

Econometrics Notes

Greg Fischer
MIT

April 25, 2005

1 Tobit and Other Censored Regression Models

- Amemiya (1973) explores demonstrates the consistency and asymptotic normality of the Tobit model under the assumption of normal errors
- Goldberger (1980) and Arabmazar & Schmidt (1982) calculate the inconsistency of the normal MLE tobit estimator for several non-normal distributions of the error term. It performs poorly (inconsistency).
- Powell (1984) shows in the “Least Absolute Deviations for the Censored Regression Model” paper that using a generalized version of LAD, one can form an estimator of such censored models that is robust to a wide class of alternative error processes and heteroscedasticity.
 - This used to be computationally burdensome to calculate in practice because the function to be minimized is not continuously differentiable. Now it’s not too bad to deal with this.
 - Since this is consistent under misspecification while MLE/Tobit is efficient under normality, we’ve got a nice set up for a Hausman test of normality in the error terms/general misspecification.
 - The basic model were trying to estimate is $y_t = \max\{0, x_t\beta_0 + u_t\}$.
 - Since the $\mathbf{1}(y > 0)$ is a monotonic transformation, $med(Y|X) = X'\beta_0$ (as long as the median is not 0). So we can estimate $\hat{\beta}$ as

$$\hat{\beta} = \arg \min_{\beta} \sum_{i=1}^n \left| Y_i - \max\{0, X_i'\beta\} \right|$$

- This estimator is CAN as shown by Powell (1984) and by Van der Vaart (1995) for a more general class of estimators that minimize a sample average.

2 Partially Linear Regression

- This is a class of semiparametric regression where the model contains both a linear and a non-parametric component, something like

$$y = X'\beta + \phi(Z) + \varepsilon$$

- The basic idea for dealing with this is to “concentrate out” the non-parametric component to find what is called a “profile” squared residual function. Whitney’s handouts on the subject (p. 22 of his Non-Parametric and Semi-Parametric Estimation notes as well as Powell’s chapter on Estimation of Semi-Parametric Models in the Handbook of Econometrics vol. 4 are both good treatments, if occasionally opaque. More practical is the

Härdle and Linton article in the same volume, p. 2329). Here's the idea: take the expectation of the above equation conditional on Z (under the assumption that $E(\varepsilon|Z) = 0$

$$E(y|Z) = E[X'|Z]\beta + \phi(Z)$$

we can estimate each of the conditional expectations non-parametrically subject to the standard problems about dimensionality. So now subtract the second equation from the first

$$\underbrace{y - E(y|Z)}_{\equiv \tilde{y}} = \underbrace{(X' - E[X'|Z])}_{\equiv \tilde{X}}\beta + \varepsilon$$

We can now run OLS (or GLS or what-have-you) on the transformed model $\tilde{y} = \tilde{X}'\beta + \varepsilon$.

- If you want to check out a few papers that actually use this, try looking at Hausman & Newey (1995) on gasoline demand, though they maximize simultaneously over the non-parametric and parametric regressors, and Stock (1991) on toxic waste]

3 Kernel Estimators

3.1 Basic Density Estimation

- The first few pages of Newey's Non-parametrics handout are a good introduction to this. I also found the Härdle and Linton chapter in the Handbook of Econometrics vol. 4 really useful. [There should be a nice JEL article too, no?, but I've yet to see one]
- [It may be helpful to work through the bias calculation in gory detail here]
- Kernel density estimators are consistent; both the bias and variance go to zero as $h \rightarrow 0$ and $nh \rightarrow \infty$ (that is, the bandwidth gets arbitrarily small and the number of elements in each bin goes to infinity). As Newey points out on p. 6 of his handout, as $h \rightarrow 0$, the MSE of the nonparametric estimator vanishes slower than $1/n$, which means the rate of convergence is slower than root- n . Why must this be so? Because to avoid bias, h must go to zero, which means the fraction of the observations used to estimate the density at each point is also going to zero.
- Bandwidth choice can be pretty important for semi-parametric models, but as far as I can tell, eyeballing plots is the best way to "optimally" choose bandwidth for simple density estimation. There are, however, some cool formulas for choosing the bandwidth that minimizes MSE.
- The Dreaded *Curse of Dimensionality*
 - Throw this phrase around a lot when discussing the use of multivariate kernels. Newey's description on p. 8 is great (although I think "radius" on p. 9 should be "diameter" as a ball with radius 1 happily covers any space of $[0, 1]^r$). The gist: the amount of data needed to estimate density goes up exponentially with the dimension of the parameter space.

