Lab 1 – Regression, endogeneity and panel data

Based on Cornwell and Rupert Panel Data. N = 595 Individuals, T = 7 periods (balanced panel). Examine various specifications based on

\[
\log\text{wage}_t = f(\text{experience, weeks worked, smsa, marital status, gender, union membership, education})_t + \epsilon_t.
\]

Preliminary setup to set some parameters for the analysis.

For most of this exercise, we will examine the impact of weeks worked on log wage, and consider the possible endogeneity of weeks worked in this equation. This exercise is based on the Cornwell and Rupert data, Cornwell-Rupert.lpj

I. Cross Section Variation

A. Data Description, get familiar with the log wages variable.
   Note the trend in the average of the log wage variable over the 7 years.
   The pattern should be evident in the box plots as well.

B. Kernel Density estimator for experience
   Kernel density estimators are used throughout the literature to describe the distribution of a variable in a sample. Here we examine experience.

C. Gender difference in log wages
   Is there a ‘gender differential’ in log wages? What does the comparison of the means of men vs. women suggest? Does the kernel density estimator suggest the same pattern?

D. Gender differential in regression model, partial effects
   Based on the finding in part C, we build the gender effect into the model.
   (1) Examine the regression results. Does there appear to be a gender effect in the regression?
   (2) The first set of partial effects shows the overall average partial effect of experience. Since the model is nonlinear in experience, the partial effect varies with experience. What do the results suggest about the partial effect of experience on log wages?
   (3) Since the model contains an interaction between gender and experience, there will be different effects for experience for the two genders. What do we find?
   (4) There are different estimates for the average partial effect of years of education for men and women in the sample. What are the results?
E. Endogeneity of weeks worked

Is weeks worked endogenous in the log wage equation? We approach the examination in two ways. For purposes of the illustration, we slightly respecify the model. First, ‘X’ defines the right hand side variables of the model, including WKS; second, ‘Z’ defines the list of all exogenous variables. This includes X as well as four ‘instruments,’ occ, ind, south, ms. To carry out the examination, we do the following:

(1) use OLS ignoring the problem, for the basis of comparison.
(2) Control function. Obtain residuals from regression of WKS on Z, and add the residual to the model. A ‘test’ of endogeneity is a test of whether the coefficient on ’u’ is significant. Is it?
(3) 2SLS (IV) estimation of the model, X as regressors, Z, as instruments. (Notice that the coefficients on X in this regression are the same as those on X in the control function regression. Are the standard errors? Why not? Which are right?

F. Repeating the examination in part C. for marital status and union membership

I’m interested in whether the gender effect I’ve observed in part C is repeated with respect to marital status and union membership. Back in part C of the script, change FEM to MS and repeat the analysis. Do likewise with UNION. What do you find?

II. Panel Data

A. Ignoring panel effects in a pooled estimator

It might be tempting to ignore the panel data effects in analyzing the effects of WKS on log wage, hoping that they will go away. Will they? Compare the results of the ‘pooled’ model to the ‘fixed effects’ model. Does the effect go away if we ignore it?

B. The implication of correcting for clustering

Maybe the problem with the pooled estimator is that we haven’t corrected for the clustering in the data. Is that the problem? Does the correction fix the problem?

C. The implication of including control for fixed effects

Does controlling for fixed effects change the estimated effect of WKS on LWAGE? We already computed the regression in part A. Take another look.

D. Testing for the presence of fixed effects (vs. no effects).

Testing for the presence of fixed effects is a simple F test. The needed results are generated automatically when the model is fit. What is the result? Is the null hypothesis of no individual effects rejected?

E. Time invariant variables in a fixed effects regression model

This is a perennial problem in fitting fixed effects models. In our case, education, ED, and gender, FEM, are interesting variables. Note how the estimator reacts when we attempt to include them in the model. The random effects model is an alternative specification that allows time invariant variables, but makes a strong assumption. The Breusch and Pagan test is the standard device for testing the REM against the no effects null. It is reported with the REM results. What is the outcome of the test?
F. The Hausman test for fixed vs. random effects
   The $H$ statistic is based on the difference between the RE estimates and the FE estimates. It often fails, however, because of a built in problem of the rank of the difference of the two covariance matrices. By trimming the model, we get our result. What is the result of the test? (This part of the exercise includes a side trip to do some matrix algebra and uncover the problem with the Hausman test in the larger model.)

G. Using a variable addition test instead of the Hausman test
   The Wu test, which is a variable addition test, is a convenient way to carry out the Hausman test – it builds on the idea of Mundlak’s specification of the FE model.

H. Variable addition test in Mundlak’s specification of the random effects model
   The more natural, appropriate place to carry out the Wu test is in the random effects model. Here, we carry out the test in the context of a full specification of the model. But, ignoring the possible endogeneity of WKS. Leaving that aside for now, what is the result of the Wu test?

I. Reexamine endogeneity of weeks worked after fixed effects are added to the model
   Finally, we reconsider the endogeneity of WKS in the model. Here we reconsider a fairly small model (for this example, of course). Oddly, to do a FE version of 2SLS here, we need to drop a second component of the set of time dummy variables, because of a multicollinearity problem. That done, we consider simple FE vs. FE, 2SLS estimation of the equation. Based on our results, what becomes of the endogeneity of WKS in the model? What becomes of the significance of WKS in the model?