What's New in Econometrics? **NBER**, **Summer 2007**
Lecture 14, **Wednesday**, **August 1st**, **3**:**15**-**4**:**15 pm**

Quantile Methods

These notes review quantile estimation in a variety of situations, including models with endogenous explanatory variables – including endogenous treatment effects – and panel data models with unobserved heterogeneity. Recent work on interpreting quantile estimators when the quantile is misspecified is also covered.

1. Reminders About Means, Medians, and Quantiles

Consider the standard linear model in a population, with intercept α and $K \times 1$ slopes $\boldsymbol{\beta}$:

$$y = \alpha + \mathbf{x}\boldsymbol{\beta} + u. \tag{1.1}$$

Assume $E(u^2) < \infty$, so that the distribution of u is not too spread out. Given a large random sample, when should we expect ordinary least squares, which solves

$$\min_{a,\mathbf{b}} \sum_{i=1}^{N} (y_i - a - \mathbf{x}_i\mathbf{b})^2, \tag{1.2}$$

and least absolute deviations (LAD), which solves

$$\min_{a,\mathbf{b}} \sum_{i=1}^{N} |y_i - a - \mathbf{x}_i\mathbf{b}|, \tag{1.3}$$

to provide similar parameter estimates? There are two important cases. If

$$D(u|\mathbf{x}) \text{ is symmetric about zero} \tag{1.4}$$

then OLS and LAD both consistently estimate α and $\boldsymbol{\beta}$. If

$$u \text{ is independent of } \mathbf{x} \text{ with } E(u) = 0, \tag{1.5}$$

where $E(u) = 0$ is the normalization that identifies α, then OLS and LAD both consistently estimate the slopes, $\boldsymbol{\beta}$. If u has an asymmetric distribution, then $Med(u) \equiv \eta \neq 0$, and $\hat{\alpha}_{LAD}$ converges to $\alpha + \eta$ because $Med(y|\mathbf{x}) = \alpha + \mathbf{x}\boldsymbol{\beta} + Med(u|\mathbf{x}) = \alpha + \mathbf{x}\boldsymbol{\beta} + \eta$. Of course, independence between u and $\mathbf{x}$ rules out heteroskedasticity in $Var(u|\mathbf{x})$.

In many applications, neither (1.4) nor (1.5) is likely to be true. For example, y may be a measure of wealth, in which case the error distribution is probably asymmetric and $Var(u|\mathbf{x})$ not constant. Therefore, it is important to remember that if $D(u|\mathbf{x})$ is asymmetric and changes with $\mathbf{x}$, then we should not expect OLS and LAD to deliver similar estimates of $\boldsymbol{\beta}$, even for "thin-tailed" distributions. In other words, it is important to separate discussions of resiliency to outliers from the different quantities identified by least squares (the conditional mean,

$E(y|\mathbf{x})$) and least absolution deviations (the conditional median, $Med(y|\mathbf{x})$). Of course, it is true that LAD is much more resilient to changes in extreme values because, as a measure of central tendency, the median is much less sensitive than the mean to changes in extreme values. But a significant difference between OLS and LAD should not lead one to somehow prefer LAD. It is possible that $E(y|\mathbf{x}) = \alpha + \mathbf{x}\boldsymbol{\beta}$, $Med(y|\mathbf{x})$ is not linear, and therefore LAD does not consistently estimate $\boldsymbol{\beta}$. Generally, if we just use linear models as approximations to underlying nonlinear functions, we should not be surprised if the linear approximation to the conditional mean, and that for the median, can be very different. (Warning: Other so-called "robust" estimators, which are intended to be insensitive to outliers or influential data, usually require symmetry of the error distribution for consistent estimation. Thus, they are not "robust" in the sense of delivering consistency under a wide range of assumptions.)

Sometimes one can use a transformation to ensure conditional symmetry or the independence assumption in (1.5). When $y_i > 0$, the most common transformation is the natural log. Often, the linear model $\log(y) = \alpha + \mathbf{x}\boldsymbol{\beta} + u$ is more likely to satisfy symmetry or independence. Suppose that symmetry about zero holds in the linear model for $\log(y)$. Then, because the median passes through monotonic functions (unlike the expectation), $\text{Med}(y|\mathbf{x}) = \exp(\text{Med}[\log(y)|\mathbf{x}]) = \exp(\alpha + \mathbf{x}\boldsymbol{\beta})$, and so we can easily recover the partial effects on the median of y itself. By contrast, we cannot generally find $\text{E}(y|\mathbf{x}) = \exp(\alpha + \mathbf{x}\boldsymbol{\beta})\text{E}[\exp(u)|\mathbf{x}]$. If, instead, we assume $D(u|\mathbf{x}) = D(u)$, then $\text{Med}(y|\mathbf{x})$ and $\text{E}(y|\mathbf{x})$ are both exponential functions of $\mathbf{x}\boldsymbol{\beta}$, but with different "intercepts" inside the exponential function.

The fact that the median passes through monotonic functions is very handy for applying LAD to a variety of problems, particularly corner solution responses where an outcome has nonnegative support and a mass point at zero. But the expectation operator has useful properties that the median does not: linearity and the law of iterated expectations. To see how these help to identify interesting quantities, suppose we begin with a random coefficient model

$$y_i = a_i + \mathbf{x}_i\mathbf{b}_i, \tag{1.6}$$

where a_i is the heterogeneous intercept and $\mathbf{b}_i$ is a $1 \times K$ matrix of heterogeneous slopes ("random coefficients"). If we assume that $(a_i, \mathbf{b}_i)$ is independent of $\mathbf{x}_i$, then

$$\text{E}(y_i|\mathbf{x}_i) = \text{E}(a_i|\mathbf{x}_i) + \mathbf{x}_i\text{E}(\mathbf{b}_i|\mathbf{x}_i) \equiv \alpha + \mathbf{x}_i\boldsymbol{\beta}, \tag{1.7}$$

where $\alpha = \text{E}(a_i)$ and $\boldsymbol{\beta} = \text{E}(\mathbf{b}_i)$. Because OLS consistently estimates the parameters of a

conditional mean linear in those parameters, OLS consistently estimates the population averaged effects, or average partial effects, $\boldsymbol{\beta}$. Even under independence, there is no way to derive $\text{Med}(y_i|\mathbf{x}_i)$ without imposing more restrictions. In general, LAD of y_i on $1, \mathbf{x}_i$ does not consistently estimate $\boldsymbol{\beta}$ or the medians of the elements of b_{ij}. Are there any reasonable assumptions that imply LAD consistenty estimates something of interest in (1.7)? Yes, although multivariate symmetry is involved. With multivariate distributions there is no unique definition of symmetry. A fairly strong restriction is the notion of a *centrally symmetric* distribution (Serfling (2006)). If $\mathbf{u}_i$ is a vector, then its distribution conditional on $\mathbf{x}_i$ is centrally symmetric if

$$\text{D}(\mathbf{u}_i|\mathbf{x}_i) = \text{D}(-\mathbf{u}_i|\mathbf{x}_i). \tag{1.8}$$

This condition implies that, for any $\mathbf{g}_i$ a function of $\mathbf{x}_i$, $\text{D}(\mathbf{g}_i'\mathbf{u}_i|\mathbf{x}_i)$ has a univariate distribution that is symmetric about zero. Of course, (1.8) implies that $E(\mathbf{u}_i|\mathbf{x}_i) = \mathbf{0}$.

