SECTION 11

Least Absolute Deviations Estimators for Censored Regression

Suppose random variables $y_1, y_2, \ldots$ are generated by a regression $y_i = x_i'\theta_0 + u_i$, with θ_0 an unknown d-dimensional vector of parameters, $\{x_i\}$ a sequence of observed vectors, and $\{u_i\}$ unobserved errors. The method of least absolute deviations would estimate θ_0 by the θ that minimized the convex function

$$\sum_{i \le n} |y_i - x_i'\theta|.$$

Convexity in θ makes the asymptotic analysis not too difficult (Pollard 1990). Much more challenging is a related problem, analyzed by Powell (1984), in which the value of y_i is observed only if $y_i \ge 0$ and otherwise only the information that $y_i < 0$ is available. That is, only y_i^+ is observed. In the econometrics literature this is called a Tobit model (at least when the $\{u_i\}$ are independent normals).

Powell proposed an interesting variation on the least absolute deviations estimation; he studied the $\widehat{\theta}_n$ that minimizes

$$\sum_{i \le n} |y_i^+ - (x_i'\theta)^+|$$

over a subset Θ of $\mathbb{R}^d$. This function is not convex in θ; analysis of $\widehat{\theta}_n$ is quite difficult. However, by extending a technique due to Huber (1967), Powell was able to give conditions under which $\sqrt{n}(\widehat{\theta}_n - \theta_0)$ has an asymptotic normal distribution.

With the help of the maximal inequalities developed in these notes, we can relax Powell's assumptions and simplify the analysis a little. The strategy will be to develop a uniformly good quadratic approximation to the criterion function, then show that $\widehat{\theta}_n$ comes close to maximizing the approximation. Powell's consistency argument was based on the same idea, but for asymptotic normality he sought

an approximate zero for a vector of partial derivatives, a method that is slightly complicated by the lack of smoothness of the criterion function.

Assumptions. Let us assume that the $\{x_i\}$ vectors are deterministic. Results for random $\{x_i\}$ could also be obtained by a conditioning argument. The following assumptions would be satisfied by a typical realization of independent, identically distributed random vectors $\{X_i\}$ with finite second moments and $\mathbb{P}\{X_i'\theta_0 = 0\} = 0$ and $\mathbb{P}X_iX_i'\{X_i'\theta_0 > 0\}$ nonsingular. The assumptions on the errors $\{u_i\}$ are the usual ones for least absolute deviations estimation. They could be relaxed slightly at the cost of increased notational complexity.

(i) The $\{u_i\}$ are independent, identically distributed random variables each having zero median and a continuous, strictly positive density $p(\cdot)$ near zero.

(ii) For each $\epsilon > 0$ there is a finite K such that

$$\frac{1}{n}\sum_{i\le n}|x_i|^2\{|x_i| > K\} < \epsilon \qquad \text{for all } n \text{ large enough.}$$

(iii) For each $\epsilon > 0$ there is a $\delta > 0$ such that

$$\frac{1}{n}\sum_{i\le n}|x_i|^2\{|x_i'\theta_0| \le \delta\} < \epsilon \qquad \text{for all } n \text{ large enough.}$$

(iv) The sequence of smallest eigenvalues of the matrices

$$\frac{1}{n}\sum_{i\le n}x_ix_i'\{x_i'\theta_0 > 0\}$$

is bounded away from zero, for n large enough.

Powell required slightly more smooothness for $p(\cdot)$, and a more awkward moment condition analogous to (iii), in order to fit his analysis into the framework of Huber's method.

(11.1) THEOREM. *Suppose θ_0 is an interior point of a Θ, a bounded subset of $\mathbb{R}^d$. Then, under assumptions (i) to (iv),*

$$2p(0)\sqrt{n}V_n(\widehat{\theta}_n - \theta_0) \rightsquigarrow N(0, I_d),$$

where V_n is the positive definite square root of the matrix from assumption (iv).

