应用概率统计　第二十七卷
第一期　2011年2月

Chinese Journal of Applied Probability
and Statistics Vol.27 No.1 Feb. 2011

Estimation of Parameters of Linear Regression Model under Contaminated and Interval Censored Response Variable *

He Qixiang
(*Department of Applied Mathematics, Shanghai University of Finance and Economics, Shanghai, 200433*)
Zheng Ming
(*Department of Statistics, Fudan University, Shanghai, 200433*)

Abstract

The estimation of regression parameters and the contamination coefficient of a linear model are studied when its response variables are contaminated and interval censored. Under some suitable conditions it is proved that the estimators which are established in this paper are strongly consistent. Some simulation results indicate that our method performs very well even though the data both contaminated and interval censored.

Keywords: Contaminated data, interval censored data, linear model.

AMS Subject Classification: 62N01, 62J05.

§1. Introduction

Consider the linear regression model

$$Y_i = \alpha + \beta x_i + \varepsilon_i, \qquad i = 1, 2, \cdots, n, \tag{1.1}$$

where ε_i's are i.i.d., $\mathsf{E}\varepsilon_i = 0$, $\mathsf{Var}(\varepsilon_i) = \sigma_1^2$ and σ_1^2 is known. $\{Y_i\}$ are contaminated by another random variables sequence $\{t_i\}$.

$$Y_i^\star = (1-\nu)Y_i + \nu t_i, \qquad i = 1, 2, \cdots, n. \tag{1.2}$$

Here t_i's are i.i.d., $\mathsf{E}t_i = 0$, $\mathsf{Var}(t_i) = \sigma_2^2$, σ_2^2 is known. $\{t_i\}$ and $\{Y_i\}$ are independent. $\nu\,(0 \le \nu < 1)$ is called contamination coefficient. (1.2) can be changed to the following:

$$Y_i^\star = (1-\nu)(\alpha + \beta x_i + \varepsilon_i) + \nu t_i,$$

or

$$Y_i^\star = (1-\nu)\alpha + (1-\nu)\beta x_i + \eta_i, \tag{1.3}$$

*This work was supported by National Natural Science Foundation of China (10971033) and Leading Academic Discipline Program, 211 Project for Shanghai University of Finance and Economics (the 3rd phase).
Received September 24, 2008. Revised December 22, 2009.

where $\eta_i = (1-\nu)\varepsilon_i + \nu t_i$.

$$\mathsf{E}Y_i^\star = (1-\nu)\alpha + (1-\nu)\beta x_i, \qquad \mathsf{Var}\,(Y_i^\star) = (1-\nu)^2\sigma_1^2 + \nu^2\sigma_2^2 \widehat{=} \sigma^2. \tag{1.4}$$

Suppose that $\sigma_1^2/\sigma_2^2 > \nu/(1-\nu)$. It is easily got that $\{Y_i^\star\}$ are independent. Let $F_i(\cdot)$ be the distribution function of $Y_i^\star$, suppose $\mathsf{E}Y_i^{\star r} < \infty$, $r \geq 1$. Zheng, Ding and Yang[1] introduced the estimators of α, β and contamination parameter ν in the model (1.3).

In this paper, we assume the response variables $Y_i^\star$'s are interval censored. The data we observed are

$$(U_i, V_i, \delta_{1i}, \delta_{2i}, x_i), \qquad i = 1, 2, \cdots, n,$$

where $(U_i, V_i)\,(i = 1, 2, \cdots, n)$ are nonnegative i.i.d. r.v's. Their densities $g(u, v)$ are positive, the distribution functions are $G(u, v)$. $\delta_{1i} = I_{Y_i^\star \leq U_i}$, $\delta_{2i} = I_{U_i < Y_i^\star \leq V_i}$. $\{Y_i^\star\}$ and $\{(U_i, V_i)\}$ are independent.

The problem of interval censored data is an important topic of the statistics. There are lots of backgrounds of the application in the epidemiology studies[2, 3], AIDS studies[4, 5], demography studies[6, 7], etc.

