Joel Cabrera

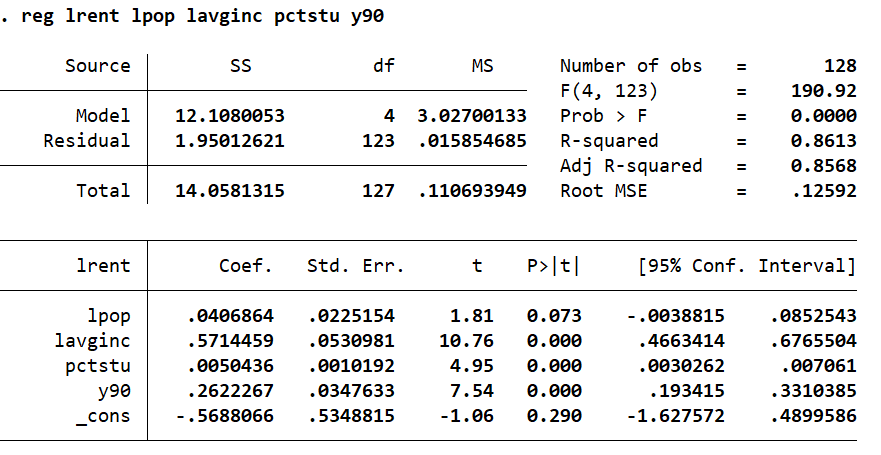
Advanced Cross-Sectional and Panel Econometrics (01:220:401)

Professor Piehl

30 October 2019

**Chapter 14 – Homework Problems & Answers**

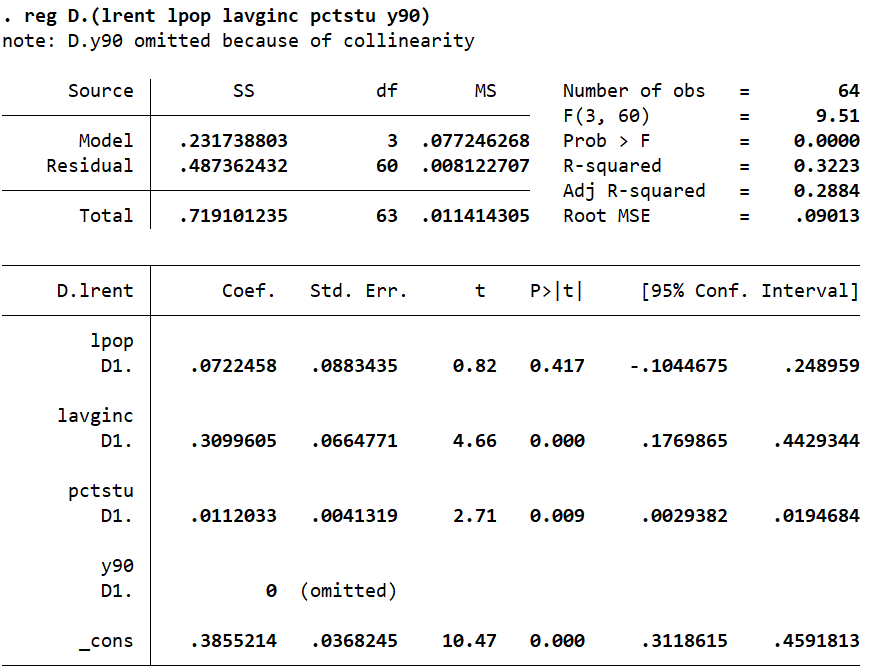
Computer Exercises

C1i. Estimating the equation using pooled OLS in Stata, we obtain the following output:  


Interpreting the coefficient on the 1990 dummy variable: 1990, relative to 1980, is associated with a 26% increase in rental rates, controlling for all other independent variables. The estimated beta coefficient on student population (pctstu) is 0.0050436.