The proof of this result is quite a challenge. Let us begin with some heuristic arguments, which will develop notation and focus attention on the main technical difficulties.

Heuristics. The assumptions (ii), (iii), and (iv) control how much influence any single x_i can have over V_n. If $x_i'\theta_0 < 0$ then, for θ near θ_0, the term $|y_i^+ - (x_i'\theta)^+|$ reduces to y_i^+; it should not greatly affect the local minimization; it should not have an effect on the limiting distribution of $\widehat{\theta}_n$; it should not contribute to V_n. Assumption (iv) captures this idea. Assumption (ii) prevents a single very large

$|x_i|$ from dominating V_n; the least absolute deviations criterion prevents it from having a dominating influence on the minimization. The x_i with $x_i'\theta_0 \approx 0$ are the most troublesome, because their contribution to the criterion function is nonlinear in θ, even when θ is close to θ_0; assumption (iii) will allow us to ignore the combined effect of all such troublesome x_i.

The assumption of boundedness for the parameter set Θ is aesthetically irksome, even if it does have little practical significance. I would be pleased to learn how to dispose of it.

As a notational convenience, reparametrize by putting $t = V_n(\theta - \theta_0)$. Then define $z_{ni} = V_n^{-1}x_i$ and $x_i'\theta_0 = \mu_i$. Define

$$f_{ni}(\omega, t) = |y_i^+ - (\mu_i + z_{ni}'t)^+| - |y_i^+ - \mu_i^+|.$$

The centering ensures that

$$|f_{ni}(\omega, t)| \le |z_{ni}'t|,$$

and hence $f_{ni}(\cdot, t)$ has a finite expectation for each t. The centering does not affect the minimization; the standardized estimator $\widehat{t}_n = V_n(\widehat{\theta}_n - \theta_0)$ minimizes the process

$$G_n(\omega, t) = \frac{1}{n}\sum_{i \le n} f_{ni}(\omega, t).$$

Assumptions (ii) and (iv) imply existence of positive constants C' and C'' for which, when n is large enough,

$$C'|x_i| \ge |z_{ni}| \ge C''|x_i| \qquad \text{for } i \le n,$$

which lets us translate the assumptions on the $\{x_i\}$ into:

(ii)* For each $\epsilon > 0$ there is a finite K such that

$$\frac{1}{n}\sum_{i \le n} |z_{ni}|^2\{|z_{ni}| > K\} < \epsilon \qquad \text{for all } n \text{ large enough.}$$

(iii)* For each $\epsilon > 0$ there is a $\delta > 0$ such that

$$\frac{1}{n}\sum_{i \le n} |z_{ni}|^2\{|\mu_i| \le \delta\} < \epsilon \qquad \text{for all } n \text{ large enough.}$$

(iv)*

$$\frac{1}{n}\sum_{i \le n} z_{ni}z_{ni}'\{\mu_i > 0\} = I_d \qquad \text{for all } n \text{ large enough.}$$

For convenience, let us ignore from now on the finitely many n excluded by these three conditions.

As a first approximation, $G_n(\cdot, t)$ should be close to its expected value, $\Gamma_n(t)$. If we define

$$H_i(s) = \mathbb{P}\left(|y_i^+ - (\mu_i + s)^+| - |y_i^+ - \mu_i^+|\right) \qquad \text{for } s \in \mathbb{R},$$

then

$$\Gamma_n(t) = \mathbb{P}G_n(\cdot, t) = \frac{1}{n}\sum_{i \le n} H_i(z_{ni}'t).$$

