

The 'Variance of the Gradient'

In the last lecture, we had " $E_0\{ [\sqrt{n}\bar{g}(\theta_0)][\sqrt{n}\bar{g}(\theta_0)]'\} = \mathcal{J}$, where \mathcal{J} is something else that is the variance of the gradient at θ_0 ." We want to take this expression apart so that we understand the characterization of \mathcal{J} as the "outer product of the gradient". The point we will be making arises again in GMM estimation.

$\bar{g}(\theta_0)$ is the mean of the gradient of the log likelihood function at θ_0 . So it is

$$\frac{1}{n} \left\{ \frac{\partial l_1}{\partial \theta} + \frac{\partial l_2}{\partial \theta} + \dots + \frac{\partial l_n}{\partial \theta} \right\},$$

where each $\frac{\partial l_i}{\partial \theta}$ is a $k \times 1$ object. (Since θ has k elements.) Thus

$$\begin{aligned} & [\sqrt{n}\bar{g}(\theta_0)][\sqrt{n}\bar{g}(\theta_0)]' \\ = & \sqrt{n} \frac{1}{n} \left\{ \frac{\partial l_1}{\partial \theta} + \frac{\partial l_2}{\partial \theta} + \dots + \frac{\partial l_n}{\partial \theta} \right\} \sqrt{n} \frac{1}{n} \left\{ \frac{\partial l_1}{\partial \theta} + \frac{\partial l_2}{\partial \theta} + \dots + \frac{\partial l_n}{\partial \theta} \right\}' \\ = & \frac{1}{n} \left\{ \frac{\partial l_1}{\partial \theta} + \frac{\partial l_2}{\partial \theta} + \dots + \frac{\partial l_n}{\partial \theta} \right\} \left\{ \frac{\partial l_1}{\partial \theta} + \frac{\partial l_2}{\partial \theta} + \dots + \frac{\partial l_n}{\partial \theta} \right\}' \end{aligned}$$

The 'Variance of the Gradient' (2)

$$[\sqrt{n}\bar{g}(\theta_0)][\sqrt{n}\bar{g}(\theta_0)]' = \frac{1}{n} \left\{ \frac{\partial l_1}{\partial \theta} + \frac{\partial l_2}{\partial \theta} + \dots + \frac{\partial l_n}{\partial \theta} \right\} \left\{ \frac{\partial l_1}{\partial \theta} + \frac{\partial l_2}{\partial \theta} + \dots + \frac{\partial l_n}{\partial \theta} \right\}'$$

Doing the multiplication:

$$\begin{aligned} & \frac{1}{n} \left[\left\{ \frac{\partial l_1}{\partial \theta} \frac{\partial l_1'}{\partial \theta} + \frac{\partial l_1}{\partial \theta} \frac{\partial l_2'}{\partial \theta} + \dots + \frac{\partial l_1}{\partial \theta} \frac{\partial l_n'}{\partial \theta} \right\} + \right. \\ & \left. \left\{ \frac{\partial l_2}{\partial \theta} \frac{\partial l_1'}{\partial \theta} + \frac{\partial l_2}{\partial \theta} \frac{\partial l_2'}{\partial \theta} + \dots + \frac{\partial l_2}{\partial \theta} \frac{\partial l_n'}{\partial \theta} \right\} + \dots \right. \\ & \left. \left\{ \frac{\partial l_n}{\partial \theta} \frac{\partial l_1'}{\partial \theta} + \frac{\partial l_n}{\partial \theta} \frac{\partial l_2'}{\partial \theta} + \dots + \frac{\partial l_n}{\partial \theta} \frac{\partial l_n'}{\partial \theta} \right\} \right] \end{aligned}$$

Remember that each of the n^2 'cross-products' is a $(k \times 1) * (1 \times k)$ operation, thus producing $k \times k$ matrixes. The next step is to take the expectation of this. And here's the point: when observations are independent, $\frac{\partial l_i}{\partial \theta} \frac{\partial l_j'}{\partial \theta}$ has zero expectation for $i \neq j$.