3.2 Kernel Regression

[Should write something about this]

4 Other Non-Parametric Regression Techniques

4.1 Series Regression

[help]

4.2 Spline Regression

[help help]

5 GMM

- What were going to do: think about a basis method of moments estimator.
 - Suppose you knew that $E(y) = \mu$ and you wanted to estimate μ . So what do you think a reasonable way of estimating μ might be?
 - Well, we know from any of our favorite laws of large numbers, that the sample average should get close to the true mean, so we'd probably come up with something like $\hat{\mu} = \frac{1}{n} \sum y_i$.
 - But let's say we had another bit of information, like $E(y^2) = \mu$. We could just throw it away and stick with our original estimator. We could just use this new moment condition. But finding some estimator that uses both restriction is likely to be the best (in terms of minimum variance). We'll try to pick a $\hat{\beta}$ such that our sample moments are as close to zero, their population counterpart, as possible.
 - But while both moment restrictions hold in the population (we're assuming they do after all) it will only be by happenstance that we'll be able to satisfy both of them in our sample. So what we do is come up with some weighting scheme (a way to think about distance) and try to get these weighted moments as close to zero. That's the idea behind GMM.
- So here's the set up: you've got yourself some moment restrictions. They could be something like $E(x'\varepsilon) = 0$ or $E(y) = \mu$, or even *kurtosis*(y) = κ . More generally, let's say $E[g_i(\beta)] = 0$. Note, Newey uses β_0 to denote the *true* parameter. All the extra notation confuses me so I leave it out. You get the idea. Let's say we have m of these restrictions.
- Form the sample analog to these

$$\hat{g}(\beta) \equiv \frac{1}{n} \sum_{i=1}^n g_i(\beta)$$

and pick yourself some p.s.d. matrix $\hat{\mathbf{A}}$ to use as a weighting matrix. The identity matrix will do, we'll show in a bit the the optimal weighting matrix is the inverse of the variance of the sample moment restrictions (this will start to look a bit like GLS, the idea being you want to put more weight on moment restrictions that aren't varying all over the place).

- So our GMM estimator is

$$\hat{\beta}_{GMM} = \arg \min_{\beta} \hat{g}(\beta)' \hat{\mathbf{A}} \hat{g}(\beta).$$

- GMM is a pretty general formulation. You can quite easily show that 2SLS is just a special case. What's our moment restriction: $E(Z_i \varepsilon_i) = 0$. Run from there.

5.1 Asymptotic Properties

5.1.1 Consistency

Under a handful of regularity conditions:

1. iid data
2. Identification: $E[g_i(\beta)] = 0$ iff $\beta = \beta_0$, i.e. the moment conditions only hold at the true parameter value.
3. GMM minimization takes place over a compact parameter space B , containing β (note that this isn't necessary if $g_i(\beta)$ is linear).
4. Some technical conditions on the continuity and sup of g_i
5. And $\hat{A} \rightarrow_p A$ which is p.s.d

Then we get that $\hat{\beta} \rightarrow_p \beta_0$

Proof. The idea of the proof is that continuity and identification force the population moment conditions to be bounded away from zero out a neighborhood of the true β . By some LLN, we can get the sample analog of the moment distance measure arbitrarily close to zero, so $\hat{\beta}$ must get close to β . [check lecture notes for more on this]. ■

5.1.2 Asymptotic Normality

Once you have consistency, asymptotic normality follows more or less along similar lines as other asymptotic normality proofs.

Proof. Start with the following definitions: $G \equiv E[\partial g_i(\beta_0)/\partial \beta]$; $\Omega = E[g_i g_i']$; and $\hat{G} = \partial \hat{g}(\hat{\beta})/\partial \beta$.