We can apply this to the random coefficient model as follows. Write $\mathbf{c}_i = (a_i, \mathbf{b}_i)$ with $\boldsymbol{\gamma} = E(\mathbf{c}_i)$, and let $\mathbf{d}_i = \mathbf{c}_i - \boldsymbol{\gamma}$. Then we can write

$$\begin{aligned} y_i &= \alpha + \mathbf{x}_i\boldsymbol{\beta} + (a_i - \alpha) + \mathbf{x}_i(\mathbf{b}_i - \boldsymbol{\beta}) \\ &\equiv \alpha + \mathbf{x}_i\boldsymbol{\beta} + \mathbf{g}_i'\mathbf{d}_i \end{aligned} \tag{1.9}$$

with $\mathbf{g}_i = (1, \mathbf{x}_i)$. Therefore, if $\mathbf{c}_i$ has a centrally symmetric distribution about $\boldsymbol{\gamma}$, then $Med(\mathbf{g}_i'\mathbf{d}_i|\mathbf{x}_i) = 0$, and LAD applied to the usual model $y_i = \alpha + \mathbf{x}_i\boldsymbol{\beta} + u_i$ consistently estimates α and $\boldsymbol{\beta}$. Because a_i and $\mathbf{b}_i$ have centrally symmetric distributions about their α and $\boldsymbol{\beta}$, respectively, it is clear that these are the only sensible measures of central tendency in the distribution of $\mathbf{c}_i$.

Usually, we are interested in how covariates affect quantiles other than the median, in which case quantile estimation is applied to a sequence of linear models. Write the τ^{th} quantile in the distribution $D(y_i|\mathbf{x}_i)$ as $\text{Quant}_\tau(y_i|\mathbf{x}_i)$. Under linearity,

$$\text{Quant}_\tau(y_i|\mathbf{x}_i) = \alpha(\tau) + \mathbf{x}_i\boldsymbol{\beta}(\tau) \tag{1.10}$$

where, in general, the intercept and slopes depend on the quantile, τ. Under (1.10), consistent estimators of $\alpha(\tau)$ and $\boldsymbol{\beta}(\tau)$ are obtained by minimizing the *asymmetric absolute loss function* or the "check" function:

$$\min_{\alpha \in \mathbb{R}, \boldsymbol{\beta} \in \mathbb{R}^K} \sum_{i=1}^{N} c_\tau(y_i - \alpha - \mathbf{x}_i\boldsymbol{\beta}), \tag{1.11}$$

where

$$c_\tau(u) = (\tau 1[u \geq 0] + (1-\tau)1[u < 0])|u| = (\tau - 1[u < 0])u \tag{1.12}$$

and $1[\cdot]$ is the "indicator function." Consistency is relatively easy to establish because the objective function is continuous in its parameters. Asymptotic normality is more difficult because any sensible definition of the Hessian of the objective function, away from the nondifferentiable kink, is identically zero. But it has been worked out under a variety of conditions; see Koenker (2005) for a recent treatment.

2. Some Useful Asymptotic Results

2.1. What Happens if the Quantile Function is Misspecified?

When we use OLS to estimate the parameters of a linear model, we always have a simple characterization of the plim of the OLS estimator when the mean is not linear: If α^* and $\boldsymbol{\beta}^*$ are the plims from the OLS regression y_i on $1, \mathbf{x}_i$ then these provide the smallest mean squared error approximation to $E(y|\mathbf{x}) = \mu(\mathbf{x})$. In other words, $(\alpha^*, \boldsymbol{\beta}^*)$ solves

$$\min_{a,\mathbf{b}} E[(\mu(\mathbf{x}) - \alpha - \mathbf{x}\boldsymbol{\beta})^2], \tag{2.1}$$

where, of course, the expectation is over the distribution of $\mathbf{x}$. Under some restrictions, (albeit restrictive), β_j^* is the average partial effect $E_{\mathbf{x}}[\partial\mu(\mathbf{x})/\partial x_j]$ – multivariate normality of $\mathbf{x}$ is sufficient – and under less restrictive (but still restrictive) assumptions, the β_j^* estimate the average partial effects up. These follow from the work of Chung and Goldberger (1984), Ruud (1984), and Stoker (1986).

Although the linear formulation of quantiles has been viewed by some – for example, Buchinsky (1991) and Chamberlain (1991) – as a linear approximation to the true conditional quantile, most of the the linear model is treated as being correctly specified. In some ways, this is strange because usually many quantiles are estimated. Yet assuming that different quantiles are linear in the same functions of $\mathbf{x}$ might be unrealistic.

Angrist, Chernozhukov, and Fernandez-Val (2006) provide a treatment of quantile regression under misspecification of the quantile function and characterize the probability limit of the LAD estimator. To describe the result, absorb the intercept into $\mathbf{x}$ and, rather than assume a correctly specified conditional quantile, let $\boldsymbol{\beta}(\tau)$ be the solution to the population quantile regression problem. Therefore, $\mathbf{x}\boldsymbol{\beta}(\tau)$ is the plim of the estimated quantile function. ACF have a couple of different ways to characterize $\boldsymbol{\beta}(\tau)$. One result is that $\boldsymbol{\beta}(\tau)$ solves

$$\min_{\beta} E\{w_\tau(\mathbf{x},\boldsymbol{\beta})[q_\tau(\mathbf{x}) - \mathbf{x}\boldsymbol{\beta})]^2\}, \tag{2.2}$$

where the weight function $w_\tau(\mathbf{x},\boldsymbol{\beta})$ is

$$w_\tau(\mathbf{x},\boldsymbol{\beta}) = \int_0^1 (1-u) f_{y|x}(u\mathbf{x}\boldsymbol{\beta} + (1-u)q_\tau(\mathbf{x})|\mathbf{x})du \geq 0. \tag{2.3}$$

In other words, $\boldsymbol{\beta}(\tau)$ is the best weighted mean square approximation to the true quantile function, where the weights are the average of the conditional density of y_i over a line from the candidate approximation, $\mathbf{x}\boldsymbol{\beta}$, to the true quantile function, $q_\tau(\mathbf{x})$. The multiplication of the density by $(1-u)$ gives more weight to points closer to the true conditional quantile. It is interesting that the ACF characterization is in terms of a weighted mean squared error, a concept we usually associate with conditional mean approximation. ACF also show an approximation where the weighting function does not depend on $\boldsymbol{\beta}$, and use it characterize a "partial" regression quantiles, and to characterize omitted variables bias with quantile regression.

