ECON 6511: Advanced Applied Econometrics
Homework 4 Solutions

1. (Wooldridge, Chapter 14, Problem 5) Use the data in RENTAL.dta for this exercise. The data for the years 1980 and 1990 include rental prices and other variables for college towns. The idea is to see whether a stronger presence of students affects rental rates. The unobserved effects model is:

$$\log(rent_{it}) = \beta_0 + \delta_0 y90_t + \beta_1 \log(pop_{it}) + \beta_2 \log(avginc_{it}) + \beta_3 pctstu_{it} + a_i + u_{it}$$

where *pop* is city population, *avginc* is average income, and *pctstu* is student population as a percentage of city population (during the school year).

(a) Estimate the equation by pooled OLS and report the results in equation form. What do you make of the estimate on the 1990 dummy variable? What do you get for $\hat{\beta}_{pctstu}$?
Answer: Using pooled OLS we obtain:

$$\log(rent_{it}) = -0.569 + 0.262 y90_t + 0.041 \log(pop_{it}) + 0.571 \log(avginc_{it}) + 0.005 pctstu_{it}$$

The positive and very significant coefficient on $y90$ simply means that, other things in the equation fixed, nominal rents grew by over 26% over the 10 year period. The coefficient on *pctstu* means that a one percentage point increase in *pctstu* increases rent by half a percent (.5%). The t statistic of 4.95 shows that, at least based on the usual analysis, *pctstu* is very statistically significant.

(b) Are the standard errors you report in part (a) valid? Explain.
Answer: The standard errors from part (i) are not valid, unless we think a_i does not really appear in the equation. If a_i is in the error term, the errors across the two time periods for each city are positively correlated, and this invalidates the usual OLS standard errors and t statistics.

(c) Now, difference the equation and estimate by OLS. Compare your estimate of β_{pctstu} with that from part (b). Does the relative size of the student population appear to affect rental prices?
Answer: The equation estimated in differences is:

$$\Delta \log(rent_{it}) = 0.386 + 0.072 \Delta \log(pop_{it}) + 0.310 \Delta \log(avginc_{it}) + 0.0112 \Delta pctstu_{it}$$

Interestingly, the effect of *pctstu* is over twice as large as we estimated in the pooled OLS equation. Now, a one percentage point increase in *pctstu* is estimated to increase rental rates by about 1.1%. Not surprisingly, we obtain a much less precise estimate when we difference (although the OLS standard errors from part (i) are likely to be much too small because of the positive serial correlation in the errors within each city). While we have differenced away a_i, there may be other unobservables that change over time and are correlated with $\Delta pctstu$.

(d) Estimate the model by fixed effects to verify that you get identical estimates and standard errors to those in part (c).

Answer: The fixed effects estimates, reported in equation form, are

$$\Delta \log(rent_{it}) = 0.386 y90_t + 0.072 \log(pop_{it}) + 0.310 \log(avginc_{it}) + 0.0112 pctstu_{it}$$

The coefficient on $y90_t$ is identical to the intercept from the first difference estimation, and the slope coefficients and standard errors are identical to first differencing.

2. (Wooldridge, Chapter 14, Problem 4) Papke (1994) studied the effect of the Indiana enterprise zone (EZ) program on unemployment claims. The author uses a model that allows each city to have its own time trend:

$$\log(uclms_{it}) = a_i + c_i t + \beta_1 ez_{it} + u_{it},$$

where a_i and c_i are both unobserved effects. This allows for more heterogeneity across cities.

(a) Show that, when the previous equation is first differenced, we obtain

$$\Delta \log(uclms_{it}) = c_i + \beta_1 \Delta ez_{it} + \Delta u_{it}, \quad t = 2, \ldots, T.$$

Notice that the differenced equation contains a fixed effect, c_i.

Answer: Write the equation for times t and $t-1$ as:

$$\begin{aligned} \log(uclms_{it}) =& a_i + c_i t + \beta_1 ez_{it} + u_{it} \\ \log(uclms_{i,t-1}) =& a_i + c_i(t-1) + \beta_1 ez_{i,t-1} + u_{i,t-1} \end{aligned}$$

and subtract the second equation from the first. The a_i are eliminated and $c_i t - c_i(t-1) = c_i$. So for each $t \geq 2$, we have:

$$\Delta \log(uclms_{it}) = c_i + \beta_1 \Delta ez_{it} + \Delta u_{it}$$

(b) Estimate the differenced equation by fixed effects using the data in EZUNEM.DTA. What is the estimate of β_1? Is it very different from the estimate in textbook Example 13.8 of -0.182? Is the effect of enterprise zones still statistically significant?

Answer: Because the differenced equation contains the fixed effect c_i, we estimate it by

FE. We get $\hat{\beta}_1 = -0.251$. The estimate is actually larger in magnitude than we obtain in Example 13.8 (where $\hat{\beta}_1 = -1.82$) but we have not yet included year dummies. In any case, the estimated effect of an EZ is still large and statistically significant.

(c) Add a full set of year dummies to the estimation in part (b). What happens to the estimate of β_1?

Answer: Adding the year dummies reduces the estimated EZ effect, and makes it more comparable to what we obtained with $c_i t$ in the model. Using FE on the first-differenced equation gives $\hat{\beta}_1 = -0.192$, which is fairly similar to the estimates without the city-specific trends.

3. (Wooldridge, Chapter 14, Problem 7) Use the state-level data on murder rates and executions in MURDER.dta for the following exercise

(a) Consider the unobserved model

$$mrdrte_{it} = \eta_t + \beta_1 exec_{it} + \beta_2 unem_{it} + a_i + u_{it},$$

where η_t simply denotes different year intercepts and a_i is the unobserved state effects. If past executions of convicted murderers have a deterrent effect, what should be the sign of β_1? What sign do you think β_2 should have? Explain.

