应用概率统计　第二十八卷
第二期　2012年4月

Chinese Journal of Applied Probability
and Statistics Vol.28 No.2 Apr. 2012

Regression Model with an Interval-Censored Data Covariant *

Ding Bangjun

(*School of Finance and Statistics, East China Normal University, Shanghai, 200241*)

Abstract

A likelihood approach, together with a EM-type algorithm, to jointly estimate the regression coefficient as well as the marginal distribution of the covariant in regression model with an interval-censored data covariant is developed. Under certain conditions the procedures are convergent, and the resulting estimators are asymptotically normal.

Keywords: Regression, interval data, EM-type algorithm, asymptotically normal.

AMS Subject Classification: 62N02.

§1. Introduction

Incomplete data are frequently encountered in medical follow-up studies and in reliability studies. Partially motivated by problems arising from these studies, analysis of right censored data has been one of the focal point of statistics in the past three decades. Recently, statisticians are paying more and more attention to some more complicated types of incomplete data, such as doubly censored data and interval censored data, as these data occur in important clinical trials. For instance, the former were encountered in recent studies on primary breast cancer, and the latter were encountered in AIDS research (Kim, 1993). Inference for a linear regression model with double censored data has studied by Kim (1993) and Ren (1989). This current paper is concerned with regression with an interval censored covariant.

§2. Model

Consider a simple linear regression model

$$y = \alpha + \beta X + \sigma\epsilon, \qquad (\sigma > 0), \tag{2.1}$$

*The project supported by the National Natural Science Foundation of China (10671072).
Received December 9, 2010. Revised January 19, 2011.

where Y is a continuous response variable and X an explanatory variable. We take X be scalar, discrete with distribution

$$\mathsf{P}(X = x_i) = p_i, \qquad i = 1, 2, \ldots, m$$

and linearly related to Y by (2.1), ϵ a standard normal random variable, independent of X. Let $\theta = (\alpha, \beta, \sigma^2)$ is an unknown parameter vector. Denote the conditional density function of Y given $X = x$ by $f(y|x;\theta)$. Suppose that the observable data for each subject are of the form (Y, Z), where Z is a random variable such that

$$\mathsf{P}(X \in [Z, Z + T]) = 1,$$

where T is a fix positive constant. Thus, the response Y is full observed and X is interval censored in the fixed length interval $[Z, Z + T]$. The special case $T = 0$ indicates that X is observed exactly. We shall assume that censoring occurs non-informatively in the sense that for any X, Z

$$\mathsf{P}(X = x|Z = z) = \frac{\mathsf{P}(X = x)}{\mathsf{P}(X \in [z, z + T])}, \tag{2.2}$$

and

$$f(y|X = x, Z = z;\theta) = f(y|X = x;\theta).$$

Suppose that the data consist of n independent and identically distributed realization (y_i, z_i), $i = 1, 2, \ldots, n$. It follows that the likelihood function is proportional to

$$L(\theta, p) = \prod_{i=1}^{n} \sum_{j=1}^{m} \alpha_{ij} f(y_i|x_j;\theta) p_j, \tag{2.3}$$

where $\overrightarrow{p}$ and $\alpha_{ij} = I(X_j \in [z_i, z_i + T])$, and $I(\cdot)$ is the indicator function.

In general, θ is estimable but the estimability of the individual components of X depends on the censoring process, as well as on the values of the response. For example, suppose α and β are known, in the setting, only sum of $p_1 + p_2$ is estimable if every interval $[z_i, z_i + T]$ contains both x_1 and x_2 or neither of them. It follows from Gentleman and Geyer (1994) that the p_j are estimable if $n \geq m$ and $n \times m$ matrix $A = (a_{ij})$ is of rank m, where $a_{ij} = \alpha_{ij} f(y_i|x_j;\theta)$.