2.2. Computing Standard Errors

First consider the case where we want to estimate the parameters in a linear quantile model, for a given quantile, τ. For a random draw, write

$$y_i = \mathbf{x}_i\boldsymbol{\theta} + u_i, \text{ Quant}_\tau(u_i|\mathbf{x}_i) = 0, \tag{2.4}$$

where we include unity in $\mathbf{x}_i$ so that contains an intercept and the slopes. Let $\hat{\boldsymbol{\theta}}$ be the quantile estimators, and define the quantile regression residuals, $\hat{u}_i = y_i - \mathbf{x}_i\hat{\boldsymbol{\theta}}$. Under weak conditions (see, for example, Koenker (2005)), $\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ is asymptotically normal with asymptotic variance

$$\mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}, \tag{2.5}$$

where

$$\mathbf{A} \equiv \text{E}[f_u(0|\mathbf{x}_i)\mathbf{x}_i'\mathbf{x}_i] \tag{2.6}$$

and

$$\mathbf{B} \equiv \tau(1-\tau)\text{E}(\mathbf{x}_i'\mathbf{x}_i). \tag{2.7}$$

Expression (2.5) is the now familiar standard "sandwich" form of asymptotic variances. It is fully robust in the sense that it is valid without further assumptions on $D(u_i|\mathbf{x}_i)$. The matrix $\mathbf{B}$ is simple to estimate as

$$\hat{\mathbf{B}} = \tau(1-\tau)\left(N^{-1}\sum_{i=1}^{N}\mathbf{x}_i'\mathbf{x}_i\right), \tag{2.8}$$

where $0 < \tau < 1$ is the chosen quantile. This estimator is consistent under the weak assumption of finite second moments for $\mathbf{x}_i$. The matrix $\mathbf{A}$ is harder to estimate because of the presence of $f_u(0|\mathbf{x}_i)$, and we do not have a parametrix model for the density of u_i given $\mathbf{x}_i$. But we only have to estimate this conditional density at $u = 0$, so we could use a nonparametric density estimator (based on the $\hat{u}_i$). Powell (1986, 1991) proposed a simpler approach, which leads to

$$\hat{\mathbf{A}} = (2Nh_N)^{-1}\sum_{i=1}^{N}1[|\hat{u}_i| \leq h_N]\mathbf{x}_i'\mathbf{x}_i, \tag{2.9}$$

where $\{h_N > 0\}$ is a nonrandom sequence shrinking to zero as $N \to \infty$ with $\sqrt{N}h_N \to \infty$. are sufficient for consistency. The second condition controls how quickly h_N shrinks to zero. For example, $h_N = aN^{-1/3}$ for any $a > 0$ satisfies these conditions. The practical problem in choosing a (or choosing h_N more generally) is discussed by Koenker (2005), who also discusses some related estimators. In particular, in equation (2.9), observation i does not contribute if $|\hat{u}_i| > h_N$. Other methods allow each observation to enter the sum but with a weight that declines as $|\hat{u}_i|$ increases. (As an interesting aside, the derivation of (2.9) involves the simple equality $\mathrm{E}\{(1[|u_i| \leq h_N]|\mathbf{x}_i)\mathbf{x}_i'\mathbf{x}_i\} = \mathrm{E}(1[|u_i| \leq h_N]\mathbf{x}_i'\mathbf{x}_i)$, which is analogous to the key step in the regression frameworks for justifying the heteroskedasticity-robust variance matrix estimator.)

The nonparametric bootstrap can be applied to quantile regression, but if the data set is large, the computation using several hundred bootstrap samples can be costly.

If we assume that u_i is independent of $\mathbf{x}_i$ then $f_u(0|\mathbf{x}_i) = f_u(0)$ and equation (2.5) simplifies to

$$\frac{\tau(1-\tau)}{[f_u(0)]^2}[\mathrm{E}(\mathbf{x}_i'\mathbf{x}_i)]^{-1} \tag{2.10}$$

and its estimator has the general form

$$\frac{\tau(1-\tau)}{[\hat{f}_u(0)]^2}\left(N^{-1}\sum_{i=1}^{N}\mathbf{x}_i'\mathbf{x}_i\right)^{-1}, \tag{2.11}$$

and a simple, consistent estimate of $f_u(0)$ is the histogram estimator

$$\hat{f}_u(0) = (2Nh_N)^{-1} \sum_{i=1}^{N} 1[|\hat{u}_i| \leq h_N]. \tag{2.12}$$

Of course, one can use other kernel estimators for $\hat{f}_u(0)$. This nonrobust estimator is the one commonly reported as the default be statistical packages, including Stata.

If the quantile function is misspecified, even the "robust" form of the variance matrix, based on the estimate in (2.9), is not valid. In the generalized linear models and generalized estimating equations literature, the distinction is sometimes made between a "fully robust" variance estimator and a "semi-robust" variance estimator. In the GLM and GEE literatures, the semi-robust estimator assumes $E(y_i|\mathbf{x}_i)$, or the panel version of it, is correctly specified, but does not impose restrictions on $Var(y_i|\mathbf{x}_i)$ or other features of $D(y_i|\mathbf{x}_i)$. On the other hand, a fully robust variance matrix estimator is consistent for the asymptotic variance even if the mean function is misspecified. For, say, nonlinear least squares, or quasi-MLE in the linear exponential family, one needs to include the second derivative matrix of the conditional mean function to have a fully robust estimator. For some combinations of mean functions and objective functions, the Hessian of the mean function disappears, and the fully robust and semi-robust estimators are the same. For two-step methods, such as GEE, analytical formulas for fully robust estimators are very difficult to obtain, and almost all applications use the semi-robust form. This is a long-winded way to say that there is precedent for worrying about how to estimate asymptotic variances when the main feature being estimated is misspecified. In GEE terminology, $\hat{\mathbf{A}}^{-1}\hat{\mathbf{B}}\hat{\mathbf{A}}^{-1}$ where $\hat{\mathbf{A}}$ is given by (2.9), is only semi-robust.

Kim and White (2002) and Angrist, Chernozhukov, and Fernández-Val (2006) provide a fully robust variance matrix estimator when the linear quantile function is possibly misspecified. The estimator of $\mathbf{A}$ in (2.9) is still valid, but the estimator of $\mathbf{B}$ needs to be extended. If we use the outer product of the score we obtain

$$\hat{\mathbf{B}} = \left(N^{-1} \sum_{i=1}^{N} (\tau - 1[\hat{u}_i < 0])^2 \mathbf{x}_i' \mathbf{x}_i \right), \tag{2.13}$$

where the $\hat{u}_i$ are the residuals from the (possibly) misspecified quantile regression, is generally consistent.

As shown by Hahn (1995, 1997), the nonparametric bootstrap (and the Bayesian bootstrap) generally provides consistent estimates of the fully robust variance without claims about the conditional mean being correct. It does not, however, provide asymptotic refinements for

testing and confidence intervals compared with those based on first-order asymptotics. See Horowitz (2001) for a discussion, and on how to smooth the problem so that refinements are possible.

ACF actually provide the covariance function for the process $\{\hat{\boldsymbol{\theta}}(\tau) : \varepsilon \leq \tau \leq 1 - \varepsilon\}$ for some $\varepsilon > 0$, which can be used to test hypotheses jointly across multiple quantiles (including all quantiles at once).

