Metrika (2006) 63: 329–341
DOI 10.1007/s00184-005-0023-x

ORIGINAL ARTICLE

Maria Karlsson

Estimators of regression parameters for truncated and censored data

Received: 9 January 2003 / Accepted: 28 January 2005 / Published online: 15 February 2006

Abstract Estimators of parameters in semi-parametric left truncated and right censored regression models are proposed. In contrast to the majority of existing estimators, the proposed estimators do not require the error term of the regression model to have a symmetric distribution. In addition the estimators use asymmetric "trimming" of observations. Consistency and asymptotic normality of the estimators are shown. Finite sample properties are considered in a small simulation study. For the left truncated case, an empirical application illustrates the usefulness of the estimator.

Keywords Truncated · Censored · Semi-parametric · Regression

1 Introduction

Several estimators of semi-parametric truncated and censored regression models have been suggested. Powell (1994) gives a survey on theoretical developments on estimation of semi-parametric models. Lee and Kim (1998) review some estimators of left truncated regression models and present results from a simulation study on the properties of the estimators. They find that the symmetrically trimmed least squares (STLS) estimator (Powell 1986), the quadratic mode (QME) estimator (Lee 1993), and the cosine (COS) estimator (Lee and Kim 1998) all perform well.

Suggested estimators of semi-parametric censored regression models include Buckley and James (1979), the censored least absolute deviation (CLAD) estimator (Powell 1984), the symmetrically censored least squares (SCLS) estimator (Powell 1986), the identically censored least absolute deviations (ICLAD) and the identically censored least squares (ICLS) estimator (Honoré and Powell 1994). Honoré and Powell (1994) present results from a simulation study where the CLAD,

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

SCLS, and ICLAD estimators perform best in terms of bias, mean square error, and median absolute deviation. An empirical application of the CLAD, SCLS, and ICLAD estimators is found in Chay and Powell (2001). Another estimator of censored regression models is the winsorized mean estimator (WME) (Lee 1992), which includes CLAD as a special case and is related to the SCLS. One of the advantages of the WME over the CLAD is that the asymptotic covariance matrix of the WME is easier to estimate.

Most of the above mentioned estimators of truncated and censored regression models are based on symmetry assumptions placed on the distribution of the error terms in the models. However, Laitila (2001) and Newey (2001) show that the QME of slope parameters is consistent under asymmetrically distributed errors as well. Newey (2001) also derives a similar result for the WME. Both the QME and the WME were first derived under the assumption of symmetry and the definitions of the estimators amounting to "symmetric trimming" of observations is due to this assumption. In this paper, asymmetric trimming is suggested and corresponding estimators are derived. Asymptotic results are derived using the results presented by Newey (2001), and finite sample properties are illustrated within a small simulation study and by an empirical application modelling travel distance.

The truncated and censored regression models are introduced in the next section and the proposed estimators are defined in section 3. Simulation results on the finite sample properties are presented in section 4. The estimator of truncated regression models is used in a small empirical example in section 5. A concluding discussion is given in section 6.

2 Models and assumptions

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

$$Y_i^* = X_i^T \beta_0 + \varepsilon_i, \qquad i = 1, 2, \ldots, n^*, \tag{1}$$

where X_i and β_0 are p-dimensional vectors of explanatory variables and parameters, respectively, and the ε_i are independent and identically distributed error terms.

In a left truncated regression model, observations of $\left(Y_i^*, X_i\right)$ are obtained only for the part of the population for which $Y_i^* > t_i$, where t_i is a known truncation point. For simplicity let $t_i = 0$. In a right censored regression model, the observed response variable is $\min\{s_i, Y_i^*\}$, where s_i is the known censoring point. Let $E^*[\cdot]$ and $P^*[\cdot]$ denote expectation and probability in the latent regression model (1), while $E[\cdot]$ and $P[\cdot]$ denote the counterparts under truncation and censoring.

The ordinary least squares (OLS) estimator is biased and inconsistent for estimating truncated and censored regression models because $E[\varepsilon|X]$ is a function of X and not equal to zero. Several alternative semi-parametric estimators have been proposed by Miller (1976), Buckley and James (1979), Powell (1984, 1986), Lee (1992, 1993), Honoré and Powell (1994), Lee and Kim (1998), and others.

Newey (2001) defines a class of estimators derived through a conditional moment restriction

$$E^*[m(Y^* - X^T \beta_0)|X] = E^*[m(\varepsilon)|X] = 0, \tag{2}$$

where $m(\cdot)$ is a known scalar function. The conditional moment restriction is regarded as the first order condition to a minimisation problem defining the estimator as the minimum of the corresponding objective function, obtained by "integrating back from" $m(\varepsilon)$. For instance with $m(\varepsilon) = 1[-c \leq \varepsilon \leq c] \cdot \varepsilon$ for the left truncated regression model, the QME,

$$\hat{\beta}_{QME} = \arg\min_{\beta \in B} \frac{1}{n} \sum_{i=1}^{n} 1\left[-c < Y_i - \max\left(X_i^T \beta, c\right) < c\right] \times \left(\left\{Y_i - \max\left(X_i^T \beta, c\right)\right\}^2 - c^2\right), \tag{3}$$

is obtained. A similar example for the censored regression model is $m(\varepsilon) = 1[-c \leq \varepsilon \leq c] \cdot \varepsilon + 1[\varepsilon > c] \cdot c - 1[\varepsilon < -c] \cdot c$ used by Lee (1992) to define the WME,

$$\begin{aligned} \hat{\beta}_{WME} = \arg\min_{\beta \in B} \frac{1}{n} \sum_{i=1}^{n} & 1\left[\left|Y_i - \max\left(X_i^T \beta, c\right)\right| < c\right] \\ & \times \left(0.5\left(Y_i - \max\left(X_i^T \beta, c\right)\right)^2\right) \\ & + 1\left[\left|Y_i - \max\left(X_i^T \beta, c\right)\right| \geq c\right] \\ & \times \left(c\left|Y_i - \max\left(X_i^T \beta, c\right)\right| - 0.5c^2\right) \end{aligned} \tag{4}$$

for left censoring at $s_i = 0$.