§3. Parameter Estimation

The statistical goal in current paper is the estimation of $\theta = (\alpha, \beta, \sigma^2)$ in the presence of the nuisance distribution function F of X. We develop the estimator of θ by maximizing (2.3) simultaneously for θ and X using EM algorithm as follows two steps:

Step 1: Nonparametric estimation of X assuming θ is known

Suppose first that θ is known and denote the likelihood function by $L(X|\theta)$, the conditional density function of Y given $X = x_j$ can be expressed as

$$f(y_i|X_j;\theta) = \frac{1}{\sqrt{2\pi}\sigma}\exp\Big[-\frac{\beta^2}{2\sigma^2}\Big(x_j - \frac{y_i-\alpha}{\beta}\Big)^2\Big].$$

So that

$$\begin{aligned} L(X|\theta) &= \prod_{i=1}^{n}\sum_{j=1}^{m}\alpha_{ij}f(y_i|x_j;\theta)p_j \\ &= \prod_{i=1}^{n}\sum_{j=1}^{m}\alpha_{ij}\frac{1}{\sqrt{2\pi}\sigma}\exp\Big[-\frac{\beta^2}{2\sigma^2}\Big(x_j - \frac{y_i-\alpha}{\beta}\Big)^2\Big]p_j. \end{aligned} \tag{3.1}$$

The aim is to maximize (3.1) with respect p_j, this may be done via a self-consistent algorithm, based on the idea in Turnbull (1976) and Ding (2008). For $i = 1, 2, \ldots, n$; $j = 1, 2, \ldots, m$, let $\delta_{ij} = I(X_i = x_j)$, the conditional expectation of δ_{ij}, given the observed data, say $\Delta_{ij}(\overrightarrow{p}, \theta)$ is

$$\begin{aligned} \Delta_{ij}(\overrightarrow{p},\theta) &= \mathsf{E}(\delta_{ij}|Y_i = y_i, z_i) = \mathsf{P}(X_i = x_j|Y_i = y_i, z_i) \\ &= \frac{\alpha_{ij}f(y_i|x_j;\theta)p_j}{\sum_{k=1}^{m}\alpha_{ik}f(y_i|x_k;\theta)p_k}. \end{aligned} \tag{3.2}$$

If We treat (3.2) as observed rather than expected frequencies, the proportion of individuals with the covariant equal to X_j, say $\Delta_j^*(\overrightarrow{p}, \theta)$, is

$$\Delta_j^*(\overrightarrow{p},\theta) = \frac{1}{n}\sum_{1}^{n}\Delta_{ij}(\overrightarrow{p},\theta). \tag{3.3}$$

The self-consistency equation for a fixed θ are therefore

$$p_j = \Delta_j^*(\overrightarrow{p},\theta), \qquad j = 1, 2, \ldots, m. \tag{3.4}$$

Note that $\mathsf{E}(\Delta_j^*(\overrightarrow{p},\theta)) = p_j$, but that $\Delta_j^*(\overrightarrow{p},\theta)$ is not a proper estimators for p_j because it depends on unobservable quantities. The solution $\widehat{p}$ of equations (3.4) for a given θ is the unique nonparametric maximum likelihood estimator of p.

Step 2: Estimation of θ assuming p is known

Now consider the maximum of the conditional likelihood function with respect to θ for fixed p, this likelihood can be formalized as

$$L(\theta|p) = \prod_{i=1}^{n}\sum_{j=1}^{m}\alpha_{ij}f(y_i|x_j;\theta)p_j. \tag{3.5}$$

The MLE for θ is obtained by solving the score equation, say $U(\theta)$

$$U(\theta) = \frac{\partial \log L(\theta|p)}{\partial \theta} = 0. \tag{3.6}$$

Let $f_{ij} = \alpha_{ij} f(y_i|X_j;\theta)p_j \,\widehat{=}\, a_{ij}p_j$ with $a_{ij} = \alpha_{ij} f(y_i|X_j;\theta)$, and $\pi_i(\theta) = \sum\limits_{j=1}^{m} f_{ij}(\theta)$, then $L(\theta,p) = \prod\limits_{i=1}^{n} \pi_i(\theta)$, the equation (3.6) is equivalently rewritten as

$$U(\theta) = \sum_{i=1}^{n} \frac{1}{\pi_i(\theta)} \sum_{j=1}^{m} \frac{\partial f_{ij}(\theta)}{\partial \theta} = 0, \tag{3.7}$$