Answer: If there is a deterrent effect then $\beta_1 < 0$. The sign of β_2 is not entirely obvious, although one possibility is that a better economy means less crime in general, including violent crime (such as drug dealing) that would lead to fewer murders. This would imply $\beta_2 > 0$.

(b) Using just the years 1990 and 1993, estimate the equation from part (a) by pooled OLS. Ignore the serial correlation problem in the composite errors. Do you find any evidence for a deterrent effect?

Answer: The pooled OLS estimates using 1990 and 1993 are:

$$mrdrte_{it} = -5.28 - 2.07 d98_t + 0.128 exec_{it} + 2.53 unem_{it}$$

There is no evidence of a deterrent effect, as the coefficient on $exec$ is actually positive (although it is not statistically significant).

(c) Now, using 1990 and 1993, estimate the equation by fixed effects. You may use first differencing since you are only using two years of data. Is there evidence of a deterrent effect? How strong?

Answer: The first-differenced equation is:

$$\Delta mrdrte_i = 0.413 - 0.104 \Delta exec_i - 0.067 \Delta unem_i$$

Now, there is a statistically significant deterrent effect: 10 more executions is estimated to reduce the murder rate by 1.04, or one murder per 100,000 people. Is this a large effect? Executions are relatively rare in most states, but murder rates are relatively low on average, too. In 1993, the average murder rate was about 8.7; a reduction of one would be nontrivial. For the (unknown) people whose lives might be saved via a deterrent effect, it would seem important.

(d) Compute the heteroskedasticity-robust standard error for the estimation in part (c).

Answer: The heteroskedasticity-robust standard error for $\Delta exec_i$ is .017. Somewhat surprisingly, this is well below the nonrobust standard error. If we use the robust standard error, the statistical evidence for the deterrent effect is quite strong ($t \approx ?6.1$).

(e) Find the state that has the largest number for the execution variable in 1993. (The variable *exec* is total executions in 1991, 1992, and 1993.) How much bigger is this value than the next highest value?

Answer: Texas had by far the largest value of *exec*, 34. The next highest state was Virginia, with 11. These are three-year totals.

(f) Estimate the equation using first differencing, dropping Texas from the analysis. Compute the usual and heteroskedasticity-robust standard errors. Now, what do you find? What is going on?

Answer: Without Texas in the estimation, we get the following, with heteroskedasticity-robust standard errors:

$$\Delta mrdrte_i = 0.413 - 0.067\Delta exec_i - 0.070\Delta unem_i$$

Now the estimated deterrent effect is smaller. Perhaps more importantly, the standard error on $\Delta exec_i$ has increased by a substantial amount. This happens because when we drop Texas, we lose much of the variation in the key explanatory variable, $\Delta exec_i$.

(g) Use all three years of data and estimate the model by fixed effects. Include Texas in the analysis. Discuss the size and statistical significance of the deterrent effect compared with only using 1990 and 1993.

Answer: When we apply fixed effects using all three years of data and all states we get:

$$mrdrte_{it} = 1.56d90_t + 1.73d93_t - 0.138exec_{it} + 0.221unem_{it}$$

The size of the deterrent effect is actually slightly larger than when 1987 is not used. However, the t statistic is only about $-.78$. Thus, while the magnitude of the effect is similar, the statistical significance is not. It is somewhat odd that adding another year of data causes the standard error on the exec coefficient to increase nontrivially.

4. (Wooldridge, Chapter 14, Problem 14) Use the data set in AIRFARE.dta to answer this question. The estimates can be compared with those at the end of the last lecture.

   (a) Compute the average of the variable *concen* for each route; call these *concenbar*. How many different time averages can there be? Report the smallest and the largest. Note: You can generate a variable containing the average value of *concen* for each route using the Stata command: "egen concenbar = mean(concen), by(id)"

   Answer: Because there are 1,149 routes – that is, 1,149 different cross-sectional units – there can be at most 1,149 different values of *concenbar*. The largest and smallest in the data set are, respectively, .1862 and .9997.

   (b) Estimate the equation:

$$\begin{aligned} lfare_{it} = &\beta_0 + \delta_1 y98_t + \delta_2 y99_t + \delta_3 y00_t + \beta_1 concen_{it} + \beta_2 ldist_t + \beta_3 ldistsq_i \\ &+ \gamma_1 concenbar_i + a_i + u_{it} \end{aligned}$$

   by random effects. Verify that $\hat{\beta}_1$ is identical to the FE estimate computed in class.

   Answer: Estimating the equation that includes *concenbar* by random effects gives a coefficient on *concen* equal to .168859, which agrees with the FE estimate to the reported six decimal places.

   (c) Using the equation from part (b) and the usual RE standard error, test $H_0 : \gamma_1 = 0$ against the two-sided alternative. Report the p-value. What do you conclude about RE versus FE for estimating β_1 in this application?

   Answer: The usual RE t statistic on *concenbar* in part (b) is 3.15, with two-sided p-value = .002. Thus, the RE estimator is strongly rejected (in a statistical sense). Even though the FE and RE estimates are not very different, we should go with the FE estimate.

   (d) Obtain a t-statistic (and, therefore, p-value) that is robust to arbitrary serial correlation and heteroskedasticity using ", cluster(id)". Does this change the conclusion reached in part (c)?

   Answer: The cluster-robust t statistic is 2.62. The two-sided p-value is still less than .01, and so we reject the RE estimator at the 1% significance level. The statistical evidence says we should use FE (with the cluster-robust standard error to properly account for heteroskedasticity and serial correlation in u_{it}).