Warm Up

1. Wooldridge (2015), 13.3

> **Solution:** We do not have repeated observations on the *same* cross-sectional units in each time period, and so it makes no sense to look for pairs to difference. For example, in Example 13.1, it is very unlikely that the same woman appears in more than one year, as new random samples are obtained in each year. In Example 13.3, some houses may appear in the sample for both 1978 and 1981, but the overlap is usually too small to do a true panel data analysis.

Exercises

2. *Wooldridge*, 13.6

> **Solution: Part (i):** Let FL be a binary variable equal to one if a person lives in Florida, and zero otherwise. Let $y90$ be a year dummy variable for 1990. Then, from equation (13.10), we have the linear probability model
>
> $$arrest = \beta_0 + \delta_0 y90 + \beta_1 FL + \delta_1 y90 \cdot FL + u \tag{1}$$
>
> The effect of the law is measured by δ_1, which is the change in the probability of drunk driving arrest due to the new law in Florida. Including $y90$ allows for aggregate trends in drunk driving arrests that would affect both states; including FL allows for systematic differences between Florida and Georgia in either drunk driving behavior or law enforcement.
>
> **Part (ii):** It could be that the populations of drivers in the two states change in different ways over time. For example, age, race, or gender distributions may have changed. The levels of education across the two states may have changed. As these factors might affect whether someone is arrested for drunk driving, it could be important to control for them. At a minimum, there is a possibility of obtaining a more precise estimator of ÎŽ1by reducing the error variance. Essentially, any explanatory variable that affects *arrest* can be used for this purpose. (See Section 6.3 for discussion.)
>
> **Part (iii):** The interpretation of the coefficients will differ because they represent averages across counties in a given state rather than state level averages. However, individual level data will allow to control for individual level variation more

effectively, which potentially may reduce the standard errors. A first difference method can be used because the dependent variable is in fraction and the same set of counties in both years are observed at two different time periods.

3. *Wooldridge*, 14.7. (Hint: read the last three paragraphs of Section 14.4 in the textbook)

Solution: Here, a random sample of 29,501 is obtained from the population of individuals in the United States and then grouped the individuals at the 50 states plus the District of Columbia and at the 610 geographic regions—and then treated the data as a cluster sample. The clusters are defined after the random sample is obtained. But this would be wrong and hence clustering at the *puma* level and the *state* level are a little bigger than the heteroscedasticity-robust standard error. In a true cluster sample, the clusters are first drawn from a population of clusters, and then individuals are drawn from the clusters. Therefore, for computing a confidence interval, heteroscedasticity-robust standard error is the most reliable.

Computer Exercises You should use `R` to complete these exercises. Any data set referred to in the question should be available in the `wooldridge` package in `R`. You do not need to turn in an R-script for these questions, but you are welcome to do so if you would like to.

4. *Wooldridge*, 13.C5 (Chapter 13, Computer Exercise 5)

Solution: Part (i): Using pooled OLS we obtain

$$\widehat{\log(rent)} = \underset{(.535)}{-.569} + \underset{(.035)}{.262}\,d90 + \underset{(.023)}{.041}\log(pop) + \underset{(.053)}{.571}\log(avginc) + \underset{(.0010)}{.0050}\,pctstu$$
$$n = 128, R^2 = .861$$

The positive and very significant coefficient on $d90$ 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 five shows that, at least based on the usual analysis, *pctstu* is very statistically significant.

Part (ii): 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.

Part (iii): The equation estimated in differences is

$$\widehat{\Delta \log(rent)} = \underset{(.037)}{.386} + \underset{(.088)}{.072}\Delta \log(pop) + \underset{(.066)}{.310}\Delta \log(avginc) + \underset{(.0041)}{.0112}\Delta pctstu$$
$$n = 64, R^2 = .322$$

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 $\Delta pctstu$.

Part (iv): The heteroskedasticity-robust standard error on $\Delta pctstu$ is about .0028, which is actually much smaller than the usual OLS standard error. This only makes *pctstu* even more significant (robust t statistic ≈ 4). Note that serial correlation is no longer an issue because we have no time component in the first-differenced equation.

5. *Wooldridge*, 14.C14. For part (ii), estimate by both fixed effects and random effects.

Solution: Part (i): 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. The instructor used the `group_by()` and `summarize()` commands in R to create the *concenbar* values.

Part (ii): 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.

Part (iii): Nothing happens to the estimated coefficient on *concen*; it is still the fixed effects estimate. However, the estimated coefficient on *concenbar* goes from about .214 in part (ii) to about -.709, which is a big change.

Part (iv): The usual RE t statistic on *concenbar* in part (ii) 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.

Part (v): 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}).

Cool Down

6. *Wooldridge*, 13.5.

Solution: No, we cannot include age as an explanatory variable in the original model. Each person in the panel data set is exactly two years older on January 31, 1992 than on January 31, 1990. This means that $\Delta age_i = 2$ for all i. But the equation we would estimate is of the form

$$\Delta saving_i = \delta_0 + \beta_1 \Delta age_i + \cdots , \tag{2}$$

where δ_0 is the coefficient on the year dummy for 1992 in the original model. As we know, when we have an intercept in the model, we cannot include an explanatory variable that is constant across i; this violates Assumption MLR.3. Intuitively, since age changes by the same amount for everyone, we cannot distinguish the effect of age from the aggregate time effect.

References

Wooldridge, Jeffrey M. 2015. *Introductory Econometrics: A Modern Approach*. Cengage Learning, 6 ed.