Each H_i is expressible in terms of the function

$$M(s) = \mathbb{P}\left(|u_1 - s| - |u_1|\right).$$

Indeed, by separate consideration of the contributions from the sets $\{u_i < -\mu_i\}$ and $\{u_i \geq -\mu_i\}$ one can show

$$H_i(s) = \begin{cases} M(s) & \text{if } \mu_i \geq 0 \text{ and } s > -\mu_i, \\ M(-\mu_i) & \text{if } \mu_i \geq 0 \text{ and } s \leq -\mu_i, \\ M(s) - M(-\mu_i) & \text{if } \mu_i < 0 \text{ and } s > -\mu_i, \\ 0 & \text{if } \mu_i < 0 \text{ and } s \leq -\mu_i. \end{cases} \tag{11.2}$$

From assumption (i), the expected value $M(s)$ is increasing in $|s|$. The function M has a unique minimum at the origin, and

$$M(s) = p(0)s^2 + o(s^2) \qquad \text{near the origin.}$$

Moreover, there is a constant C such that

$$H_i(s) \leq M(s) \leq C\, s^2 \qquad \text{for all } s. \tag{11.3}$$

At least when t is small, we should be able to ignore those H_i with $\mu_i < 0$ or $\mu_i \approx 0$, to get, via (iv)*,

$$\Gamma_n(t) \approx \frac{1}{n} \sum_{i \leq n} p(0)|z_{ni}'t|^2 \{\mu_i > 0\} = p(0)|t|^2.$$

If $G_n(\cdot, t)$ lies uniformly close to its expectation, the $\widehat{t}_n$ that minimizes G_n should be drawn close to zero, where the approximation to Γ_n is minimized.

With the help of a maximal inequality for $G_n - \Gamma_n$, we will even be able to force $\widehat{t}_n$ into a $O_p(1/\sqrt{n})$ neighborhood of the origin. To learn more about $\sqrt{n}\,\widehat{t}_n$ we will then need a better approximation to G_n, obtained by a sort of Taylor expansion for f_{ni}.

The random function $f_{ni}(\omega, \cdot)$ has a derivative at $t = 0$ except perhaps when $\mu_i = 0$ or $u_i(\omega) = 0$. If we ignore these cases, straightforward differentation suggests we treat

$$\Delta_{ni}(\omega) = \Big(\{u_i(\omega) < 0\} - \{u_i(\omega) \geq 0\}\Big)\{\mu_i > 0\} z_{ni}$$

as the derivative of f_{ni} at $t = 0$. The difference

$$R_{ni}(\omega, t) = f_{ni}(\omega, t) - \Delta_{ni}(\omega)'t$$

should behave like the remainder in a Taylor expansion. By direct calculation for the various pairings of inequalities, one can verify that

$$|R_{ni}(\omega, t)| \leq 2|z_{ni}'t|\Big(\{|\mu_i| \leq |z_{ni}'t|\} + \{|u_i(\omega)| \leq |z_{ni}'t|\}\Big).$$

The two indicator functions vanish when $|z_{ni}'t|$ is smaller than both $|\mu_i|$ and $|u_i(\omega)|$, which should happen with large probability when $|t|$ is small.

Write W_n for $1/\sqrt{n}$ times the sum of the Δ_{ni}. By the Lindeberg central limit

theorem it has an asymptotic $N(0, I_d)$ distribution:

$$\mathbb{P}\Delta_{ni} = \Big(\mathbb{P}\{u_i < 0\} - \mathbb{P}\{u_i \geq 0\}\Big)\{\mu_i > 0\}z_{ni} = 0;$$

$$\frac{1}{n}\sum_{i\leq n}\mathbb{P}\Delta_{ni}\Delta_{ni}' = \{\mu_i > 0\}z_{ni}z_{ni}' = I_d;$$

$$\frac{1}{n}\sum_{i\leq n}\mathbb{P}|\Delta_{ni}|^2\{|\Delta_{ni}| > \epsilon\sqrt{n}\} = \frac{1}{n}\sum_{i\leq n}\{\mu_i > 0\}|z_{ni}|^2\{|z_{ni}| > \epsilon\sqrt{n}\} \to 0.$$

