Final Review

Nathaniel Mark

May 3, 2017

1 Proving Consistency/Inconsistency of an estimator (Using classical measurement error as an example)

First to start, recall that an estimator of a parameter is a pre-determined function of the data. Therefore, before we have the data at hand, it is a *random variable*. In our class, the parameter is generally β_1 and the estimator is denoted $\hat{\beta}_1$.

An estimator is *unbiased* if the expectation of the estimator is the true parameter. Mathematically, $E[\hat{\beta}] = \beta$.

An estimator is *consistent* if the probability distribution of $\hat{\beta}$ gets closer and closer to the true parameter value as the number of observations goes to infinity. In words, the probability that $\hat{\beta}$ will be extremely close to β goes to one as n goes to infinity (no matter how strictly we define "extremely close"). Mathematically:

$$P(|\hat{\beta} - \beta| < \epsilon) \to 1$$

$\forall \epsilon$ as

$$n \to \infty$$

How do we prove consistency?

Step One: Write out the "true" model:
Say, for example,

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

$$\tilde{X}_i = X_i + \eta_i$$

$$\epsilon_i \perp \eta_i \perp X_i$$

Step Two: Write out the definition of your estimator:

$$\hat{\beta}_1 = \frac{\frac{1}{n}\sum(\tilde{X}_i - \bar{\tilde{X}})(Y_i - \bar{Y}_i)}{\frac{1}{n}(\tilde{X}_i - \bar{\tilde{X}})(\tilde{X}_i - \bar{\tilde{X}})}$$

Step Three: Use Law of Large Numbers argument. Note we need 1) iid data 2) finite fourth moments:

$$\hat{\beta}_1 \rightarrow \frac{Cov(\tilde{X}_i, Y_i)}{Cov(\tilde{X}_i, \tilde{X}_i)}$$

Step Four: Plug in our true model. This is key. Why? We know $Cov(X_i, \eta_i)$, not $Cov(X_i, Y_i)$:

$$\hat{\beta}_1 \rightarrow \frac{Cov(X_i + \eta_i, \beta_0 + \beta_1 X_i + \epsilon_i)}{Cov(X_i + \eta_i, X_i + \eta_i)}$$

Step Five: Simplify using Covariance rules:
Since Cov(c+X,Y)=Cov(X,Y):

$$\hat{\beta}_1 \rightarrow \frac{Cov(X_i + \eta_i, \beta_1 X_i + \epsilon_i)}{Cov(X_i + \eta_i, X_i + \eta_i)}$$

Since Cov(X+Y,W+Z)=Cov(X,W)+Cov(X,Z)+Cov(Y,W)+Cov(Y,Z):

$$\hat{\beta}_1 \rightarrow \frac{Cov(X_i, \beta_1 X_i) + Cov(\eta_i, \beta_1 X_i) + Cov(X_i, \epsilon_i) + Cov(\eta_i, \epsilon_i)}{Cov(X_i, X_i) + Cov(\eta_i, X_i) + Cov(X_i, \eta_i) + Cov(\eta_i, \eta_i)}$$

By independence conditions:

$$\hat{\beta}_1 \rightarrow \frac{Cov(X_i, \beta_1 X_i)}{Cov(X_i, X_i) + Cov(\eta_i, \eta_i)}$$

By Cov(X,X)=Var(X) and Cov(aX,X)=aCov(X,X):

$$\hat{\beta}_1 \rightarrow \beta_1 \frac{Var(X_i)}{Var(X_i) + Var(\eta_i)}$$

2 Proving Biasedness/Unbiasedness of an estimator (Using OVB as an example)

Step One: Write out the "true" model:
Say, for example,

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$
$$\epsilon_i = \delta Z_i + \eta_i$$

$\epsilon_i \not\perp X_i$ and $\eta_i \perp X_i$
Z is omitted variable, X is variable of interest.