Let $F_\eta(\cdot)$ be the distribution function of η_i. We can get $\widehat{\alpha}, \widehat{\beta}, \widehat{\nu}$, which are the estimators of α, β, ν respectively, by using the method of maximum likelihood. The Log-likelihood function is

$$\begin{aligned} & L(\alpha, \beta, \nu; F_\eta) \\ = & \sum_{i=1}^{n} \{\delta_{1i} \log F_\eta(U_i - (1-\nu)\alpha - (1-\nu)\beta x_i) \\ & + \delta_{2i} \log[F_\eta(V_i - (1-\nu)\alpha - (1-\nu)\beta x_i) - F_\eta(U_i - (1-\nu)\alpha - (1-\nu)\beta x_i)] \\ & + (1 - \delta_{1i} - \delta_{2i}) \log[1 - \log F_\eta(V_i - (1-\nu)\alpha - (1-\nu)\beta x_i)]\}. \end{aligned}$$

Although the process above is seemed as a concise form, the computation is too complex (Here F_η is unknown). So, in this paper, we use the method of "unbiased transform" proposed by Zheng[8] to solve the problem of parameters estimation. Let

$$\widetilde{Y}_i^{\star r} = \varphi_1^{(r)}(U_i, V_i)\delta_{1i} + \varphi_2^{(r)}(U_i, V_i)\delta_{2i} + \varphi_3^{(r)}(U_i, V_i)(1 - \delta_{1i} - \delta_{2i}),$$

where $\varphi_1^{(r)}, \varphi_2^{(r)}$ and $\varphi_3^{(r)}$ are continuous functions which are independent of $F_i(\cdot)$. But maybe they have relation with $G(\cdot, \cdot)$.

Lemma 1.1 Assume that the continuous partial derivatives $\partial\varphi_j^{(r)}/\partial u$, $\partial\varphi_j^{(r)}/\partial v$ $(j = 1, 2, 3)$ exist and that $\varphi_1^{(r)}, \varphi_2^{(r)}, \varphi_3^{(r)}$ satisfy

$$\begin{cases} \displaystyle\int_{v=0}^{+\infty} \int_{u=0}^{v} \varphi_1^{(r)}(u, v) g(u, v) \mathrm{d}u\mathrm{d}v = 0; \\ \displaystyle\int_{x}^{+\infty} [\varphi_2^{(r)}(x, v) - \varphi_1^{(r)}(x, v)] g(x, v) \mathrm{d}v + \int_{0}^{x} [\varphi_3^{(r)}(u, x) - \varphi_2^{(r)}(u, x)] g(u, x) \mathrm{d}u = r x^{r-1}, \end{cases} \tag{1.5}$$

then

$$\mathsf{E}\widetilde{Y}_i^{\star r} = \mathsf{E}Y_i^{\star r}, \qquad r = 1, 2, \cdots.$$

Proof See [9] (Theorem 2.1). □

§2. The Model and the Estimation of α, β, ν

By Lemma 1.1, replacing $Y_i^\star$ with $\widetilde{Y}_i^\star$ in (1.3), we can establish a new model as following:

$$\widetilde{Y}_i^\star = (1-\nu)\alpha + (1-\nu)\beta x_i + \mu_i. \tag{2.1}$$

It is reasonable from the view of large sample.