As an example of quantile regression, we use the data from Abadie (2003). Stata was used to do the estimation and obtain the standard errors; these are the nonrobust standard errors that use

Dependent Variable:	*nettfa*					
	(1)	(2)	(3)	(4)	(5)	(6)
Explanatory Variable	Mean (OLS)	.10 Quantile	.25 Quantile	Median (LAD)	.75 Quantile	.90 Quantile
inc	.783	−.0179	.0713	.324	.798	1.291
	(.104)	(.0177)	(.0072)	(.012)	(.025)	(.048)
age	−1.568	−.0663	.0336	−.244	−1.386	−3.579
	(1.076)	(.2307)	(.0955)	(.146)	(.287)	(.501)
age^2	.0284	.0024	.0004	.0048	.0242	.0605
	(.0138)	(.0027)	(.0011)	(.0017)	(.0034)	(.0059)
*e*401*k*	6.837	.949	1.281	2.598	4.460	6.001
	(2.173)	(.617)	(.263)	(.404)	(.801)	(1.437)
N	2,017	2,017	2,017	2,017	2,017	2,017

The effect of income is very different across quantiles, with its largest effect at upper quantiles. Similarly, eligibility for a 401(k) plan has a much larger effect on financial wealth at the upper end of the wealth distribution. The mean and median slope estimates are very different, implying that the model with an additive error that is either independent of the covariates, or has a symmetric distribution given the covariates, is not a good characterization.

3. Quantile Regression with Endogenous Explanatory Variables

Recently, there has been much interest in using quantile regression in models with endogenous explanatory variables. Some strategies are fairly simple, others are more complicated. Suppose we start with the model

$$y_1 = \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + u_1, \tag{3.1}$$

where the full vector of exogenous variables is $\mathbf{z}$ and y_2 is potential endogenous – whatever

that means in the context of quantile regression. The most straightforward case to handle is least absolute deviations, because median restrictions are easier to justify when joint distributions are involved.

Amemiya's (1982) two-stage LAD estimator, whose asymptotic properties were derived by Powell (1986), adds a reduced form for y_2, say

$$y_2 = \mathbf{z}\boldsymbol{\pi}_2 + v_2. \tag{3.2}$$

While (3.2) can be estimated by OLS to obtain $\hat{\boldsymbol{\pi}}_2$, using LAD in the first stage to estimate $\boldsymbol{\pi}_2$ is more in the spirit of 2SLAD. In the second step, the fitted values, $\hat{y}_{i2} = \mathbf{z}_i\hat{\boldsymbol{\pi}}_2$, are inserted in place of y_{i2} to given LAD of y_{i1} on $\mathbf{z}_{i1}, \hat{y}_{i2}$. By replacing $\hat{\boldsymbol{\pi}}_2$ with $\boldsymbol{\pi}_2$, it is clear that the 2SLAD estimator essentially requires symmetry of the composite error $\alpha_1 v_2 + u_1$. While the properties of 2SLAD were originally worked out for nonstochastic $\mathbf{z}_i$ – so that (u_{i1}, v_{i2}) is independent of $\mathbf{z}_i$ – it is clear that symmetry of $\alpha_1 v_2 + u_1$ given $\mathbf{z}$ is sufficient.

We might as well assume $D(u_1, v_2|\mathbf{z})$ is centrally symmetric, in which case a control function approach can be used, too. Write

$$u_1 = \rho_1 v_2 + e_1, \tag{3.3}$$

where e_1 given $\mathbf{z}$ would have a symmetric distribution. Because $Med(v_2|\mathbf{z}) = 0$, the first stage estimator can be LAD. Given the LAD residuals $\hat{v}_{i2} = y_{i2} - \mathbf{z}_i\hat{\boldsymbol{\pi}}_2$, these residuals can be added to second-stage LAD. So, we do LAD of y_{i1} on $\mathbf{z}_{i1}, y_{i2}, \hat{v}_{i2}$. It seems likely that a t test on $\hat{v}_{i2}$ is valid as a test for the null that y_2 is exogenous.

There can be problems of interpretation in just applying either 2SLAD or the CF approach. Suppose we view this as an omitted variable problem, where a_1 is the omitted variable, and interest lies in the "structual" median

$$Med(y_1|\mathbf{z}, y_2, a_1) = Med(y_1|\mathbf{z}_1, y_2, a_1) = \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + a_1. \tag{3.4}$$

Then we can write

$$y_1 = \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + a_1 + e_1 \tag{3.5}$$

$$Med(e_1|\mathbf{z}, y_2, a_1) = 0. \tag{3.6}$$

If (3.4) was stated in terms of means, then $E(e_1|\mathbf{z}) = 0$ by construction, and a very sensible exogeneity condition is $E(a_1|\mathbf{z}) = E(a_1) = 0$ (as a normalization) or that $Cov(\mathbf{z}, a_1) = \mathbf{0}$. But here we cannot even assert that $Med(e_1|\mathbf{z}) = Med(e_1)$ because (3.6) does not imply this; there is no law of iterated medians. To further compound the problem, the median of the sum is not the sum of the medians; so, even if we stated exogeneity as $Med(a_1|\mathbf{z}) = Med(a_1)$ and just

asserted $Med(e_1|\mathbf{z}) = Med(e_1)$, $a_1 + e_1 = u_1$ would not generally satisfy $Med(u_1|\mathbf{z}) = Med(u_1)$. Of course, we can make enough multivariate symmetric assumptions so that all linear combinations of errors have symmetric distributions. But then LAD methods are purely to guard against outliers; usual 2SLS will provide consisten, asymptotically normal estimates of the parameters under symmetry (and, of course, weaker assumptions).

With quantile estimation, such two-step estimators are even more difficult to justify. The Angrist, Chernozhukov, and Fernandez-Val (2006) partialling out representations can provide some sort of interpretation of netting out the control, v_2, but it is difficult to know whether the parameters are ultimately interesting.

Abadie (2003) and Abadie, Angrist, and Imbens (2002) show how to define and estimate policy parameters with a binary endogenous treatment, say D, and binary instrumental variable, say Z. The outcome is Y with observed covariates, X. The potential outcomes on Y are Y_d, $d = 0, 1$ – that is, without treatment and with treatment, respectively. The counterfactuals for treatment are D_z, $z = 0, 1$. Thus, D_0 is what treatment status would be if the instrument (often, randomized eligibility) equals zero, and D_1 is treatment status if $Z = 1$. The data we observe are $X, Z, D = (1 - Z)D_0 + ZD_1$, and $Y = (1 - D)Y_0 + DY_1$. As discussed in AAI, identification of average treatment effects, and ATE on the treated, is difficult. Instead, they focus on treatment effects for *compliers*, that is, the (unobserved) subpopulation with $D_1 > D_0$. This is the group of subjects who do not participate if ineligible but do participate if eligible.

AAI specify the linear equation

$$Quant_\tau(Y|X, D, D_1 > D_0) = \alpha_\tau D + X\beta_\tau, \tag{3.7}$$

and define α_τ as the *quantile treatment effect* (QTE) for compliers. If we observed the event $D_1 > D_0$, then (3.7) could be estimated by standard quantile regression using the subsample of compliers. But, in effect, the binary variable $1[D_1 > D_0]$ is an omitted variable. But Z is an available instrument for D. As discussed by AAI, (3.7) identifies differences in quantiles on the potential outcomes, Y_1 and Y_0, and not the quantile of the difference, $Y_1 - Y_0$. The latter effects are harder to identify. (Of course, in the case of mean effects, there is no difference in the two effects.)

