Advanced Applied Econometrics – selected OLS topics

AET and OLS with heterogeneity

By Felix Weinhardt

Question 1:

We know that for unbiasedness the most critical assumption is E(Xe)=0 and that the OLS estimator in a bivariate model can be obtained from dividing the sample covariance in X and Y by the variance in X.

Write down the formula for the OLS bias if E(Xe) does not equal 0 for the univariate case.

How do you have to modify this formula if you are interested in the same beta but include further control variables in your regression (hint: regression anatomy theorem).

Question 2:

Read the Altonji, Elter and Taber paper. Some instructions for doing this: when reading, focus in the things that we are interested in here, which roughly start in section III on page 20.

Note we do not want to reproduce everything they have done. We are only interested in the alternative key assumption that they are making to replace the normal CIA. For example, do not get distracted by the fact that they are estimating a probit models as well (we are just discussing the OLS case), and that they have some asymptotics to prove that their alternative assumption makes sense. Of course, they have to argue that what they are doing is not at hoc.

A simplified OLS version of their regression of interest is Y=a+b_1CH+b_2X+e where W is a matrix that captures both sets of independent variables, the dummy CH and the X.

Note that throughout they continue assuming E(Xe)=0 for the set of X variables within W. What we are replacing/being worried about is E(CHe)=0.

In simple words, what do CH and X and Y refer to in their paper?

What is the assumption that they use as alterative to E(CH,e)=0. Without going into the details on how they derive this, how do they justify this assumption? Do you think this makes sense?

Can you modify your results from question 1b for allow for this alternative assumption? Hint (substitute for cov(e,CH~) where CH~ comes from the auxiliary regression, and note that cov(e,CH)= cov(e,CH~))

Can we estimate the various elements of the bias-corrected beta estimate under their alternative assumption? Write down the individual elements and the specifications to estimate these. How many regressions do we need to run?

Question 3:

From the lecture notes you saw some simulated results where adding an "irrelevant" control changes the OLS estimate. This was because of violations of the i.i.d. assumption in the multiple linear regression model.

Preferably in Stata, generate a dataset with an outcome variable Y that depends on X and a random noise term only, and where effects are heterogeneous across two equally sized groups (strata). Then, generate another variable W that has no direct effects on Y but can correlate with X.

Start by typing

```
set obs 1000
```

Can you generate W in a way that adding W changes the OLS coefficient for X?

In the end, running in Stata should produce different estimates for the effect of X, even though Y is constructed not to depend on W.

```
reg Y X

reg Y X W
```

Bonus: Imagine for some reason you know which individuals belong to which strata. How can you modify the regression that you estimate so that adding W has no effect on the estimated effect of X on Y?