that is

$$\begin{cases} \sum\limits_{i=1}^{n} \dfrac{1}{\pi_i(\theta)} \sum\limits_{j=1}^{m} \alpha_{ij} f_{ij}(\theta) \Big[-\dfrac{1}{2\sigma^2}(y_i - \alpha - \beta x_j)(-2) \Big] = 0, \\ \sum\limits_{i=1}^{n} \dfrac{1}{\pi_i(\theta)} \sum\limits_{j=1}^{m} \sum\limits_{ij} f_{ij}(\theta) \Big[-\dfrac{1}{2\sigma^2}(y_i - \alpha - \beta x_j)(-2x_j) \Big] = 0, \\ \sum\limits_{i=1}^{n} \dfrac{1}{\pi_i(\theta)} \sum\limits_{j=1}^{m} \alpha_{ij} f_{ij}(\theta) \sigma^{-1} \Big[-\dfrac{1}{2}\sigma^{-2} + (y_i - \alpha - \beta x_j)^2 \dfrac{1}{2}\sigma^{-4} \Big] = 0. \end{cases}$$

Straightforward calculations yield

$$\begin{cases} \sum\limits_{i=1}^{n} (y_i - \alpha) - \beta \sum\limits_{i=1}^{n} \dfrac{1}{\pi_i(\theta)} \sum\limits_{j=1}^{m} x_j f_{ij}(\theta) = 0, \\ \sum\limits_{i=1}^{n} \dfrac{y_i - \alpha}{\pi_i(\theta)} \sum\limits_{j=1}^{m} x_j f_{ij}(\theta) - \beta \sum\limits_{i=1}^{n} \dfrac{1}{\pi_i(\theta)} \sum\limits_{j=1}^{m} x_j^2 f_{ij}(\theta) = 0, \\ \sum\limits_{i=1}^{n} \dfrac{1}{\pi_i(\theta)} \sum\limits_{j=1}^{m} (y_i - \alpha - \beta x_j)^2 f_{ij}(\theta) - n\sigma^2 = 0. \end{cases} \tag{3.7*}$$

Since the expected value of X, say u_i, and the variance of X, say v_i^2, given the observed data for the ith individual, can be express as

$$\begin{aligned} \mu_i = \mu_i(\theta,p) &= \mathsf{E}(X|\theta,p,y_i,z_i) \\ &= \sum_{j=1}^{m} \alpha_{ij} x_j \mathsf{P}(X = x_j|\theta,p,y_i,z_i) \\ &= \sum_{j=1}^{m} x_j \frac{\alpha_{ij} f(y_i|x_j;\theta)p_j}{\sum\limits_{j=1}^{n} \alpha_{ij} f(y_i|x_j;\theta)p_j} \\ &= \sum_{j=1}^{m} \frac{x_j f_{ij}(\theta)}{\pi_i(\theta)}. \end{aligned} \tag{3.8}$$

Similarly,

$$v_i^2 = \mathsf{E}[(X - \mu_i)^2|\theta,p,y_i,z_i] = \sum_{j=1}^{m} \frac{(x_j - \mu_j)^2 f_{ij}(\theta)}{\pi_i(\theta)} \tag{3.9}$$

from expressions (3.8) and (3.9), we rewrite (3.7*) as

$$\begin{cases} \sum\limits_{i=1}^{n}(y_i-\alpha)-\beta\sum\limits_{i=1}^{n}\mu_i=0, \\ \sum\limits_{i=1}^{n}(y_i-\alpha)\mu_i-\beta[\mu_i^2+v_i^2], \\ \sum\limits_{i=1}^{n}(y_i-\alpha)^2+\beta^2\sum\limits_{i=1}^{i}(\mu_i^2+v_i^2)-n\sigma^2=0. \end{cases}$$

It follows that

$$\begin{cases} \widehat{\beta}=\Big[\sum\limits_{i=1}^{n}(y_i-\overline{y})(\mu_i-\overline{\mu})\Big]\Big/\Big[\sum\limits_{i=1}^{n}v_i^2+\sum\limits_{i=1}^{n}(\mu_i-\overline{\mu})^2\Big], \\ \widehat{\alpha}=\overline{y}-\widehat{\beta}\overline{\mu}, \\ \widehat{\sigma}^2=\dfrac{1}{n}\sum\limits_{i=1}^{n}[(y_i-\widehat{\alpha})^2+\widehat{\beta}^2(\mu_i^2+v_i^2)], \end{cases} \tag{3.7**}$$

where $\overline{\mu}=(1/n)\cdot\sum\limits_{i=1}^{n}\mu_i$ and $\overline{y}=(1/n)\cdot\sum\limits_{i=1}^{n}y_i$.

