

## LEAST ABSOLUTE DEVIATIONS ESTIMATION FOR THE CENSORED REGRESSION MODEL\*

James L. POWELL

*Massachusetts Institute of Technology, Cambridge, MA 02139, USA*

Received May 1983, final version received October 1983

This paper proposes an alternative to maximum likelihood estimation of the parameters of the censored regression (or censored 'Tobit') model. The proposed estimator is a generalization of least absolute deviations estimation for the standard linear model, and, unlike estimation methods based on the assumption of normally distributed error terms, the estimator is consistent and asymptotically normal for a wide class of error distributions, and is also robust to heteroscedasticity. The paper gives the regularity conditions and proofs of these large-sample results, and proposes classes of consistent estimators of the asymptotic covariance matrix for both homoscedastic and heteroscedastic disturbances.

### 1. Introduction

Many of the important recent advances in econometric methods pertain to limited dependent variable models – that is, regression models for which the range of the dependent variable is restricted to some subset of the real line. Such prior restrictions quite commonly arise in cross-section studies of economic behavior; often, for some fraction of individuals in a sample, implicit non-negativity or other inequality constraints are binding for the variable of interest. In a regression model, an inequality constraint for the dependent variable results in a corresponding bound on the unobservable error terms, this bound being systematically related to the value of the regression function. Hence, the mean of the restricted error term is not zero, and the usual conditions for consistency of least squares estimation will not apply.

The regression model with a non-negativity constraint on the dependent variable was proposed by Tobin (1958); consistent estimation of the parameters of the regression function has been investigated by Amemiya (1973) and Heckman (1976, 1979). Amemiya demonstrates the consistency and asymptotic normality of maximum likelihood estimation for this model, termed the

\*This research was supported by National Science Foundation Grants SES79-12965 and SES79-13976 at the Institute for Mathematical Studies in the Social Sciences at Stanford University. I would like to thank Takeshi Amemiya, Theodore W. Anderson, Timothy Bresnahan, Colin Cameron, Jerry Hausman, Thomas MaCurdy, Daniel McFadden, Julio Rotemberg, and the referees for their helpful comments.

censored regression or censored 'Tobit' model, while Heckman extends least squares estimation to a more general limited dependent variable model by expressing the conditional expectation of the dependent variable as a known (nonlinear) function of the regressors and unknown parameters. A feature common to these approaches is the assumption that the error terms are normally distributed. This condition is essential to the proofs of consistency; unlike the standard linear regression model, for which normal maximum likelihood estimation (classical least squares) is consistent for a wide class of distributions of the residual, estimators based on normality in limited dependent variable models are inconsistent when the normality assumption is violated. Goldberger (1980) and Arabmazar and Schmidt (1982) illustrate this point by calculating the inconsistency of the normal maximum likelihood estimator for several common non-normal distributions of the error term.

This study extends least absolute deviations (LAD) estimation to the regression model with non-negativity of the dependent variable, and gives conditions under which this estimator is consistent and asymptotically normal. In view of the sensitivity of maximum likelihood and least squares methods to the assumption of normality in this model, it is somewhat surprising that a simple modification of LAD estimation yields a consistent estimator which does not depend upon the functional form of the distribution of the residuals.<sup>1</sup> In this sense the estimator proposed below is non-parametric, although its consistency and asymptotic normality require stronger assumptions on the behavior of the regression function than those imposed on models with normal residuals. As for the standard regression model, LAD estimation may be computationally burdensome for the censored regression model, because the function to be minimized is not continuously differentiable; nevertheless, it does provide a consistent alternative to likelihood-based procedures when prior information on the parametric form of the error density is unavailable.

In the next section, the censored regression model and the corresponding LAD estimator are defined; the sections which follow give the appropriate regularity conditions for (strong) consistency and asymptotic normality of the estimator. Section 5 shows how consistent estimators of the asymptotic covariance matrix of the estimator can be obtained. The final section summarizes the foregoing results, and points out some unresolved issues as topics for further research.

<sup>1</sup>Other non-parametric estimators of the parameters of the censored regression model have been proposed by Buckley and James (1979) and Kalbfleisch and Prentice (1980, ch. 6). The former study combines estimation based upon the 'EM algorithm' [see Dempster, Laird, and Rubin (1977)] with the Kaplan-Meier (1958) non-parametric estimator of the survival curve, while the latter is based upon rank regression techniques. It is asserted (though not rigorously demonstrated) that these estimators are also consistent and asymptotically normal. Whether they are to be preferred to the estimator proposed here, on computational or efficiency grounds, is an open question; however, unlike the censored LAD estimator, they are apparently not robust to heteroscedasticity of the error terms.

**2. Definitions and motivation**

The censored regression model considered below can be written in the form

$$y_t = \max\{0, x_t'\beta_0 + u_t\}, \quad t = 1, \dots, T, \tag{2.1}$$

where the dependent variable  $y_t$  and regression vector  $x_t$  are observed for each  $t$ , while the (conformable) parameter vector  $\beta_0$  and error term  $u_t$  are unobserved. It will be presumed throughout that estimation of  $\beta_0$  is the primary object of the statistical analysis (as opposed to, say, estimation of the conditional probability that  $y_t$  exceeds zero given  $x_t$ ).

The definition of the LAD estimator for this model will be based on the fact that, for any scalar random variable  $Z$ , the function  $E[|Z - b| - |Z|]$  is minimized by choosing  $b$  to be a median of the distribution of  $Z$ . Hence, if the median of  $y_t$  is some known function  $m(x_t, \beta_0)$  of the regressors and unknown parameters, a sample analogue to the conditional median can be defined by choosing  $\beta_T$  so that the function  $(1/T)\sum |y_t - m(x_t, \beta)|$  is minimized at the value  $\beta = \beta_T$ . But suppose the error term  $u_t$  is continuously distributed with median zero, and that the density function is positive at zero (so that the median of  $u_t$  is unique). Then it is easy to verify that the median function for  $y_t$  takes a particularly simple form, namely,  $m(x_t, \beta_0) = \max\{0, x_t'\beta_0\}$ . This fact is illustrated in fig. 1. In the top panel, for which  $x_t'\beta_0 > 0$ , the probability that  $y_t = 0$ , equivalent to the probability that  $y_t^* = x_t'\beta_0 + u_t \leq 0$ , is less than one-half, and the median of  $y_t$  is  $x_t'\beta_0$ . On the other hand, if  $x_t'\beta_0 < 0$  (the bottom panel of fig. 1), the probability that  $y_t$  equals zero exceeds one-half, and zero is the unique median of  $y_t$ .

Thus, the LAD estimator for the censored regression model minimizes the sum of absolute deviations of  $y_t$  from  $\max\{0, x_t'\beta_0\}$  over all  $\beta$  in the parameter space (denoted  $B$ ). Algebraically, the censored LAD estimator  $\hat{\beta}_T$  minimizes

$$S_T(\beta) \equiv (1/T) \sum_{t=1}^T |y_t - \max\{0, x_t'\beta\}| \tag{2.2}$$

over all  $\beta$  in  $B$ .

This minimum will always exist if the parameter space  $B$  is compact, since the function  $S_T(\beta)$  is continuous. Unfortunately, this minimization may not yield a unique value for  $\hat{\beta}_T$ . As a simple example, suppose the sample has  $y_t = 0$  for all  $t$ . In this case, any value of  $\beta$  in  $B$  which has  $x_t'\beta \leq 0$  for all  $t$  will yield the same minimizing value of  $S_T(\beta)$ , zero. Of course, this case can be ruled out for large samples, as long as the probability that  $u_t > -x_t'\beta_0$  is positive for a substantial proportion of the regressors  $\{x_t\}$ ; nonetheless, it will be necessary to restrict the possible behavior of the regression function  $x_t'\beta_0$  to ensure the uniqueness of the censored LAD estimator  $\hat{\beta}_T$  for large samples. The precise condition to be imposed is given by Assumption R.1 of the