Ignoring the contributions from the $\{R_{ni}\}$, we get an improved approximation to G_n:

$$G_n(\omega, t) = \frac{1}{\sqrt{n}}W_n't + \Gamma_n(t) + \frac{1}{n}\sum_{i\leq n}\Big(R_{ni}(\omega, t) - \mathbb{P}\,R_{ni}(\cdot, t)\Big)$$

$$\simeq \frac{1}{\sqrt{n}}W_n't + p(0)|t|^2 \qquad \text{for small } |t|.$$

The random vector $\widehat{t}_n$ should be close to the vector $-W_n/2\sqrt{n}p(0)$ that minimizes the approximating quadratic, which leads us to the limit distribution asserted by Theorem 11.1.

Now let us make these arguments precise. The technical challenge in the proof will come from the two approximations to G_n. To obtain the necessary uniform bounds on the errors we will make use of maximal inequalities for processes with finite pseudodimension.

Behavior of Γ_n. From (11.2), it follows that $\Gamma_n(t) = \mathbb{P}\,G_n(\cdot, t)$ is an increasing function of $|t|$, taking its minimum value uniquely at $t = 0$. Given $\epsilon > 0$, choose K and δ according to (ii)* and (iii)*; then put $r = \delta/K$. From (11.3), the terms where $|\mu_i| \leq \delta$ or $|z_{ni}| > K$ contribute at most $2C\epsilon|t|^2$ to $\Gamma_n(t)$. For the other terms we have $|z_{ni}'t| \leq \delta$ if $|t| \leq r$. If $\mu_i \leq -\delta$ this makes $H_i(z_{ni}'t)$ zero. Within an error of $2C\epsilon|t|^2$, the expectation equals

$$\sum_i \{\mu_i > \delta, |z_{ni}| \leq K\}\Big(p(0)|z_{ni}'t|^2 + o(|z_{ni}'t|^2)\Big),$$

the $o(\cdot)$ being uniform in n and i. Adding back contributions from the terms where $|\mu_i| \leq \delta$ or $|z_{ni}| > K$, we then get via (iv)* that

$$\Gamma_n(t) = p(0)|t|^2 + o(|t|^2) \qquad \text{uniformly in } n. \tag{11.4}$$

In particular, if r is small enough,

$$\Gamma_n(t) \geq \frac{1}{2}p(0)|t|^2 \qquad \text{for all } n, \text{ all } |t| \leq r.$$

This local lower bound implies a global lower bound,

$$\liminf_{n\to\infty}\ \inf_{|t|\geq r}\ \Gamma_n(t) > 0 \qquad \text{for each } r > 0, \tag{11.5}$$

because $\Gamma_n(t)$ is an increasing function of $|t|$. The last inequality together with a uniform law of large numbers will imply consistency of $\widehat{t}_n$.

Manageability. We will need maximal inequalities for both $\{f_{ni}\}$ and $\{R_{ni}\}$. Let us verify that both processes generate random subsets of $\mathbb{R}^n$ with a pseudo-dimension determined by d. From the results in Section 4, this will imply that both processes are manageable.

Consider first the set $\mathcal{F}_{n\omega}$ of all points in $\mathbb{R}^n$ with coordinates $f_{ni}(\omega, t)$, as t ranges over the set

$$T_n = \{t \in \mathbb{R}^d : \theta_0 + V_n^{-1} t \in \Theta\}.$$

We need to find a dimension V such that, for every $\boldsymbol{\beta}$ in $\mathbb{R}^{V+1}$, no $(V+1)$-dimensional coordinate projection of $\mathcal{F}_{n\omega}$ can surround $\boldsymbol{\beta}$. This property is not affected if we translate $\mathcal{F}_{n\omega}$ by a fixed amount; it is good enough to establish the property for the set of points with coordinates

$$|y_i^+ - (\mu_i + z_{ni}' t)^+| = \max\left[y_i^+ - (\mu_i + z_{ni}' t)^+, (\mu_i + z_{ni}' t)^+ - y_i^+\right].$$