Remark 1 It is easy seen that if $T=0$ and Z is non-random variable, then the expressions in (3.7**) will become

$$\begin{cases} \widehat{\beta}=\Big[\sum\limits_{i=1}^{n}(y_i-\overline{y})(x_i-\overline{x})\Big]\Big/\Big[\sum\limits_{i=1}^{n}(x_i-\overline{x})^2\Big], \\ \widehat{\alpha}=\overline{y}-\widehat{\beta}\overline{x}, \\ \widehat{\sigma}^2=\dfrac{1}{n}\sum\limits_{i=1}^{n}[(y_i-\overline{y})^2. \end{cases}$$

This is usually estimation of parameters for simple linear regression.

The EM algorithm can be progressed as follows:

(1) Take initial estimator for $\overrightarrow{p}^{(0)}$, for example $\mathscr{U}$ $\overrightarrow{p}^{(0)}=(1/m,1/m,\ldots,1/m)$.

(2) Take initial estimator for μ_i, v_i and θ, for example $\mu_i(0)=z_i+T/2$, $v_i^{(0)^2}=[T-z_i]^2/2$ and $\theta^{(0)}=\theta^{(0)}(\mu^{(0)},v^{(0)},z^{(0)})$, and evaluate Δ_{ij} and Δ_{ij}^*.

(3) Update $\overrightarrow{p}^{(i+1)}$ use formula (3.4).

§4. Asymptotic Normality Distribution for the Estimator

From the score function (3.7)

$$U(\theta)=\sum_{i=1}^{n}\frac{1}{\pi_i(\theta)}\sum_{j=1}^{m}\frac{\partial f_{ij}(\theta)}{\partial\theta}=0. \tag{4.1}$$

It follows that the observed information matrix is

$$I(\widehat{\theta})=-\mathsf{E}\Big(\frac{\partial^2U(\theta)}{\partial\theta_i\partial\theta_j}\Big)\Big|_{\theta=\widehat{\theta}}, \tag{4.2}$$

where $I(\theta)$ is 3×3 matrix:

$$I(\theta)=-\mathsf{E}\begin{pmatrix}\dfrac{\partial^2 U}{\partial^2\alpha} & \dfrac{\partial^2 U}{\partial\alpha\partial\beta} & \dfrac{\partial^2 U}{\partial\alpha\partial\sigma^2}\\ & \dfrac{\partial^2 U}{\partial^2\beta} & \dfrac{\partial^2 U}{\partial\beta\partial\sigma^2}\\ & & \dfrac{\partial^2 U}{\partial^2\sigma^2}\end{pmatrix}.$$

Let

$$a_{kl}^i=\sum_{j=1}^m\Big[\frac{1}{\pi_i}f_{ij}(y_j-\alpha-\beta x_j)^k x_j^l\Big]=\mathsf{E}_i((y_j-\alpha-\beta x_j)^k x_j^l),$$

straightforward calculations yield

$$\frac{\partial^2 U}{\partial^2\alpha}=-\frac{n}{\sigma^2}+\frac{\beta}{\sigma^4}\sum_{i=1}^n a_{10}^i a_{10}^i, \tag{4.3}$$

$$\frac{\partial^2 U}{\partial\alpha\partial\beta}=-\frac{1}{\sigma^4}\sum_{i=1}^n[a_{01}^i\sigma^2+\beta(a_{12}^i-a_{01}^i a_{11}^i)], \tag{4.4}$$

$$\frac{\partial^2 U}{\partial\alpha\partial\sigma^2}=-\frac{1}{2\sigma^6}\sum_{i=1}^n[a_{20}^i a_{10}^i-a_{30}^i], \tag{4.5}$$

$$\frac{\partial^2 U}{\partial^2\beta}=-\frac{1}{\sigma^4}\sum_{i=1}^n[a_{02}^i\sigma^2+(a_{11}^i)^2-a_{22}^i], \tag{4.6}$$

$$\frac{\partial^2 U}{\partial\beta\partial\sigma^2}=-\frac{1}{2\sigma^6}\sum_{i=1}^n[a_{20}^i a_{11}^i-a_{31}^i], \tag{4.7}$$

$$\frac{\partial^2 U}{\partial^2\sigma^2}=-\frac{1}{4\sigma^6}\sum_{i=1}^n[(a_{20}^i)^2-a_{40}^i]. \tag{4.8}$$