The idea of using a symmetric "window" $\pm c$ when defining the QME and the WME, is that a symmetric window together with a unimodal and symmetric density, for the error term, implies the moment condition $E[m(\varepsilon)|X] = 0$. However, the QME and the WME are both consistent for the slope parameters under asymmetrically distributed errors as well (Laitila 2001; Newey 2001). Thus, the use of a symmetric window $\pm c$ does not seem to be necessary for defining consistent estimators.

Estimators based on asymmetric windows are suggested in the next section. To show consistency and asymptotic normality of the proposed estimators the following assumptions, which correspond to the assumptions of Theorem 5.1 in Newey (2001), are used:

(A1) Let $\wp_X$ denote the probability distribution of X and U denote Lebesgue measure. (ε_i, X_{iT}) is independent and identically distributed (i.i.d.) with distribution that is absolutely continuous with respect to $U \times \wp_X$. ε_i and $X_i = (1, x_i^T)^T$ are independent and X_i belongs to a bounded set.
(A2) $\beta_0 \in \text{interior}(B)$, B is compact.
(A3) There is $f_\varepsilon(\varepsilon)$ such that the density function of the error term $f(\varepsilon) = \int_{-\infty}^{\varepsilon} f_\varepsilon(u)du$ and $\int \left[(f_\varepsilon(\varepsilon))^2 / f(\varepsilon)\right] d\varepsilon < \infty$.
(A4) $f(\varepsilon)$ is such that it exists a unique μ satisfying $E^*[m(\varepsilon - \mu)] = 0$ and such that the derivative $d^*(X) = \frac{\partial}{\partial \alpha} E^*[m(\varepsilon - \mu + \alpha)]|_{\alpha=0}$ is positive.

For the left truncated regression model the following additional assumptions are used:

(A5a) $P(X^T\beta_0 + \mu = -\ell) = 0$, where μ satisfies $E^*[m(\varepsilon - \mu)] = 0$ and $\ell = \sup\{\bar{\varepsilon} : m(\varepsilon) = 0 \quad \forall \quad \varepsilon \leq \bar{\varepsilon}\}$,
(A6a) $E^*[1[X^T\beta_0 + \mu > -\ell]XX^T]$ exists and is non-singular, and
(A7a) $P^*(Y^* > 0|X) \geq \tau > 0$.

For the right censored regression model

(A5b) $P(X^T\beta_0 + \mu = -\ell + s) = 0$, where $\ell = \inf\{\bar{\varepsilon} : m(\varepsilon) = m(\bar{\varepsilon}) \quad \forall \quad \varepsilon \geq \bar{\varepsilon}\}$, and
(A6b) $E^*[1[X^T\beta_0 + \mu < -\ell + s]XX^T]$ exists and is non-singular

are assumed in addition to Assumptions A1–A4.

3 Estimators

In this section, estimators based on asymmetric windows are suggested for the estimation of the left truncated regression model and the right censored regression model. The scalar function used in the truncated case is $m(\varepsilon) = 1[-c_L \leq \varepsilon \leq c_U] \cdot \varepsilon$ while $m(\varepsilon) = 1[-c_L < \varepsilon < c_U] \cdot \varepsilon + 1[\varepsilon \geq c_U] \cdot c_U - 1[\varepsilon \leq -c_L] \cdot c_L$ is used in the censored case. Here c_L and c_U are finite positive constants chosen by the researcher. The functions used to obtain the QME and the WME are special cases of these functions with $c_L = c_U = c$. However, the use of an asymmetric window makes these estimators more flexible than the QME and the WME. More observations can, with the use of an asymmetric window, contribute with possible information to the estimator instead of being trimmed. This might lead to estimators that are more efficient.

For the left truncated regression model let

$$E^*[m(\varepsilon)|X] = E^*[1[-c_L \leq \varepsilon \leq c_U] \cdot \varepsilon|X] = 0 \tag{5}$$

define the conditional moment restriction in the latent regression. The corresponding estimator, defined as the minimum of the objective function $q(\varepsilon) = \int_0^\varepsilon m(u)du + C$, where C can be any constant (Newey 2001), is

$$\begin{aligned}\hat{\beta}_{LT} = \arg\min_{\beta\in B} \sum_{i=1}^{n} & 1[X_i^T\beta > c_L] \cdot \{1[-c_L \leq \varepsilon_i \leq c_U] \cdot \frac{1}{2}\varepsilon_i^2 \\ & + 1[\varepsilon_i < -c_L] \cdot \frac{1}{2}c_L^2 + 1[\varepsilon_i > c_U] \cdot \frac{1}{2}c_U^2\} \\ & + [X_i^T\beta \leq c_L] \cdot \{1[-c_L \leq Y_i - c_L \leq c_U] \cdot \frac{1}{2}(Y_i - c_L)^2 \\ & + 1[Y_i - c_L < -c_L] \cdot \frac{1}{2}c_L^2 + 1[Y_i - c_L > c_U] \cdot \frac{1}{2}c_U^2\}\end{aligned} \tag{6}$$

If Assumptions (A1)–(A4) and (A5a)–(A7a) are satisfied, then the function $m(\varepsilon)$ in expression (5) has the following properties:

(P1) $E^*[m(\varepsilon - \mu + \alpha)] \geq (\leq)0$ for $\alpha \geq (\leq)0$.
(P2) $m(\varepsilon)$ is bounded and continuous almost everywhere.

