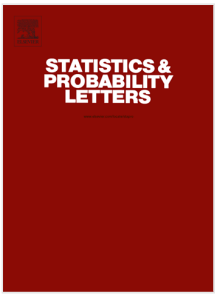
A semiparametric regression estimator under left truncation and right censoring

Maria Karlsson, Thomas Laitila



A semiparametric regression estimator under left truncation and right censoring

Maria Karlsson, Thomas Laitila

PII: S0167-7152(08)00309-X
DOI: 10.1016/j.spl.2008.06.009
Reference: STAPRO 5122

To appear in: *Statistics and Probability Letters*

Received date: 28 February 2006
Revised date: 17 April 2008
Accepted date: 2 June 2008

Please cite this article as: Karlsson, M., Laitila, T., A semiparametric regression estimator under left truncation and right censoring. *Statistics and Probability Letters* (2008), doi:10.1016/j.spl.2008.06.009

A semiparametric regression estimator under left truncation and right censoring

Maria Karlsson *
Department of Statistics, Umeå University, Sweden
Thomas Laitila
Statistics Sweden and Department of Statistics,
Örebro University, Sweden

Abstract

An estimator is proposed for semiparametric linear regression models with left truncated and right censored dependent variables. The estimator is derived from a moment condition following the principles of Newey (2001) on conditional moment conditions. Consistency of the estimator is shown and simulation is used for illustration of the small sample properties.

Key Words: LTRC data; duration; consistency; moment condition.

*Corresponding author: Maria Karlsson, Department of Statistics, Umeå University, SE-901 87 Umeå, Sweden. E-mail: maria.karlsson@stat.umu.se

1 Introduction

Consider a linear regression model for the response variable Y_i^*

$$Y_i^* = X_i^T \beta + \varepsilon_i, \qquad i = 1, 2, ..., n^*, \tag{1}$$

where X_i and β are p-dimensional vectors of explanatory variables and parameters, respectively, and the errors ε_i, defined as $\varepsilon_i = Y_i^* - X_i^T \beta$, are independent and identically distributed random terms with mean zero and constant finite variance.

If the observations of (Y_i^*, X_i^T) are obtained only for the part of the population for which $Y_i^* > t$ and if the observed response variable is $\min\{s, Y_i^*\}$, then data is left truncated and right censored (LTRC). Here t is the known truncation point and s is the known censoring point. For simplicity let $t = 0$ and let Y_i, $i = 1, \ldots, n$, denote the observed response variable, i.e., when $Y_i^* > 0$, $Y_i = \min(s, Y_i^*)$ is observed.

The least squares estimator (LSE) is biased and inconsistent for estimation of β in (1) with LTRC data. The reason is that $E[Y - X^T \beta | X]$ is a function of X and not equal to zero in general. Past research has mostly focused on estimators for data which are either censored or truncated but not both. Reviews of estimators for truncated and censored regression models are found in Lee and Kim (1998) and Honoré and Powell (1994), respectively. It is, however, desirable to develop a new, alternative estimator for LTRC data as well. LTRC data is frequently encountered in studies of durations such as survival or failure times. Duration studies are used in disciplines such as biometrics, epidemiology, and econometrics.

The most widely used regression method for survival data is based on Cox's proportional hazards model (Cox, 1972), which is also applicable for analysing LTRC data. An alternative model to Cox's model for survival analysis is the accelerated failure time (AFT) model, where the logarithm (or another increasing function) of the survival or failure time is regressed on the explanatory variables (Kalbfleisch and Prentice, 1980). The AFT model is therefore a special case of (1). By specifying a distribution, up to unknown parameters, for ε_i in (1), a

maximum likelihood (ML) estimator of β can be defined for LTRC data.

Without any assumption of the parametric distribution of the error term in (1) the model is said to be semiparametric, and an estimator not utilising any parametric assumption is called a semiparametric estimator. Karlsson (2006) considered generalisations of the quadratic mode regression estimator (QME) of Lee (1993) and the winsorized mean estimator (WME) of Lee (1992), two semiparametric estimators for the truncated and the censored regression models, respectively. The purpose of this paper is to propose a semiparametric estimator of regression models using a similar approach as the one used in Karlsson (2006). The formulation of the estimator builds on the principles on moment conditions as presented in Newey (2001), who considers moment conditions for estimation of truncated and censored regressions, respectively. In this paper a moment condition is formulated for the case of LTRC regression data. In effect, the proposed estimator can be interpreted as a combination of the generalizations of the QME and the WME estimators.

Model assumptions and the derivation of the estimator are presented in the next section. Consistency properties are considered in Section 3, and Section 4 includes results from a simulation study. A discussion of the results and topics for future research are given in the final section.

