**Exercise 9**

Use the data in RENTAL.dta for this exercise. The data on rental prices and other variables for college towns are for the years 1980 and 1990. The idea is to see whether a stronger presence of students affects rental rates. The unobserved effects model is

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

(i) Estimate the equation by pooled OLS and report the results in standard form. What do you make of the estimate on the 1990 dummy variable? What do you get for *pctstu*?

**ols lrent const y90 lpop lavginc pctstu**

****

(ii) Are the standard errors you report in part (i) valid? Explain.

**The standard errors from part (i) are not valid, unless we think *ai* does not really appear in**

**the equation. If *ai* 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.**

(iii) Now, difference the equation and estimate by OLS. Compare your estimate of with that from part (i). Does the relative size of the student population appear to affect rental prices?

**diff lrent**

**diff lpop**

**diff lavginc**

**diff pctstu**

**ols d\_lrent const d\_lpop d\_lavginc d\_pctstu**



(iv) Estimate the model by fixed effects to verify that you get identical estimates and

standard errors to those in part (iii).

**panel lrent const y90 lpop lavginc pctstu --fixed-effects**



**The coefficient on *y90t* is identical to the intercept from the first difference estimation, and the slope coefficients and standard errors are identical to first differencing. We do not report an *R*-squared because none is comparable to the *R*-squared obtained from first differencing. The constant term can be ignored because some packages display it. it is usually the average of the estimated intercepts for the cross-sectional units, and it is not especially informative.**

1. Suppose that, for one semester, you can collect the following data on a random sample

of college juniors and seniors for each class taken: a standardized final exam score, percentage of lectures attended, a dummy variable indicating whether the class is within the student’s major, cumulative grade point average prior to the start of the semester, and SAT score

1. Why would you classify this data set as a cluster sample? Roughly, how many observations would you expect for the typical student?

**For each student we have several measures of performance, typically three or four, the number of classes taken by a student that have final exams. When we specify an equation for each standardized final exam score, the errors in the different equations for the same student are certain to be correlated: students who have more (unobserved) ability tend to do better on all tests.**

1. Write a model that explains final exam performance in terms of attendance and the other characteristics. Use s to subscript student and c to subscript class. Which variables do not change within a student?

**An unobserved effects model is**

****

**where *as* is the unobserved student effect. Because SAT score and cumulative GPA depend only on the student, and not on the particular class he/she is taking, these do not have a *c* subscript. The attendance rates do generally vary across class, as does the indicator for whether a class is in the student’s major. The term denotes different intercepts for different classes. Unlike with a panel data set, where time is the natural ordering of the data within each cross-sectional unit, and the aggregate time effects apply to all units, intercepts for the different classes may not be needed. If all students took the same set of classes then this is similar to a panel data set, and we would want to put in different class intercepts. But with students taking different courses, the class we label as “1” for student A need have nothing to do with class “1” for student B. Thus, the different class intercepts based on arbitrarily ordering the classes for each student probably are not needed. We can replace with , an intercept constant across classes.**

1. If you pool all of the data and use OLS, what are you assuming about unobserved student characteristics that affect performance and attendance rate? What roles do SAT score and prior GPA play in this regard?

**Maintaining the assumption that the idiosyncratic error, *usc*, is uncorrelated with all**

**explanatory variables, we need the unobserved student heterogeneity, *as*, to be uncorrelated with *atndrtesc*. The inclusion of SAT score and cumulative GPA should help in this regard, as *as*, is the part of ability that is not captured by *SATs* and *cumGPAs*. In other words, controlling for *SATs* and *cumGPAs* could be enough to obtain the ceteris paribus effect of class attendance.**

1. If you think SAT score and prior GPA do not adequately capture student ability, how would you estimate the effect of attendance on final exam performance?

**If *SATs* and *cumGPAs* are not sufficient controls for student ability and motivation, *as* is**

**correlated with *atndrtesc*, and this would cause pooled OLS to be biased and inconsistent. We could use fixed effects instead. Within each student we compute the demeaned data, where, for each student, the means are computed across classes. The variables *SATs* and *cumGPAs* drop out of the analysis.**