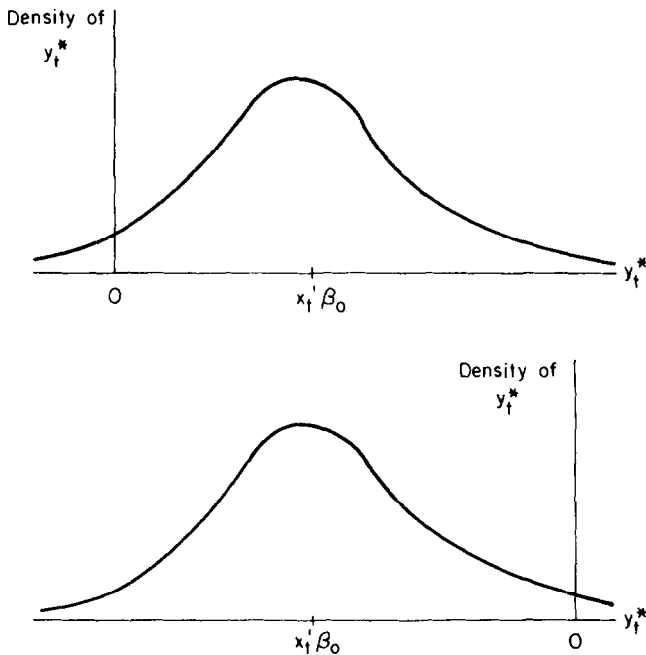


Fig. 1. Density of  $y_t^* \equiv x_t' \beta_0 + u_t$ , when  $x_t' \beta_0 > 0$  and  $x_t' \beta_0 < 0$ .

following section; it is, loosely speaking, a requirement that  $x_t' \beta_0 \geq 0$  for 'enough' sample points, and that the regressors  $x_t$  are not collinear for these observations. This restriction is easily motivated. For those values of  $x_t$  for which  $x_t' \beta_0 < 0$ , there is no one-to-one relationship between the median of  $y_t$  (zero) and the value of  $x_t' \beta_0$ , so knowledge of the median of  $y_t$  for these regressors yields no information concerning the unknown  $\beta_0$ . This is reflected in the definition of the minimand (2.2); for terms in the sum with  $x_t' \beta < 0$ , the value of the term,  $y_t$ , is unrelated to  $\beta$ . Thus minimization of (2.2) can be interpreted as performing LAD estimation only for those points with  $x_t' \beta \geq 0$ . If this procedure is to be consistent, the true regression function  $x_t' \beta_0$  must be non-negative for a substantial proportion of the sample as well.

Unlike the usual conditions for consistency in linear regression models, which only need to rule out multicollinearity of the regressors for large samples, consistency for censored LAD estimation requires a joint condition on the regressors and parameter vector, a condition which may be difficult to verify *a priori*. Fortunately, this assumption seems likely to hold in practice. If the regressors can be chosen freely by the researcher, the experiment can be designed so that the condition will hold for all points in the parameter space. For economic applications, in which the individual characteristics  $x_t$  are

generally not under the control of the researcher, the condition amounts to the assumption that the ‘typical’ (i.e., median) value of the ‘true’ dependent variable  $y_i^* = x_i'\beta_0 + u_i$  is non-negative for a positive fraction of individuals in the sample, which is not unreasonable for many populations sampled.

The assumptions on the distribution of the error term  $u_i$  required for consistency are much weaker than those for maximum likelihood or least squares estimators for the censored regression model. It is the fact that the median of the censored variable  $y_i$  does not depend upon the functional form of the density of the errors that makes censored LAD a ‘distribution-free’ estimator, a property not shared by the mean (if it exists) of  $y_i$ . Hence least absolute deviations estimation is a natural approach for censored data when the assumption of normality of the errors is suspect.

### 3. Strong consistency of the censored LAD estimator

To show the large-sample (almost sure) convergence of  $\hat{\beta}_T$  to the true value  $\beta_0$ , the following conditions on  $\beta_0$ ,  $u_i$ , and  $x_i$  are imposed:

*Assumption P.1.* The parameter vector  $\beta_0$  is an element of a compact parameter space  $B$ .

*Assumption E.1.* The error terms  $\{u_i\}$  are independently and identically distributed random variables, are independent of the regressors, and have median zero, and the distribution function of  $u_i$  is continuously differentiable with density which is bounded above and positive at zero. Hence, defining  $F(\lambda) \equiv \Pr(u_i \leq \lambda)$ ,  $F(0) = \frac{1}{2}$  and  $dF(\lambda) = f(\lambda) d\lambda$  with  $f(\lambda) < f_0$ , some  $f_0 > 0$ , and  $f(0) > 0$ .

*Assumption R.1.* The regressors  $\{x_i\}$  are independently distributed random vectors with  $E\|x_i\|^3 < K_0$  for all  $i$  and some positive  $K_0$ , and  $\nu_T$ , the smallest characteristic root of the matrix

$$E\left[\left(\frac{1}{T}\right) \sum_i 1(x_i'\beta_0 \geq \epsilon_0) x_i x_i'\right],$$

has  $\nu_T > \nu_0$  whenever  $T > T_0$ , some positive  $\epsilon_0$ ,  $\nu_0$ , and  $T_0$ .<sup>2</sup>

A few remarks concerning the necessity and generality of these assumptions (beyond those made in the previous section) are in order. The compactness of  $B$  would be difficult to relax in this context, since it ensures the existence and

<sup>2</sup>The symbol ‘1(A)’ denotes the indicator function of the event ‘A’, i.e., it is a function which takes the value one if A is true and is zero otherwise.

measurability of  $\hat{\beta}_T$  [by Lemma 2 of Jennrich (1969)] and the uniformity of the (almost sure) convergence of the minimand over  $B$ , as required by Lemma A.2 of appendix 1 below.

As usual, the requirement that the errors have median zero is not restrictive, provided there is an unknown constant term in the regression function (although the condition on the matrix in Assumption R.1 presumes that the regression function is written so that  $u_t$  has zero median). Also, the conditions that the error terms are continuously and identically distributed are made for convenience, and can easily be relaxed. It suffices that the conditional distribution of  $u_t$  given  $x_t$  has median zero for all  $t$ , and the corresponding distribution functions for the  $\{u_t\}$  only need to be continuously differentiable in a uniform neighborhood of zero, with density functions  $\{f_t(\lambda|x_t)\}$ , which are uniformly bounded away from zero, i.e.,

$$f_t(\lambda|x_t) > k > 0, \quad (3.1)$$

whenever  $|\lambda| < k$ , some  $k > 0$ , all  $t$ .

Thus, while Assumption E.1 presumes homoscedasticity of the residuals, this condition is not needed for strong consistency of  $\hat{\beta}_T$  (nor for its asymptotic normality, as discussed below), since the conditional median of the dependent variable will still be of the form  $\max\{0, x_t'\beta_0\}$ . This robustness of the censored LAD estimator to (bounded) heteroscedasticity makes it attractive even when the error terms are Gaussian; as Maddala and Nelson (1974) and Arabmazar and Schmidt (1981) have shown, heteroscedastic errors also cause inconsistency of likelihood-based estimators for Tobit models.

Finally, the assumption that the regression vectors  $\{x_t\}$  are stochastic and mutually independent is imposed to simplify the additional condition to be imposed below to obtain asymptotic normality of  $\hat{\beta}_T$ ; however, since the  $\{x_t\}$  are not presumed to be identically distributed, this assumption is not as restrictive as it might first appear. In particular, this setup will accommodate a fixed design matrix, where the  $\{x_t\}$  take fixed (and uniformly bounded) values with probability one. Nevertheless, there may be samples for which the assumption of independence of the  $\{x_t\}$  over  $t$  is inappropriate (e.g., in panel studies, where the regressors represent several measurements of individual characteristics over time); while independence of the regressors does not appear to be essential to the large-sample results given below, relaxation of this assumption would greatly complicate the discussion of asymptotic normality, and thus will not be pursued here.<sup>3</sup>

<sup>3</sup>Robinson (1982) has recently shown Tobit maximum likelihood estimation to be strongly consistent and asymptotically normal under weaker conditions on the dependence of the error terms, and it seems likely that his approach could be extended to the estimator considered here.

*Theorem 1.* For the censored regression model (2.1) under Assumptions P.1, R.1, and E.1, the censored LAD estimator  $\hat{\beta}_T$  is strongly consistent; that is, if

$$\min_{\beta \in B} S_T(\beta) = S_T(\hat{\beta}_T),$$

where  $S_T(\beta)$  is defined in (2.2) above, then  $\hat{\beta}_T$  converges to  $\beta_0$  almost surely.

The proof of this result, as well as the proofs for Theorems 2 and 3 below, are given in appendix 2.

