Problem Set 4

Econometrics I
NYU Stern, Fall 2018

Professor Paul T. Scott
Email: ptscott@stern.nyu.edu

Problem 1 - A Binary Instrument

Suppose we have the following linear model of test scores:

$$Y_i = \beta_0 + \beta_1 T_i + X\boldsymbol{\beta}_2 + \varepsilon_i$$

where Y_i is the test score for student i and T_i is the hours of tutoring time received by student i (per week), and X is a vector of controls variables.

Suppose that students were randomly offered highly subsidized tutoring services. The means and standard deviations of test scores and tutoring hours for the two groups follow:

Group	n	$\bar{Y}$	$\hat{\sigma}_Y$	$\bar{T}$	$\hat{\sigma}_T$
Treated	400	29	3.8	4.5	.7
Not treated	1200	25	4	1.2	1.8

For each of the following, compute the answer if you can. If you can't, explain what additional information or assumptions you would need to do so.

(a) Provide an IV estimate of β_1.

(b) Compute standard errors for your estimate to (a).

(c) Compute the first stage F-statistic (actually a t-statistic here) for the instrument in this example.

Problem 2 - Panel Data Theory

Suppose we have a balanced panel with n individuals and T time periods. We have the following linear model:

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \gamma_i + \varepsilon_{it}$$

(a) Show that the first differences and fixed effects estimators are equivalent when $T = 2$.

(b) If $T = 2$ and the only regressor is a dummy variable with $\forall i: \quad X_{i1} = 0$, show that the fixed effects estimate of β_1 can be expressed as a difference-in-differences.

Problem 3 - The Wage Equation, Revisited

The Cornwell and Rupert data for this problem can be downloaded from

http://ptscott.com/teaching/data/cornwell-rupert-withid.csv

Source: Cornwell and Rupert, (1988) "Efficient estimation with panel data: an empirical comparison of instrumental variables estimators"

Consider the following wage equation:

$$lwage = \beta_1 + \beta_2 fem + \beta_3 ed + \beta_4 occ + \beta_5 ind + \beta_6 exp + \beta_7 blk + \beta_8 union + \varepsilon$$

(a) Consider the OLS pooled estimator for this model. The data is a balanced panel with individuals indexed by "id" and years given by "year". It's plausible that the error term is correlated across years for a given individual, and this might affect the precision of the estimates. Compute the OLS estimates and compare the standard OLS standard errors to standard errors allowing for clustering by individual.

(b) With the correlated error terms, OLS is potentially no longer the most efficient estimator. Compute the random effects estimate and compare to OLS.

(c) Compute the fixed effects estimator. Explain which variables no longer have an identified effect and why.

(d) Discuss which model (RE or FE) you prefer. Perform a test of the RE exogeneity assumption, or discuss why such a test is not practical.

(e) Put aside the panel structure and individual effects for the moment. As the experience variable might be endogenous, estimate the model using 2SLS with (wks,south,smsa) as instruments. Compute both the OLS and 2SLS estimates and report your results. Perform a test for the OLS exogeneity assumption, or discuss why such a test is not practical.

(f) Consider the possibility of estimating the model with both fixed effects and instruments. Would the above instruments still be relevant? Estimate this model, or discuss why doing so is not practical.