Under the independent interval censoring, standard asymptotic arguments apply. The score components in (4.1) are independent and a central limit theorem applies provided that the censoring mechanism and the covariant are such that the Lindeberge condition holds for the variance of the independent score components. From the expression of variance in (3.9), this condition is

$$\frac{1}{n^{1+c}}\sum_{i=1}^n v_i^{2+\delta}\to 0\qquad(n\to\infty) \tag{4.9}$$

for some constant $c>0$ and $\delta>0$. It is reasonable to assume that as n become large, the average information converges to a positive-definite covariant matrix M. In other words $n^{-1}I(\widehat{\theta})\to M$ as $n\to\infty$, it follows that

$$n^{-1/2}U(\widehat{\theta})\to N(0,M) \tag{4.10}$$

in law, and

$$\sqrt{n}\begin{pmatrix}\widehat{\alpha}-\alpha\\ \widehat{\beta}-\beta\\ \widehat{\sigma^2}-\sigma^2\end{pmatrix}\to N(0,M^{-1}) \tag{4.11}$$

in law, where M can be estimated by $n^{-1}I(\widehat{\theta})$.

References

[1] Ding, B.J., The rate of convergence for the estimation of a distribution with interval censored data, *Chinese Journal of Apllied Probability and Statistics*, **24(5)**(2008), 531–540.

[2] Ding, B.J., The rate of convergence for the estimation of a distribution with interval censored data, *Journal of Systems Science and Complexity*, **28(6)**(2008), 641–648.

[3] Efron, B. and Petrosian, V., Nonparametric methods for double truncated data, *Journal of the American Statistical Association*, **94**(1999), 824–834.

[4] Frydman, H., A note on nonparametric estimation of the distribution function from interval-censored and truncated observation, *Journal of the Royal Statistical Society: Series B*, **56(1)**(1994), 71–74.

[5] Geskus, R. and Groeneboom, P., Asymptotically optimal estimation of smooth functional for interval-censoring, Case 2, *The Annals of Statistics*, **27**(1999), 627–674.

[6] Gentleman, R. and Geyer, C.J., Maximum likelihood for interval censored data: consistency and computation, *Biometrika*, **81(3)**(1994), 618–623.

[7] Groeneboom, P. and Wellner, J.A., *Information Bounds and Nonparametric Maximum Likelihood Estimation*, DMV Seminar Band 19, Birkhauser, Basel, 1992.

[8] Kim, P., The calculation of posterior distributions data augmentation, *Journal of the American Statistical Association*, **27**(1993), 871–878.

[9] Lee, C., An urn model in the simulation of interval censored failure time data, *Statistics and Probability Letter*, **45**(1999), 131–139.

[10] Ren, J., Maximum likelihod for interval censored data, *Statistics in Medicine*, **23**(1989), 3838–3842.

[11] Topp, R. and Gomez, G., Residual analysis in linear regression models with an interval-censored covariate, *Statistics in Medicine*, **23**(2004), 3377–3391.

[12] Turnbull, B.W., The empirical distribution function with arbitrarily grouped, censored and truncated data, *Journal of the Royal Statistical Society: Series B*, **38**(1976), 290–295.

变量为区间截断数据时回归模型的参数估计

丁 帮 俊

(华东师范大学金融与统计学院, 上海, 200241)

在假设自变量X的分布为离散未知分布且样本为区间截断数据而因变量Y是可观察的情况下, 利用EM方法得到了回归参数的极大似然估计, 在一定的条件下估计量的分布为渐近正态的.

关键词: 回归, 区间数据, EM算法, 渐近正态.

学科分类号: O213.7.