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

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

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

(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 $\widehat{\beta_3}$ *pctstu*?

ols lrent const y90 lpop lavginc pctstu

$$\widehat{\log(rent)} = -.569 + .262\, d90 + .041 \log(pop) + .571 \log(avginc) + .0050\, pctstu$$
$$(.535) \quad (.035) \qquad (.023) \qquad (.053) \qquad (.0010)$$

$$n = 128, \quad R^2 = .861.$$

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.

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

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.

(iii) Now, difference the equation and estimate by OLS. Compare your estimate of β_{pctsu} 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

```
Model 8: Pooled OLS, using 64 observations
Included 64 cross-sectional units
Time-series length = 1
Dependent variable: d_lrent

              coefficient   std. error   t-ratio    p-value
  ---------------------------------------------------------
  const         0.385521     0.0368245    10.47     3.66e-015 ***
  d_lpop        0.0722458    0.0883435     0.8178   0.4167
  d_lavginc     0.309960     0.0664771     4.663    1.79e-05  ***
  d_pctstu      0.0112033    0.00413194    2.711    0.0087    ***

Mean dependent var   0.559677   S.D. dependent var   0.106838
Sum squared resid    0.487362   S.E. of regression   0.090126
R-squared            0.322262   Adjusted R-squared   0.288375
F(3, 60)             9.509917   P-value(F)           0.000031
Log-likelihood       65.27210   Akaike criterion    -122.5442
Schwarz criterion   -113.9087   Hannan-Quinn        -119.1422

Excluding the constant, p-value was highest for variable 25 (d_lpop)
```

(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

```
Model 9: Fixed-effects, using 128 observations
Included 64 cross-sectional units
Time-series length = 2
Dependent variable: lrent

             coefficient   std. error   t-ratio    p-value
  ---------------------------------------------------------
  const        1.40938      1.16725       1.207    0.2320
  y90          0.385521     0.0368245    10.47     3.66e-015 ***
  lpop         0.0722458    0.0883435     0.8178   0.4167
  lavginc      0.309960     0.0664771     4.663    1.79e-05  ***
  pctstu       0.0112033    0.00413194    2.711    0.0087    ***

Mean dependent var   5.746195   S.D. dependent var   0.332707
Sum squared resid    0.243681   S.E. of regression   0.063729
LSDV R-squared       0.982666   Within R-squared     0.976531
LSDV F(67, 60)       50.76776   P-value(F)           5.29e-36
Log-likelihood       219.2670   Akaike criterion    -302.5341
Schwarz criterion   -108.5960   Hannan-Quinn        -223.7361
rho                 -1.000000   Durbin-Watson        1.976360
```

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. 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.

(2) 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

(i) 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.

(ii) 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

$$score_{sc} = \theta_c + \beta_1 atndrte_{sc} + \beta_2 major_{sc} + \beta_3 SAT_s + \beta_4 cumGPA_s + a_s + u_{sc},$$

where a_s 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 θ_c 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 θ_c with β_0, an intercept constant across classes.

(iii) 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, u_{sc}, is uncorrelated with all explanatory variables, we need the unobserved student heterogeneity, a_s, to be uncorrelated with $atndrte_{sc}$. The inclusion of SAT score and cumulative GPA should help in this regard, as a_s, is the part of ability that is not captured by SAT_s and $cumGPA_s$. In other words, controlling for SAT_s and $cumGPA_s$ could be enough to obtain the ceteris paribus effect of class attendance.

(iv) 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 SAT_s and $cumGPA_s$ are not sufficient controls for student ability and motivation, a_s is correlated with $atndrte_{sc}$, 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 SAT_s and $cumGPA_s$ drop out of the analysis.