1. So a central limit theorem gives us: $\sqrt{n}\hat{g} \rightarrow_d N(0, \Omega)$
2. From some LLN we get: $\partial \hat{g}(\bar{\beta})/\partial \beta \rightarrow_p G$, for any $\bar{\beta} \rightarrow_p \beta$.
3. The first order conditions for a minimum give us $0 = \hat{G}' \hat{A} \hat{g}(\hat{\beta})$
4. Take a first order Taylor series expansion about the true β

$$0 = \hat{G}' \hat{A} \left\{ \hat{g}(\beta_0) + \frac{\partial \hat{g}(\bar{\beta})}{\partial \beta} (\hat{\beta} - \beta_0) \right\}$$

5. Rearrange and solve for $\sqrt{n}(\hat{\beta} - \beta_0) = -[\hat{G}' \hat{A} \frac{\partial \hat{g}(\bar{\beta})}{\partial \beta}]^{-1} \hat{G}' \hat{A} \sqrt{n} \hat{g}(\beta_0)$
6. All the muck in front of the root-n term converges: $-[\hat{G}' \hat{A} \frac{\partial \hat{g}(\bar{\beta})}{\partial \beta}]^{-1} \hat{G}' \hat{A} \rightarrow_p -(G' A G)^{-1} G' A$.
7. Our central limit theorem gives us $\sqrt{n}\hat{g} \rightarrow_d N(0, \Omega)$.
8. And the Slutsky Theorem gives us the distribution of the GMM estimator:

$$\sqrt{n}(\hat{\beta}_{GMM} - \beta_0) \rightarrow_d N(0, (G' A G)^{-1} G' A \Omega A G (G' A G)^{-1}).$$

■
Note that just looking at this and considering Yaccine's Law, you'd think that $A = \Omega^{-1}$ would likely be the optimal weighting matrix. Someone has, I'm sure, written an *Econometrica* paper on why this is so, but intuition is also right.

5.1.3 Actually Calculating the Variance

So the variance of GMM (in fact, the variance of any general minimum distance estimator where $\hat{G} \equiv \partial \hat{g}(\hat{\theta}) / \partial \theta$ and $\hat{\Omega} \rightarrow_p \Omega \equiv Avar(\hat{g}(\theta))$) is

$$\hat{V} = (\hat{G}' \hat{A} \hat{G})^{-1} \hat{G}' \hat{A} \hat{\Omega} \hat{A} \hat{G} (\hat{G}' \hat{A} \hat{G})^{-1}$$

Calculating this generally means finding $\hat{\Omega}$. For GMM we can use $\hat{\Omega} = \frac{1}{n} \sum_{i=1}^n g(z_i, \hat{\theta}) g(z_i, \hat{\theta})'$. [I'm not sure what were supposed to use for \hat{G} . Do we first take the derivatives of the moment conditions and then just plug in our parameter moments? This seems to make sense, but as discussed in the MLE section below, there may be other options.

5.1.4 Aside: Proving Yaccine's Law

This is as good of a place as any to work through the proof of what Jerry call's Yaccine's law. [This can parallel what's in Newey's GMM handout, p. 6 (though unnumbered), but with a bit more explanation]. Chapter 1 of Hayashi has a good write-up, I think.

5.2 Miscellaneous

- You can iterate GMM to convergence, but it turns out that this has little effect on either asymptotics or small sample properties. As long as estimator of $\hat{\Omega}$ is consistent, it's just the same as 2-step GMM.
- Continuously Updated GMM
 - However, optimizing simultaneously over moments and weights does have an effect

$$\hat{\beta}_{CU-GMM} = \arg \min_{\beta} \hat{g}(\beta)' \hat{\Omega}(\beta) \hat{g}(\beta)$$

So the $\frac{\partial \hat{\Omega}(\beta)}{\partial \beta}$ term appears in the first order conditions.

- This has the same asymptotic properties but has *smaller bias* and *larger variance* than 2-step [in finite samples. Does this only hold for IV or for all GMM?]

6 Panel Data

6.1 Dynamic Panel Data

[Truth be told, I can't make heads or tails of Whitney's write up of dynamic panel data in his GMM handout. It'd be good to understand both his handout and the general idea of d.p.d]

6.2 Whitney's Recent Favorite Panel Data Problem

This problem appeared on the final exam for 14.385 in fall 2004 and on at least one prior general.

Add the problem. How the hell do you do it?

