Notes On Median and Quantile Regression

James L. Powell
Department of Economics
University of California, Berkeley

Conditional Median Restrictions and Least Absolute Deviations

It is well-known that the expected value of a random variable Y minimizes the expected squared deviation between Y and a constant; that is,

$$\begin{aligned} \mu_Y &\equiv E[Y] \\ &= \arg\min_c E(Y-c)^2, \end{aligned}$$

assuming $E||Y||^2$ is finite. (In fact, it is only necessary to assume $E||Y||$ is finite, if the minimand is normalized by subtracting Y^2, i.e.,

$$\begin{aligned} \mu_Y &\equiv E[Y] \\ &= \arg\min_c E[(Y-c)^2 - Y^2] \\ &= \arg\min_c [c^2 - 2cE[Y]], \end{aligned}$$

and this normalization has no effect on the solution if the stronger condition holds.) It is less well-known that a *median* of Y, defined as any number η_Y for which

$$\begin{aligned} \Pr\{Y &\leq \eta_Y\} \geq \frac{1}{2} \qquad \textit{and} \\ \Pr\{Y &\geq \eta_Y\} \geq \frac{1}{2}, \end{aligned}$$

minimizes an expected absolute deviation criterion,

$$\eta_Y = \arg\min_c E[|Y-c| - |Y|],$$

though the solution to this minimization problem need not be unique. When the c.d.f. F_Y is strictly increasing everywhere (i.e., Y is continuously distributed with positive density), then uniqueness is not an issue, and

$$\eta_Y = F_Y^{-1}(1/2).$$

In this case, the first-order condition for minimization of $E[|Y-c|-|Y|]$ is

$$0 = -E[sgn(Y-c)],$$

for $sgn(u)$ the "sign" (or "signum") function

$$sgn(u) \equiv 1 - 2 \cdot 1\{u < 0\},$$

here defined to be right-continuous.

Thus, just as least squares (LS) estimation is the natural generalization of the sample mean to estimation of regression coefficients, *least absolute deviations (LAD)* estimation is the generalization of the sample median to the linear regression context. For the linear structural equation

$$y_i = x_i'\beta_0 + \varepsilon_i,$$

if the error terms ε_i are assumed to have (unique) conditional median zero given the regressors x_i, i.e.

$$\begin{aligned} E[sgn(\varepsilon_i)|x_i] &= 0, \\ E[sgn(\varepsilon_i - c)|x_i] &\neq 0 \qquad if \qquad c = c(x_i) \neq 0, \end{aligned}$$

then the true regression coefficients β_0 are identified by

$$\beta_0 = \arg\min_{\beta} E[|y_i - x_i'\beta| - |\varepsilon_i|],$$

and, given an i.i.d. sample of size n from this model, a natural sample analogue to β_0 is

$$\begin{aligned} \hat{\beta} &\equiv \arg\min_{\beta} \frac{1}{n}\sum_{i=1}^{n} |y_i - x_i'\beta| \\ &\equiv \arg\min_{\beta} S_n(\beta), \end{aligned}$$

a slightly-nonstandard extremum estimator (because the minimand is not twice continuously differentiable for all β.

Consistency of LAD

Demonstration of consistency of $\hat{\beta}$ is straightforward, because the LAD minimand $S_n(\beta)$ is clearly continuous in β with probability one; in fact, $S_n(\beta)$ is convex in β, so consistency follows if S_n can be shown to converge *pointwise* to a function that is uniquely minimized at the true value β_0. (Typically we

need to show *uniform* convergence, but pointwise convergence of convex functions implies their uniform convergence on compact subsets.) To prove consistency, we need to impose some conditions on the model; here are some conditions that will suffice:

A1. The data $\{(y_i, x_i')'\}_{i=1}^n$ are independent and identically distributed across i;

A2. The regressors have bounded second moment, i.e., $E[||x_i||^2] < \infty$.