Step Two: Write out the definition of your estimator:

$$\hat{\beta}_1 = \frac{\frac{1}{n}\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\frac{1}{n}\sum(X_i - \bar{X})(X_i - \bar{X})}$$

Step Three: Plug in the true model::

$$\hat{\beta_1} = \frac{\frac{1}{n}\sum(X_i - \bar{X})(\beta_0 + \beta_1 X_i + \delta Z_i + \eta_i - \beta_0 - \beta_1\bar{X} - \delta\bar{Z} + \bar{\eta})}{\frac{1}{n}\sum(X_i - \bar{X})(X_i - \bar{X})}$$

Step Four: Simplify:

$$\hat{\beta_1} = \frac{\frac{1}{n}\sum(X_i - \bar{X})(\beta_1(X_i - \bar{X}) + (\delta Z_i - \bar{Z}) + (\eta_i - \bar{\eta})}{\frac{1}{n}\sum(X_i - \bar{X})(X_i - \bar{X})}$$

$$= \beta_1 + \delta\frac{\frac{1}{n}\sum((X_i - \bar{X})Z_i - \bar{Z})}{\frac{1}{n}\sum(X_i - \bar{X})(X_i - \bar{X})} + \frac{\frac{1}{n}\sum((X_i - \bar{X})(\eta_i - \bar{\eta})}{\frac{1}{n}\sum(X_i - \bar{X})(X_i - \bar{X})}$$

Step Five: Take Expectations and simplify:

$$E[\hat{\beta_1}] = E[\beta_1] + E[\delta\frac{\frac{1}{n}\sum((X_i - \bar{X})Z_i - \bar{Z})}{\frac{1}{n}\sum(X_i - \bar{X})(X_i - \bar{X})}] + E[\frac{\frac{1}{n}\sum((X_i - \bar{X})(\eta_i - \bar{\eta})}{\frac{1}{n}\sum(X_i - \bar{X})(X_i - \bar{X})}]$$

$$= \beta_1 + \delta E[\frac{\frac{1}{n}\sum((X_i - \bar{X})Z_i - \bar{Z})}{\frac{1}{n}\sum(X_i - \bar{X})(X_i - \bar{X})}]$$

$$= \beta_1 + \delta\frac{Cov(X,Z)}{Var(X)}$$

Note: Sign of bias is sign of δ times sign of cov(X,Z)

3 TSLS in potential outcomes framework - LATE (interpretation n'at).

Recall that our standard, simple, TSLS (IV) TRUE model is of the following form:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

$$X_i = \pi_0 + \pi_1 Z_i + \eta_i$$

With Z_i being a valid instrument: 1)Exogeneity $Z_i \perp \epsilon_i$, 2) Relevance $\pi_1 \neq 0$. These can also be written as 1) $Corr(Z_i, \epsilon_i) = 0$ and 2) $Corr(Z_i, X_i) \neq 0$. And the LSA #1 does not hold in the first (second stage) equation.

How do we estimate this?

Step One:
Run the (first stage) regression

$$X_i = \pi_0 + \pi_1 Z_i + \eta_i$$

Step Two:
Estimate the predicted X_i's:

$$\hat{X}_i = \hat{\pi}_0 + \hat{\pi}_1 Z_i$$

Step Three:
Run the (second stage) regression:

$$Y_i = \beta_0 + \beta_1 \hat{X}_i + \epsilon_i$$

The resulting $\hat{\beta_1}$ is our TSLS estimator of β_1.
In the special case with one endogenous variable and one instrument, we have an equation for this TSLS estimator:

$$\hat{\beta_1} = \frac{\frac{1}{n}\sum(Z_i - \bar{Z})(Y_i - \bar{Y})}{\frac{1}{n}\sum(Z_i - \bar{Z})(X_i - \bar{X})}$$