Note $\zeta = (1-\nu)\alpha$, $\xi = (1-\nu)\beta$. Using the method of linear regression model, the estimators of ζ and ξ have the following construction.

$$\begin{cases} \widehat{\xi} = (1-\widehat{\nu})\widehat{\beta} = \dfrac{n\sum\limits_{i=1}^n x_i\widetilde{Y}_i^\star - \Big(\sum\limits_{i=1}^n x_i\Big)\Big(\sum\limits_{i=1}^n \widetilde{Y}_i^\star\Big)}{n\sum\limits_{i=1}^n x_i^2 - \Big(\sum\limits_{i=1}^n x_i\Big)^2}; \\ \widehat{\zeta} = (1-\widehat{\nu})\widehat{\alpha} = \dfrac{\Big(\sum\limits_{i=1}^n x_i^2\Big)\Big(\sum\limits_{i=1}^n \widetilde{Y}_i^\star\Big) - \Big(\sum\limits_{i=1}^n x_i\Big)\Big(\sum\limits_{i=1}^n x_i\widetilde{Y}_i^\star\Big)}{n\sum\limits_{i=1}^n x_i^2 - \Big(\sum\limits_{i=1}^n x_i\Big)^2}. \end{cases} \tag{2.2}$$

We have $\mathsf{E}\widehat{\xi} = \xi$, $\mathsf{E}\widehat{\zeta} = \zeta$.

To get the estimator $\widehat{\nu}$, we estimate σ^2 first. Let

$$\widetilde{Y}_i^{\star 2} = \varphi_1^{(2)}(U_i, V_i)\delta_{1i} + \varphi_2^{(2)}(U_i, V_i)\delta_{2i} + \varphi_3^{(2)}(U_i, V_i)(1-\delta_{1i}-\delta_{2i}),$$

where $\varphi_1^{(2)}$, $\varphi_2^{(2)}$ and $\varphi_3^{(2)}$ satisfy the assumption of Lemma 1.1 for $r = 2$.

Considering of $\mathsf{E}\widetilde{Y}_i^{\star 2} = \mathsf{E}Y_i^{\star 2}$, $i = 1, 2, \cdots, n$, we use

$$\widehat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^n \widetilde{Y}_i^{\star 2} - (\widehat{\zeta}^2 + 2\widehat{\zeta}\widehat{\xi}\overline{x} + \widehat{\xi}^2\overline{x^2}) \tag{2.3}$$

to estimate σ^2, where

$$\overline{x} = \frac{1}{n}\sum_{i=1}^n x_i, \qquad \overline{x^2} = \frac{1}{n}\sum_{i=1}^n x_i^2.$$

To get the strong consistency of the estimators of α, β, ν, we need the following lemma.

Lemma 2.1 For the linear model (2.1), assume that the following conditions hold:

(C1) $\sup\limits_{1\le i\le n}[\mathsf{Var}\,(\widetilde{Y}_i^\star) - \sigma^2] = R < \infty$;

(C2) $\sup\limits_{1\le i\le n}|x_i| = M < \infty$;

(C3) $\lim\limits_{n\to\infty} S_n^2 = \lim\limits_{n\to\infty}\sum\limits_{i=1}^{n}(x_i-\overline{x})^2 = \infty$, $\lim\limits_{n\to\infty}(\overline{x})^2/S_n^2 = 0$;

(C4) $\int_0^{+\infty} \sup\limits_{1\le i\le n} \mathsf{P}(|\varphi_1^{(2)}(U_i,V_i)\delta_{1i}+\varphi_2^{(2)}(U_i,V_i)\delta_{2i}+\varphi_3^{(2)}(U_i,V_i)(1-\delta_{1i}-\delta_{2i})|\ge t)\mathrm{d}t<\infty$,

then $\widehat{\sigma}_n^2 \xrightarrow{\text{a.s.}} \sigma^2$.