The 'Variance of the Gradient' (3)

Thus

$$E \left\{ \frac{\partial l_1}{\partial \theta} \frac{\partial l_1'}{\partial \theta} + \frac{\partial l_1}{\partial \theta} \frac{\partial l_2'}{\partial \theta} + \dots + \frac{\partial l_1}{\partial \theta} \frac{\partial l_n'}{\partial \theta} \right\} = E \frac{\partial l_1}{\partial \theta} \frac{\partial l_1'}{\partial \theta}$$

$$E \left\{ \frac{\partial l_2}{\partial \theta} \frac{\partial l_1'}{\partial \theta} + \frac{\partial l_2}{\partial \theta} \frac{\partial l_2'}{\partial \theta} + \dots + \frac{\partial l_2}{\partial \theta} \frac{\partial l_n'}{\partial \theta} \right\} = E \frac{\partial l_2}{\partial \theta} \frac{\partial l_2'}{\partial \theta}, \dots$$

and so

$$E[\sqrt{n}\bar{g}(\theta_0)][\sqrt{n}\bar{g}(\theta_0)]' = E\left[\frac{1}{n} \left\{ \frac{\partial l_1}{\partial \theta} + \frac{\partial l_2}{\partial \theta} + \dots + \frac{\partial l_n}{\partial \theta} \right\} \left\{ \frac{\partial l_1}{\partial \theta} + \frac{\partial l_2}{\partial \theta} + \dots + \frac{\partial l_n}{\partial \theta} \right\}'\right]$$

$$= \frac{1}{n} E \left\{ \frac{\partial l_1}{\partial \theta} \frac{\partial l_1'}{\partial \theta} + \frac{\partial l_2}{\partial \theta} \frac{\partial l_2'}{\partial \theta} + \dots + \frac{\partial l_n}{\partial \theta} \frac{\partial l_n'}{\partial \theta} \right\},$$

which pretty clearly suggests estimating \mathcal{J} by averaging up the n sample 'outer products' of the observation gradients. BUT trouble might be brewing if $\frac{\partial l_i}{\partial \theta} \frac{\partial l_j'}{\partial \theta}$ had non-zero expectation for $i \neq j$.

Three 'Principles' of Testing in the ML Framework

We face the situation of testing the null hypothesis that a parametric restriction of a model is true. Suppose this is expressed as $H_0 : c(\beta) = 0$. We can

1. Estimate the model in both the restricted and unrestricted forms, and test the hypothesis by comparing the log likelihoods of the two models. In particular, twice the difference in the log likelihoods is $\chi^2(J)$, where J is the difference in the dimensions of the parameter space of the two models.
2. Estimate the model in the unrestricted form only and then use the estimated $V(\hat{\beta})$ matrix to see if $c(\beta) = 0$ is plausible. This also gives rise to a $\chi^2(J)$ test. When $c(\beta)$ is nonlinear, this requires the 'delta method'; otherwise we are computing the variance of a linear combination of normally distributed objects, and testing the hypothesis that these are zero.

3. Estimate the model in restricted form only. Heuristically, we then proceed to see whether the relaxing the constraint will increase the log likelihood a lot, basing this computation on the quadratic approximation to likelihood implied by the model estimated under the null hypothesis. This also results in $\chi^2(J)$ test.

A classic reference is Engle's Chapter in the Handbook of Econometrics, volume 2, particularly pp. 776-788, the first 13 pages. There he expounds the intuition that if we faced a (unknown) quadratic likelihood function, expressible as

$$\begin{aligned} L(\theta) &= L(\theta^*) + (\theta - \theta^*)^T L'(\theta^*) + \frac{1}{2}(\theta - \theta^*)^T L''(\theta^*)(\theta - \theta^*) \\ &= L(\theta^*) + (\theta - \theta^*)^T g(\theta^*) + \frac{1}{2}(\theta - \theta^*)^T H(\theta^*)(\theta - \theta^*), \end{aligned}$$

we could equally well calculate $L(\tilde{\theta}) - L(\theta^*)$ in three different ways: directly (the LR test), by computing the value and the first and second derivatives at θ^* and computing $L(\tilde{\theta})$ on the basis of the quadratic function (a Wald test), or by computing the value and the first and second derivatives at $\tilde{\theta}$ and computing $L(\theta^*)$ on the basis of the quadratic model (the score or Lagrange Multiplier test.)

The Delta Method

The LR test is in many ways simple: it requires computing both the restricted and unrestricted model, and is invariant to reparameterizations since each of the individual computations is invariant to reparameterization. The score test is too complicated to cover in this course.

This leaves the Wald test for further exposition: it is not invariant to reparameterization, but it offers the advantage of having a natural 'robustified' version, which is not true of the LR test.

Suppose $\hat{\theta} \sim N(\theta_0, V)$, then linear combinations of elements of θ are also normal, in particular $R\theta$, where R is $h \times k$, is $N(R\theta_0, RVR')$. Consequently the hypothesis that $R\theta_0 - r = 0$ is tested by computing $(R\theta_0 - r)'[RVR']^{-1}(R\theta_0 - r)$ and referring it to $\chi^2(h)$. 'Robustification' is accommodated by using a robust or sandwich estimate for V .