This value, under the usual finite fourth moments and iid assumptions, converges in probability to its population counterpart:

$$\hat{\beta_1} \to \frac{Cov(Z_i, Y_i)}{Cov(Z_i, X_i)}$$

$$= \frac{Cov(Z_i, Y_i)Cov(Z_i, Z_i)}{Cov(Z_i, X_i)Cov(Z_i, Z_i)}$$

By $\frac{Cov(Z_i,X_i)}{Cov(Z_i,Z_i)} = E[\pi_1] = \pi_1$: (Note: this is a good definition to know!)[1]

$$= \frac{Cov(Z_i, Y_i)}{\pi_1 Cov(Z_i, Z_i)}$$

Plugging in Y_i:

$$= \frac{Cov(Z_i, \beta_0 + \beta_1 X_i + \epsilon_i)}{\pi_1 Cov(Z_i, Z_i)}$$

By exogeneity condition:

$$= \frac{Cov(Z_i, \beta_1 X_i)}{\pi_1 Cov(Z_i, Z_i)}$$

[1] Proof:

$$\frac{Cov(Z_i, X_i)}{Cov(Z_i, Z_i)} = \frac{E[(Z_i - E(Z_i)](X_i - E(X_i))}{E[Z_i - E(Z_i)]^2}$$

$$= \frac{E[(Z_i - E(Z_i)](\pi_0 + \pi_1 Z_i + \eta_i - E(\pi_0 + \pi_1 Z_i + \eta_i))}{E[Z_i - E(Z_i)]^2}$$

$$= \frac{E[(Z_i - E(Z_i)](\pi_1 Z_i - E(\pi_1 Z_i))}{E[Z_i - E(Z_i)]^2}$$

$$= \frac{E[\pi_1 (Z_i - E(Z_i))^2]}{E[Z_i - E(Z_i)]^2}$$

IF π_1 is independent of Z_i

$$= E[\pi_1] \frac{E[(Z_i - E(Z_i))^2]}{E[Z_i - E(Z_i)]^2}$$

$$= E[\pi_1]$$

$$= \frac{\beta_1 Cov(Z_i, X_i)}{\pi_1 Cov(Z_i, Z_i)}$$

By $\frac{Cov(Z_i,X_i)}{Cov(Z_i,Z_i)} = \pi_1$: (Same definition!)

$$= \frac{\beta_1 \pi_1}{\pi_1} = \beta_1$$

So we just found consistency!

Now, what is different when we are in the heterogeneity framework?

First, recall that the heterogeneity framework is when the effect of X_i on Y_i is different for different people and the relationship between Z_i and X_i is different for different people. Mathematically,

$$Y_i = \beta_0 + \beta_{i,1} X_i + \epsilon_i$$
$$X_i = \pi_0 + \pi_{i,1} Z_i + \eta_i$$

A little bit of lingo:
1) ATE (Average Treatment Effect), the average effect on Y_i of increasing X_i by one unit: $ATE = E[\beta_{i,1}]$

2) LATE, (Local Average Treatment Effect), the weighted average effect on Y_i of increasing X_i by one unit where people who are effected more by the instrument have higher weights. $LATE = \frac{E[\pi_{1,i}\beta_{1,i}]}{E[\pi_{1,i}]}$

Example: Say two types of people. For half, $\pi_i = 0$, for the others, $\pi_i = 1$. Then

$$LATE = \frac{1}{2}\frac{E[0\beta_{1,i}|\pi_{i,1} = 0]}{E[.5]}] + \frac{1}{2}\frac{E[1\beta_{1,i}|\pi_{i,1} = 1]}{E[.5]} = E[1\beta_{1,i}|\pi_{i,1} = 1]$$

In this extreme case LATE is simply the average effect on those who the instrument effects.