#### 4. Asymptotic normality of the censored LAD estimator

Having established the consistency of the censored LAD estimator  $\hat{\beta}_T$ , determination of its asymptotic distribution is the next logical step in the analysis of its sampling properties. However, for estimation methods based upon minimization of an absolute value criterion, distribution theory is complicated by the lack of differentiability of the minimand. Hence the standard approach to the demonstration of asymptotic normality, based on a Taylor's series expansion of the objective function, is not directly applicable in the present context. Several methods have been used to obtain the large-sample distributions of LAD estimators for the standard linear regression model. Bassett and Koenker (1978) prove the asymptotic normality of LAD in its alternative representation as the solution of a linear programming problem, an approach suggested by Taylor (1974); unfortunately, their proof would be difficult to generalize in the present context, since the minimization problem which defines  $\hat{\beta}_T$  cannot be rewritten as a linear program, except possibly in a neighborhood of  $\hat{\beta}_T$ .<sup>4</sup> An alternative proof of the limiting normality of standardized LAD regression coefficients has been given by Amemiya (1982), who approximates the sum of absolute deviations by a twice-differentiable function that tends to the same limit function; the LAD estimator is then shown to have the same large-sample behavior as the estimator that minimizes this differentiable function. While this method appears to be more applicable in this case, it is algebraically quite complicated, and thus is not adopted here.

The proofs of asymptotic normality given below are based on an approach used by Huber (1967), Bickel (1975), Jurečková (1977), Ruppert and Carroll (1980), Koenker and Bassett (1982a), and Powell (1983). The proof of asymptotic normality given in appendix 2 uses a simple modification of Huber's version of this technique, the modification being required because Huber's

<sup>4</sup>Minimization of  $S_T(\beta)$  cannot be restated as a linear programming problem because the minimand, while piecewise linear, is not convex in  $\beta$ . However, if at the respective solution values there are no data points for which  $x_i'\hat{\beta}_T = 0$ , then for small variations about  $\hat{\beta}_T$  the function  $S_T(\beta)$  will behave just like a sum of absolute deviations for those data points with  $x_i'\hat{\beta}_T > 0$ .

discussion involves identically distributed random variables. The formal statement of this theorem, along with the conditions for its validity, are given in appendix 1 (Lemma A.3). In short, the lemma approximates the (discontinuous) subgradient of the objective function by its continuously differentiable expectation, and then relates this approximation to the asymptotic 'first-order condition' defining the estimator (which sets the subgradient equal to zero asymptotically).

For this method to be applicable to the censored LAD estimator, additional regularity conditions are required; that is, conditions P.1, E.1, and R.1 are not sufficient in general to ensure that  $\sqrt{T}(\hat{\beta}_T - \beta_0)$  has a limiting normal distribution. One additional condition is needed because the conditional median function  $\max\{0, x_i'\beta\}$  of  $y_i$  is not well-behaved for values of  $x_i$  which are nearly orthogonal to  $\beta$ , since the function is not differentiable in  $\beta$  when  $x_i'\beta = 0$ . To ensure asymptotic normality of  $\hat{\beta}_T$ , sequences of values of  $x_i$  which are orthogonal to  $\beta_0$  with positive frequency must be ruled out. For this purpose, the following condition will suffice:

*Assumption R.2.* In addition to the conditions of Assumption R.1, defining

$$G_t(z, \beta, r) \equiv E\left[1(|x_i'\beta| \leq \|x_i\| \cdot z) \|x_i\|^r\right],$$

the function  $G_t$  is  $o(z)$  for  $z$  near zero,  $\beta$  near  $\beta_0$ , and  $r = 0, 1, 2$ , uniformly in  $t$ , i.e.,

$$G_t(z, \beta, r) \leq K_1 \cdot z \quad \text{if} \quad 0 \leq z < \zeta_0, \quad \|\beta - \beta_0\| < \zeta_0, \quad r = 0, 1, 2,$$

for some positive  $K_1$  and  $\zeta_0$ .

To interpret this condition, first suppose for simplicity that the regressors have bounded support (i.e.,  $\|x_{it}\| \leq K$  with probability one for all  $t$  and some  $K > 0$ ). Then it suffices that  $\Pr(|x_i'\beta| \leq z) = o(z)$  uniformly in  $t$  for  $\beta$  near  $\beta_0$  and  $z$  near zero, since  $\Pr(|x_i'\beta| \leq \|x_i\| \cdot z)$  is no greater than  $\Pr(|x_i'\beta| \leq K \cdot z)$  in this case. When the  $\{x_i\}$  are preassigned vectors, this condition would require that  $|x_i'\beta_0| > k$  (with probability one) for some  $k > 0$ ; for many applications, though, such an assumption would be too restrictive, and this is the reason for the more general condition R.2. Heuristically, the assumption implies that the regression function  $x_i'\beta$  is distributed much like the corresponding error term  $u_i$ , at least for  $\beta$  near  $\beta_0$  and  $x_i'\beta$  near the censoring point (zero). Obviously, this condition does not exclude deterministic components of the vector  $x_i$ , nor components which have discrete distributions; only the linear combination  $x_i'\beta$  must have a Lipschitz continuous distribution function near zero.

For regressors with unbounded support, it is the 'normalized' regression function  $\|x_i\|^{-1}x_i'\beta$  which must be smoothly distributed near zero. In addition,



condition R.2 requires the values of  $x_t$  nearly orthogonal to  $\beta$  to be well behaved, in the sense that the conditional moments of  $\|x_t\|$  given the event  $\|x_t\|^{-1}|x_t'\beta| \leq x$  (if these moments are well defined) are uniformly bounded in  $t$  for  $z$  and  $\beta$  in the given neighborhoods. For example, the regressors will satisfy this latter condition if they are spherically distributed about zero, i.e., the  $\{x_t\}$  are continuously distributed with density functions of the form  $h_t(x_t'x_t)$ ; in this case,  $\|x_t\|$  and  $\|x_t\|^{-1}x_t$  are independent, the latter being distributed uniformly on the unit sphere, so R.2 will be satisfied (by the uniform boundedness of  $E\|x_t\|^3$  imposed in R.1). Of course, in practical applications the regressors will have much more complicated distributions, but it is still reasonable to expect in practice that the moments of  $\|x_t\|$  will be well behaved (in the sense of R.2) when  $x_t'\beta/\|x_t\|$  is near zero.

Two additional technical conditions on the parameter vector and error terms are imposed to obtain asymptotic normality of  $\hat{\beta}_T$ ; the second can be relaxed when the support of the regressors is uniformly bounded.

*Assumption P.2.* The parameter vector  $\beta_0$  is an interior point of the compact parameter space  $B$ .

*Assumption E.2.* In addition to the conditions of Assumption E.1, the density function  $f(\lambda)$  of the  $\{u_t\}$  is Lipschitz continuous:

$$|f(\lambda_1) - f(\lambda_2)| \leq L_0 \cdot |\lambda_1 - \lambda_2|, \quad \text{some } L_0 > 0.$$

With these additional conditions, the following result can be established:

*Theorem 2.* For the censored regression model (2.1) under Assumptions P.2, E.2, and R.2, the censored LAD estimator is asymptotically normal,

$$[2f(0)] M_T^{\frac{1}{2}} \cdot \sqrt{T} (\hat{\beta}_T - \beta_0) \xrightarrow{d} N(0, I),$$

where  $M_T^{\frac{1}{2}}$  is any square root of the matrix

$$M_T \equiv E \left[ (1/T) \sum_t 1(x_t'\beta_0 > 0) x_t x_t' \right].^5$$

Before turning to a discussion of consistent estimation of  $C_0$ , a simple extension of the preceding result should be indicated. Asymptotic normality of

<sup>5</sup>If  $\lim M_T \equiv M_\infty$  exists, the conclusion of the theorem can be written in the more suggestive form  $\sqrt{T}(\hat{\beta}_T - \beta_0) \xrightarrow{d} N(0, [2f(0)]^{-2} M_\infty^{-1})$ .

$\hat{\beta}_T$  will still hold if the error terms  $\{u_i\}$  are (boundedly) heteroscedastic, although the asymptotic covariance matrix  $[2f(0)]^{-2}M_T^{-1}$  will no longer be appropriate. Specifically, suppose that the conditional density functions of the  $\{u_i\}$  are of the form given in (3.1), and that ' $f(\lambda)$ ' is replaced by ' $f_i(\lambda|x_i)$ ' in the statement of E.2; then the proof of Theorem 2 can easily be modified (provided  $E\|x_i\|^5 < K_0$ , some  $K_0 > 0$ ) to show that  $M_T^{-\frac{1}{2}}C_T \cdot \sqrt{T}(\hat{\beta}_T - \beta_0)$  has a limiting  $N(0, I)$  distribution, where

$$C_T \equiv E \left[ (1/T) \sum_i 2f_i(0|x_i) \cdot 1(x_i'\beta_0 > 0) x_i x_i' \right]; \quad (4.1)$$

this clearly reduces to the result of Theorem 2 when  $f_i(0|x_i) \equiv f(0)$ .