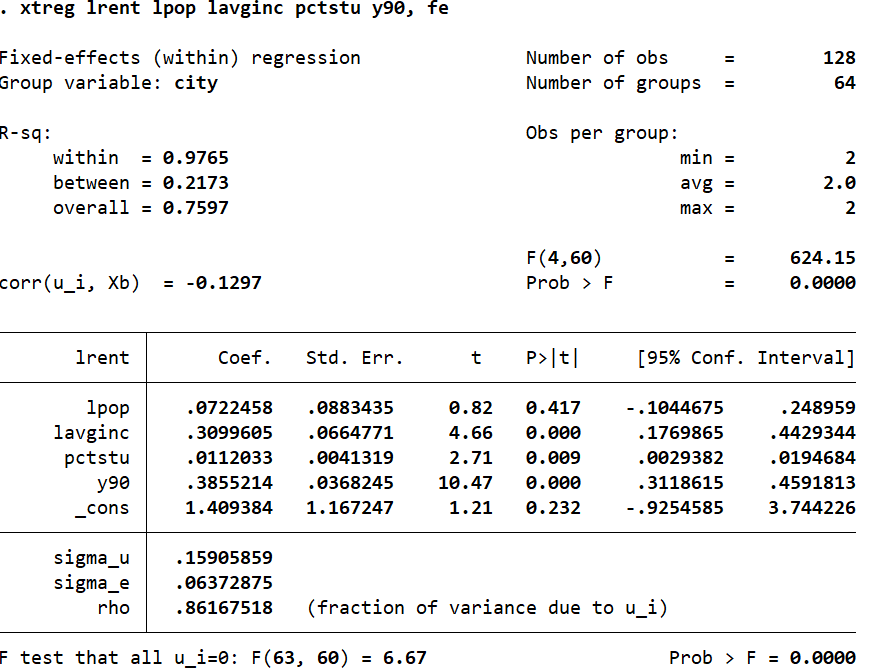
C1ii. The standard errors reported in the previous problem are not valid. This is because the pooled OLS computation does not recognize the data as a panel one, resulting in inflated standard errors for the beta coefficients.

C1iii. Performing the first difference and estimating by OLS, we receive the following Stata output:



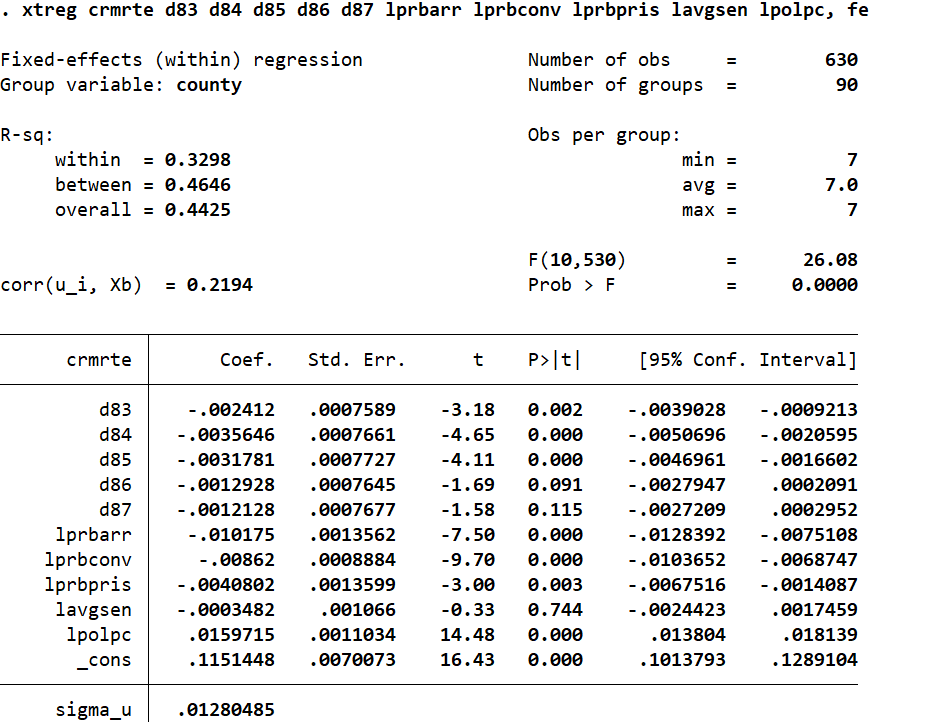
Comparing the beta estimate between pctstu here and that from C1i., it seems that the relative size of the student population does appear to affect rental prices; it seems to have a more than twice as large effect on lrent now (0.0112, as opposed to 0.005).

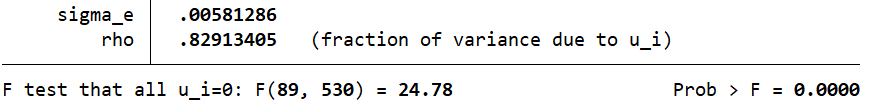
C1iv. Estimating the model by fixed effects, we receive the following Stata output:



As we can see from the output pictured above, we see that we get the same beta estimates and standard errors as those from C1iii.