Finally, the notes discuss how the IV estimator estimates LATE in the heterogenous treatment effect environment. To do so, we **need to assume** IV assumptions **and** $\pi_{i,1}$ is independent of Z_i. In the heterogenous treatment effect environment, what is different with the consistency derivation above? Just that we cannot say $E[\pi_{1,i}] = \pi_1$ and $E[\beta_{1,i}] = \beta_1$ or $E[\pi_{1,i}\beta_{1,i}] = E[\beta_{1,i}]E[\pi_{1,i}]$! Everything else is that same! So, our TSLS estimator goes to LATE:

$$\hat{\beta}_1 \rightarrow \frac{E[\pi_{1,i}\beta_{1,i}]}{E[\pi_{1,i}]}$$

LATE=ATE if one of the three discussed above conditions holds:
1)$E[\pi_{1,i}] = \pi_1$, 2)$E[\beta_{1,i}] = \beta_1$, or 3)$E[\pi_{1,i}\beta_{1,i}] = E[\beta_{1,i}]E[\pi_{1,i}]$ (independence of $\pi_{1,i}$ and $\beta_{1,i}$)

4 Panel Data Models and Clustered Standard Errors

Remember our fixed effects true model:

$$y_{it} = \beta_0 + \beta_1 X_{it} + \alpha_i + \gamma_t + \epsilon_{it}$$

Recall we can estimate in three ways:

1) First differencing– Usually only used when T=2. This removes both omitted variables bias due to variables that change only over time (in γ_t) and those that change only over entity (in α_i).

2) Dummy Variables – Dummies included for (all but one) entities and/or for (all but one) time periods. Adding entity dummies removes omitted variable bias from variables that change over time but not over entity while adding entity dummies removes omitted variable bias from variables that change over entity but not over time.

3)Demeaning – means within entity are subtracted from all variables and/or means within time are subtracted from all variables (it is slightly more complicated if both are). Entity demeaning removes omitted variable bias from variables that change over time but not over entity and Time demeaning removes omitted variable bias from variables that change over entity but not over time.

In all of these models, and in panel data models that we just ignore the α and γs, we need to use *clustered standard errors*.

What you need to know and what you don't need to know about clustered standard errors:

You should know:

1) We use clustered standard errors in fixed effects instead of the usual SEs because the iid assumption does not hold: $\epsilon_{i,t}$ is correlated with $\epsilon_{i,t+1}$.

2) The clustered standard error of the mean estimator of $Y_{i,t}$ is given by estimating the variance *of the within entity means* and using that variance estimate instead of the usual one in the SE formula. i.e. the true model is

$$Y_{i,t} = \beta_0 + \epsilon_{i,t}$$

Define

$$\bar{Y_i} = \frac{1}{T}\sum_{t=1}^{T} Y_{i,t}$$

Then, the clustered standard error of $\hat{\beta}_0 = \frac{1}{nT}\sum_{\forall i,t} Y_{i,t}$ is

$$SE(\hat{\beta}_0) = \sqrt{\frac{s^2_{\bar{Y_i}}}{n}}$$

3) The clustered standard error for β_1 is analogous to the simple mean model above. i.e. it uses a similar idea.

You do not need to know:
1) OLS estimator clustered SE equation.

5 NLLS and MLE of binary models

We assume the true model looks like:

$$P(Y_i = 1|X_{i,1}, X_{i,2}, ...) = F(\beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2})$$

OR, equivalently,

$$Y_i = F(\beta_0 + \beta_1 X_1 + \beta_2 X_2) + u_i$$

If this F is the normal CDF, this is a probit model, it this F is the logit CDF, this is a logit model.

We want estimates of β_0,β_1,... How do we estimate?

METHOD ONE: NON-LINEAR LEAST SQUARES
We let a computer find the solution to:

$$min_{\beta_0,\beta_1,...} \sum (Y_i - F(\beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2}))^2$$

We then set $\hat{\beta_{1,NLLS}}$ to the β_1 the computer found minimized this function.

Properties of the NLLS estimators:
1) Consistent
2) Asymptotically normal
3) **Inefficient**
4) Almost no-one uses it.