A3. The error terms ε_i are continously distributed given x_i, with conditional density $f(\varepsilon|x_i)$ satisfying the conditional median restriction, i.e.

$$\int_{-\infty}^{0} f(\lambda|x_i)d\lambda = \frac{1}{2}.$$

A4. The regressors and error density satisfy a "local identification" condition – namely, the matrix

$$C \equiv E[f(0|x_i)x_i x_i']$$

is positive definite.

Note that moments of y_i or ε_i need not exist under these assumptions, which is why LAD estimation is attractive for heavy-tailed error distributions. Condition **A4**. combines a "unique median" assumption (implied by positivity of the conditional density $f(\varepsilon|x_i)$ at $\varepsilon = 0$) with the usual full-rank assumption on the second moments of the regressors.

Imposing these conditions, the first step in the consistency proof is to calculate the probability limit of the minimand. To avoid assuming $E[|y_i|] < \infty$, it is convenient to normalize the minimand $S_n(\beta)$ by subtracting off its value at the true parameter β_0, which clearly does not affect the minimizing value $\hat{\beta}$. That is,

$$\begin{aligned}
\hat{\beta} &= \arg\min_{\beta} S_n(\beta) - S_n(\beta_0) \\
&= \arg\min_{\beta} \frac{1}{n}\sum_{i=1}^{n}[|y_i - x_i'\beta| - |y_i - x_i'\beta_0|] \\
&= \arg\min_{\beta} \frac{1}{n}\sum_{i=1}^{n}[|\varepsilon_i - x_i'\delta| - |\varepsilon_i|],
\end{aligned}$$

where

$$\delta \equiv \beta - \beta_0.$$

But since

$$-||x_i|| \cdot ||\delta|| \leq |\varepsilon_i - x_i'\delta| - |\varepsilon_i| \leq ||x_i|| \cdot ||\delta||$$

by the triangle and Cauchy-Schwarz inequalities, the normalized minimand is a sample average of i.i.d. random variables with finite first (and even second) moments under condition **A2**, so by Khintchine's Law of Large Numbers,

$$\begin{aligned}
S_n(\beta) - S_n(\beta_0) &\rightarrow {}^p\bar{S}(\delta) \\
&\equiv E[S_n(\beta) - S_n(\beta_0)] \\
&= E[|\varepsilon_i - x_i'\delta| - |\varepsilon_i|] \\
&= E\left[(\varepsilon_i - x_i'\delta)sgn\{\varepsilon_i - x_i'\delta\} - \varepsilon_i sgn\{\varepsilon_i\}\right] \\
&= E\left[(\varepsilon_i - x_i'\delta)(sgn\{\varepsilon_i - x_i'\delta\} - sgn\{\varepsilon_i\})\right] \\
&= E\left[2\int_{x_i'\delta}^{0}[\lambda - (x_i'\delta)]f(\lambda|x_i)d\lambda\right],
\end{aligned}$$

where the second-to-last equality uses the fact that $E[(x_i'\delta)sgn\{\varepsilon_i\}] = E[E[(x_i'\delta)sgn\{\varepsilon_i\}|x_i]] = 0$. (The integral in the last equality is well-defined for both positive and negative values of $x_i'\delta$, under the standard convention $\int_a^b dF = -\int_b^a dF$.)

By inspection, the limit $\bar{S}(\delta)$ equals zero at $\delta = \beta - \beta_0 = 0$, and is non-negative otherwise (since the sign of the integrand is the same as the sign of the lower limit $x_i'\delta$). Furthermore, since $S_n(\beta) - S_n(\beta_0)$ is convex for all n, so is its probability limit $\bar{S}(\beta - \beta_0)$; thus, if $\beta = \beta_0$ is a unique local minimizer, it is also a global minimizer, implying consistency of $\hat{\beta}$. But by Leibnitz' rule,