### 5. Estimation of the asymptotic covariance matrix

In order for the result of Theorem 3 to be useful in constructing large-sample Wald-type hypothesis tests concerning the unknown parameter vector  $\beta_0$ , a consistent estimate of the asymptotic covariance matrix of  $\hat{\beta}_T$  must be obtained.<sup>6</sup> The most difficult problem this poses – one which is generic to estimation methods based on least absolute deviations – is the estimation of the density function  $f(\cdot)$  of the underlying error terms  $\{u_i\}$ . While there is no lack of reasonable estimators for  $f(0)$ , there is no unique 'natural' sample counterpart; any estimator which is the derivative of a 'smoothed' version of the empirical distribution function of the residuals will necessarily be sensitive to the nature and amount of 'smoothing' for finite samples.

To be more specific, for estimation of  $[2f(0)]^{-2}M_T^{-1}$ , a 'natural' estimator for  $M_T$ , the matrix given in Theorem 2, is

$$\hat{M}_T \equiv (1/T) \sum_i 1(x_i'\hat{\beta}_T > 0) x_i x_i'. \quad (5.1)$$

Now suppose  $f(0)$  is estimated by

$$\begin{aligned} \hat{f}_T(0) &\equiv \left[ \sum_i 1(x_i'\hat{\beta}_T > 0) \right]^{-1} \left[ \hat{c}_T^{-1} \sum_i 1(x_i'\hat{\beta}_T > 0) \cdot 1(0 \leq \hat{u}_i \leq \hat{c}_i) \right] \\ &\equiv \hat{f}_T(0; \hat{\beta}_T, \hat{c}_T), \end{aligned} \quad (5.2)$$

where  $\hat{u}_i \equiv y_i - x_i'\hat{\beta}_T$  and where  $\hat{c}_T$  is an appropriately chosen function of the

<sup>6</sup>Koenker and Bassett (1982b) have shown that, for tests of linear hypotheses in the standard linear model, estimation of the density function of the residuals is not needed for calculation of Lagrange multiplier test statistics based on LAD estimates. While their analysis does not immediately generalize to the censored regression model, similar results for censored LAD estimation can undoubtedly be obtained.

data. It is assumed that there is some non-stochastic sequence  $\{c_T\}$  such that

$$\text{plim}_{T \rightarrow \infty} \hat{c}_T/c_T = 1, \quad c_T = o(1), \quad c_T^{-1} = o(\sqrt{T}); \tag{5.3}$$

that is,  $\hat{c}_T$  tends to zero in probability, but at a rate slower than  $T^{-\frac{1}{2}}$ . To interpret this estimator, let  $\hat{F}_T(\lambda)$  denote the empirical c.d.f. of the set of residuals  $\{\hat{u}_i : x_i' \hat{\beta}_T > 0\}$ ; then

$$\hat{f}_T(0) \equiv [\hat{F}_T(\hat{c}_T) - \hat{F}_T(0-)]/\hat{c}_T, \tag{5.4}$$

so the density function is estimated by the fraction of residuals (with positive regression functions) in the interval  $[0, \hat{c}_T]$  divided by the width of this interval.<sup>7</sup> The function  $\hat{c}_T$  might be defined to be scale equivariant; one such example would be

$$\hat{c}_T = c_0 T^{-\gamma} \text{median}\{\hat{u}_i : \hat{u}_i > 0, x_i' \hat{\beta}_T > 0\}, \quad c_0 > 0, \quad \gamma \in (0, \frac{1}{2}). \tag{5.5}$$

Yet even with this sequence of interval widths, there will be a wide range of possible estimates of  $f(0)$  corresponding to different choices of  $c_0$  and  $\gamma$  (i.e., different degrees of ‘smoothing’). The following result shows that  $\hat{f}_T(0)$  is consistent for  $f(0)$  under (5.3) and the conditions of Theorem 2 without further specifying the sequence  $\{\hat{c}_T\}$ .

*Theorem 3. Under Assumptions P.2, E.2, and R.2, and condition (5.3), the estimator  $[2\hat{f}_T(0)]^2 \hat{M}_T$  of  $[2f(0)]^2 M_T$ , where  $\hat{M}_T$  and  $\hat{f}_T(0)$  are defined by (5.1) and (5.2) above, is (weakly) consistent, i.e.,  $[2\hat{f}_T(0)]^2 \hat{M}_T - [2f(0)]^2 M_T$  converges in probability to the zero matrix.*

Of course, more elaborate smoothing schemes for the estimation of  $f(0)$  could be devised for this problem, but there seems to be no *a priori* reason to prefer an alternate estimator. Even if the estimator  $\hat{f}_T(0)$  of  $f(0)$  is adopted, with a definition of  $\hat{c}_T$  like that in (5.5), the constants  $c_0$  and  $\gamma$  must be specified. With additional smoothness conditions on  $f(\lambda)$ , Parzen (1962) showed that  $\gamma = 0.2$  would be optimal (in the sense of minimizing mean squared error) for such a density estimator in an i.i.d. sample, so this may be a reasonable choice in the present circumstance; the optimal value of  $c_0$  would depend upon the form of the density function, but may be set in (5.5) to be best for some nominal (e.g., Gaussian) density.

The estimator  $[2\hat{f}_T(0)]^{-2} \hat{M}_T^{-1}$  will in general be inconsistent for the true asymptotic covariance matrix of  $\sqrt{T}(\hat{\beta}_T - \beta_0)$  if the error terms are hetero-

<sup>7</sup>This interval is centered at  $\frac{1}{2}\hat{c}_T$  rather than at zero to ensure that no data points with  $\hat{u}_i = x_i' \hat{\beta}_T$  (corresponding to points with  $y_i = 0$ ) are used to estimate the density at zero.

scedastic. Following White (1980b), the correct expression for the asymptotic covariance matrix, given immediately after the statement of Theorem 2, can be consistently estimated by  $\hat{C}_T^{-1}\hat{M}_T\hat{C}_T^{-1}$ , where  $\hat{M}_T$  is given in (5.1) and

$$\hat{C}_T \equiv 2(\hat{c}_T T)^{-1} \sum_i 1(x_i' \hat{\beta}_T > 0) 1(0 \leq \hat{u}_i \leq \hat{c}_T) x_i x_i'. \quad (5.6)$$

It is possible to show that  $\hat{C}_T$  is consistent for  $C_T$ , defined in (4.1), by an argument analogous to that given for Theorem 3.

## 6. Conclusion

The results given above complete the investigation of the first-order asymptotic properties of the censored LAD estimator; since  $\hat{\beta}_T$  is consistent and asymptotically normal, and since its asymptotic covariance matrix can be consistently estimated, tests of hypotheses concerning the unknown regression coefficients  $\beta_0$  can be constructed which are valid in large samples. Because neither the computation nor the large-sample properties of  $\hat{\beta}_T$  require the distribution of the error terms to be specified, the associated hypothesis tests are *a fortiori* robust to misspecification of the likelihood function (although proper specification of the regression function is still essential for the validity of the asymptotic results). Indeed, the censored LAD estimator may be used to test whether the distribution of the errors has a particular (e.g., Gaussian) parametric form of interest, by comparison of  $\hat{\beta}_T$  with the corresponding maximum likelihood estimator, as suggested by Hausman (1978). A statistically significant difference between the maximum likelihood estimator and the censored LAD estimator may indicate failure of the assumptions of homoscedasticity or normality, or may be due to some other misspecification of the model.

From a practical point of view, computation of the estimator is an important consideration. While the usual gradient methods of optimization are not applicable to the minimization of  $S_T(\beta)$ , the censored LAD estimator can be computed using 'direct search' methods developed for nonlinear programming. Among these are M.J.D. Powell's (1964) 'conjugate directions' method and the 'flexible polygon' search of Nelder and Mead (1965); the latter method has been successfully used to compute  $\hat{\beta}_T$  for test cases involving 200 observations and three regression parameters. Description of these methods and Fortran programs which implement them are given by Himmelblau (1972); the methods are also available in the GQOPT statistical package.