(P3) As a function of α, $E^*[(m(\varepsilon-\mu+\alpha))^2]$ is bounded in a neighbourhood of every α.
(P4) $E^*[(m(\varepsilon-\mu))^2]>0$.
(P5) $E^*[||X||^2(d(X))^2/E^*[(m(\varepsilon-\mu))^2]]<\infty$, where $d(X)=\frac{\partial}{\partial\alpha}E[m(\varepsilon-\mu+\alpha)]|_{\alpha=0}$.
(P6) $Q=E^*[d(X)1[X^T\beta_0+\mu>c_L]XX^T]$ exists and is non-singular, where $d(X)=\frac{\partial}{\partial\alpha}E[m(\varepsilon-\mu+\alpha)]|_{\alpha=0}$.

Proofs of (P1)–(P6) are given in Appendix A.

By Theorem 5.1 in Newey (2001), since (P1)–(P6) are satisfied, the LT estimator, $\hat{\beta}_{\mathrm{LT}}$, is $\sqrt{n}$-consistent for the slope parameters and $\sqrt{n}(\hat{\beta}_{\mathrm{LT}}-\tilde{\beta}_0)$ converges in distribution to $N(0,V_{\mathrm{LT}})$, where $\tilde{\beta}_0$ denotes the true parameter vector with the constant μ added to the intercept. The estimator is also $\sqrt{n}$-consistent for the intercept plus μ. The asymptotic covariance matrix is $V_{\mathrm{LT}}=\frac{1}{d^2}\sigma^2M^{-1}$, where $M=E[\frac{1[X^T\beta_0+\mu>c_L]XX^T}{\Pi(X)}]$, $d^*=\frac{\partial}{\partial\alpha}E^*[m(\varepsilon-\mu+\alpha)]|_{\alpha=0}$, $\sigma^2=E^*[(m(\varepsilon-\mu))^2]$, and $\Pi(X)=E^*[1[Y^*>0]|X]$.

For the right censored regression, consider the conditional moment restriction

$$E^*\left[1[-c_L<\varepsilon<c_U]\cdot\varepsilon+1[\varepsilon\geq c_U]\cdot c_U-1[\varepsilon\leq -c_L]\cdot c_L|X\right]=0 \tag{7}$$

The estimator obtained based on moment restriction (7) is

$$\begin{aligned}\hat{\beta}_{RC}=\ &\arg\min_{\beta\in B}\sum_{i=1}^{n}1[X_i^T\beta<-c_U+s]\cdot\left\{1[-c_L<\varepsilon_i<c_U]\cdot\frac{1}{2}\varepsilon_i^2\right.\\
&+1[\varepsilon_i\geq c_U]\cdot\left(\varepsilon_i\cdot c_U-\frac{1}{2}c_U^2\right)-1[\varepsilon_i\leq -c_L]\cdot\left.\left(\varepsilon_i\cdot c_L+\frac{1}{2}c_L^2\right)\right\}\\
&+[X_i^T\beta\geq -c_U+s]\cdot\left\{1[-c_L<Y_i+c_U-s<c_U]\cdot\frac{1}{2}(Y_i+c_U-s)^2\right.\\
&+1[Y_i+c_U-s\geq c_U]\cdot\left((Y_i+c_U-s)\cdot c_U-\frac{1}{2}c_U^2\right)\\
&\left.-1[Y_i+c_U-s\leq -c_L]\cdot\left((Y_i+c_U-s)\cdot c_L+\frac{1}{2}c_L^2\right)\right\}\end{aligned} \tag{8}$$

If the Assumptions (A1)–(A4) and (A5b)–(A6b) are satisfied then the function $m(\varepsilon)$ in (7) has the properties (P1)–(P6), with Q in (P6) defined as $Q=E^*[d(X)1[X^T\beta_0+\mu<-c_U+s]XX^T]$. Proofs are in Appendix B.

By Theorem 5.1 in Newey (2001) the RC estimator, $\hat{\beta}_{\mathrm{RC}}$, is $\sqrt{n}$-consistent for the slope parameters and the intercept parameter plus μ and $\sqrt{n}(\hat{\beta}_{\mathrm{RC}}-\tilde{\beta}_0)$ converges in distribution to $N(0,V_{\mathrm{RC}})$. The asymptotic covariance matrix is $V_{\mathrm{RC}}=\frac{1}{d^2}\sigma^2M^{-1}$, where $M=E[1[X^T\beta_0+\mu<-c_U+s]XX^T]$, $d^*=\frac{\partial}{\partial\alpha}E^*[m(\varepsilon-\mu+\alpha)]|_{\alpha=0}$, and $\sigma^2=E^*[(m(\varepsilon-\mu))^2]$.