$$\begin{aligned}
\frac{\partial\bar{S}(\delta)}{\partial\delta} &= -2E[x_i \cdot \int_{x_i'\delta}^{0} f(\lambda|x_i)d\lambda], \\
\frac{\partial\bar{S}(0)}{\partial\delta} &= 0,
\end{aligned}$$

and

$$\begin{aligned}
\frac{\partial^2\bar{S}(\delta)}{\partial\delta\partial\delta'} &= 2E[x_i x_i' \cdot f(x_i'\delta|x_i)], \\
\frac{\partial^2\bar{S}(0)}{\partial\delta\partial\delta'} &= 2E[x_i x_i' \cdot f(0|x_i)] \equiv 2C,
\end{aligned}$$

which is positive definite by condition **A4**. So $\delta = 0 = \beta - \beta_0$ is indeed a unique local (and global) minimizer of $\bar{S}(\delta) = \bar{S}(\beta - \beta_0)$, and thus

$$\hat{\beta} \rightarrow^p \beta_0.\blacksquare$$

To generalize this consistency result to nonlinear median regression models

$$\begin{aligned} y_i &= g(x_i, \beta_0) + \varepsilon_i, \\ 0 &= E[sgn\{\varepsilon_i\}|x_i], \end{aligned}$$

the regularity conditions would have to be strengthened, since convexity of the corresponding LAD minimand $n^{-1}\sum_i |y_i - g(x_i, \beta)|$ in β is no longer assured. Standard conditions would include the assumption that the LAD criterion is minimized over a compact parameter space B (and not over all of R^p), and a uniform Lipschitz continuity condition on the median regression function $g(x_i, \beta)$ would typically be imposed, with the Lipschitz term assumed to have finite moments. Finally, the identification condition **A4** would have to be strengthened to a global identification condition, such as:

A4.' For some $\tau > 0$, the conditional density $f(\lambda|x_i) > \tau$ if $|\lambda| < \tau$, and $\Pr\{|g(x_i, \beta) - g(x_i, \beta_0)| \geq \tau\} > 0$ if $\beta \neq \beta_0$.

Asymptotic Normality of LAD

Returning to the linear LAD estimator, while demonstration of consistency of $\hat{\beta}$ involves routine application of asymptotic arguments for extremum estimators, demonstration of $\sqrt{n}$-consistency and asymptotic normality is complicated by the fact that the LAD criterion $\bar{S}(\beta)$ is not continuously differentiable in β. For comparison, consider the "standard" theory for extremum estimators, where the estimator $\hat{\theta}$ is defined to minimize (or maximize) a twice-differentiable criterion, e.g.,

$$\hat{\theta} = \arg\min_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \rho(z_i, \theta),$$

and (for large n) to satisfy a first-order condition for an interior minimum,

$$\begin{aligned} 0 &= \frac{1}{n} \sum_{i=1}^{n} \frac{\partial \rho(z_i, \hat{\theta})}{\partial \theta} \\ &\equiv \frac{1}{n} \sum_{i=1}^{n} \psi_i(\hat{\theta}), \end{aligned}$$

assuming consistency of $\hat{\theta}$ for an (interior) parameter θ_0 has been established. The true value θ_0 satisfies the corresponding population first-order condition

$$0 = E[\psi_i(\theta_0)];$$

derivation of the asymptotic distribution of $\hat{\theta}$ is based upon a Taylor's series expansion of the sample first-order condition for $\hat{\theta}$ around $\hat{\theta} = \theta_0$:

$$\begin{aligned} 0 &\equiv \frac{1}{n}\sum_{i=1}^{n}\psi_i(\hat{\theta}) \qquad (*) \\ &\equiv \frac{1}{n}\sum_{i=1}^{n}\psi_i(\theta_0) + \left[\frac{1}{n}\sum_{i=1}^{n}\frac{\partial\psi_i(\theta_0)}{\partial\theta'}\right](\hat{\theta}-\theta_0) + o_p(||\hat{\theta}-\theta_0||), \end{aligned}$$