From the stability results in Section 5 for pseudodimension, it is good enough to treat the two terms in the maximum separately. Consider, for example, the set of points with coordinate $y_i^+ - (\mu_i + z_{ni}' t)^+$. Again we translate to eliminate the y_i^+. We now must determine, for $I = \{i_1, \ldots, i_k\}$, with k suitably large, and $\boldsymbol{\beta}$ a point in $\mathbb{R}^k$, whether it is possible to find for each $J \subseteq I$ a t in T_n for which

$$(\mu_i + z_{ni}' t)^+ \quad \begin{cases} > \beta_i & \text{for } i \in J, \\ < \beta_i & \text{for } i \in I \backslash J. \end{cases}$$

The inequalities when J is the empty set show that every β_i would have to be strictly positive, so the problem is unchanged if we replace $(\mu_i + z_{ni}' t)^+$ by $\mu_i + z_{ni}' t$, which is linear in t. Even if t ranges over the whole of $\mathbb{R}^d$, the points with these linear coordinate functions can at best trace out an affine subspace of dimension d. If $k = d + 1$, it is impossible to find for each J a t that satisfies the stated inequalities. By Lemma 5.1, the set $\mathcal{F}_{n\omega}$ has pseudodimension less than $10d$. (The bound could be improved, but there is no point in doing so; it matters only that the pseudodimension is the same for all n.)

Similar arguments serve to bound the pseudodimension for the set of points $\mathcal{R}_{n\omega}$ with coordinates $R_{ni}(\omega, t)/|t|$, as t ranges over the nonzero points in T_n. Indeed, inequalities

$$R_{ni}(\omega, t)/|t| \quad \begin{cases} > \beta_i & \text{for } i \in J, \\ < \beta_i & \text{for } i \in I \backslash J, \end{cases}$$

are equivalent to

$$|y_i^+ - (\mu_i + z_{ni}' t)^+| - |y_i^+ - \mu_i^+| - \Delta_{n_i}(\omega)' t - \beta_i |t| \quad \begin{cases} > 0 & \text{for } i \in J, \\ < 0 & \text{for } i \in I \backslash J. \end{cases}$$

Again several translations and appeals to the stability property for maxima reduces the problem to the result for affine subspaces of dimension d. The sets $\mathcal{R}_{n\omega}$ have pseudodimension less than $1000d$ (or something like that).

Maximal Inequalities. The sets T_n all lie within some bounded region of $\mathbb{R}^d$; there is a constant κ such that $|t| \le \kappa$ for every t in every T_n. It follows that

$$|f_{ni}(\omega, t)| \le \kappa |z_{ni}| \qquad \text{for all } t.$$

The maximal inequality (7.10) for manageable processes provides a constant C for which

$$\mathbb{P}\sup_t |G_n(\cdot,t) - \Gamma_n(t)|^2 \le \frac{C}{n^2}\sum_i |z_{ni}|^2.$$

Condition (ii)* bounds the sum on the right-hand side by a multiple of $1/n$. We deduce that

$$\sup_t |G_n(\cdot,t) - \Gamma_n(t)| = o_p(1). \tag{11.6}$$

[Actually we get $O_p(1/\sqrt{n})$, but $o_p(1)$ will suffice for our later purposes.] For the remainder terms we have a slightly more complicated envelope for t in a small neighborhood of the origin,

$$\sup_{|t|\le r}\frac{|R_{ni}(\omega,t)|}{|t|} \le 2|z_{ni}|\Big(\{|\mu_i| \le r|z_{ni}|\} + \{|u_i(\omega)| \le r|z_{ni}|\}\Big).$$