7 Quantile Regression

7.1 The Basic Idea

It turns out that

7.2 Standard Errors

The normal sandwich formula for standard errors doesn't work because the check function $\rho_\tau(u)$ is not twice differentiable. At least two options are available:

1. Replace the expected Hessian of the objective function (which doesn't exist because $\rho_\tau(u)$ doesn't have second derivatives) with the Hessian of the expected value (this effectively smooths the check function, and gives us derivatives, by taking expected values). See Van Der Vaart (1998) *Asymptotic Statistics* for details.
2. Bootstrap the standard errors. See

7.3 Worth Reading

The following papers are a nice introduction to quantiles and worth a look

1. Koenker & Hallock (2001) in *Journal of Economic Perspectives*
2. Koenker (???) *Econometrica*

8 Extremum Estimators

8.1 Consistency

The proof of the consistency of extremum estimators is beyond the scope of my brain, but the basic line of proof is similar to that outlined above for GMM. Here's the statement of the theorem:

Theorem 1 (Consistency of m-estimators) *If we have some estimator $\hat{\theta} = \arg \max_{\theta \in \Theta} \hat{Q}(\theta)$ such that*

1. *The function $Q(\theta)$ attains a unique maximum at θ_0 , the true parameter value (the Identification Condition)*
2. *Θ is compact (or $\hat{Q}(\theta)$ is convex and Θ is convex set)*
3. *$Q(\cdot)$ is continuous over Θ*
4. *An uniform convergence of the sample analog, $\sup_{\theta \in \Theta} |\hat{Q}(\theta) - Q(\theta)| \rightarrow_p 0$*

Then huzzah, huzzah! $\hat{\theta} \rightarrow_p \theta$, that is our extremum estimator of θ is consistent.

Newey-McFadden (1984) show that concavity of Q can replace compactness in the consistency proof, so we'll usually use this result (it's pretty friendly and nice to cite).

So when we're trying to show that some m-estimator (like MLE, GMM, Probit, or another special case of this near catch all class) is consistent, all we need to do is appeal to the theorem, show that the four requirements hold, and get on with our lives. Some examples of how such a proof would look (note the happy hand waving):

Example 2 (Consistency of Probit) *If $E(xx')$ exists and is non-singular, the the MLE probit estimator is consistent, i.e., $\hat{\theta} \rightarrow_p \theta$.*

- *The derivative of the log likelihood $\frac{\partial \ln \Phi(v)}{\partial v} = \frac{\phi(v)}{\Phi(v)}$ which is decreasing globally (you can take this on faith).*
- *Since $\ln \Phi(v)$ is concave, we can use Newey-McFadden (1984)*
- *[... need to finish this]*

9 Duration Models

[I must have dozed off when he was covering these in class. What should we know?]

10 MLE

10.1 Variance Estimation

So we know that $\sqrt{n}(\hat{\theta}_{MLE} - \theta) \sim N(0, I(\theta)^{-1})$, where I is the information matrix, i.e. $I = E[\partial^2 \ln L / \partial \theta_0 \partial \theta_0']$, the second derivative of the log likelihood evaluated at the true parameter. But how do we estimate this? We've got three options for the estimation of $I(\theta)$.

1. The empirical information: calculate the form of the information as a function of the true parameter, θ_0 , based on the distribution. Plug in the sample analog $\hat{\theta}$.

$$\hat{J}_1 = n^{-1} \sum_{i=1}^n J(x_i, \hat{\theta}), \text{ where } J(x, \theta) = - \int [\partial^2 \ln f(y|x, \theta) / \partial \theta \partial \theta'] f(y|x, \theta) dy$$

Greene makes the point (p. 481) that although this is a wonderful estimator (OK, he doesn't actually say "wonderful" but as Whitney points out, it's got good asymptotic properties), this estimator will rarely be available. The second derivatives of the log-likelihood will almost always be complicated nonlinear functions of the data whose exact expected values will be unknown.

2. The empirical mean of minus the Hessian.

$$\hat{J}_2 = -n^{-1} \sum_{i=1}^n \partial^2 \ln f(y_i|x_i, \hat{\theta}) / \partial \theta \partial \theta'$$

[So I'm not sure how we'd actually calculate this. We know what our distribution is, but I don't see how we can plug in our estimate for θ and then take the second derivative. It would make more sense to me if we said $\hat{J}_2 = -n^{-1} \sum_{i=1}^n \partial^2 \ln f(y_i|x_i, \theta) / \partial \theta \partial \theta' |_{\theta=\hat{\theta}}$. How does this work?]