An open question is the efficiency loss of the proposed estimator relative to maximum likelihood estimation if the error terms have a known (Gaussian) distribution; since the asymptotic covariance matrices of the LAD and maximum likelihood estimators depend in a non-trivial way on the distribution of

the regressors as well as the error terms in the model considered, efficiency comparisons will depend in general on the particular design matrix chosen. It would also be useful to know when a sample size is large enough for the asymptotic distribution to be a good approximation; a simulation study, based on sample sizes and designs encountered in practice, would be valuable in addressing the latter question.

**Appendix 1: Preliminary lemmas**

*A.1. Consistency*

The proof of consistency given in appendix 2 is similar to the approach taken by Obenhofer (1982) to prove (weak) consistency of the LAD estimator for the nonlinear regression model [although his regularity condition B4 does not hold for model (2.1)].<sup>8</sup> The proof here uses the following lemmas by White (1980a):

*Lemma A.1* [Lemma 2.3 of White (1980)]. *Let  $\{Z_t\}$  be independent random variables which assume values in some set  $\Xi$  endowed with the  $\sigma$ -field  $A$ . Let  $q_t: \Xi \times \Theta \rightarrow R^1$ , where  $\Theta \in R^p$  is compact. Assume:*

- (a) *For each  $\theta$  in  $\Theta$ ,  $q_t(z, \theta)$  is  $A$ -measurable.*
- (b)  *$q_t(Z_t, \theta)$  is continuous on  $\Theta$ , uniformly in  $t$ , a.s.*
- (c) *There exist measurable  $m_t: \Xi \rightarrow R^1$  for which  $|q_t(z, \theta)| < m_t(z)$  for all  $\theta$  in  $\Theta$ , and for all  $t$ ,  $E|m_t(Z_t)|^{1+\delta} \leq M < \infty$ , some  $\delta > 0$ . Then*
  - (i)  *$E[q_t(Z_t, \theta)]$  is continuous on  $\Theta$ , uniformly in  $t$ .*
  - (ii)  *$\sup_{\theta \in \Theta} |(1/T)\sum_{t=1}^T q_t(Z_t, \theta) - \bar{Q}_T(\theta)| \rightarrow 0$  a.s., where  $\bar{Q}_T(\theta) \equiv (1/T)\sum_{t=1}^T E[q_t(Z_t, \theta)]$ .*

*Lemma A.2* [Lemma 2.2 of White (1980)]. *Let  $Q_T(\omega, \theta)$  be a measurable function on a measurable space  $\Omega$  and for each  $\omega$  in  $\Omega$  a continuous function of  $\theta$  on a compact set  $\Theta$ . Then there exists a measurable function  $\hat{\theta}_T(\omega)$  such that*

$$Q_T[\omega, \hat{\theta}_T(\omega)] = \inf_{\theta \in \Theta} Q_T(\omega, \theta),$$

*for all  $\omega$  in  $\Omega$ . If  $|Q_T(\omega, \theta) - \bar{Q}_T(\theta)| \rightarrow 0$  a.e. uniformly for all  $\theta$  in  $\Theta$  and if  $\bar{Q}_T(\theta)$  has a minimum at  $\theta_0$  which is identifiably unique (i.e., for any  $\epsilon > 0$ ,*

$$\inf_{T \geq T_0} \min_{\theta} \{ \bar{Q}_T(\theta) - \bar{Q}_T(\theta_0) : \theta \in \Theta, \|\theta - \theta_0\| > \epsilon \} > 0,$$

*for some positive integer  $T_0$ ), then  $\hat{\theta}_T$  converges to  $\theta_0$  almost surely ( $\omega$ ).*

<sup>8</sup>I am grateful to a referee for pointing out Obenhofer's paper.

### A.2. Asymptotic normality

The proof of asymptotic normality of  $\hat{\beta}_T$  is based upon a simple extension of a theorem proved by Huber (1967) for maximum-likelihood-type estimation for i.i.d. samples. This theorem gives sufficient conditions under which a sequence of consistent estimators  $\{\hat{\theta}_T\}$  which satisfy some 'first-order condition'

$$(1/\sqrt{T}) \sum_{t=1}^T \psi(Z_t, \hat{\theta}_T) = o_p(1),$$

is asymptotically normal, where the  $\{Z_t\}$  are i.i.d. random variables taking values in some sample space  $\Xi$ , and where  $\psi: \Xi \times \Theta \rightarrow R^p$  is some given function, for  $\Theta$  an open subset of Euclidean space  $R^p$ .

For the estimator considered above, the corresponding random variables  $\{Z_t\}$  are not identically distributed (although their mutual independence is assumed throughout), and in other applications the appropriate functions  $\{\psi_t(Z_t, \theta)\}$  may vary systematically with the index  $t$ ; nonetheless, Huber's conditions can be easily modified to apply in such cases. Define

$$\lambda_T(\theta) \equiv (1/T) \sum_t E[\psi_t(Z_t, \theta)],$$

where the expectations here and throughout are taken with respect to the true distributions of the  $\{Z_t\}$ , and let

$$\mu_t(Z_t, \theta, d) \equiv \sup_{\tau} \{ \|\psi_t(Z_t, \tau) - \psi_t(Z_t, \theta)\| : \|\tau - \theta\| \leq d \}.$$

The following conditions are imposed on  $\psi_t$ ,  $\lambda_T$ , and  $\mu_t$ :

*Assumption N.1.* For each fixed  $\theta \in \Theta$  and each  $t$ ,  $\phi_t(Z_t, \theta)$  is measurable, and  $\psi_t(Z_t, \theta)$  is separable in the sense of Doob for all  $t$ : i.e., there is a countable subset  $\Theta' \subset \Theta$  such that for every open set  $U \subset \Theta$  and every closed interval  $A$ , the sets  $\{Z_t : \|\psi_t(Z_t, \theta)\| \in A, \text{ for all } \theta \in U\}$  and  $\{Z_t : \|\phi_t(Z_t, \theta)\| \in A, \text{ for all } \theta \in U \cap \Theta'\}$  differ at most by a set of (probability) measure zero.

*Assumption N.2.* There is some  $\theta_0 \in \Theta$  such that  $\lambda_T(\theta_0) = 0$  for  $T$ .

*Assumption N.3.* There are (strictly positive) numbers  $a, b, c, d_0$ , and  $T_0$  such that, for all  $t$ ,

- (i)  $\|\lambda_T(\theta)\| \geq a \cdot \|\theta - \theta_0\|$  for  $\|\theta - \theta_0\| \leq d_0, T \geq T_0$ ,
- (ii)  $E[\mu_t(Z_t, \theta, d)] \leq b \cdot d$  for  $\|\theta - \theta_0\| + d \leq d_0, d \geq 0$ ,
- (iii)  $E[\mu_t(Z_t, \theta, d)]^2 \leq c \cdot d$  for  $\|\theta - \theta_0\| + d \leq d_0, d \geq 0$ .

*Assumption N.4.*  $E\|\phi_t(Z_t, \theta_0)\|^2 \leq K$  for all  $t$  and some  $K > 0$ .

With these conditions, a simple asymptotic expression for  $\lambda_T(\hat{\theta}_T)$  can be obtained.

*Lemma A.3.* Assume that Assumptions N.1 to N.4 hold and that

$$(1/\sqrt{T}) \sum_t \psi_t(Z_t, \hat{\theta}_T) = o_p(1) \quad \text{for} \quad \hat{\theta}_T \equiv \hat{\theta}_T(Z_1, \dots, Z_T).$$

If  $\hat{\theta}_T$  is consistent for  $\theta_0$  (i.e.,  $\Pr\{\|\hat{\theta}_T - \theta_0\| > \eta\} = o(1)$  for any  $\eta > 0$ ), then

$$\sqrt{T} \cdot \lambda_T(\theta)|_{\theta=\hat{\theta}_T} = -(1/\sqrt{T}) \sum_t \psi_t(Z_t, \theta_0) + o_p(1).$$

*Proof.* The method of proof is identical to that in Section 4 of Huber (1967). In each step of the proof of his Lemma 3, terms of the form

$$T \cdot \lambda(\theta), \quad T \cdot E\mu(Z, \theta, d), \quad \text{etc.},$$

can be replaced by their ‘averaged’ counterparts

$$T \cdot \lambda_T(\theta), \quad \sum E\mu_t(Z_t, \theta, d), \quad \text{etc.},$$

without affecting the validity of the argument if  $T$  is sufficiently large.