2 The estimator

Consider the conditional moment restriction

$$E[m(Y^* - X^T\beta)|X] = E[m(\varepsilon)|X] = 0, \tag{2}$$

where $m(\cdot)$ is a known scalar function. In the latent model (1) the conditional moment restriction (2) implies the moment condition $E[Xm(Y^* - X^T\beta)] = 0$. This can be interpreted as a first order condition for $E[q(Y^* - X^T\beta)]$, where $q(\varepsilon) = \int_0^\varepsilon m(u)du$, to have a minimum at the true parameter vector β_0. Minimisation of the sample analogue to this expectation yields an estimator of β (cf. Newey, 2001).

Newey (2001) presented conditions on the function $m(\cdot)$ for consistent estimation of semiparametric censored and truncated regression models. For the right censored regression model $m(\varepsilon)$ must be constant for all ε which are large enough and for the left truncated regression model $m(\varepsilon)$ must be zero for all ε which are small enough.

The conditional moment restriction proposed here for linear regression models with LTRC data is

$$E[m(\varepsilon)|X] = E[1[-c_L \leq \varepsilon \leq c_U]\varepsilon + 1[\varepsilon > c_U]c_U|X] = 0, \tag{3}$$

where $c_L > 0$ and $c_U > 0$ are constants chosen by the researcher and $1[A]$ denotes an indicator function such that $1[A] = 1$ if condition A holds, otherwise it is equal to 0.

This moment restriction is a combination of the two conditional moment restrictions used by Karlsson (2006) for estimation of slope parameters of left truncated and right censored regression models. Those estimators are generalizations of the quadratic mode estimator (Lee, 1993) and the winsorized mean estimator (Lee, 1992).

For LTRC data, the moment condition

$$E\left[X1[c_L < X^T\beta < s - c_U]m(Y - X^T\beta)\right] = 0,$$

where $m(\cdot)$ is defined in (3), is a first order condition for a solution to the minimisation of

$$E\left[q\left(Y - \min\left(\max\left(X^T\beta, c_L\right), s - c_U\right)\right)\right] \tag{4}$$

where $q(\cdot)$ is obtained by intergration of $m(\cdot)$ defined in (3) (cf. Newey, 2001). Note that $1[c_L < \alpha < s - c_U]m(Y - \alpha) = -\frac{d}{d\alpha}q(Y - \min\left(\max\left(\alpha, c_L\right), s - c_U\right)$, for a scalar α, except at $\alpha = c_L$ and $\alpha = s - c_U$. An estimator, $\hat{\beta}_{LTRC}$, and objective function $Q(\beta)$ can be derived from the sample analogue to (4). In the present case, the estimator of β and the objective function are

$$\hat{\beta}_{LTRC} = \arg\min_{\beta \in B} Q_n(\beta) \tag{5}$$

where

$$\begin{aligned} Q_n(\beta) &= \sum_{i=1}^{n} Q_i(\beta) \\ &= \arg\min_{\beta\in B} \frac{1}{n}\sum_{i=1}^{n}\{1[c_L < X_i^T\beta < s - c_U]q(Y_i - X_i^T\beta) \\ &+ 1[X_i^T\beta \le c_L]q(Y_i - c_L) + 1[X_i^T\beta \ge s - c_U]q(Y_i - (s - c_U))\}, \end{aligned}$$

and

$$\begin{aligned} q(z) &= \int_0^z m(u)du = 1[-c_L \le z \le c_U]\frac{1}{2}z^2 \\ &+ 1[z < -c_L]\frac{1}{2}c_L^2 + 1[z > c_U](c_U z - \frac{1}{2}c_U^2) \end{aligned}$$

3 Consistency

To establish consistency of the estimator $\hat{\beta}_{LTRC}$ defined in (5), the following assumptions are made:

(A.1) The true unknown parameter vector is denoted β_0. The vector of explanatory variables X_i includes a constant, i.e., β_0 includes a intercept. There exist a unique constant μ such that, when added to the intercept of β_0, the conditional moment restriction (3) with $\varepsilon - \mu = Y - (X^T\beta_0 + \mu)$ as argument of the function $m(\cdot)$ is satisfied. The true parameter vector with the constant μ added to the intercept is denoted β_μ. The vector β_μ belongs to the interior of a compact parameter set $\mathcal{B} \subset \mathbb{R}^p$.

(A.2) The censoring point, s, is larger than $c_L + c_U$.

(A.3) The probability that a latent observation is untruncated is positive, i.e., $1 - F(-X\beta_0) \ge \varphi > 0$, where $F(\cdot)$ denotes the conditional cumulative distribution function of the error term given X.

(A.4) $E[||X_i||^4] < K < \infty$.