The assumptions used by AAI to identify α_τ are

$$(Y_1, Y_0, D_1, D_0) \text{ is independent of } Z \text{ conditional on } X \tag{3.8}$$

$$0 < P(Z = 1|X) < 1 \tag{3.9}$$

$$P(D_1 = 1|X) \neq P(D_0 = 1|X) \tag{3.10}$$

$$P(D_1 \geq D_0|X) = 1. \tag{3.11}$$

Under these assumptions, AAI show that a weighted quantile estimation identifies α_τ. The estimator that is computationally most convenient is obtained as follows. Define

$$\kappa_v(U) = 1 - \frac{D(1 - v(U))}{1 - \pi(X)} - \frac{(1 - D)v(U)}{\pi(X)}, \tag{3.12}$$

where $U = (Y,D,X)$, $v(U) = P(Z = 1|U)$, and $\pi(X) = P(Z = 1|X)$. AAI show that $\kappa_\tau(u) = P(D_1 > D_0|U = u)$, and so this weighting function is nonnegative. They also show that α_τ and β_τ in (3.7) solve

$$\min_{\alpha,\beta} E[\kappa_\tau(U)c_\tau(Y - \alpha D - X\beta)], \tag{3.13}$$

where $c_\tau(\cdot)$ is the check function defined earlier. To operationalize the estimate, $\kappa_\tau(\cdot)$ needs to be estimated, which means estimating $P(Z = 1|Y,D,X)$ and $P(Z = 1|X)$. AAI use linear series estimators to approximate $P(Z = 1|Y,D,X)$ and $P(Z = 1|X)$, and derive the asymptotic variance of the two-step estimator that solves

$$\min_{\delta} \sum_{i=1}^{N} 1[\hat{\kappa}_v(U_i) \geq 0]\hat{\kappa}_v(U_i)c_\tau(Y_i - W_i\delta), \tag{3.14}$$

where $W_i = (D_i,X_i)$ and δ contains α and β. The indicator function $1[\hat{\kappa}_v(U_i) \geq 0]$ ensures that only observations with nonnegative weights are used. Asymptotically, $\hat{\kappa}_v(u) \geq 0$, and this trimming of observations becomes less and less necessary. To ensure that $\hat{v}(u)$ and $\hat{\pi}(x)$ act like probabilities, series estimation using logit functions, as in Hirano, Imbens, and Ridder (2003), might be preferred (although that still would not ensure nonnegativity of $\hat{\kappa}_v(U_i)$ for all i).

Other recent work has looked at quantile estimation with endogenous treatment effects. Chernozhukov and Hansen (2005, 2006) consider indentification and estimation of QTEs in a model with endogenous treatment and without imposing functional form restrictions. Let $q(d,x,\tau)$ denote the τ^{th} quantile function for treatment level $D = d$ and covariates x. In the binary case, CH define the QTE as

$$QTE_\tau(x) = q(1,x,\tau) - q(0,x,\tau). \tag{3.15}$$

Using a basic result from probability, the average treatment effect, again conditional on x, can be obtained by integrating (3.15) over $0 < \tau < 1$.

The critical representation used by CH is that each potential outcome, Y_d, conditional on $X = x$, can be expressed as

$$Y_d = q(d, x, U_d) \tag{3.16}$$

where

$$U_d | Z \sim \text{Uniform}(0, 1), \tag{3.17}$$

and Z is the instrumental variable for treatment assignment, D. Thus, D is allowed to be correlated with U_d. Key assumptions are that $q(d, x, u)$ is strictly increasing in u and a "rank invariance" condition. The simplest form of the condition is that , conditional on $X = x$ and $Z = z$, U_d does not depend on d. The CH show that, with the observed Y defined as $Y = q(D, X, U_D)$,

$$P[Y \leq q(D, X, \tau) | X, Z] = P[Y < q(D, X, \tau) | X, Z] = \tau. \tag{3.18}$$

Equation (3.18) acts as a nonparametric conditional moment condition which, under certain assumptions, allows identification of $q(d, x, \tau)$. If we define $R = Y - q(D, X, \tau)$, then (3.18) implies that the τ^{th} quantile of R, conditional on (X, Z), is zero. This is similar to the more common situation where we have a conditional moment condition of the form $E(R|X, Z) = 0$. See Chernozhukov and Hansen (2005) for details concerning identification – they apply results of Newey and Powell (2003) – and Chernozhukov and Hansen (2005) for estimation methods, where they assume a linear form for $q(d, x, \tau)$ and obtain what they call the *quantile regression instrumental variables estimator*.

Other work that uses monotonicity assumptions and identifies structural quantile functions is Chesher (2003) and Imbens and Newey (2006).

4. Quantile Regression for Panel Data

Quantile regression methods can be applied to panel data, too. For a given quantile $0 < \tau < 1$, suppose we specify

$$\text{Quant}_\tau(y_{it} | \mathbf{x}_{it}) = \mathbf{x}_{it}\boldsymbol{\theta}, \; t = 1, \ldots, T, \tag{4.1}$$

where $\mathbf{x}_{it}$ probably allows for a full set of time period intercepts. Of course, we can write $y_{it} = \mathbf{x}_{it}\boldsymbol{\theta} + u_{it}$ where $\text{Quant}_\tau(u_{it} | \mathbf{x}_{it}) = 0$. The natural estimator of $\boldsymbol{\theta}_o$ is the pooled quantile regression estimator. Unless we assume that (3.1) has correctly specified dynamics, the variance matrix needs to be adjusted for serial correlation in the resulting score of the objective

function. These scores have the form

$$\mathbf{s}_{it}(\boldsymbol{\theta}) = -\mathbf{x}_{it}'\{\tau 1[y_{it} - \mathbf{x}_{it}\boldsymbol{\theta} \geq 0] - (1-\tau)1[y_{it} - \mathbf{x}_{it}\boldsymbol{\theta} < 0]\}, \tag{4.2}$$

which can be shown to have zero mean (at the "true" parameter), conditional on $\mathbf{x}_{it}$, under (4.1). The serial dependence properties are not restricted, nor is heterogeneity in the distributions across t. A consistent estimator of $\mathbf{B}$ (with T fixed and $N \to \infty$) is

$$\hat{\mathbf{B}} = N^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{r=1}^{T}\mathbf{s}_{it}(\hat{\boldsymbol{\theta}})\mathbf{s}_{ir}(\hat{\boldsymbol{\theta}})'. \tag{4.3}$$

This estimator is not robust to misspecification of the conditional quantiles, but the extension of Angrist, Chernozhukov, and Fernandez-Val (2006) should work in the pooled panel data case as well..