## Appendix 2: Proofs of theorems in text

*Proof of Theorem 1.* Let

$$u_t^* \equiv y_t - \max\{0, x_t' \beta_0\}, \tag{A.1}$$

and define

$$h_t \equiv \max\{0, x_t' \beta\} - \max\{0, x_t' \beta_0\} \equiv h_t(\beta, \beta_0), \tag{A.2}$$

where

$$\delta \equiv \beta - \beta_0, \tag{A.3}$$

here and throughout. Now minimization of  $S_T(\beta)$  is equivalent to minimization of

$$Q_T(\beta) \equiv S_T(\beta) - S_T(\beta_0), \tag{A.4}$$

since  $S_T(\beta_0)$  does not depend upon  $\beta$ ; because the sequence of minimizing values  $\{\hat{\beta}_T\}$  is not affected by this normalization, the function  $Q_T(\beta)$  will be shown to satisfy the conditions of Lemma A.1.<sup>9</sup> Writing this minimand in terms of (A.1) and (A.2),

$$Q_T(\beta) = (1/T) \sum_i [|u_i^* - h_i| - |u_i^*|]. \tag{A.5}$$

Each of the terms in this summation is bounded (in magnitude) by  $|h_i(\beta, \beta_0)| \leq 2[\|\beta\| + \|\beta_0\|]\|x_i\|$ , which is  $O(\|x_i\|)$  by the compactness of  $B$ . Assumptions R.1 and E.1 and Lemma A.1 thus yield

$$\lim_{T \rightarrow \infty} \|Q_T(\beta) - \bar{Q}_T(\beta)\| = 0, \text{ a.s., uniformly in } \beta, \tag{A.6}$$

where

$$\bar{Q}_T(\beta) \equiv E[S_T(\beta) - S_T(\beta_0)]. \tag{A.7}$$

By Lemma A.2, and since  $\bar{Q}_T(\beta_0) \equiv 0$  and  $\bar{Q}_T(\beta) \geq 0$ ,  $\hat{\beta}_T$  will be strongly consistent if, for any  $\epsilon > 0$ ,  $\bar{Q}_T(\beta)$  is bounded away from zero, uniformly in  $\beta$ , for  $\|\beta - \beta_0\| \geq \epsilon$  and all  $T$  sufficiently large.

The expected value of  $Q_T(\beta)$ , conditional on the regressors  $\{x_i\}$ , is

$$\begin{aligned} & E[Q_T(\beta) | x_1, \dots, x_T] \\ &= 2(1/T) \sum_i 1(x_i'\beta_0 \geq 0, x_i'\beta \geq 0) \int_0^{x_i'\delta} (x_i'\delta - \lambda) f(\lambda) d\lambda \\ &+ 2(1/T) \sum_i 1(x_i'\beta_0 \geq 0, x_i'\beta < 0) \int_{-x_i'\beta_0}^0 (\lambda + x_i'\beta_0) f(\lambda) d\lambda \\ &+ 2(1/T) \sum_i 1(x_i'\beta_0 < 0, x_i'\beta \geq 0) \int_0^{-x_i'\beta_0} (x_i'\beta) f(\lambda) d\lambda \\ &+ 2(1/T) \sum_i 1(x_i'\beta_0 < 0, x_i'\beta < 0) \int_{-x_i'\beta_0}^{x_i'\delta} (x_i'\delta - \lambda) f(\lambda) d\lambda. \end{aligned} \tag{A.8}$$

All terms in this expression are positive, so

$$\begin{aligned} & E[Q_T(\beta) | x_1, \dots, x_T] \\ &\geq 2(1/T) \sum_i 1(x_i'\beta_0 \geq 0, x_i'\beta \geq 0) \int_0^{x_i'\delta} (x_i'\delta - \lambda) f(\lambda) d\lambda \\ &+ 2(1/T) \sum_i 1(x_i'\beta_0 \geq 0, x_i'\beta < 0) \int_{-x_i'\beta_0}^0 (\lambda + x_i'\beta_0) f(\lambda) d\lambda. \end{aligned} \tag{A.9}$$

<sup>9</sup>This normalization is used by Huber (1963, 1967), Amemiya (1982), and Obenhofer (1982), among others; its purpose is to eliminate the requirement that  $u_i$  have finite first moment.



Now by Assumption E.1, there exists an  $f_1 > 0$  such that  $f(\lambda) > f_1$  whenever  $|\lambda| < f_1$ . Then, for any number  $\tau$  such that  $0 < \tau < \min\{\varepsilon_0, f_1\}$  (where  $\varepsilon_0$  is defined in Assumption R.1), inequality (A.9) yields

$$\begin{aligned}
 & E[Q_T(\beta) | x_1, \dots, x_T] \\
 & \geq 2(1/T) \sum_t 1(x'_t \beta_0 \geq \varepsilon_0, x'_t \beta \geq 0) \cdot 1(|x'_t \delta| \geq \tau) \int_0^\tau (\tau - \lambda) f_1 d\lambda \\
 & \quad + 2(1/T) \sum_t 1(x'_t \beta_0 \geq \varepsilon_0, x'_t \beta < 0) \int_{-\tau}^0 (\lambda + \tau) f_1 d\lambda \tag{A.10} \\
 & \geq (\tau^2 f_1)(1/T) \sum_t 1(x'_t \beta_0 \geq \varepsilon_0) \cdot 1(|x'_t \delta| \geq \tau).
 \end{aligned}$$

Taking expectations with respect to the distribution of the  $\{x_t\}$ , and applying Hölder's and Jensen's inequalities, implies

$$\begin{aligned}
 \bar{Q}_T(\beta) & \equiv E[Q_T(\beta)] \\
 & \geq (\tau^2 f_1)(1/T) \sum_t E[1(x'_t \beta_0 \geq \varepsilon_0) \cdot 1(|x'_t \delta| > \tau)] \\
 & \geq (\tau^2 f_1)(1/T) \sum_t E[1(x'_t \beta_0 \geq \varepsilon_0) \cdot 1(|x'_t \delta| > \tau)]^3 [E\|x_t\|^3 / K_0]^2 \\
 & \geq (\tau^2 f_1 K_0^{-2})(1/T) \sum_t \{E[1(x'_t \beta_0 \geq \varepsilon_0) \cdot 1(|x'_t \delta| > \tau) \|x_t\|^2]\}^3 \tag{A.11} \\
 & \geq (\tau^2 f_1 K_0^{-2}) \left\{ (1/T) \sum_t E[1(x'_t \beta_0 \geq \varepsilon_0) \cdot 1(|x'_t \delta| > \tau) (x'_t \delta)^2 \|\delta\|^{-2}] \right\}^3 \\
 & \geq (\tau^2 f_1 K_0^{-2}) [(\nu_T \|\delta\|^2 - \tau^2) \|\delta\|^{-2}]^3,
 \end{aligned}$$

where  $\nu_T$  is the smallest characteristic root of  $E[(1/T) \sum_t 1(x'_t \beta_0 \geq \varepsilon_0) x_t x'_t]$ , which is strictly positive for large  $T$  by Assumption R.1. Thus, choosing  $\tau^2 < \nu_T \varepsilon^2$  shows that  $\bar{Q}_T(\beta)$  is uniformly bounded away from zero for large  $T$  and  $\|\delta\| \equiv \|\beta - \beta_0\| > \varepsilon$ , as required.