hich is solved for $\hat{\theta}$ to yield the asymptotic linearity expression

$$\hat{\theta} = \theta_0 + H_0^{-1}\frac{1}{n}\sum_{i=1}^{n}\psi_i(\theta_0) + o_p\left(\frac{1}{\sqrt{n}}\right),$$

where

$$\begin{aligned} H_0 &\equiv -E\left[\frac{\partial\psi_i(\theta_0)}{\partial\theta'}\right] \\ &= -E\left[\frac{\partial^2\rho(z_i,\theta_0)}{\partial\theta\partial\theta'}\right] \end{aligned}$$

is minus one times the expected Hessian of the original minimand. From the linearity expression, it follows that

$$\sqrt{n}(\hat{\theta}-\theta_0) \rightarrow^d \mathcal{N}(0, H_0^{-1}V_0H_0^{-1}),$$

where V_0 is the asymptotic covariance matrix of the sample average of $\psi_i(\theta_0)$, i.e.,

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_i(\theta_0) \rightarrow^d \mathcal{N}(0, V_0),$$

which is established by appeal to a suitable central limit theorem.

In the LAD case (where $\theta \equiv \beta$), the criterion function

$$\rho(z_i,\theta) = |y_i - x_i'\beta|$$

is not continuously differentiable at values of β for which $y_i = x_i'\beta$; furthermore, the (discontinuous) subgradient

$$\frac{\partial\rho(z_i,\theta)}{\partial\theta} = sgn\{y_i - x_i'\beta\}x_i$$

itself has a derivative that is identically zero wherever it is defined. Thus the Taylor's expansion (*) is not applicable to this problem, even though an approximate first-order condition

$$\frac{1}{n}\sum_{i=1}^{n}sgn\{y_i - x_i'\hat{\beta}\}x_i = o_p\left(\frac{1}{\sqrt{n}}\right)$$

can be established for this problem. This condition can be shown to hold by showing that each element of the subgradient of the LAD criterion, when evaluated at the minimizing value $\hat{\beta}$, is bounded in magnitude by the difference between the right and left derivatives of the criterion, so that

$$\begin{aligned}\left|\frac{1}{n}\sum_{i=1}^{n} sgn\{y_i - x_i'\hat{\beta}\}x_i\right| &\leq \left|\frac{1}{n}\sum_{i=1}^{n} 1\{y_i = x_i'\hat{\beta}\}x_i\right| \\ &\leq \left[\sum_{i=1}^{n} 1\{y_i = x_i'\hat{\beta}\}\right] \max_i \frac{||x_i||}{n} \\ &= K \cdot o_p\left(\frac{1}{\sqrt{n}}\right),\end{aligned}$$

where $K \equiv \dim\{\beta\}$ and $E[||x_i||^2] < \infty$ by **A2**.

Though the subgradient for the LAD minimization is not differentiable, its expected value

$$\begin{aligned}E\left[\frac{\partial \rho(z_i,\theta)}{\partial \theta}\right] &= E\left[sgn\{y_i - x_i'\beta\}x_i\right] \\ &= E[sgn\{\varepsilon_i - x_i'\delta\}x_i] \\ &= 2E\left[\left(\int_0^{x_i'\delta} f(\lambda|x_i)d\lambda\right)x_i\right]\end{aligned}$$

is differentiable in $\delta = \beta - \beta_0$. The Taylor's series expansion would thus be applicable if the order of the expectation (over y_i and x_i) and differentiation (over θ) could somehow be interchanged. To do this rigorously, a *stochastic equicontinuity* condition on the sample average moment function

$$\bar{\Psi}_n(\beta) \equiv \frac{1}{n}\sum_{i=1}^{n} sgn\{y_i - x_i'\beta\}x_i$$