(A.5) $\frac{1}{n}\sum_{i=1}^{n} E[1[c_L + \lambda_1 < X^T\beta_\mu < s - c_U - \lambda_2]X_i X_i^T]$ is a positive definite matrix with the smallest eigenvalue bounded from below by $v > 0$ for sufficiently large n and some $\lambda_j > 0, j = 1, 2$.

Most of the assumptions made here are similar to those made by, for example, Powell (1986) and Lee (1993). A.3 and A.4 are included to ensure the existence of finite moments of certain functions used in the proof of consistency. A.5 plays a similar role as the assumption of a non-singular design matrix $X^T X$ in least squares estimation.

A somewhat stringent assumption is the assumption of existence of a unique μ satisfying the moment condition (3). This is similar to the "single crossing" property of the moment conditions considered by Newey (2001). The uniqueness can be obtained from appropriate assumptions on the thresholds and the density function $f(\cdot)$.

Theorem 1 (Strong Consistency) *If (A.1)-(A.5) are satisfied then $\hat{\beta}_{LTRC}$ defined in (5) is a consistent estimator of β_μ, i.e., $\hat{\beta}_{LTRC} \overset{a.s.}{\to} \beta_\mu$.*

The proof of theorem is based on Lemmas 2.2 and 2.3 in White (1980), by first showing convergence of $Q_n(\beta) = \frac{1}{n}\sum_{i=1}^{n} Q_i(\beta)$ in (5) to its expectation, $E[Q_n(\beta)]$, uniformly over $\mathcal{B}$ and then that $E[Q_n(\beta)]$ attains a unique minimum at $\beta = \beta_\mu$. (Details of the proof will be sent by the authors upon request.)

4 Small sample properties

Consistency of an estimator is a desirable property. However, asymptotic properties do not guarantee desirable finite sample behaviour. It is therefore necessary to also study the properties of the estimator in small sample situations. This is done here by means of simulations.

In the simulation study 5000 samples were generated from the latent model

$$Y_i^* = \beta_I + \beta_1 X_{1i} + \beta_2 X_{2i} + 10\varepsilon_i, \tag{6}$$

where X_1 is Uniform(-2.5, 2.5), X_2 is Uniform(0,10) and the slope parameters are 2 and 3. The intercept, β_I, and the censoring point, s, were varied to achieve the same rate of truncation and censoring regardless of error distribution. Two error distributions were considered: the standard normal and the standard Gumbel distributions. For the standard normal distribution $\beta_I = 2$

and $s = 35$ and for the standard Gumbel distribution $\beta_I = -2$ and $s = 40$. The rates of truncation and censoring were both about 10 percent. The subroutine *DUMPOL* of the *IMSL Fortran 90 MP Library* was used to calculate (5).

Results from using two different combinations of thresholds are reported in Table 1. The lower threshold was chosen to equal the error standard deviation in the normal case, adapting to the suggestion by Lee (1993) for the QME estimator. The upper threshold was chosen to be less than the lower threshold by experience from initial simulations. The Mean square error (MSE) and relative bias are given in the table.

The results show that both bias and MSE of the estimator decrease as the sample size increases. This is in accordance with the consistency property of the estimator. For the first set of threshold values ($c_L = 10$ and $c_U = 2$), bias and MSE results are similar between distributions. For the second set of threshold values ($c_L = 10$ and $c_U = 5$), results are much better in the normal case compared with the Gumbel distribution case.

For comparison, the parameters of (6) were also estimated using the maximum likelihood estimator assuming a normal distribution for the dependent variable. The ML estimator under left truncation at 0 and right censoring at s is

$$\hat{\beta}_{ML} = \arg\max_{\beta \in B} L(\beta, \eta | y, X)$$

where

$$L(\beta, \eta | y, X) = \prod_{i=1}^{n} \frac{\left(f(y_i - X_i^T \beta | \eta)\right)^{z_i} \left(1 - F(s - X_i^T \beta | \eta)\right)^{(1-z_i)}}{\left(1 - F(-X_i^T \beta | \eta)\right)} \tag{7}$$

and z_i equals 1 if observation i is uncensored and 0 if it is censored, and $f(\cdot)$ and $F(\cdot)$ denote the probability density and cumulative density function of the error term, respectively.

The relative bias and MSE of the ML estimator are also found in Table 1. As could be expected, the bias and MSE are much smaller for the ML estimator than the LTRC estimator when the likelihood function in (7) is correctly specified, i.e., the normal error case. Also, both bias and MSE of the ML estimator decreases when the sample size increases. However, with a misspecified likeli-

hood function, i.e., the Gumbel case, the bias of the LTRC estimator is lower for the first set, but not for the second set, of threshold values compared to the bias of the ML estimator. When the sample size is 10000 this holds for second set of threshold values too. Note also that for both sets of threshold values the bias of the LTRC decreases when the sample size increases while the bias of the ML estimator does not. For sample sizes $n = 500$ and $n = 1000$ the MSE of the ML estimator is lower than the MSE of the LTRC estimator for both sets of threshold values. The MSE of both the ML and the LTRC estimators decreases but the relative decrease is larger for the LTRC estimator and when $n = 10000$ the MSE is lower for the LTRC estimator, at least for the first set of threshold values.