Note that the asymptotic results of the LT and RC estimators are based on the assumption that there exists a unique μ satisfying $E^*[m(\varepsilon-\mu)]=0$ (Assumption A4). This can be made valid by placing appropriate restrictions on c_U, c_L, and the density $f(\varepsilon)$. For instance, in the LT estimator case, suppose $f(\varepsilon)$ is strict unimodal and positive for all ε, and define μ' by the restriction $f(\mu')=f(-c_L+\mu')$. If

there is a constant $\theta \in (0, 1)$ such that $f(-\theta c_L + \mu)/f(\theta c_L + \mu) > \gamma > 1$ for all $\mu > \mu'$, then $E^*[m(\varepsilon - \mu)] = 0$ holds if $c_L < c_U < c_L \cdot \sqrt{1 + (1 - \theta^2)(\gamma - 1)}$.

For the function $m(\varepsilon)$ used to derive the RC estimator a sufficient condition for Assumption A4 is, $F(c_U + \mu) - F(-c_L + \mu) > 0$ for all μ.

4 Finite sample properties

Finite sample properties of the LT and RC estimators were studied by means of simulation. Samples of sizes $n = 200$, 500, and 1,000 were generated from the latent model

$$Y_i^* = \beta_I + \beta_1 X_{1i} + \beta_2 X_{2i} + \sigma\varepsilon_i, \tag{9}$$

where $X_1 \sim \text{Uniform}(-2.5, 2.5)$, $X_2 \sim \text{Uniform}(0, 10)$, $\beta_1 = 2, \beta_2 = 3$. The intercept β_I was varied to achieve the same levels of truncation or censoring for the different error distributions used. The parameter σ was adjusted such that the coefficient of determination, R^2, is 0.45 regardless of error distribution. Three error distributions were considered: the standard normal distribution, the double exponential distribution, and the extreme value distribution (standard Gumbel distribution for maxima) for the truncated regressions. For the censored regressions the Gumbel distribution for minima was used instead of the Gumbel distribution for maxima to achieve a larger effect of the asymmetric density. The truncation point, t, was set to zero and the censoring point, s, was set to 35. The QME and the WME were included for comparison.

Lee and Kim (1998) suggest that the threshold c in expression (3) for QME is assigned the value of the observed standard deviation of the response variable. Here the observed standard deviation of the response variable conditional on the explanatory variables (the standard deviation of the residuals) $std(\hat{\varepsilon})$ was used. The same choice of c was used in (4) for WME. c_L and c_U in expression (6) for the LT estimator were assigned $std(\hat{\varepsilon})$ and $2 \cdot std(\hat{\varepsilon})$ respectively and $c_L = 2 \cdot std(\hat{\varepsilon})$ and $c_U = std(\hat{\varepsilon})$ in (8) for the RC estimator. The $std(\hat{\varepsilon})$ was based on the OLS residuals. The subroutine *DUMPOL* of the *IMSL Fortran 90 MP Library* was used to calculate (6) and (8).

Table 1 reports the results on relative bias and relative root mean square error (MSE) of the estimators of the slope parameters of model (9). The bias and the MSE of both the LT and RC estimators decrease for all error distributions when the sample size increase; this was expected from the asymptotic results.

For normal distributed errors the LT estimator has smaller bias than the QME estimator and for the two largest sample sizes the QME has a MSE (in absolute numbers) almost twice the size of the MSE of the LT estimator. For double exponential distributed errors better results are found for QME than LT but the difference is rather small. For the extreme value distributed errors better results are found for QME than LT. The bias and the MSE of the RC estimator are smaller than the bias and the MSE of the WME estimator for both normal and gumbel distributed errors, with some few exceptions. For double exponential distributed errors the difference in results between the two estimators is small. Noteworthy is that the bias, especially of $\hat{\beta}_2$, of all four estimators are high for the extreme value distributed errors when $n = 200$.

Table 1 Average relative bias and average relative root mean square error (MSE) of estimators of slope parameters of model (9)

			QME		LT		WME		RC	
	n		Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE
Normal	200	$\hat{\beta}_1$	0.061	0.497	0.041	0.466	0.030	0.311	0.022	0.293
		$\hat{\beta}_2$	0.075	0.206	0.107	0.210	0.050	0.171	0.036	0.149
	500	$\hat{\beta}_1$	0.032	0.362	0.017	0.297	0.012	0.189	0.008	0.177
		$\hat{\beta}_2$	0.048	0.161	0.043	0.131	0.021	0.103	0.015	0.092
	1,000	$\hat{\beta}_1$	0.022	0.286	0.010	0.208	0.008	0.134	0.004	0.126
		$\hat{\beta}_2$	0.030	0.124	0.017	0.089	0.010	0.072	0.007	0.064
Double exp.	200	$\hat{\beta}_1$	0.024	0.249	0.028	0.293	0.015	0.237	0.017	0.245
		$\hat{\beta}_2$	0.026	0.114	0.050	0.135	0.032	0.128	0.029	0.127
	500	$\hat{\beta}_1$	0.011	0.164	0.008	0.175	0.005	0.148	0.006	0.153
		$\hat{\beta}_2$	0.012	0.076	0.019	0.085	0.012	0.079	0.012	0.079
	1,000	$\hat{\beta}_1$	0.003	0.114	0.004	0.122	0.003	0.103	0.003	0.107
		$\hat{\beta}_2$	0.005	0.056	0.008	0.059	0.005	0.056	0.005	0.055
Gumbel	200	$\hat{\beta}_1$	0.167	0.618	0.138	0.860	0.092	0.433	0.088	0.452
		$\hat{\beta}_2$	0.208	0.342	0.363	0.489	0.123	0.283	0.117	0.270
	500	$\hat{\beta}_1$	0.086	0.401	0.090	0.558	0.030	0.245	0.025	0.255
		$\hat{\beta}_2$	0.119	0.233	0.198	0.316	0.043	0.161	0.040	0.153
	1,000	$\hat{\beta}_1$	0.046	0.285	0.056	0.396	0.012	0.172	0.012	0.179
		$\hat{\beta}_2$	0.072	0.165	0.117	0.225	0.021	0.111	0.019	0.104