Maximal inequality (7.10) provides another constant C for which

$$\frac{1}{n}\mathbb{P}\sup_{0<|t|\le r}\left||t|^{-1}\sum_{i\le n} R_{n_i}(\omega,t) - \mathbb{P}R_{n_i}(\cdot,t)\right|^2 \le \frac{C}{n}\sum_{i\le n}|z_{ni}|^2\Big(\{|\mu_i| \le r|z_{ni}| + \mathbb{P}\{|u_i| \le r|z_{ni}|\}\Big).$$

By condition (ii)* the summands where $|z_{ni}| > K$ contribute at most $C\epsilon$ to the right-hand side. The remaining summands contribute at most

$$\frac{C}{n}\sum_{i\le n}|z_{ni}|^2\{|\mu_i| \le Kr\} + \frac{C}{n}\mathbb{P}\{|u_1| \le Kr\}\sum_{i\le n}|z_{ni}|^2.$$

Conditions (iii)* and (i) ensure that this contribution converges to zero, uniformly in n, as $r \to 0$. It follows that

$$\left|\frac{1}{n}\sum_{i\le n} R_{ni}(\omega,t) - \mathbb{P}R_n(\cdot,t)\right| = o_p(|t|/\sqrt{n})$$

uniformly in n and uniformly over t in shrinking neighborhoods of the origin. That is,

$$\begin{aligned} G_n(\omega,t) &= \Gamma_n(t) + \frac{1}{\sqrt{n}}W_n't + o_p(|t|/\sqrt{n}) \\ &= p(0)|t|^2 + o(|t|^2) + \frac{1}{\sqrt{n}}W_n't + o_p(|t|/\sqrt{n}) \qquad \text{uniformly,} \end{aligned} \tag{11.7}$$

where the uniformity is over all n and all t in a neighborhood $\{|t| \le r_n\}$, for every sequence $\{r_n\}$ of positive real numbers converging to zero.

Proof of the Theorem. Drop the ω from the notation. It will be enough if we show that $\hat{t}_n = o_p(1/\sqrt{n}) - W_n/2\sqrt{n}p(0)$. First establish consistency, by means of the inequality

$$G_n(\hat{t}_n) = \inf_t G_n(t) \le G_n(0) = 0.$$

The random point $\widehat{t}_n$ lies in the range over which the $o_p(1)$ bound from (11.6) holds. Approximating G_n by Γ_n we get

$$\Gamma_n(\widehat{t}_n) \le o_p(1).$$

Using (11.5) deduce that $\widehat{t}_n = o_p(1)$. If r_n tends to zero slowly enough,

$$\mathbb{P}\{|\widehat{t}_n| > r_n\} \to 0.$$

This brings $\widehat{t}_n$ into the range where we can appeal to (11.7) to deduce

$$G_n(\widehat{t}_n) = \Big(p(0) + o_p(1)\Big)\left|\widehat{t}_n + \frac{W_n + o_p(1)}{2\sqrt{n}\,p(0)}\right|^2 - \frac{|W_n|^2}{4np(0)} + o_p(1/n).$$

When $-W_n/2\sqrt{n}p(0)$ lies in T_n, which happens with probability tending to one because θ_0 is an interior point of Θ and $W_n = O_p(1/\sqrt{n})$, we can again invoke approximation (11.7) to get

$$G_n\big(-W_n/2\sqrt{n}p(0)\big) = -\frac{|W_n|^2}{4np(0)} + o_p(1/n).$$

From the comparison

$$G_n(\widehat{t}_n) \le G_n\big(-W_n/2\sqrt{n}p(0)\big),$$

deduce that

$$\left|\widehat{t}_n + \frac{W_n + o_p(1)}{2\sqrt{n}p(0)}\right|^2 = o_p(1/n),$$

from which the desired approximation to $\widehat{t}_n$ follows.

REMARKS. For the theory of (uncensored) least absolute deviations estimators see Bloomfield and Steiger (1983). A central limit theorem for such estimators was derived using elementary convexity arguments (which will reappear in Section 14) by Pollard (1990).

Chapter 10 of Amemiya (1985) describes many different approaches to estimation for Tobit models.