Estimation of $\mathbf{A}$ is similar to the cross section case. A robust estimator, that does not assume independence between u_{it} and $\mathbf{x}_{it}$, and allows the distribution of u_{it} to change across t, is

$$\hat{\mathbf{A}} = (2Nh_N)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}1[|\hat{u}_{it}| \leq h_N]\mathbf{x}_{it}'\mathbf{x}_{it}, \tag{4.4}$$

or, we can replace the indicator function with a smoothed version. Rather than using $\hat{\mathbf{A}}^{-1}\hat{\mathbf{B}}\hat{\mathbf{A}}^{-1}/N$ as the estimate of the asymptotic variance of $\hat{\boldsymbol{\theta}}$, the bootstrap can be applied by resampling cross section units.

Allowing explicitly for unobserved effects in quantile regression is trickier. For a given quantile $0 < \tau < 1$, a natural specification, which incorporates strict exogeneity conditional on c_i, is

$$\text{Quant}_\tau(y_{it}|\mathbf{x}_i, c_i) = \text{Quant}_\tau(y_{it}|\mathbf{x}_{it}, c_i) = \mathbf{x}_{it}\boldsymbol{\theta} + c_i,\ t = 1, \ldots, T, \tag{4.5}$$

which is reminiscent of the way we specified the conditional mean in Chapter 10. Equivalently, we can write

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\theta} + c_i + u_{it},\ \text{Quant}_\tau(u_{it}|\mathbf{x}_i, c_i) = 0,\ t = 1, \ldots, T. \tag{4.6}$$

Unfortunately, unlike in the case of estimating effects on the conditional mean, we cannot proceed without further assumptions. A "fixed effects" approach, where we allow $\text{D}(c_i|\mathbf{x}_i)$ to be unrestricted, is attractive. Generally, there are no simple transformations to eliminate c_i and estimate $\boldsymbol{\theta}$. If we treat the c_i as parameters to estimate along with $\boldsymbol{\theta}$, the resulting estimator generally suffers from an incidental parameters problem. Briefly, if we try to estimate c_i for

each i then, with large N and small T, the poor quality of the estimates of c_i causes the accompanying estimate of $\boldsymbol{\theta}$ to be badly behaved. Recall that this was *not* the case when we used the FE estimator for a conditional mean: treating the c_i as parameters led us to the within estimator. Koenker (2004) derives asymptotic properties of this estimation procedure when T grows along with N, but also adds the assumptions that the regressors are fixed and $\{u_{it} : t = 1, \ldots, T\}$ is serially independent.

An alternative approach is suggested by Abrevaya and Dahl (2006) for $T = 2$. They are motivated by Chamberlain's correlated random effects linear model. In the $T = 2$ case, Chamberlain (1982) specifies

$$E(y_t|\mathbf{x}_1, \mathbf{x}_2) = \psi_t + \mathbf{x}_t\boldsymbol{\beta} + \mathbf{x}_1\boldsymbol{\xi}_1 + \mathbf{x}_2\boldsymbol{\xi}_2, \; t = 1, 2. \tag{4.7}$$

Notice that $\partial E(y_1|\mathbf{x})/\partial \mathbf{x}_1 = \boldsymbol{\beta} + \boldsymbol{\xi}_1$ and $\partial E(y_2|\mathbf{x})/\partial \mathbf{x}_1 = \boldsymbol{\xi}_1$. Therefore,

$$\boldsymbol{\beta} = \frac{\partial E(y_1|\mathbf{x})}{\partial \mathbf{x}_1} - \frac{\partial E(y_2|\mathbf{x})}{\partial \mathbf{x}_1}, \tag{4.8}$$

and similarly if we reverse the roles of $\mathbf{x}_1$ and $\mathbf{x}_2$. Abrevaya and Dahl use this motivation to estimate separate linear quantile regressions $\text{Quant}_\tau(y_t|\mathbf{x}_1, \mathbf{x}_2)$ – reminiscent of Chamberlain's method – and then define the partial effect as

$$\boldsymbol{\beta}_\tau = \frac{\partial \text{Quant}_\tau(y_1|\mathbf{x})}{\partial \mathbf{x}_1} - \frac{\partial \text{Quant}_\tau(y_2|\mathbf{x})}{\partial \mathbf{x}_1}. \tag{4.9}$$

For quantile regression, CRE approaches are generically hampered because finding quantiles of sums of random variables is difficult. For example, suppose we impose the Mundlak representation $c_i = \psi_o + \bar{\mathbf{x}}_i\boldsymbol{\xi}_o + a_i$. Then we can write $y_{it} = \psi_o + \mathbf{x}_{it}\boldsymbol{\theta}_o + \bar{\mathbf{x}}_i\boldsymbol{\xi}_o + a_i + u_{it} \equiv y_{it} = \psi_o + \mathbf{x}_{it}\boldsymbol{\theta}_o + \bar{\mathbf{x}}_i\boldsymbol{\xi}_o + v_{it}$, where v_{it} is the composite error. Now, if we assume v_{it} is independent of $\mathbf{x}_i$, then we can estimate $\boldsymbol{\theta}_o$ and $\boldsymbol{\xi}_o$ using pooled quantile regression of y_{it} on $1, \mathbf{x}_{it}$, and $\bar{\mathbf{x}}_i$. (The intercept does not estimate a quantity of particular interest.) But independence is very strong, and, if we truly believe it, then we probably believe all quantile functions are parallel. Of course, we can always just assert that the effect of interest is the set of coefficients on $\mathbf{x}_{it}$ in the pooled quantile estimation, and we allow these, along with the intercept and coefficients on $\bar{\mathbf{x}}_i$, to change across quantile. The asymptotic variance matrix estimator discussed for pooled quantile regression applies directly once we define the explanatory variables at time t to be $(1, \mathbf{x}_{it}, \bar{\mathbf{x}}_i)$.

We have more flexibility if we are interested in the median, and a few simple approaches suggest themselves. Write the model $\text{Med}(y_{it}|\mathbf{x}_i, c_i) = \text{Med}(y_{it}|\mathbf{x}_{it}, c_i) = \mathbf{x}_{it}\boldsymbol{\theta} + c_i$ in error form

as

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\theta} + c_i + u_{it}, \text{Med}(u_{it}|\mathbf{x}_i, c_i) = 0, \quad t = 1, .., T \tag{4.10}$$

and consider the multivariate conditional distribution $D(\mathbf{u}_i|\mathbf{x}_i)$. Above we discussed the centrally symmetric assumption, conditional on $\mathbf{x}_i$: $D(\mathbf{u}_i|\mathbf{x}_i) = D(-\mathbf{u}_i|\mathbf{x}_i)$. If we make this assumption, then the time-demeaned errors $\ddot{u}_{it}$ have (univariate) conditional (on $\mathbf{x}_i$) distributions symmetric about zero, which means we can consistently estimate $\boldsymbol{\theta}$ by applying pooled least absolute deviations to the time-demeaned equation $\ddot{y}_{it} = \ddot{\mathbf{x}}_{it}\boldsymbol{\theta} + \ddot{u}_{it}$, being sure to obtain fully robust standard errors for pooled LAD.