METHOD TWO: MAXIMUM LIKELIHOOD ESTIMATOR
Estimating by MLE is efficient and is also consistent and asymptotically normal. How do we do it? We use the fact that we know these are binary outcome data.

The steps to solving for the MLEs are (note I am moving to the 1 variable case for simplicity):

Step 1) Determine the probability of observing the data pair (Y_i, X_i):

$$F(\beta_0 + \beta_1 X_i)^{Y_i}(1 - F(\beta_0 + \beta_1 X_i))^{1-Y_i}$$

Step 2) By the IID data assumption, the probability of observing all of your data (i.e. the **likelihood**) is simply the product of all those probabilities.

$$L(\beta_0, \beta_1|X, Y) = \prod_{i=1}^{n} F(\beta_0 + \beta_1 X_i)^{Y_i}(1 - F(\beta_0 + \beta_1 X_i))^{1-Y_i}$$

Step 3) To attain the log-likelihood (what we use for estimation), take the log:

$$\mathcal{L} = \sum_{i=1}^{n} Y_i log(F(\beta_0 + \beta_1 X_i)) + \sum_{i=1}^{n} (1 - Y_i) log(1 - F(\beta_0 + \beta_1 X_i))$$

Step 4) To find the parameters that maximize the likelihood, take the first order conditions of the log-likelihood:
FOC wrt β_0:

$$\frac{\partial \mathcal{L}}{\partial \beta_0} = \sum_{i=1}^{n} (Y_i) \frac{f(\beta_0 + \beta_1 X_i)}{F(\beta_0 + \beta_1 X_i)} + \sum_{i=1}^{n} (1 - Y_i) \frac{-f(\beta_0 + \beta_1 X_i)}{1 - F(\beta_0 - \beta_1 X_i)} = 0$$

FOC wrt β_1:

$$\frac{\partial \mathcal{L}}{\partial \beta_1} = \sum_{i=1}^{n} X_i (Y_i) \frac{f(\beta_0 + \beta_1 X_i)}{F(\beta_0 + \beta_1 X_i)} + \sum_{i=1}^{n} X_i (1 - Y_i) \frac{-f(\beta_0 + \beta_1 X_i)}{1 - F(\beta_0 - \beta_1 X_i)} = 0$$

where $\frac{\partial F(x)}{\partial x} = f(x)$.

You will not need to go further than this to solve, except for the possible exception of the logit model (i.e. your homework).

6 Final Notes

J-Test:

The J-test is a test of instrument exogeneity. If the test rejects the null, then we cannot say we have exogeneity. It does not tell us which instrument fails.
m=# of instruments.
k=# of endogenous variables.

In order to use the J-test, there must be more instruments than endogenous variables (i.e. we must be **overidentified**).
Step 1) Run the standard IV model with all instruments. We end up with a second stage estimated model of

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 \hat{X}_{1i} + ... + \hat{\beta}_k \hat{X}_{ki} + \hat{\beta}_{k+1} W_{1i} + ... + u_i$$

Step 2) Estimate the residuals from that IV model:

$$\hat{u}_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 \hat{X}_{1i} + ... + \hat{\beta}_k \hat{X}_{ki} + \hat{\beta}_{k+1} W_{1i} + ...)$$

Step 3) Run a regression of these estimated residuals on instruments and exogenous variables. ie a regression of the model

$$\hat{u}_i = \gamma_0 + \gamma_1 Z_{1i} + ... + \gamma_m Z_{mi} + \hat{\beta}_{m+1} W_{1i} + ...)$$

Step 4) Run an F-test using the null hypothesis that the instruments have no effect on the estimated residuals. Save the estimated F stat. ($H_0 : \gamma_1 = ... = \gamma_m = 0$) Step 5) Ask is m times the F-stat is larger than the chi-square with m-k degrees of freedom critical value. i.e. Reject if $m\hat{F} > \chi^2_{m-k}(.95)$

Any other questions??