The Delta Method (2)

(Following Cameron and Trivedi, pp. 223-227.) A nonlinear hypothesis can also be tested. Suppose we want to test $h(\theta_0) = 0$ against $h(\theta_0) \neq 0$ where $h(\cdot)$ is an $h \times 1$ vector function of θ , and that $h(\theta_0) = 0$ allows θ_0 to be in the interior of the parameter space. (This rules out testing $\theta_0 = 0$ when $\theta_0 \geq 0$ is assumed.) Using an exact first-order Taylor series expansion around θ_0 , we have:

$$h(\hat{\theta}) = h(\theta_0) + \frac{\partial h(\bar{\theta})}{\partial \theta'} (\hat{\theta} - \theta_0),$$

so that $h(\hat{\theta})$ is approximated by a locally linear function (and $\bar{\theta}$ is between $\hat{\theta}$ and θ_0). It follows that (by appeal to the 'Product limit normal rule', Cameron and Trivedi Theorem A.17, p.952; which follows from the Transformation Theorem (briefly, $a_n \xrightarrow{d} a$, $b_n \xrightarrow{p} b$, a a random variable and b a constant, then $a_n b_n \xrightarrow{p} ab$)) that

$$(h(\hat{\theta}) - h(\theta_0)) \xrightarrow{d} N\left(0, \left[\frac{\partial h(\theta_0)}{\partial \theta'} \right] [V] \left[\frac{\partial h(\theta_0)}{\partial \theta'} \right]'\right)$$

The Delta Method (3)

We have

$$(h(\hat{\theta}) - h(\theta_0)) \xrightarrow{d} N\left(0, \left[\frac{\partial h(\theta_0)}{\partial \theta'}\right] [V] \left[\frac{\partial h(\theta_0)}{\partial \theta'}\right]'\right);$$

writing $\frac{\partial h(\theta_0)}{\partial \theta'} = R$ and recognizing the null is $h(\theta_0) = 0$, this simplifies to

$$h(\hat{\theta}) \xrightarrow{d} N(0, RVR')$$

so our test statistic is

$$h(\hat{\theta})'[RVR']^{-1}h(\hat{\theta})$$

(I have handled the factor in n differently than the text since I started from $\hat{\theta} \sim N(\theta_0, V)$ rather than $\sqrt{n}(\hat{\theta} - \theta_0) \sim N(0, C)$.)

Quantile Regression

The median is the .5 quantile. The .9 quantile of the random variable y is the value q such that $\text{prob}(y \leq q) = .9$. Or it's the number q satisfying $F_y(q) = .9$ or $q = F_y^{-1}(.9)$. A .5 quantile regression of y on X gives us the conditional median; if it is linear we can write

$$\text{med}(y|X) = X\beta$$

An object of some importance in 'advanced' contexts is the 'residual' from a quantile regression, which for the median regression is:

$$y_i - \text{med}(y|X_i)$$

Our informal investigation of quantile regression starts from thinking about how to do a median regression, how to generalize that to other quantiles, and how to interpret the resulting collection of regressions.

Median Regression

The sample median m minimizes the sum of the absolute value of the 'errors' or 'residuals', $u_i = (y_i - m)$. To see this, start with the objective function

$$\min_m \sum_{i=1}^n |u_i| = \sum_{i=1}^n |(y_i - m)|$$

The proper way to solve this and the more complicated problems to follow is by linear programming, but I really don't know how to do that. So let's plough ahead with calculus.

If we just try to differentiate the above function, the absolute value sign gives us a little trouble, but it is easy to see that when $(y_i - m)$ is positive the derivative of $(y_i - m)$ is -1 and when $(y_i - m)$ is negative it is $+1$. So we need a notation that will make handling this easier. So let $\mathbf{1}(x)$, called the indicator function, have the value 1 when x is true and 0 otherwise.