*Proof of Theorem 2.* Defining

$$\psi(u_t, x_t, \beta) \equiv 1(x'_t \beta > 0) \left[ \frac{1}{2} - 1(u_t < x'_t(\beta - \beta_0)) \right] x_t, \tag{A.12}$$

the condition

$$\Psi_T(\hat{\beta}_T) \equiv (1/\sqrt{T}) \sum_t \psi(u_t, x_t, \hat{\beta}_T) = o_p(1) \quad (\text{A.13})$$

must be established. Noting that the function  $\psi$  can be written as

$$\begin{aligned} \psi(u_t, x_t, \beta) &= 1(x'_t\beta > 0) \left[ \frac{1}{2} - 1(\max\{0, x'_t\beta_0 + u_t\} < x'_t\beta) \right] x_t \\ &= 1(x'_t\beta > 0) \left[ \frac{1}{2} - 1(y_t < x'_t\beta) \right] x_t, \end{aligned} \quad (\text{A.14})$$

it is easily verified that  $-2T^{-\frac{1}{2}}\Psi_T(\beta)$  is the vector of left partial derivatives of the objective function  $S_T(\beta)$  defined in (2.2) above. Because  $\hat{\beta}_T$  minimizes  $S_T(\beta)$ , it is straightforward to show that  $\Psi_T(\hat{\beta}_T)$  satisfies

$$\begin{aligned} |\Psi_{T,j}(\hat{\beta}_T)| &\leq (1/\sqrt{T}) \sum_t |x_{tj}| \cdot \left[ 1(y_t = x'_t\hat{\beta}_T) + \frac{1}{2} \cdot 1(x'_t\hat{\beta}_T = 0) \right] \\ &\leq \left[ (1/\sqrt{T}) \max_{t \leq T} \{ \|x_t\| \} \right] \\ &\quad \times \sum_t \left[ 1(0 < y_t = x'_t\hat{\beta}_T) + \frac{1}{2} \cdot 1(x'_t\hat{\beta}_T = 0) \right], \end{aligned} \quad (\text{A.15})$$

where the subscript  $j$  denotes the  $j$ th component of the corresponding vector [see Ruppert and Carroll (1980), proof of Lemma A.2, for demonstration of a similar result]. But  $\max\{\|x_t\| : t \leq T\} = o(\sqrt{T})$  almost surely by Assumption R.1, and the sum on the right-hand side of (A.15) is finite with probability one for all  $T$  suitably large by Assumptions E.1, R.2, and the strong consistency of  $\hat{\beta}_T$ . Thus condition (A.13) holds.

Next, conditions N.1 though N.4 of Lemma A.3 must be established; only condition N.3 will be verified here, since the remaining conditions are easily checked. To show N.3(i) holds, it is convenient to write

$$\begin{aligned} \lambda_T(\beta) &= E_x \left[ (1/T) \sum_t 1(x'_t\beta > 0) x_t \cdot \int_{x'_t(\beta - \beta_0)}^0 f(\lambda) d\lambda \right] \\ &= E_x \left[ (1/T) \sum_t -1(x'_t\beta > 0) \cdot f(\lambda_t^*) x_t x'_t \right] (\beta - \beta_0), \quad (\text{A.16}) \\ &\equiv [-\Lambda_T(\beta)] (\beta - \beta_0), \end{aligned}$$

where 'E<sub>x</sub>' denotes the expectation operator for the marginal distribution of  $x_t$  and the  $\{\lambda_t^*\}$  are mean-value coefficients lying between zero and  $x'_t(\beta - \beta_0)$ .

The second equality of (A.16) can be used to show

$$\Lambda_T(\beta) - f(0) \cdot M_T = O(\|\beta - \beta_0\|), \tag{A.17}$$

uniformly in  $T$  for  $\beta$  near  $\beta_0$ , where  $M_T$  is the matrix given in the statement of the theorem. This holds because each element of the difference satisfies

$$\begin{aligned} & \left| [\Lambda_T(\beta) - f(0) \cdot M_T]_{jk} \right| \\ &= \left| (1/T) \sum_i E_x [1(x'_i \beta > 0) [f(\lambda_i^*) - f(0)] x_{ij} x_{ik} \right. \\ &\quad \left. + [f(0)] (1/T) \sum_i E_x [\{1(x'_i \beta > 0) - 1(x'_i \beta_0 > 0)\} x_{ij} x_{ik}] \right| \\ &\leq (1/T) \sum_i E_x [1(x'_i \beta > 0) |f(\lambda_i^*) - f(0)| \|x_i\|^2] \\ &\quad + [f(0)] (1/T) \sum_i E_x [1(|x'_i \beta_0| \leq \|x_i\| \|\beta - \beta_0\|) \|x_i\|^2] \\ &\leq [K_0 L_0 + [f(0)] K_1] \|\beta - \beta_0\|, \end{aligned} \tag{A.18}$$

for  $\|\beta - \beta_0\| < \zeta_0$  by Assumptions R.2 and E.2. Thus

$$\lambda_T(\beta) = f(0) \cdot M_T (\beta - \beta_0) + O(\|\beta - \beta_0\|^2), \tag{A.19}$$

and since the minimum characteristic root of  $M_T$  is bounded above zero for all  $T$  sufficiently large (by Assumption R.1), inequality N.3(i) will be satisfied for all  $\beta$  in a sufficiently small neighborhood of  $\beta_0$ .

To verify N.3(ii) and (iii), define

$$\mu_t(\beta, d) \equiv \sup_{\|\gamma - \beta\| < d} \|\psi(u_t, x_t, \beta) - \psi(u_t, x_t, \gamma)\|. \tag{A.20}$$

Then the definition of  $\psi$  and some manipulation yields

$$\begin{aligned} \mu_t(\beta, d) &\leq \sup_{\|\gamma - \beta\| < d} \frac{1}{2} \|x_t\| \cdot 1(|x'_t \beta| \leq |x'_t (\beta - \gamma)|) \\ &\quad + \sup_{\|\gamma - \beta\| < d} \|x_t\| \cdot 1(|u_t - x'_t \beta| \leq |x'_t (\beta - \gamma)|) \\ &\leq \|x_t\| \left[ \frac{1}{2} \cdot 1(|x'_t \beta| \leq \|x_t\| d) + 1(|u_t - x'_t \beta| \leq \|x_t\| d) \right], \end{aligned} \tag{A.21}$$

so

$$E[\mu_t(\beta, d)] \leq \left[ \frac{1}{2}K_1 + f_0(K_0)^{2/3} \right] \cdot d, \quad (\text{A.22})$$

when  $\|\beta - \beta_0\| \leq \zeta_0$ , by Assumptions R.2 and E.1. Similarly,

$$E[\mu_t(\beta, d)]^2 \leq [K_1 + 2f_0K_0] \cdot d, \quad (\text{A.23})$$

when  $\|\beta - \beta_0\| \leq \zeta_0$ , since  $\mu_t < 2\|x_t\|$  by (A.21).

Hence N.3 and the other conditions of Lemma A.3 hold, so

$$\begin{aligned} -\sqrt{T}[\lambda_T(\hat{\beta}_T)] &\equiv [\Lambda_T(\hat{\beta}_T)]\sqrt{T}(\hat{\beta}_T - \beta_0) \\ &= (1/\sqrt{T})\sum_t \psi(u_t, x_t, \beta_0) + o_p(1). \end{aligned} \quad (\text{A.24})$$

Conditions (A.19), (A.24), and the consistency of  $\hat{\beta}_T$  imply

$$\begin{aligned} \sqrt{T}(\hat{\beta}_T - \beta_0) &= [f(0)M_T]^{-1}(1/\sqrt{T})\sum_t 1(x'_t\beta_0 > 0) \\ &\quad \times \left[ \frac{1}{2} - 1(u_t > 0) \right] x_t + o_p(1). \end{aligned} \quad (\text{A.25})$$

The asymptotic normality of  $\hat{\beta}_T$  then follows from the application of Liapunov's central limit theorem to (A.25).

*Proof of Theorem 3.* Only the weak consistency of  $\hat{f}_T(0)$  will be demonstrated here; the convergence of  $\hat{M}_0$  to  $M_0$  can be proved in a similar fashion. First, the denominator of  $\hat{f}_T(0)$ , when multiplied by  $1/T$ , can be replaced by  $(1/T)\sum \Pr\{x'_t\beta_0 > 0\}$ , since

$$\begin{aligned} &\left| (1/T)\sum_t [1(x'_t\hat{\beta}_T > 0) - \Pr\{x'_t\beta_0 > 0\}] \right| \\ &\leq (1/T)\sum_t 1(|x'_t\beta_0| \leq \|x_t\| \cdot \|\hat{\beta}_T - \beta_0\|) \\ &\quad + \left| (1/T)\sum_t [1(x'_t\beta_0 > 0) - \Pr\{x'_t\beta_0 > 0\}] \right|. \end{aligned} \quad (\text{A.26})$$

The second term on the right-hand side is clearly  $o_p(1)$ ; to show the first term

also converges to zero in probability, note that, for any  $\eta > 0$ ,

$$\begin{aligned} & \Pr\left\{(1/T) \sum_t 1(|x'_t \beta_0| \leq \|x_t\| \cdot \|\hat{\beta}_T - \beta_0\|) > \eta\right\} \\ & \leq \Pr\left\{(1/T) \sum_t 1(|x'_t \beta_0| \leq \|x_t\| z) > \eta\right\} + \Pr\{\|\hat{\beta}_T - \beta_0\| > z\} \\ & \leq \eta^{-1} (1/T) \sum_t \Pr\{|x'_t \beta_0| \leq \|x_t\| z\} + \Pr\{\|\hat{\beta}_T - \beta_0\| > z\} \\ & \leq \eta^{-1} K_1 z + \Pr\{\|\hat{\beta}_T - \beta_0\| > z\}, \end{aligned} \tag{A.27}$$

when  $z < \zeta_0$ , which follows from Markov's inequality and Assumption R.2(i). Thus, by choosing  $z$  sufficiently small, the right-hand side of (A.27) can be made arbitrarily small for large  $T$  by the consistency of  $\hat{\beta}_T$ .