Proof

$$\begin{aligned}\widehat{\sigma}_n^2-\sigma^2 &= \frac{1}{n}\sum_{i=1}^{n}(\widetilde{Y}_i^{\star 2}-\mathsf{E}\widetilde{Y}_i^{\star 2})+\frac{1}{n}\sum_{i=1}^{n}\mathsf{E}\widetilde{Y}_i^{\star 2}-\sigma^2-(\widehat{\zeta}^2+2\widehat{\zeta}\widehat{\xi}\overline{x}+\widehat{\xi}^2\overline{x^2})\\ &= \frac{1}{n}\sum_{i=1}^{n}(\widetilde{Y}_i^{\star 2}-\mathsf{E}\widetilde{Y}_i^{\star 2})+\frac{1}{n}\sum_{i=1}^{n}(\zeta+\xi x_i)^2-(\widehat{\zeta}^2+2\widehat{\zeta}\widehat{\xi}\overline{x}+\widehat{\xi}^2\overline{x^2})\\ &= \frac{1}{n}\sum_{i=1}^{n}(\widetilde{Y}_i^{\star 2}-\mathsf{E}\widetilde{Y}_i^{\star 2})+(\zeta^2-\widehat{\zeta}^2)+\overline{x^2}(\xi^2-\widehat{\xi}^2)+2\overline{x}(\zeta\xi-\widehat{\zeta}\widehat{\xi}).\end{aligned}$$

By the result of Zheng[8], if $\lim\limits_{n\to\infty}\sum\limits_{i=1}^{n}(x_i-\overline{x})^2=\infty$, $\lim\limits_{n\to\infty}(\overline{x})^2/S_n^2=0$, we have $\widehat{\zeta}\xrightarrow{\text{a.s.}}\zeta$, $\widehat{\xi}\xrightarrow{\text{a.s.}}\xi$ as $n\to\infty$. So

$$\widehat{\zeta}^2-\zeta^2\xrightarrow{\text{a.s.}}0,\qquad \widehat{\xi}^2-\xi^2\xrightarrow{\text{a.s.}}0,\qquad \widehat{\zeta}\widehat{\xi}-\zeta\xi\xrightarrow{\text{a.s.}}0.$$

We need the following conclusion. Let V_i' be i.i.d. r.v's. If W is a random variable which satisfies

$$\mathsf{P}(|W|\ge t)\ge\mathsf{P}(|V_i|\ge t),\qquad \forall\, t\ge 0$$

and $\mathsf{E}|W|<\infty$, then

$$\frac{1}{n}\sum_{i=1}^{n}(V_i-\mathsf{E}V_i)\longrightarrow 0,\qquad \text{a.s.}.$$

From the condition (C4), we have

$$\begin{aligned}&\int_0^{\infty}\sup_{1\le i\le n}\mathsf{P}(|\widetilde{Y}_i^{\star 2}|\ge t)\mathrm{d}t\\ =\ &\int_0^{\infty}\sup_{1\le i\le n}\mathsf{P}(|\varphi_1^{(2)}(U_i,V_i)\delta_{1i}+\varphi_2^{(2)}(U_i,V_i)\delta_{2i}+\varphi_3^{(2)}(U_i,V_i)(1-\delta_{1i}-\delta_{2i})|\ge t)\mathrm{d}t<\infty.\end{aligned}$$

So

$$\lim_{n\to\infty}\sum_{i=1}^{n}(\widetilde{Y}_i^{\star 2}-\mathsf{E}\widetilde{Y}_i^{\star 2})=0,\qquad \text{a.s.}.\qquad \square$$

From (1.4),

$$\widehat{\sigma}_n^2=(1-\widehat{\nu})^2\sigma_1^2+\widehat{\nu}^2\sigma_2^2,$$

then

$$\widehat{\nu}=\frac{\sigma_1^2\pm\sqrt{(\sigma_1^2+\sigma_2^2)\widehat{\sigma}_n^2-\sigma_1^2\sigma_2^2}}{\sigma_1^2+\sigma_2^2}.$$

Since $\sigma_1^2/\sigma_2^2 > \nu/(1-\nu)$,

$$\widehat{\nu} = \frac{\sigma_1^2 - \sqrt{(\sigma_1^2+\sigma_2^2)\widehat{\sigma}_n^2 - \sigma_1^2\sigma_2^2}}{\sigma_1^2+\sigma_2^2}. \tag{2.4}$$