Alternatively, under the centrally symmetric assumption, the difference in the errors, $\Delta u_{it} = u_{it} - u_{i,t-1}$ have symmetric distributions about zero, so one can apply pooled LAD to $\Delta y_{it} = \Delta\mathbf{x}_{it}\boldsymbol{\theta} + \Delta u_{it}, \ t = 2, \ldots, T$. From Honoré (1992) applied to the uncensored case, LAD on the first differences is consistent when $\{u_{it} : t = 1, \ldots, T\}$ is an i.i.d. sequence conditional on $(\mathbf{x}_i, c_i)$, even if the common distribution is not symmetric – and this may afford robustness for LAD on the first differences rather than on the time-demeaned data. Interestingly, it follows from the discussion in Honoré (1992, Appendix 1) that when $T = 2$, applying LAD on the first differences is equivalent to estimating the c_i along with $\boldsymbol{\theta}_o$. So, in this case, there is no incidental parameters problem in estimating the c_i as long as $u_{i2} - u_{i1}$ has a symmetric distribution. Although not an especially weak assumption, central symmetry of $D(\mathbf{u}_i|\mathbf{x}_i)$ allows for serial dependence and heteroskedasticity in the u_{it} (both of which can depend on $\mathbf{x}_i$ or on t). As always, we should be cautious in comparing the pooled OLS and pooled LAD estimates of $\boldsymbol{\theta}$ on the demeaned or differenced data because they are only expected to be similar under the conditional symmetry assumption.

If we impose the Mundlak-Chamberlain device, we can get by with conditional symmetry of a sequence of bivariate distributions. Write $y_{it} = \psi_t + \mathbf{x}_{it}\boldsymbol{\theta} + \bar{\mathbf{x}}_i\boldsymbol{\xi} + a_i + u_{it}$, where $\text{Med}(u_{it}|\mathbf{x}_i, a_i) = 0$. If $D(a_i, u_{it}|\mathbf{x}_i)$ has a symmetric distribution around zero then $D(a_i + u_{it}|\mathbf{x}_i)$ is symmetric about zero, and, if this holds for each t, pooled LAD of y_{it} on $1, \mathbf{x}_{it}$, and $\bar{\mathbf{x}}_i$ consistently estimates $(\psi_t, \boldsymbol{\theta}, \boldsymbol{\xi})$. (Therefore, we can estimate the partial effects on $\text{Med}(y_{it}|\mathbf{x}_{it}, c_i)$ and also test if c_i is correlated with $\bar{\mathbf{x}}_i$.) The assumptions used for this approach are not as weak as we would like, but, like using pooled LAD on the time-demeaned data, adding $\bar{\mathbf{x}}_i$ to pooled LAD gives a way to compare with the usual FE estimate of $\boldsymbol{\theta}$.. (Remember, if we use pooled OLS with $\bar{\mathbf{x}}_i$ included, we obtain the FE estimate.) Fully robust inference can be obtained by computing $\hat{\mathbf{B}}$ and $\hat{\mathbf{A}}$ in (4.3) and (4.4).

5. Quantile Methods for "Censored" Data

As is well known, the statistical structure of parametric models for data that have truly been censored – such as top-coded wealth, or a right-censored duration – is essentially the same as models for corner solution responses – that is, variables that have a mass point, or pile up, at one or couple of values (usually, zero). Examples are labor supply, charitable contributions, and amount of life insurance. But an important point is that the interpretation of the estimates is different in these two cases. In the data censoring case, there is an underlying linear model (usually) whose coefficients we are interested in. For example, we are interested in the conditional distribution of wealth given covariates. That wealth has been top-coded means that we do not observe underlying wealth over its entire range. In effect, it is a missing data problem. The same is true with duration models.

In the corner solution case, we observe the response of interest over its entire range. We use models such as Tobit simply because we want to recognize the mass point or points. Linear functional forms for the mean, say, can miss important nonlinearities. When we apply, say, standard Tobit to a corner solution, y, we are interested in features of $D(y|\mathbf{x})$, such as $P(y > 0|\mathbf{x})$, $E(y|\mathbf{x}, y > 0)$, and $E(y|\mathbf{x})$. While the parameters in the model are important, they do not directly provide partial effects on the quantities of interest. Of course, if we use a linear model approximation for, say, $E(y|\mathbf{x})$, then the coefficients are approximate partial effects. A related point is: if we modify standard models for corner responses, say, consider heteroskedasticity in the latent error of a Tobit, we should consider how it affects $D(y|\mathbf{x})$, and not just the paramter estimates. In the case of censored data, it is the parameters of the underlying linear model we are interested in, and then it makes much more sense to focus on parameter sensitivity.

In applying LAD methods to "censored" outcomes, we should also be aware of the difference between true data censoring and corner solution responses. With true data censoring we clearly have an interest in obtaining estimates of, say,

$$y_i^* = \mathbf{x}_i\boldsymbol{\beta} + u_i, \tag{5.1}$$

where y_i^* is the variable we would like to explain. If y_i^* is top coded at, say, r_i, then we observe $y_i = \min(y_i^*, r_i)$. If we assume $D(u_i|\mathbf{x}_i, r_i) = Normal(0, \sigma^2)$, then we can apply censored normal regression (also called type I Tobit). This method applies even if r_i is observed only when y_i^* has been censored, which happens sometimes in duration studies. As shown by Powell (1986), we can estimate (5.1) under much weaker assumptions than normality:

$$Med(u_i|\mathbf{x}_i, r_i) = 0 \tag{5.2}$$

suffices, provided the censoring values value, r_i, are always observed. Because the median passes through monotonic functions,

$$\begin{aligned} Med(y_i|\mathbf{x}_i, r_i) &= Med[\min(\mathbf{x}_i\boldsymbol{\beta} + u_i, r_i)|\mathbf{x}_i, r_i] \\ &= \min[Med(\mathbf{x}_i\boldsymbol{\beta} + u_i|\mathbf{x}_i, r_i), r_i] \\ &= \min(\mathbf{x}_i\boldsymbol{\beta}, r_i). \end{aligned} \tag{5.3}$$

Because LAD consistently estimates the parameters of a conditional median, at least under certain regularity conditions, (5.3) suggest estimate $\boldsymbol{\beta}$ as the solution to

$$\min_{\mathbf{b}} \sum_{i=1}^{N} |y_i - \min(\mathbf{x}_i\mathbf{b}, r_i)|. \tag{5.4}$$

Powell (1986) showed that, even though the objective function has a corner it it, the *censored least absolute deviations* (CLAD) estimator is $\sqrt{N}$-asymptotically normal. Honoré, Khan, and Powell (2002) provide methods that can be used when r_i is not always observed.

CLAD can also be applied to corner solution responses. Suppose the variable of interest, y_i, has a corner at zero, and is determined by

$$y = \max(0, \mathbf{x}\boldsymbol{\beta} + u). \tag{5.5}$$

If $D(u|\mathbf{x})$ is Normal$(0, \sigma^2)$, then the MLE is the type I Tobit. Given $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$, we can compute partial effects on the mean and various probabilities. The partial effects on $Med(y|\mathbf{x})$ depend only on $\boldsymbol{\beta}$, because

$$Med(y|\mathbf{x}) = \max(0, \mathbf{x}\boldsymbol{\beta}). \tag{5.6}$$

Of course, (5.6) provides a way to estimate $\boldsymbol{\beta}$ by CLAD under just

$$Med(u|\mathbf{x}) = 0. \tag{5.7}$$

The β_j measure the partial effects on $Med(y|\mathbf{x})$ once $Med(y|\mathbf{x}) > 0$.

