\tilde{Y} \to 0 \text{ a.s.,}$$

hence

$$\begin{pmatrix} \hat{\beta}_n \\ \hat{\psi}_n \end{pmatrix} \to \begin{pmatrix} \beta \\ \psi \end{pmatrix}, \text{ a.s.}$$

Thus proof is completed.

§4. Some examples

Example 1 Let $\begin{pmatrix} e_1 \\ u_1 \end{pmatrix} \sim N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & r\sigma \\ r\sigma & 1 \end{pmatrix}\right)$; $\{e_i\}$ iid. $\phi(x/\sigma)$; $\{u_i\}$ iid. $\phi(x)$ and $\{(e_i, u_i)'\}$iid., where r and are unknown, $\phi(x)$ is standard normal $d.f, \phi(x)$ is density function of $\phi(x)$. Furthermore, we assumed that (e_1, u_1) has a joint density function $f(x, y)$. Since

$$P(e_1 \le x | u_1 \ge y) = [\phi(x/\sigma) - P(e \le x, u \le y)]/(1 - \phi(y)].$$

Under given $u_1 \ge y, e_1$ has conditional density function

$$f(x|u \ge y) = [(1/\sigma)\phi(x/\sigma) - \int f(x,t)dt]/(1 - \phi(y))$$

and

$$\begin{aligned} E(e_1|u_1 \geq y) &= \int_{-\infty}^{+\infty} xf(x/u \geq y)dx \\ &= (\int_{-\infty}^{+\infty}(x/\sigma)\phi(x/\sigma)dx - \int_{-\infty}^{+\infty}\int_{-\infty}^{y} xf(x,t)dtdx)/(1-\Phi(y)) \\ \int_{-\infty}^{+\infty}\int_{-\infty}^{y} xf(x,t)dtdx &= \psi_2(\sigma,r)\phi(y) + \psi_3(\sigma,r)\Phi(y). \end{aligned}$$

let

$$\psi_1(\sigma,r) = \int_{-\infty}^{+\infty}(x/\sigma)\Phi(x/\sigma)dx,$$

then

$$E(e_1|u_1 \geq y) = (\psi_1(\sigma,r) - \psi_2(\sigma,r)\phi(y) - \psi_3(\sigma,r)\Phi(y))/(1-\Phi(y)).$$

Example 2 Let $\log z_{1i}, i = 0,1,\cdots,n$ iid $N(\mu/2,1)$ and $e_{1i} = z_{1i}z_{10} - e^{\mu+1}, e_{2i} = (z_{1i}/z_{10}) - e$. Then $\{e_{1i}\}$ iid. $F(x/e^{\mu})$ and $\{e_{2i}\}$ iid. $F(x)$, where $F(x-e)$ is logarithm normal d.f. with parametric (0,2) and $f(x-e)$ is density of $F(x-e)$. Denote $\sigma = e^{\mu}$, then (e_{1i}, e_{2i}) has a joint density

$$f(x,y) = \exp\{-(1/4)((\log(x+e^{\sigma})-\mu)^2 + (\log(y+e))^2\}/(4\pi(x+e\sigma)(y+e)).$$

Under the condition of given $e_{2i} \geq y, e_{1i}$ has conditional expectation

$$E(e_{1i}|e_{2i} \geq y) = (\psi_1(\sigma) - \psi_2(\sigma)F(y))/(1-F(y)).$$

From example 2, we can conclude that the regular condition (iii) is also satisfied for non-normal distributions. So our results improve [2] essentially. However, the condition (iii) can be weaken, it is still an open problem.

References

[1] Tobin J. *Estimation of relationships for limited dependent variables,* Econometrica, **26**(1958), 24-36.

[2] Takeshi, Amemiya, *Regression analysis when the dependent variable is truncted normal,* Econometrica, **41**(1973), 997-1016.

[3] P.K. Bhattacharya et. al., *Nonparametric estimation of the slope of a truncted regression,* The Annals of Statistics, **11**(1983), 505-514.

[4] C.F.J. Wu, et. al., *A nonparametric approach to the truncted regression problem,* Journal of American Statistical Association, **83**(1988), 183-192.

[5] Chai Genxiang, et. al., *Consistent nonparametric estimation of error distributions in linear model,* Acta Math. Appl. Sinica **7(3)**(1991), 245-256.

[6] Chen Xiru, et. al., *Theory of Parametric Estimation in Linear Model,* Science Press, Peijng, 1985, p92.

截断回归模型参数的非参数最小二乘估计

林华珍　　　柴根象
(成都科技大学应用数学系，610065)　(四川大学数学系，成都610064)

摘　要

本文考虑截断回归模型，给出了基于截断数据估计回归参数的一种新方法，此处并不设定残差分布．我们使用早先的关于误差分布非参数估计的结果，在某些正则条件下建立了估计量的相合性．并给出实例说明我们的结果是Heckman(1979)-项工作的本质改进．