Under the assumption of Lemma 2.1, $\widehat{\sigma}_n^2 \xrightarrow{\text{a.s.}} (1-\nu)^2\sigma_1^2 + \nu^2\sigma_2^2$, it follows that

$$\widehat{\nu} \xrightarrow{\text{a.s.}} \nu.$$

The expression (2.4) with (2.2), we get

$$\begin{cases} \widehat{\beta} = \dfrac{n\sum\limits_{i=1}^n x_i\widetilde{Y}_i^\star - \Big(\sum\limits_{i=1}^n x_i\Big)\Big(\sum\limits_{i=1}^n \widetilde{Y}_i^\star\Big)}{n\sum\limits_{i=1}^n x_i^2 - \Big(\sum\limits_{i=1}^n x_i\Big)^2} \cdot \dfrac{1}{1-\widehat{\nu}}; \\ \widehat{\alpha} = \dfrac{\Big(\sum\limits_{i=1}^n x_i^2\Big)\Big(\sum\limits_{i=1}^n \widetilde{Y}_i^\star\Big) - \Big(\sum\limits_{i=1}^n x_i\Big)\Big(\sum\limits_{i=1}^n x_i\widetilde{Y}_i^\star\Big)}{n\sum\limits_{i=1}^n x_i^2 - \Big(\sum\limits_{i=1}^n x_i\Big)^2} \cdot \dfrac{1}{1-\widehat{\nu}}. \end{cases} \tag{2.5}$$

If the conditions of Lemma 2.1 hold, we have

$$\widehat{\alpha} \xrightarrow{\text{a.s.}} \alpha, \qquad \widehat{\beta} \xrightarrow{\text{a.s.}} \beta.$$

Finally, we can get the strong consistency of $\widehat{\alpha}$, $\widehat{\beta}$, $\widehat{\nu}$ as state below.

Theorem 2.1 For the linear model (2.1), suppose that the conditions (C1)-(C4) in Lemma 2.1 are satisfied, then

$$\widehat{\alpha} \xrightarrow{\text{a.s.}} \alpha, \qquad \widehat{\beta} \xrightarrow{\text{a.s.}} \beta, \qquad \widehat{\nu} \xrightarrow{\text{a.s.}} \nu,$$

where $\widehat{\alpha}$, $\widehat{\beta}$, $\widehat{\nu}$ are defined by (2.3), (2.4) and (2.5).

§3. Some Simulation Studies

In this section, we shall describe some simulation studies to examine the properties of the estimators $\widehat{\alpha}, \widehat{\beta}, \widehat{\nu}$.

The following linear model is treated:

$$\begin{aligned} &Y_i^\star = (1-\nu)(\alpha + \beta x_i + \varepsilon_i) + \nu t_i, \\ &\alpha = 3, \quad \beta = 2, \quad \nu = 0.1, \quad G(u,v) = \int_v^u \int_0^v \frac{1}{t}\lambda e^{-\lambda t}\mathrm{d}s\mathrm{d}t, \quad \lambda = \frac{1}{5}. \end{aligned}$$

The data set x_i's is generated from the uniform distribution $U[0,2]$, and the errors ε_i's are assigned $N(0,1)$.

The following two patterns are considered:

$$\text{P1:}\begin{cases}\varphi_1(u,v)=u-\dfrac{1}{2\lambda};\\ \varphi_2(u,v)=u-\dfrac{1}{2\lambda};\\ \varphi_3(u,v)=u-\dfrac{1}{2\lambda}+\dfrac{1}{\lambda}e^{\lambda v},\end{cases}\qquad \begin{cases}\varphi_1^{(2)}(u,v)=u-\dfrac{1}{2\lambda};\\ \varphi_2^{(2)}(u,v)=u-\dfrac{1}{2\lambda};\\ \varphi_3^{(2)}(u,v)=u-\dfrac{1}{2\lambda}+\dfrac{2v}{\lambda}e^{\lambda v}.\end{cases}$$

$$\text{P2:}\begin{cases}\varphi_1(u,v)=0;\\ \varphi_2(u,v)=v;\\ \varphi_3(u,v)=ve^{\lambda(v-u)}+v,\end{cases}\qquad \begin{cases}\varphi_1^{(2)}(u,v)=0;\\ \varphi_2^{(2)}(u,v)=v;\\ \varphi_3^{(2)}(u,v)=v+\dfrac{1}{\lambda}(e^{\lambda v}-1).\end{cases}$$

For each of the two patterns, we chose the sample sizes $n=200$, 500, 800, 1000, 2000. The results based on 300 simulations are displayed in Table 1 (P1) and Table 2 (P2).