5,000 replicates. 20–25% truncation or censoring

To study the affect of the number of explanatory variables on the performance of the estimators samples of sizes $n = 500$ and $n = 1{,}000$ were generated from model (9) expanded with one and two explanatory variables, respectively, under normal distributed errors. These models were

$$Y_i^* = \beta_I + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \sigma\varepsilon_i, \tag{10}$$

$$Y_i^* = \beta_I + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_4 X_{4i} + \sigma\varepsilon_i, \tag{11}$$

$$Y_i^* = \beta_I + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \sigma\varepsilon_i, \tag{12}$$

where $X_3 \sim \text{Normal}(0, \sqrt{25/12})$ and X_4 is an indicator variable with $P(X_4 = 1) = 0.6$. The variance of X_3 was chosen equal to the variance of X_1 and $\beta_3 = \beta_1 = 2$ for comparison. Similarly, $\beta_4 = 5.89256$ was chosen such that the variance of $\beta_4 X_4$ is equal to the variance of $\beta_1 X_1$. The intercept β_I was varied such that the same levels of truncation or censoring were achieved and σ was adjusted such that $R^2 = 0.45$ for all models.

Results on relative bias and relative root MSE of the estimators of the slope parameters are reported in Tables 2–4. For models (10) and (11) the bias and MSE of the LT estimator are smaller than the bias and MSE of the QME estimator for all but a few cells. For model (12) the LT estimator has larger bias but smaller MSE

Table 2 Average relative bias and average relative root MSE of estimators of slope parameters of model (10)

			QME		LT		WME		RC	
	n		Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE
Normal	500	$\hat{\beta}_1$	0.036	0.376	0.034	0.322	0.009	0.199	0.005	0.188
		$\hat{\beta}_2$	0.040	0.143	0.061	0.142	0.023	0.101	0.017	0.090
		$\hat{\beta}_3$	0.033	0.374	0.030	0.317	0.014	0.197	0.011	0.185
	1,000	$\hat{\beta}_1$	0.029	0.290	0.015	0.218	0.004	0.136	0.002	0.129
		$\hat{\beta}_2$	0.025	0.114	0.023	0.094	0.011	0.070	0.008	0.063
		$\hat{\beta}_3$	0.021	0.280	0.015	0.219	0.009	0.139	0.008	0.132

5,000 replicates. 20–25% truncation or censoring

Table 3 Average relative bias and average relative root MSE of estimators of slope parameters of model (11)

			QME		LT		WME		RC	
	n		Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE
Normal	500	$\hat{\beta}_1$	0.051	0.388	0.021	0.312	0.011	0.197	0.008	0.188
		$\hat{\beta}_2$	0.041	0.147	0.054	0.135	0.022	0.099	0.017	0.088
		$\hat{\beta}_3$	0.035	0.350	0.021	0.316	0.014	0.188	0.012	0.180
	1,000	$\hat{\beta}_1$	0.034	0.294	0.014	0.217	0.006	0.138	0.003	0.131
		$\hat{\beta}_2$	0.029	0.117	0.022	0.090	0.012	0.070	0.010	0.063
		$\hat{\beta}_3$	0.034	0.274	0.011	0.220	0.007	0.134	0.006	0.126

5,000 replicates. 20–25% truncation or censoring

Table 4 Average relative bias and average relative root MSE of estimators of slope parameters of model (12)

			QME		LT		WME		RC	
	n		Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE
Normal	500	$\hat{\beta}_1$	0.035	0.394	0.047	0.325	0.015	0.209	0.011	0.195
		$\hat{\beta}_2$	0.035	0.138	0.065	0.141	0.025	0.103	0.018	0.091
		$\hat{\beta}_3$	0.037	0.392	0.041	0.332	0.021	0.207	0.015	0.192
		$\hat{\beta}_4$	0.037	0.336	0.024	0.319	0.015	0.194	0.013	0.184
	1,000	$\hat{\beta}_1$	0.019	0.301	0.022	0.224	0.007	0.144	0.003	0.135
		$\hat{\beta}_2$	0.026	0.110	0.027	0.091	0.011	0.070	0.008	0.063
		$\hat{\beta}_3$	0.030	0.300	0.018	0.226	0.011	0.143	0.008	0.134
		$\hat{\beta}_4$	0.029	0.266	0.007	0.230	0.007	0.138	0.006	0.133

5,000 replicates. 20–25% truncation or censoring

than the QME estimator. Both bias and MSE of the RC estimator are smaller than bias and MSE of the WME estimator for all three models.

The results are similar to the results in Table 1 for model (9) and the bias and the MSE decrease when the sample sizes increases. Thus, the number of explanatory variables does not appear to affect the performance of the estimator.