5 Discussion

This paper contributes with a proposal for a semiparametric estimator of linear regression models with LTRC data. Such an estimator is a valuable complement to parametrically defined estimators since semiparametric estimators are based on milder assumptions on the error distribution. Therefore the estimator can be used, for instance, as an alternative to ML estimators when information on the distribution is scarce.

The estimator proposed is shown to be strongly consistent using mild assumptions on the error distribution and the distribution of the regressors. The finite sample properties of the estimator are studied in a small simulation study. The results obtained on bias and MSE show that the estimator also works in finite samples. However, it is also concluded that the choice of threshold values are important for the properties of the estimator. The first set of thresholds used is associated with much lower bias and MSE measures compared with the second set of thresholds.

The selection of threshold values is rather ad hoc. Here the lower threshold was set based on a suggestion by Lee (1993). The upper thresholds were decided from a small set of simulations in the normal case. It is reasonable to assume

that the efficiency of the estimator can be improved by better choice of threshold values. Thus, additional results on how to select the threshold values are needed.

For future applications it is also necessary to study the asymptotic distribution of the estimator. Histograms of the estimates of the slope parameters show that the sampling distributions converge to bell-shaped forms as the sample size increases in the simulation study. These result indicate that the small sample distribution can be approximated with a normal distribution. It is expected that the estimator can be shown to have an asymptotic normal distribution.

References

Cox, D. R. (1972), Regression Models and Life Tables (with discussion), *Journal of the Royal Statistics Society* **B 34**, 187–220.

Honoré, B. E., Powell, J. L. (1994), Pairwise difference estimators for censored and truncated regression models, *Journal of Econometrics* **64**, 241–278.

Kalbfleisch, J. D., Prentice, R. L. (1980), *The Statistical Analysis of Failure Time Data* (Wiley, New York.).

Karlsson, M. (2006), Estimators of regression parameters for truncated and censored data, *Metrika* **63**, 329-341.

Lee, M.J. (1992), Winsorized mean estimator for censored regression, *Econometric Theory* **8**, 368–382.

Lee, M.J. (1993), Quadratic mode regression, *Journal of Econometrics* **57**, 1–19.

Lee, M. J., Kim, H. (1998), Semiparametric econometric estimation for a truncated regression model: a review with an extension, *Statistica Neerlandica* **52**, 200–225.

Newey, W.K. (2001), Conditional moment restrictions in censored and truncated regression models, *Econometric Theory* **17**, 863–888.

Powell, J.L. (1986), Symmetrically trimmed least squares estimation for tobit

models, *Econometrica* **54**, 1435–1460.

White, H. (1980), Nonlinear regression on cross-section data, *Econometrica* **48**, 721–746.

Table 1: Average MSE and relative bias of estimates of the slope parameters. 5000 replicates.

		n		MSE	Rel. bias
Normal	LTRC, $c_L = 10, c_U = 2$	500	β_1	0.441	0.019
			β_2	0.272	0.022
		1000	β_1	0.211	0.014
			β_2	0.123	0.013
		10000	β_1	0.020	0.002
			β_2	0.012	0.002
	LTRC, $c_L = 10, c_U = 5$	500	β_1	0.549	0.058
			β_2	0.263	0.055
		1000	β_1	0.251	0.035
			β_2	0.119	0.031
		10000	β_1	0.022	0.002
			β_2	0.011	0.002
	ML	500	β_1	0.136	0.002
			β_2	0.046	0.001
		1000	β_1	0.067	0.002
			β_2	0.022	0.001
		10000	β_1	0.007	0.001
			β_2	0.001	-0.003
Gumbel	LTRC, $c_L = 10, c_U = 2$	500	β_1	0.431	0.018
			β_2	0.263	0.028
		1000	β_1	0.209	0.011
			β_2	0.123	0.013
		10000	β_1	0.019	0.001
			β_2	0.013	-0.001
	LTRC, $c_L = 10, c_U = 5$	500	β_1	1.687	0.121
			β_2	0.956	0.145
		1000	β_1	0.717	0.071
			β_2	0.365	0.074
		10000	β_1	0.048	0.002
			β_2	0.018	0.002
	ML	500	β_1	0.285	0.054
			β_2	0.153	0.073
		1000	β_1	0.147	0.055
			β_2	0.103	0.073
		10000	β_1	0.025	0.051
			β_2	0.048	0.068