Median Regression (2)

Armed with the indicator function, and multiplying our objective function by 1/2 for reasons that will become apparent, we have:

$$\hat{m} = \arg \min_m S(y, m) = \frac{1}{2} \sum_{i=1}^n |u_i| = \frac{1}{2} \sum_{i=1}^n |(y_i - m)|$$

$$\begin{aligned} \frac{\partial S(y, m)}{\partial m} &= \frac{1}{2} \sum_{i=1}^n \{ \mathbf{1}(u_i > 0)(+1) + \mathbf{1}(u_i < 0)(-1) \} \\ &= \frac{1}{2} \sum_{i=1}^n \{ \mathbf{1}(u_i > 0) - \mathbf{1}(u_i < 0) \} \end{aligned}$$

So if we just regarded this as an ordinary first order condition, we would pick m so that half the u_i 's would be positive and half the u_i 's would be negative. And that would correspond to half the observed y_i 's being above m and half the observed y_i 's being below m .

Median Regression (3): Covariates

Now let's let $m = X\beta$, so that the median depends on covariates X . Then

$$\hat{\beta} = \arg \min_{\hat{\beta}} S(y, X, \hat{\beta}) = \frac{1}{2} \sum_{i=1}^n |u_i| = \frac{1}{2} \sum_{i=1}^n |(y_i - X_i \hat{\beta})| \quad \text{and}$$
$$\frac{\partial S(y, X, \hat{\beta})}{\partial \hat{\beta}} = \frac{1}{2} \sum_{i=1}^n \{ \mathbf{1}(u_i > 0) - \mathbf{1}(u_i < 0) \} X_i' = 0$$

is our pseudo first order condition. One way of looking at this is that X 's are orthogonal not to the magnitude-weighted residuals as in OLS, but to the signs of the residuals. Another way to look at this is to see what happens as I get repeated realizations of particular values of X (a very asymptotic notion considering most X processes)—asymptotically, for any value of X , half the y 's corresponding to that X should be above $X\hat{\beta}$ and half below.

Quantile Regressions for General Quantiles

Median or .5 quantile regressions divide things in half. τ quantiles divide things in proportions τ and $(1 - \tau)$. To help us along the way we introduce the check function $\rho_\tau(u)$:

$$\rho_\tau(u) = (\tau - \mathbf{1}(u < 0)) \cdot u$$

So when $\tau = .5$, u is either multiplied by .5 (if $u \geq 0$) or $-.5$. Similarly, if $\tau = .9$, u is multiplied either by .9 or $-.1$. If I want 90% to fall below $q_{.9}$, then I weight the things that fall above $q_{.9}$ 9 times as much as those that fall below $q_{.9}$. (Should have a picture; Koenker book, p.6). Notice that if u is negative it is multiplied by a negative number so that $\rho_\tau(u) \geq 0$.

The median regression objective function can thus be seen to be the $\tau = .5$ case of:

$$\min_{\hat{\beta}} S(y, X, \hat{\beta}, \tau) = \sum_{i=1}^n \rho_\tau(u_i) = \sum_{i=1}^n \rho_\tau(y_i - X_i \hat{\beta})$$

Implications of Quantile Regressions and Linear Quantile Regressions

From the foregoing, we can see that different values of τ in a quantile regression give different quantiles of y conditional upon X . Now here is the important thing: if I have the regressions for every value of τ , then I have the entire distribution of y conditional upon X . And, while this gift horse deserves further inspection, this has been ostensibly achieved without recourse to any assumptions on the disturbance process.

Let's look at the collection of quantile regressions, writing them as:

$$Q_y(\tau|x) = \beta(\tau)x,$$

where (exceptionally) $\beta(\tau)$ is $(1 \times k)$ and x is $(k \times 1)$. This form makes it obvious that if we simply run linear quantile regressions for every value of τ , we will get a different coefficient vector for every value of τ . Under the assumptions of the classical linear regression model, the (population) values of $\beta(\tau)$ are just β , except for the intercept, where (using α for the first component of β), $\alpha(.5) = \alpha_{CLRM}$ if ε is symmetrically distributed.

The Asymptotic Distribution of Quantile Regression Coefficients

Following Koenker book p. 74. Though it is not immediately obvious that M -estimation theory is applicable, it basically is and

$$\sqrt{n}(\hat{\beta}(\tau) - \beta(\tau)) \rightsquigarrow N(0, \tau(1 - \tau)H_n^{-1}J_nH_n^{-1})$$

where

$$J_n(\tau) = n^{-1} \sum_{i=1}^n x_i x_i^T$$

$$H_n(\tau) = \lim_{n \rightarrow \infty} \sum_{i=1}^n x_i x_i^T f_i(\xi_i(\tau))$$

The term $f_i(\xi_i(\tau))$ denotes the conditional density of the response, y_i , evaluated at the τ^{th} conditional quantile. For *i.i.d.* errors this formula collapses to

$$\frac{1}{[2f(0)]^2} X'X$$

in the case of the median.