Therefore, despite that the choice of threshold values was made more or less roughly, the proposed estimators with asymmetric windows are in general better compared to the corresponding estimators with symmetric windows. However, there are exceptions such as the truncated case with Gumbel distributed errors where the QME which may be less sensitive to deviating observations, performed better.

5 Travel distance modelling

To illustrate the estimation methods, this section presents a problem concerning the analysis of left truncated data. Data was collected in the recurrent Swedish Travel Habit Survey and consists of 3,824 shopping trips by car.

The response variable, Y, is the logarithm of self reported travelling distance. To diminish the risk that travellers forget reporting short travels the travelling distances are truncated at 2 km. The response is transformed such that the truncation point, t equals zero. A linear regression model with explanatory variables; X_1: age of traveller; X_2: sex of traveller (0=woman, 1=man); X_3: logarithm of annual income (in thousands of Swedish krona) of traveller; X_4 and X_5, indicators for travellers living in urban areas and less densely populated areas (reference group is travellers living in large cities), are estimated with the LT estimator and, for comparison, the QME estimator and a maximum likelihood (ML) estimator assuming normal errors. Values on c_L, c_U, and c are based on the same rules as in the simulation study. The choice of model is in accordance with a model studied by Brännäs and Laitila (1991).

In the survey, some travellers reported no income. Therefore, an approach suggested by Battese (1997) is used for the modelling of income effects. An indicator variable, X_6, for having a positive income is included in the model and X_3 is set to zero for individuals without income.

Estimated parameters are presented in Table 5 along with standard errors. The standard errors of the LT and QME estimates are estimated using a bootstrap procedure because their covariance matrices depend on the density of the error distribution and are difficult to estimate. Bootstrap have been used by Buchinsky (1995) and Hahn (1995) in different but similar situations to avoid difficulties of estimation of covariance matrices with density components. Results in Karlsson (2004) support the use of the bootstrap technique for estimation of the QME covariance matrix.

The estimated parameters for age, sex, and income have the same sign regardless of estimator and the difference in parameter estimates are small between the

Table 5 Parameter estimates (standard errors in parenthesis) for the travel distance model

	Intercept	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$
LT	2.538	−0.006	0.087	−0.260	−0.037	0.579	0.199
	(0.191)	(0.003)	(0.095)	(0.143)	(0.122)	(0.244)	(0.172)
QME	1.651	−0.004	0.136	−0.535	−0.126	−0.381	0.074
	(0.142)	(0.003)	(0.098)	(0.276)	(0.110)	(0.489)	(0.151)
ML	1.968	−0.004	0.094	−0.277	−0.064	0.064	0.044
	(0.075)	(0.001)	(0.043)	(0.073)	(0.048)	(0.066)	(0.075)

ML and LT estimators. The estimated impact of age is negative, i.e. a higher age reduce the distance travelled for shopping purposes. Men have longer shopping travel distances than women. The income effect is negative. Brännäs and Laitila (1991) suggest that the negative income effect on shopping trip distance is due to a time restriction when income has a positive effect on work and business trips. The ML estimator has the smallest estimated standard errors, but the standard error is relatively small for the LT and QME estimators too.

The sign of the effects of the indicator variables X_4 and X_5 are as expected. Travellers living in urban areas travel in general shorter distances than travellers living in the largest cities. Living in the least densely populated areas prolongs the distance travelled for shopping purposes compared to living in large cities, according to the LT and ML estimates. The QME estimate indicated the opposite but has a large standard error.

6 Discussion

The new estimators of the left truncated and right censored regression models have desirable asymptotic properties such as consistency and asymptotic normality. The estimators proposed are not based on symmetry assumptions on the error distributions, whereby they are of more general applicability than many earlier estimators proposed. In addition, the results of the simulation study show that the estimators behave well in finite samples with respect to bias and MSE. The results also indicate that the LT and RC estimators with asymmetric windows have a potential to be more efficient and have smaller bias than the QME and WME estimators with symmetric windows, although the choice of threshold values, c_L and c_U, in the simulation study were made more or less roughly.

The results of the simulation study indicate that rather large sample sizes are necessary for good performance of the estimators since the bias of the estimators were rather high, when the sample size was only 200. This was especially clear for extreme value distributed errors. Perhaps, this result would improved if c_L and c_U were chosen more carefully. As pointed out in section 3, the choice of c_L and c_U might in some situations be important for the single crossing property of the conditional moment restriction, e.g. c_U should be bounded.

In the example in section 5 the estimated parameters are reasonable on the basis of theory on travel distances.

The covariance matrix estimation for the LT estimator is complicated by the dependence on the density of the error distribution. This is also the case for the QME estimator. The solution used here is the use of bootstrap methods. Simulation results in Karlsson (2004) support the use of the bootstrap technique for covariance matrix estimation for the QME. The empirical example indicate that such a solution might also work for the LT estimator. However, more research on the properties of the method is needed. The asymptotic covariance matrix for the RC estimator do not include density components and is easier to estimate.

Finally, data is often both left truncated and right censored (LTRC). A semiparametric estimator of linear regression models with LTRC data might be obtained by combining the two conditional moment restrictions used to obtain the LT and RC estimators. Studies on the properties of such an estimator are desirable.

Acknowledgements The author is thankful to the referees for their suggestions and helpful comments.