Table 1 (P1)

n	$\widehat{\alpha}$	$\widehat{\beta}$	$\widehat{\nu}$
200	3.6142	2.4386	0.2123
500	2.7953	2.1343	0.1561
800	3.1533	1.8738	0.1311
1000	3.1081	2.0531	0.1211
2000	3.0353	2.0509	0.09806

Table 2 (P2)

n	$\widehat{\alpha}$	$\widehat{\beta}$	$\widehat{\nu}$
200	3.9698	2.6595	0.2680
500	3.4484	2.2947	0.1834
800	3.2596	2.2210	0.1464
1000	2.8629	2.1432	0.8547
2000	3.1045	1.9762	0.1064

The following conclusions may be drawn from the results summarized in Table 1 and Table 2.

(1) The results are good when the sample size n increase to 500 and become better and better when the sample size n increases.

(2) There are no significant difference between the results even if $(\varphi_1,\varphi_2,\varphi_3,\varphi_1^{(2)},\varphi_2^{(2)},\varphi_3^{(2)})$ take different cases.

References

[1] Zheng, Z.K., Ding, B.J., Yang, Y. and Ying, H., Parameter estimation in regression analysis for two kinds of contamination data, *Applied Mathematics-A Journal of Chinese University*, **11A(1)**(1996), 31–40.

[2] Finkelstein, D.M. and Wolfe, R.A., A semiparametric model for regression analysis of interval-censored failure time data, *Biometrics*, **41**(1985), 933–945.

[3] Finkelstein, D.M., A proportional harzard model for interval-censored failure time data, *Biometrics*, **42**(1986), 845–854.

[4] De Gruttola, V. and Lagakos, S.W., Analysis of doubly-censored survival data with application to AIDS, *Biometrics*, **45**(1989), 1–11.

[5] Shiboski, S.C. and Jewell, N.P., Statistical analysis of the time dependence of HIV infectivity based on partner study data, *J. Amer. Statist. Assoc.*, **87**(1992), 360–372.

[6] Diamond, I.D. and Mcdonald, J.W., Analysis of current status data, In: J. Trussell, R. Hankinson, J. Tilton, eds, *Demographic Applications of Event History Analysis*, Oxford, U.K., Oxford University Press, 1991, 231–252.

[7] Diamond, I.D., Mcdonald, J.W. and Shah, I.H., Proportional harzard models for current stuts data: application to the study of differentials in age at weaning in Pakistein, *Demography*, **23**(1986), 607–620.

[8] Zheng, Z.K., A class of estimators of the mean survival time from interval censored data with application to linear regression, manuscript.

[9] Deng, W.L. and Zheng, Z.K., The estimation of the r-th original moment of interval censored data, *Chinese Journal of Applied Probability and Statistics*, **22(4)**(2006), 419–428.

污染数据区间截断下线性模型的参数估计

何其祥 郑 明

(上海财经大学应用数学系, 上海, 200433) (复旦大学统计学系, 上海, 200433)

本文研究了线性模型响应变量被污染且被区间截断下的参数估计问题, 借助于区间数据的无偏转换, 得到了回归系数和污染系数的估计, 并在一定的条件下得到了这些估计的强相合性. 通过若干模拟例子说明, 尽管数据经过污染和区间截断的双重信息损失, 但用本文提出的方法得到的估计, 仍能取得良好的估计效果.

关键词: 污染数据, 区间数据, 线性模型.

学科分类号: O212.7.