Similarly, the numerator, normalized by  $1/T$ , can be replaced by  $(c_T T)^{-1} \sum 1(x'_t \beta_0 > 0) \cdot 1(0 \leq u_t \leq c_T)$ , because  $\text{plim } \hat{c}_T / c_T = 1$  and

$$\begin{aligned} & (c_T T)^{-1} \left| \sum_t [1(x'_t \hat{\beta}_T > 0) \cdot 1(0 \leq \hat{u}_t \leq \hat{c}_T) \right. \\ & \quad \left. - 1(x'_t \beta_0 > 0) \cdot 1(0 \leq u_t \leq c_T)] \right| \\ & \leq (c_T T)^{-1} \sum_t 1(|x'_t \beta_0| \leq \|x_t\| \cdot \|\hat{\beta}_T - \beta_0\|) \\ & \quad + (c_T T)^{-1} \sum_t 1(|u_t| \leq \|x_t\| \cdot \|\hat{\beta}_T - \beta_0\|) \\ & \quad + (c_T T)^{-1} \sum_t 1(|u_t - c_T| \leq \|x_t\| \cdot \|\hat{\beta}_T - \beta_0\| + |\hat{c}_T - c_T|). \end{aligned} \tag{A.28}$$

These terms can each be shown to converge to zero in probability in the same fashion as in (A.27), using Assumptions R.2 and E.2 and the facts that  $c_T^{-1} \|\hat{\beta}_T - \beta_0\| = o_p(1)$  and that  $c_T^{-1} |\hat{c}_T - c_T| = o_p(1)$ . Now

$$\begin{aligned} & E \left[ (c_T T)^{-1} \sum_t 1(x'_t \beta_0 > 0) \cdot 1(0 \leq u_t \leq c_T) \right] \\ & = \left[ (1/T) \sum_t \Pr\{x'_t \beta_0 > 0\} \right] \int_0^{c_T} c_T^{-1} f(\lambda) d\lambda \\ & = \left[ (1/T) \sum_t \Pr\{x'_t \beta_0 > 0\} \right] \int_0^1 f(c_T \lambda) d\lambda \\ & = \left[ (1/T) \sum_t \Pr\{x'_t \beta_0 > 0\} \right] f(0) + o(1), \end{aligned} \tag{A.29}$$

by the dominated convergence theorem, and

$$\begin{aligned} & \text{var} \left[ (c_T T)^{-1} \sum_t 1(x'_t \beta_0 > 0) \cdot 1(0 \leq u_t \leq c_T) \right] \\ &= (c_T T)^{-2} \sum_t \text{var} [1(x'_t \beta_0 > 0) \cdot 1(0 \leq u_t \leq c_T)] \\ &\leq (c_T T)^{-1} c_T^{-1} \Pr\{0 \leq u_t \leq c_T\} = o(1/\sqrt{T}). \end{aligned} \tag{A.30}$$

Hence

$$\begin{aligned} \hat{f}_T(0) &= \left[ (1/T) \sum_t \Pr\{x'_t \beta_0 > 0\} + o_p(1) \right]^{-1} \\ &\quad \times \left[ f(0)(1/T) \sum_t \Pr\{x'_t \beta_0 > 0\} + o_p(1) \right] \\ &= f(0) + o_p(1), \end{aligned} \tag{A.31}$$

as asserted.

## References

- Amemiya, T., 1973, Regression analysis when the dependent variable is truncated normal, *Econometrica* 41, 997–1016.
- Amemiya, T., 1982, Two stage least absolute deviations estimators, *Econometrica* 50, 689–711.
- Arabmazar, A. and P. Schmidt, 1981, Further evidence on the robustness of the Tobit estimator to heteroskedasticity, *Journal of Econometrics* 17, 253–258.
- Arabmazar, A. and P. Schmidt, 1982, An investigation of the robustness of the Tobit estimator to non-normality, *Econometrica* 50, 1055–1063.
- Bassett, G. and R. Koenker, 1978, Asymptotic theory of least absolute error regression, *Journal of the American Statistical Association* 73, 667–677.
- Bickel, P.J., 1975, One-step Huber estimation in the linear model, *Journal of the American Statistical Association* 70, 428–433.
- Buckley, J. and I. James, 1979, Linear regression with censored data, *Biometrika* 66, 429–436.
- Dempster, A.P., N.M. Laird and D.B. Rubin, 1977, Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society B* 39, 1–22.
- Goldberger, A.S., 1980, Abnormal selection bias, Workshop paper no. 8006 (Social Systems Research Institute, University of Wisconsin, Madison, WI).
- Hausman, J.A., 1978, Specification tests in econometrics, *Econometrica* 46, 1251–1271.
- Heckman, J., 1976, The common structure of statistical models of truncation, sample selection, and limited dependent variables and a simple estimator for such models, *Annals of Economic and Social Measurement* 5, 475–492.
- Heckman, J., 1979, Sample bias as a specification error, *Econometrica* 47, 153–162.
- Himmelblau, D.M., 1972, *Applied nonlinear programming* (McGraw-Hill, New York).
- Huber, P.J., 1964, Robust estimation of a location parameter, *Annals of Mathematical Statistics* 35, 73–101.
- Huber, P.J., 1965, The behavior of maximum likelihood estimates under nonstandard conditions, *Proceedings of the Fifth Berkeley Symposium* 1, 221–233.

- Jurečková, J., 1977, Asymptotic relations of  $M$ -estimates and  $R$ -estimates in the linear model, *Annals of Statistics* 5, 464–472.
- Kalbfleisch, J.O. and R.L. Prentice, 1980, *The statistical analysis of failure time data* (Wiley, New York).
- Kaplan, E.L. and P. Meier, 1958, Nonparametric estimation from incomplete observations, *Journal of the American Statistical Association* 53, 457–481.
- Koenker, R. and G. Bassett, 1982a, Robust tests for heteroscedasticity based on regression quantiles, *Econometrica* 50, 43–61.
- Koenker, R. and G. Bassett, 1982b, Tests of linear hypotheses and  $L_1$  estimation, *Econometrica* 50, 1577–1583.
- Maddala, G.S. and F.D. Nelson, 1975, Specification errors in limited dependent variable models, National Bureau for Economic Research working paper no. 96.
- Nelder, J.A. and R. Mead, 1965, A simplex method for function minimization, *The Computer Journal* 7, 308–313.
- Obenhofer, W., 1982, The consistency of nonlinear regression minimizing the  $L_1$  norm, *Annals of Statistics* 10, 316–319.
- Parzen, E., 1962, On estimation of a probability density function and its mode, *Annals of Mathematical Statistics* 33, 1065–1076.
- Powell, J.L., 1983, The asymptotic normality of two-stage least absolute deviations estimators, *Econometrica* 51, 1569–1575.
- Powell, M.J.D., 1964, An efficient method for finding the minimum of a function of several variables without calculating derivatives, *The Computer Journal* 7, 155–162.
- Robinson, P.M., 1982, On the asymptotic properties of estimators of models containing limited dependent variables, *Econometrica* 50, 27–41.
- Ruppert, D. and R.J. Carroll, 1980, Trimmed least squares estimation in the linear model, *Journal of the American Statistical Association* 75, 828–838.
- Taylor, L.D., 1974, Estimation by minimizing the sum of absolute errors, in: P. Zarembka, ed., *Frontiers in econometrics* (Academic Press, New York).
- Tobin, J., 1958, Estimation of relationships for limited dependent variables, *Econometrica* 26, 24–36.
- White, H., 1980a, Nonlinear regression on cross-section data, *Econometrica* 48, 721–746.
- White, H., 1980b, A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity, *Econometrica* 48, 817–838.