Once we recognize in corner solution applications that it is features of $D(y|\mathbf{x})$ that are of interest, (5.6) becomes just a particular feature of $D(y|\mathbf{x})$ that we can identify, and it is no better or worse than other features of $D(y|\mathbf{x})$ that we might want to estimate, such as a quantile other than the median, or the mean $E(y|\mathbf{x})$, or the "conditional" mean, $E(y|\mathbf{x}, y > 0)$. Emphasis is often given on the fact that the functional form for the median in (5.6) holds very generally when (5.5) holds; other than (5.7), no restrictions are made on the shape of the distribution $D(u|\mathbf{x})$ or of its dependence on $\mathbf{x}$. But for corner solution responses, there is nothing sacred

about (5.5). In fact, it is pretty restrictive because y depends on only one unobservable, u. Two-part models, summarized recently in Wooldridge (2007), allow more flexibility.

A model that is no more or less restrictive than (5.5) is

$$y = a \cdot \exp(\mathbf{x}\boldsymbol{\beta}), \tag{5.8}$$

where the only assumption we make is

$$E(a|\mathbf{x}) = 1, \tag{5.9}$$

where $D(a|\mathbf{x})$ is otherwise unrestricted. In particular, we do not know $P(a = 0|\mathbf{x})$, which is positive if y has mass point at zero, or $Med(a|\mathbf{x})$. Under (5.9),

$$E(y|\mathbf{x}) = \exp(\mathbf{x}\boldsymbol{\beta}), \tag{5.10}$$

which means we can consistently estimate $\boldsymbol{\beta}$ using nonlinear regression or a quasi-MLE in the linear exponential family (such as Poisson or Gamma); it does not matter that y is a corner if its mean is given by (5.10). The point here is that, if we simply focus on assumptions and what can be identified under those assumptions, the model in (5.8) and (5.9) identifies just as many features of $D(y|\mathbf{x})$ as the model in (5.5) and (5.7). They are different features, but neither is inherently better than the other.

Continuing with this point, we can modify (5.5) rather simply and see that CLAD breaks down. Suppose we add multiplicative heterogeneity:

$$y = a \cdot \max(0, \mathbf{x}\boldsymbol{\beta} + u), \tag{5.11}$$

where $a \geq 0$, and even make the strong assumption that a is independent of $(\mathbf{x}, u)$. The distribution $D(y|\mathbf{x})$ now depends on the distribution of a, and does not follow a type I Tobit model; generally, finding its distribution would be difficult, even if we specify a simple distribution for a. Nevertheless, if we normalize $\mathrm{E}(a) = 1$, then $\mathrm{E}(y|\mathbf{x}, u) = \mathrm{E}(a|\mathbf{x}, u) \cdot \max(0, \mathbf{x}\boldsymbol{\beta} + u) = \max(0, \mathbf{x}\boldsymbol{\beta} + u)$ (because $\mathrm{E}(a|\mathbf{x}, u) = 1$). It follows immediately by iterated expectations that if assumption (17.3) holds, then $E(y|\mathbf{x})$ has exactly the same form as the type I Tobit model:

$$E(y|\mathbf{x}) = \Phi(\mathbf{x}\boldsymbol{\beta}/\sigma)\mathbf{x}\boldsymbol{\beta} + \sigma\phi(\mathbf{x}\boldsymbol{\beta}/\sigma). \tag{5.12}$$

Therefore, the parameters $\boldsymbol{\beta}$ and σ^2 are identified and could be estimate by nonlinear least squares or weighted NLS, or a quasi-MLE using the mean function (5.12). Note that $D(y|\mathbf{x})$ does not follow the type I Tobit distribution, so MLE is not available.

On the other hand, if we focus on the median, we have

$$Med(y|\mathbf{x}, a) = a \cdot \max(0, \mathbf{x}\boldsymbol{\beta}). \tag{5.13}$$

But there is no "law of iterated median," so, generaly, we cannot determine $Med(y|\mathbf{x})$ without further assumptions. One might argue that we are still interested in the β_j because they measure the average partial effects on the median. But the do not appear to be generally identified under this variation on the standard Tobit model.

The issues in applying CLAD to corners gets even trickier in panel data applications. Suppose

$$y_{it} = \max(0, \mathbf{x}_{it}\boldsymbol{\beta} + c_i + u_{it}) \tag{5.14}$$

$$Med(u_{it}|\mathbf{x}_i, c_i) = 0, \tag{5.15}$$

so that (5.15) embodies strict exogeneity of $\mathbf{x}_{it}$ conditional on c_i. Under (5.14) and (5.15),

$$Med(y_{it}|\mathbf{x}_i, c_i) = \max(0, \mathbf{x}_{it}\boldsymbol{\beta} + c_i). \tag{5.16}$$

Honoré (1992) and Honoré and Hu (2004) provide methods of estimating $\boldsymbol{\beta}$ without making any assumptions about the distribution of c_i, or restricting its dependence on $\mathbf{x}_i$. They do assume conditional exchangeability assumptions on the u_{it}; sufficient is independence with $\mathbf{x}_i$ and $\{u_{it}\}$ i.i.d. over t. Given estimates of the β_j, we can estimate the partial effects of the x_{tj} on $Med(y_t|\mathbf{x}_t, c)$ for $Med(y_t|\mathbf{x}_t, c) > 0$. Unfortunately, because we do not observe c_i, or know anything about its distribution, we do not know when the nonzero effect kicks in. We can write the partial effect of x_{tj} as

$$\theta_{tj}(\mathbf{x}_t, c) = 1[\mathbf{x}_t\boldsymbol{\beta} + c > 0]\beta_j. \tag{5.17}$$

We might be interested in averaging these across the distribution of unobserved heterogeneity, but this distribution is not identified. (Interesting, if c_i has a $Normal(\mu_c, \sigma_c^2)$ distribution, then it is easy to show that the average of (5.17) across the heterogeneity is $E_{c_i}[\theta_{tj}(\mathbf{x}_t, c_i)] = \Phi[(\mu_c - \mathbf{x}_t\boldsymbol{\beta})/\sigma_c]\beta_j$, and we can see immediatly that it depends on the location and scale of c_i.)

We can compare the situation of the median with the mean. Using the Altonji and Matkin (2005), suppose we assume $D(c_i|\mathbf{x}_i) = D(c_i|\bar{\mathbf{x}}_i)$. Then $E(y_{it}|\mathbf{x}_i) = g_t(\mathbf{x}_{it}\boldsymbol{\beta}, \bar{\mathbf{x}}_i)$ for some unknown function $g_t(\cdot, \cdot)$, and $\boldsymbol{\beta}$ is identified (usually only up to scale) and the average partial effects on the mean are generally identified.

References

(To be added.)