3. The empirical variance of the score.

$$\hat{J}_3 = n^{-1} \sum_{i=1}^n \{ \partial \ln f(y_i|x_i, \hat{\theta}) / \partial \theta \} \{ \partial \ln f(y_i|x_i, \hat{\theta}) / \partial \theta' \}$$

[I've got the same questions regarding calculating this one.] I think this is the BHHH estimator. One nice advantage of it: in most cases it's easy to compute since we've calculated the scores to solve the likelihood. Moreover, it's always positive semidefinite.

According to Whitney's notes (see the last page of Asymptotic Theory for Nonlinear Estimators), these numbers correspond to their asymptotic efficiency. Apparently \hat{J}_3 has pretty crappy properties. But Greene (again, p. 481) says that none of the three estimators is preferable to the others on statistical grounds (I would trust Whitney on this). Greene does present an empirical example showing that the three estimators can give very different results [perhaps Whitney's point was about finite sample not asymptotic properties, but my notes clearly state asymptotics].

10.2 The Information Matrix Equality

There's a terse proof of this on p. 476 of Greene, version 5; Ruud p. 403 is a bit nicer and more informative. [Check to see if the stuff from 381 was more informative/rigorous]

The idea is that the variance of the score equals minus the expectation of the Hessian:

$$\text{Var}\left[\frac{\partial \ln L(\theta)}{\partial \theta}\right] = -E\left[\frac{d^2 \ln L(\theta)}{\partial \theta \partial \theta'}\right]$$

11 Hypothesis Testing

11.1 The Trinity of Hypothesis Testing

These three tests are all asymptotically equivalent for testing some hypothesis $H_0 : q(\theta) = 0$. All are distributed asymptotically χ_p^2 , where p is the number of restrictions being tested.

- **Likelihood ratio test.** If the restriction $q(\theta) = 0$ is valid, imposing it shouldn't lead to a big reduction in the log-likelihood function, so we form our test statistic by looking at the difference between these two values:

$$LR = 2(\hat{L}_U - \hat{L}_R)$$

One annoying feature of the likelihood ratio test is that it requires us to calculate both the restricted and the unrestricted estimators. Often, one (or both) of these will be painful to calculate.

- **Wald Test.** If the restriction is valid, then $q(\hat{\theta})$ should be close to zero. When the restrictions are linear (such as $\beta_1 = 0$, or $\beta_1 + \beta_2 = 1$) we can rewrite them as $R\beta = r$. Under the null, $R\hat{\beta} - r$ will be an asymptotically normal mean zero vector so we can form

$$W = [R\hat{\beta} - r]' [R \text{Var}(\hat{\beta}) R']^{-1} [R\hat{\beta} - r]$$

This also generalizes to testing nonlinear hypotheses. Recall from the delta method that for $\hat{\theta} \sim N(\mu, \Sigma)$, that for continuous function $g(\theta)$, with $G \equiv \partial g / \partial \theta$, that $g(\hat{\theta}) \sim N(g(\mu), G\Sigma G')$. So the Wald statistic

$$W = q' \left(\frac{\partial q}{\partial \theta} \text{Var}(\hat{\theta}) \frac{\partial q'}{\partial \theta} \right) q \sim \chi_p^2$$

Note that for calculating the Wald Test, you only need calculate the unrestricted estimator. The restriction is imposed in the test, not in the parameter estimation.

- **Lagrange Multiplier (aka Score) Test.** If the restriction is valid, then the restricted estimator should be near the point that maximizes the log-likelihood. Therefore, the slope of the log-likelihood function (the score) should be close to zero at the point where it's maximized subject to the restriction.

$$LM = \left(\frac{\partial \ln L(\hat{\theta}_R)}{\partial \hat{\theta}_R} \right)' [I(\hat{\theta}_R)]^{-1} \left(\frac{\partial \ln L(\hat{\theta}_R)}{\partial \hat{\theta}_R} \right)'$$

For the LM test, you only need to calculate the restricted estimator. In some cases, this might be much easier than calculating the unrestricted estimates. For the curious, Greene (p. 489) works through the Lagrangian (it's brief and easy).