Inference in Complex Models

Tyler Ransom
Duke University

June 23, 2012

Outline

- Review inference
- Talk about clustering
- Talk about bootstrapping

Why all the fuss over standard errors?

- Empiricists are in the business of estimating models and assigning meaning to those model estimates
- Coupled with point estimates, standard errors are the most crucial piece of an empiricists findings
- If the standard errors are wrong, then the results will be wrong, too, because false inferences will be made
- This is why so much energy is spent deriving correct standard errors under a variety of assumptions (e.g. Hayashis textbook)

Standard errors

Economists prefer to have models with parameter estimates that are *consistent* and *asymptotically normal*, i.e.

$$\sqrt{n}\left(\hat{\beta}-\beta\right) \xrightarrow{d} N\left(0, \text{Avar}\right)$$

(Hayashi, p. 113)

- We want to know the variance of our parameter estimates so that we can tell whether or not we got a lucky draw, or if they are actually different from some number (typically zero)
- The standard errors of the $\hat{\beta}$ vector are embedded in the $\widehat{\text{Avar}}$ matrix (our estimate of the true Avar matrix)
- In the next few slides, we'll go over what the formula is for $\widehat{\text{Avar}}$ in more advanced econometric models

Corrected standard errors in clustered models

Suppose we want to estimate standard errors for models in which there are multiple dimensions of agents

- e.g. in development, we might have data on individuals in families, across villages, and across countries
- We want to get "correct" standard errors for the estimates of our models, and we know that there is some amount of correlation (i.e. heteroskedasticity) within families, villages, and countries
- By "correct" we mean "consistent in the presence of heteroskedasticity"

Other examples of clustering

IO:

- Car data might be classified by vehicle class (e.g. sedan), make (e.g. Ford), and country of origin (i.e. Mitsubishi is different than Toyota, but both come from Japan)

Health:

- Data on patients for a specific doctor and across hospitals for a similarly specialized doctor (i.e. patients for Doctor X, who is an orthopedic surgeon at Hospital Y; and patients for Doctor Z who is an orthopedic surgeon at hospital W)

Education:

- Test scores for students in a certain grade, across teachers, across schools in the same district

Labor:

- Workers in a specific occupation across different firms in different cities

Robust clustered standard errors for GMM models

The formula for robust clustered standard errors is

$$\widehat{\text{Avar}} = \left(\mathbf{X}'\mathbf{X} \left(\sum_{j=1}^{N_c} u_j' u_j \right)^{-1} \mathbf{X}'\mathbf{X} \right)^{-1}$$

where

$$u_j = \sum_{i \in j} \varepsilon_i \mathbf{x}_i$$

and N_c is the number of clusters. When each cluster has just one element, this formula collapses to the "robust" variance matrix [next slide]

Robust standard errors for GMM models

Stata's `robust` option produces the following Avar matrix:

$$\widehat{\text{Avar}} = \left(\mathbf{X}'\mathbf{X} \left(\mathbf{X}'\mathbf{B}\mathbf{X} \right)^{-1} \mathbf{X}'\mathbf{X} \right)^{-1}$$

where

- $\mathbf{X}$ is a matrix of regressors
- $\mathbf{B}_{i,i} = \left(y_i - x_i \hat{\beta} \right)^2$

Cluster vs. Robust

When should I use cluster-corrected standard errors, and when should I use robust standard errors?

- If you have any sort of cross-unit correlation (i.e. panel data for people over time, cross sectional data for people over space), you should use cluster-corrected standard errors
- The formula collapses to the regular robust formula if there is no clustering
- If using fixed effects, cluster at the same level as your fixed effects
- If all else fails, you can always bootstrap

Standard errors for other models

- All classical econometric models can be considered as either GMM or MLE. (Furthermore, GMM and MLE are asymptotically equivalent in some situations)
- Most structural econometric models, however, incorporate simulation methods, multiple estimation stages, and/or contraction mapping/fixed point algorithms (in addition to a main GMM/MLE optimization) – we'll cover these in more detail next class
- In these more complex estimation algorithms, it's unclear how the standard errors of the parameter estimates are affected by the use of previous stage estimates and/or simulation draws
- The solution to this problem is bootstrapping, which we will cover today

Bootstrapping

- Suppose we have a crazy model and we have no idea what the variance of our estimates looks like
- The solution to this problem is bootstrapping:
- We can recover the variance matrix by inducing randomness into our estimates by sampling observations *with replacement*
- We can then repeat this a large number of times and look at the distribution of our estimates
- e.g. the 90% confidence interval will be the 5th and 95th percentiles of our bootstrap distribution

Bootstrap Formula

After running B bootstrap iterations of the program, the bootstrap approximation to the asymptotic variance is calculated according to

$$\widehat{\text{Avar}} = \frac{1}{B-1} \sum_{b=1}^{B} \left[\hat{\theta}_m(b) - \overline{\hat{\theta}} \right] \left[\hat{\theta}_m(b) - \overline{\hat{\theta}} \right]'$$

where $\hat{\theta}_m(b)$ refers to the parameter estimates taken from the subsample of the population (from draw b, with replacement), and $\overline{\hat{\theta}}$ is the average parameter estimate, computed over the B bootstrap draws.

Example

- Simple example: Suppose we have a random variable Z which is a function of two other random variables X and Y (i.e. $Z = \frac{X}{Y}$), and we want to estimate $\mathbb{E}[Z] = \mathbb{E}\left[\frac{X}{Y}\right]$, as well as the standard error of our estimate
  1. Draw a sample with replacement from the data
  2. Calculate `mean(Z)`
  3. Repeat B times
  4. Apply the formula from the previous slide
- This same process holds for any estimation routine. Just replace step 2 with whatever estimation you're doing.

Bootstrapping in Matlab

- Matlab has a useful command `bootstrp` which will automatically sample with replacement and perform bootstrap iterations on any function desired
- For more open-ended problems, `randsample` is a function that will sample data with replacement so users can bootstrap an estimation procedure which may not fit well with `bootstrp`
- Syntax:

```
bootstat = bootstrp(nboot,bootfun,d1,...)

y = randsample(population,k,replacement)
```