Appendix A

Proofs of (P1)–(P6) for $m(\varepsilon) = 1[-c_L \le \varepsilon \le c_U] \cdot \varepsilon$

(P1) $E^*[m(\varepsilon - \mu + \alpha)] \ge (\le) 0$ for $\alpha \ge (\le) 0$

Proof $E^*[m(\varepsilon - \mu)]$ is a continuous function by Assumption A3 and μ satisfying $E^*[m(\varepsilon - \mu)] = 0$ is unique by A4. P1 follows.

(P2) $m(\varepsilon - \mu)$ is bounded and continuous almost everywhere.

Proof $m(\varepsilon - \mu) = 1[-c_L \le \varepsilon - \mu \le c_U] \cdot (\varepsilon - \mu)$ is larger than or equal to $-c_L$ and smaller than or equal to c_U (hence bounded). $m(\varepsilon)$ is continuous for all ε except at $\varepsilon = -c_L$ and $\varepsilon = c_U$. The set of discontinuities is of Lebesgue measure zero.

(P3) As a function of α, $E^*[(m(\varepsilon - \mu + \alpha))^2]$ is bounded in a neighbourhood of every α.

Proof

$$E^*[(m(\varepsilon - \mu + \alpha))^2] = E^*\left[1[-c_L + \mu - \alpha \le \varepsilon \le c_U + \mu - \alpha] \cdot (\varepsilon - \mu + \alpha)^2\right] \le \max\left\{c_U^2, c_L^2\right\} \qquad \text{for all } \alpha$$

(P4) $E^*[(m(\varepsilon - \mu))^2] > 0$.

Proof

$$\begin{aligned} E^*[(m(\varepsilon - \mu))^2] = {} & E^*[1[-c_L \le \varepsilon - \mu \le c_U] \cdot (\varepsilon - \mu)^2] \\ = {} & E^*[1[-c_L \le \varepsilon - \mu \le 0] \cdot (\varepsilon - \mu)^2] \\ & + E^*[1[0 < \varepsilon - \mu \le c_U] \cdot (\varepsilon - \mu)^2] \\ & \ge \{E^*[1[-c_L \le \varepsilon - \mu \le 0] \cdot (\varepsilon - \mu)]\}^2 \\ & + \{E^*[1[0 < \varepsilon - \mu \le c_U] \cdot (\varepsilon - \mu)]\}^2 \end{aligned}$$

by Jensen's inequality. $E^*[1[-c_L \le \varepsilon - \mu \le 0] \cdot (\varepsilon - \mu)] < 0$ and $E^*[1[0 < \varepsilon - \mu \le c_U] \cdot (\varepsilon - \mu)] > 0$ because μ satisfying $E^*[m(\varepsilon - \mu)] = 0$ is unique and $E^*[m(\varepsilon - \mu + \alpha)] \ge (\le) 0$ for $\alpha \ge (\le) 0$ (see P1). Hence, $E^*[(m(\varepsilon - \mu))^2] > 0$.

(P5) $E^*[||X||^2 (d(X))^2 / E^*[(m(\varepsilon - \mu))^2]] < \infty$, where $d(X) = \frac{\partial}{\partial \alpha} E[m(\varepsilon - \mu + \alpha)]|_{\alpha = 0}$

Proof $E^*[(m(\varepsilon - \mu))^2] \ge r > 0$ (see P4). $E^*[||X||^4] < \infty$ because X belongs to a bounded set (Assumption A1). $E^*[(d(X))^4] < \infty$ by Assumptions (A4) and (A7a). Then, by Hölder´s inequality,

$$\begin{aligned} E^*[||X||^2 (d(X))^2 / E^*[(m(\varepsilon - \mu))^2]] & \le \frac{1}{r} E^*[||X||^2 (d(X))^2] \\ & \le \frac{1}{r} \left(E^*[||X||^{2 \cdot 2}]\right)^{1/2} \cdot \left(E^*[(d(X))^{2 \cdot 2}]\right)^{1/2} \\ & < \infty \end{aligned}$$

□

(P6) $Q = E^*[d(X)1[X^T\beta_0 + \mu > c_L]XX^T]$ exists and is non-singular, where $d(X) = \frac{\partial}{\partial\alpha}E[m(\varepsilon - \mu + \alpha)]|_{\alpha=0}$

Proof For all $a \neq 0$,

$$\begin{aligned} a^T Qa = \; & a^T E^*[1[X^T\beta_0 + \mu > c_L] \cdot \frac{1}{1 - F_{\varepsilon^*}(-X\beta_0)} \cdot [(-c_U) \cdot f(c_U + \mu) \\ & + (-c_L) \cdot f(-c_L + \mu) + F(c_U + \mu) - F(-c_L + \mu)] \cdot XX^T]a \\ & \geq \kappa \cdot a^T E^*[1[X^T\beta_0 + \mu > c_L]XX^T]a > 0, \end{aligned}$$

where κ is a positive constant, because by Assumption A7a

$$1 \leq \frac{1}{1 - F_{\varepsilon^*}(-X\beta_0)} = \frac{1}{P^*(Y^* > 0|X)} \leq \frac{1}{\tau}, \quad \tau > 0,$$

by (A4), $F(c_U + \mu) - F(-c_L + \mu) - c_U \cdot f(c_U + \mu) - c_L \cdot f(-c_L + \mu) = d^*(X)$ is positive, and by (A6a) $E^*[1[X^T\beta_0 + \mu > c_L]XX^T]$ is a positive definite matrix. Hence, Q is positive definite and non-singular, because the determinant of a positive definite matrix is positive. □

Appendix B

Proofs of (P1)–(P6) for $m(\varepsilon) = 1[-c_L < \varepsilon < c_U] \cdot \varepsilon + 1[\varepsilon \geq c_U]c_U - 1[\varepsilon \leq -c_L]c_L$

(P1) and (P2) $E^*[m(\varepsilon - \mu + \alpha)] \geq (\leq)0$ for $\alpha \geq (\leq)0$ and $m(\varepsilon - \mu)$ is bounded and continuous almost everywhere.