must be established; specifically, the stochastic equicontinuity condition is that, for any $\hat{\beta} \to^p \beta$,

$$\sqrt{n}\left[\bar{\Psi}_n(\hat{\beta}) - \bar{\Psi}_n(\beta_0) - E\left[\bar{\Psi}_n(\beta) - \bar{\Psi}_n(\beta_0)\right]|_{\beta=\hat{\beta}}\right] \to^p 0,$$

or, written alternatively,

$$\sqrt{n}\left[\left(\bar{\Psi}_n(\beta) - E[\bar{\Psi}_n(\beta)]\right)|_{\beta=\hat{\beta}} - \left(\bar{\Psi}_n(\beta_0) - E[\bar{\Psi}_n(\beta_0)]\right)\right] \to^p 0. \qquad (**)$$

Intuitively, while we would expect the normalized difference $\sqrt{n}\left(\bar{\Psi}_n(\beta) - E[\bar{\Psi}_n(\beta)]\right)$ to have a limiting normal distribution for each fixed value of β by a central limit theorem, the stochastic equicontinuity condition specifies that the normalized difference, evaluated at the consistent estimator $\hat{\beta}$, is asymptotically equivalent to its value evaluated at $\beta_0 = \text{plim } \hat{\beta}$.

Such a condition can be established for this LAD (and related quantile regression) problems using *empirical process* theory; once it has been established, it can be used to derive the asymptotic normal distribution of $\hat{\beta}$. Inserting the previous results that

$$\bar{\Psi}_n(\hat{\beta}) \equiv \frac{1}{n}\sum_{i=1}^{n} sgn\{y_i - x_i'\hat{\beta}\}x_i = o_p\left(\frac{1}{\sqrt{n}}\right)$$

and

$$\begin{aligned} E[\bar{\Psi}_n(\beta_0)] &\equiv E\left[sgn\{y_i - x_i'\beta_0\}x_i\right] \\ &= E[sgn\{\varepsilon_i\}x_i] \\ &= 0 \end{aligned}$$

into (**), it follows that

$$\sqrt{n}\left[\bar{\Psi}_n(\beta_0) - E[\bar{\Psi}_n(\beta)]|_{\beta=\hat{\beta}}\right] \to^p 0,$$

and a mean-value expansion of $E[\bar{\Psi}_n(\beta)]|_{\beta=\hat{\beta}}$ around $\hat{\beta} = \beta_0$ yields

$$\begin{aligned} \sqrt{n}\left(\hat{\beta} - \beta_0\right) &= H_0^{-1}\sqrt{n}\bar{\Psi}_n(\beta_0) + o_p(1) \\ &= H_0^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} sgn\{\varepsilon_i\}x_i + o_p(1), \end{aligned}$$

where now

$$\begin{aligned} H_0 &\equiv \frac{\partial E[\bar{\Psi}_n(\beta)]}{\partial \beta'}|_{\beta=\beta_0} \\ &= \frac{\partial E\left[sgn\{y_i - x_i'\beta\}x_i\right]}{\partial \beta'}|_{\beta=\beta_0} \\ &= 2E[f(0|x_i)x_i x_i'] \\ &\equiv 2C, \end{aligned}$$

assumed positive definite in **A4** above. Application of the Lindeberg-Levy central limit theorem to $\sqrt{n}\bar{\Psi}_n(\beta_0)$ yields the asymptotic distribution of the LAD estimator $\hat{\beta}$ as

$$\sqrt{n}(\hat{\beta} - \beta_0) \to^d \mathcal{N}(0, \frac{1}{4}C^{-1}DC^{-1}),$$

for

$$\begin{aligned} D &= E\left(\left[sgn\{y_i - x_i'\beta\}x_i\right] \cdot \left[sgn\{y_i - x_i'\beta\}x_i\right]'\right) \\ &= E\left(\left[sgn\{y_i - x_i'\beta\}\right]^2 \cdot x_i x_i'\right) \\ &= E[x_i x_i']. \end{aligned}$$

In the special case where ε_i is independent of x_i, so the conditional density $f(0|x_i)$ equals the marginal density $f(0)$, the asymptotic distribution simplifies to

$$\sqrt{n}(\hat{\beta} - \beta_0) \rightarrow^d \mathcal{N}(0, \frac{1}{[2f(0)]^2} D^{-1}).$$

Alternatively, for the nonlinear median regression model

$$\begin{aligned} y_i &= g(x_i, \beta_0) + \varepsilon_i, \\ 0 &= E[sgn\{\varepsilon_i\}|x_i], \end{aligned}$$

the relevant matrices C and D would be defined as

$$\begin{aligned} C &\equiv E\left[\frac{\partial g(x_i, \beta_0)}{\partial \beta} \frac{\partial g(x_i, \beta_0)}{\partial \beta'}\right], \\ D &\equiv E\left[f(0|x_i) \frac{\partial g(x_i, \beta_0)}{\partial \beta} \frac{\partial g(x_i, \beta_0)}{\partial \beta'}\right], \end{aligned}$$

which reduce to the previous definitions when $g(x_i, \beta) \equiv x_i'\beta$.

Asymptotic Covariance Matrix Estimation

To use the asymptotic normality of $\hat{\beta}$ to do the usual large-sample inference on β_0, consistent estimators of the matrices $C \equiv E[f(0|x_i)x_i x_i']$ and $D \equiv E[x_i x_i']$ must be constructed. The latter is easy; clearly

$$\begin{aligned} \hat{D} &\equiv \frac{1}{n} \sum_{i=1}^{n} x_i x_i' \\ &\rightarrow {}^p D. \end{aligned}$$

Consistently estimating the matrix C is trickier, due to the presence of the unknown conditional density function $f(0|x_i)$; while the error density might be parametrized, and its (finite-dimensional) parameter vector consistently estimated using standard methods, this would run counter to the spirit of the LAD theory so far, which does not rely upon a parametric form for the error terms. An alternative, nonparametric estimation strategy can be based upon kernel estimation methods for density functions. A specific form for an estimator is

$$\hat{C} \equiv \frac{1}{n} \sum_{i=1}^{n} \left[h^{-1} 1\{\left|y_i - x_i'\hat{\beta}\right| \leq h/2\}\right] x_i x_i',$$

where $h = h_n$ is a user-chosen "bandwidth" term that is assumed to tend to zero as the sample size n increases. The term $h^{-1}1\{|u| \leq h/2\}$ (which is evaluated at $u = y_i - x_i'\hat{\beta}$) is essentially a numerical derivative of the function $sgn\{u\}$, based upon the small perturbation h. It can be shown that $\hat{C} \rightarrow^p C$ as

$n \to \infty$, provided that $h = h_n \to 0$ in such a way that $n\sqrt{h_n} \to \infty$, using the sorts of mean and variance calculations used to demonstrate consistency of the standard kernel density estimator. A generalization of this estimator would be

$$\hat{C}^* \equiv \frac{1}{n} \sum_{i=1}^{n} \left[h^{-1} K \left(\frac{y_i - x_i' \hat{\beta}}{h} \right) \right] x_i x_i',$$

where the kernel function $K(\cdot)$ satisfies

$$\int K(u) du = 1$$

(for example, $K(u)$ could be a density function for a continuous random variable). The estimator $\hat{C}$ is a special case with $K(u) = 1\{|u| \leq 1/2\}$, the density for a uniform random variable on the interval $[-1/2, 1/2]$. And both the estimators for C and D can be extended to the nonlinear median regression model by replacing the terms "$x_i x_i'$" with the more general "$[\partial g(x_i, \hat{\beta})/\partial \beta] \left[\partial g(x_i, \hat{\beta})/\partial \beta \right]'$.