Proof See corresponding proofs in Appendix A. □

(P3) As a function of α, $E^*[(m(\varepsilon - \mu + \alpha))^2]$ is bounded in a neighbourhood of every α.

Proof For all α, $E^*[(m(\varepsilon - \mu + \alpha))^2] \leq \max\left\{c_U^2, c_L^2\right\} + c_U^2 + c_L^2$ □

(P4) $E^*[(m(\varepsilon - \mu))^2] > 0$.

Proof

$$\begin{aligned} E^*[(m(\varepsilon - \mu))^2] = \; & E^*[1[-c_L < \varepsilon - \mu < c_U] \cdot (\varepsilon - \mu)^2 \\ & + 1[\varepsilon - \mu \geq c_U] \cdot c_U^2 + 1[\varepsilon - \mu \leq -c_L] \cdot c_L^2] \\ = \; & E^*[1[-c_L < \varepsilon - \mu \leq 0] \cdot (\varepsilon - \mu)^2] \\ & + E^*[1[0 < \varepsilon - \mu < c_U] \cdot (\varepsilon - \mu)^2] \\ & + P^*[\varepsilon - \mu \geq c_U] \cdot c_U^2 + P^*[\varepsilon - \mu \leq -c_L] \cdot c_L^2 \\ & \geq \{E^*[1[-c_L < \varepsilon - \mu \leq 0] \cdot (\varepsilon - \mu)]\}^2 \\ & + \{E^*[1[0 < \varepsilon - \mu < c_U] \cdot (\varepsilon - \mu)]\}^2 \\ & + P^*[\varepsilon - \mu \geq c_U] \cdot c_U^2 + P^*[\varepsilon - \mu \leq -c_L] \cdot c_L^2 > 0 \end{aligned}$$

by Jensen's inequality and Assumption A4. □

(P5) $E^*[||X||^2(d(X))^2/E^*[(m(\varepsilon-\mu))^2]] < \infty$, where $d(X) = \frac{\partial}{\partial\alpha}E[m(\varepsilon-\mu+\alpha)]|_{\alpha=0}$

Proof (P5) follows by Hölders inequality because $E^*[(m(\varepsilon-\mu))^2]] \geq r > 0$ (see P4), $E^*[||X||^4] < \infty$ (by Assumption A1), and $E^*[(d(X))^4] \leq 1 < \infty$.

(P6) $Q = E^*[1[X^T\beta_0+\mu < -c_U+s]XX^T]$ exists and is non-singular, where $d(X) = \frac{\partial}{\partial\alpha}E[m(\varepsilon-\mu+\alpha)]|_{\alpha=0}$

Proof $E^*[1[X^T\beta_0+\mu < -c_U+s]XX^T]$ is positive definite by Assumption A6b. Q is also a positive definite matrix, because for all $a \neq 0$

$$a^TQa \geq \kappa a^TE^*[1[X^T\beta_0+\mu < -c_U+s]XX^T]a > 0,$$

where κ is a positive constant. Q is non-singular because the determinant of a positive definite matrix is positive.

References

Battese GE (1997) A note on the estimation of Cobb–Douglas production functions when some explanatory variables have zero values. J Agric Econ 48:250–252

Brännäs K, Laitila T (1991) Modelling and prediction of travel distances by car. Transportation Plann Technol 16:129–143

Buchinsky M (1995) Estimating the asymptotic covariance matrix for quantile regression models: a Monte Carlo study. J Econom 68:303–338

Buckley J, James I (1979) Linear regression with censored data. Biometrika 66:429–436

Chay KY, Powell JL (2001) Semiparametric censored regression models. J Econ Perspect 15:29–42

Hahn, J (1995) Bootstrapping quantile regression estimators. Econom Theory 11:105–121

Honoré BE, Powell JL (1994) Pairwise difference estimators for censored and truncated regression models. J Econom 64:241–278

Karlsson M (2004) Finite sample properties of the QME. Commun Stat Sim Comput 33:567–583

Laitila T (2001) Properties of the QME under asymmetrically distributed disturbances. Stat Probab Lett 52:347–352

Lee MJ (1992) Winsorized mean estimator for censored regression. Econom Theory 8:368–382

Lee MJ (1993) Quadratic mode regression. J Econom 57:1–19

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

Miller RG (1976) Least squares regression with censored data. Biometrika 63:449–464

Newey WK (2001) Conditional moment restrictions in censored and truncated regression models. Econom Theory 17:863–888

Powell JL (1984) Least absolute deviation estimation for the censored regression model. J Econom 25:303–325

Powell JL (1986) Symmetrically trimmed least squares estimation for tobit models. Econometrica 54:1435–1460

Powell JL (1994) Estimation of semiparametric models. In: Engel RF, McFadden DL (eds) Handbook of econometrics, vol 4. North-Holland, Amsterdam, pp 2444–2521