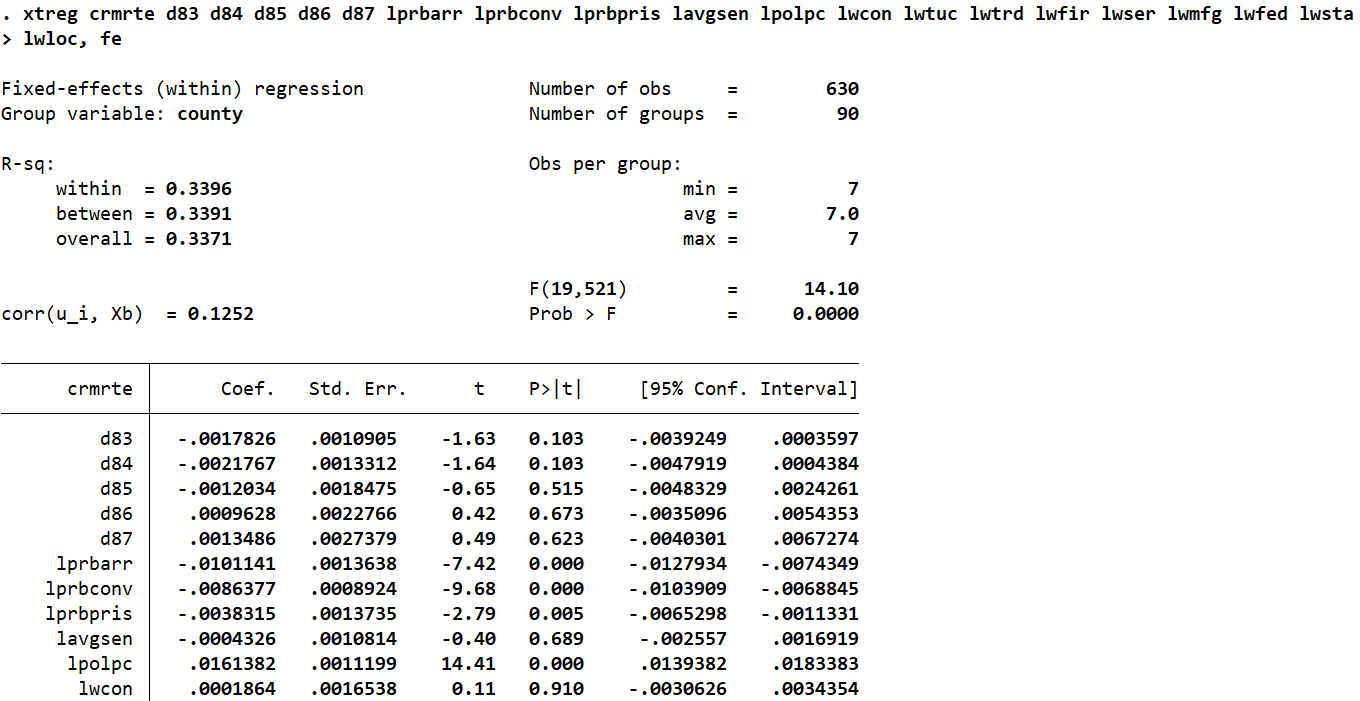
C2i. Re-estimating the unobserved effects model for crime in Example 13.9 using fixed effects instead of differencing, we receive the following Stata output:

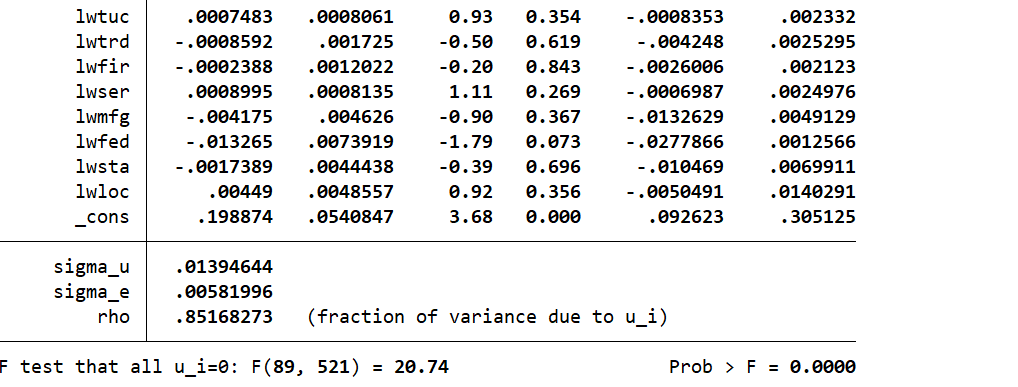




Comparing the estimated beta coefficients pictured above to those in Example 13.9, we can see that there are notable sign changes for the estimated coefficients on d86 and d87. These two independent variables are now negative instead of positive. However, for all other independent variables, they all seem to have increased (but may still seem negative or positive). As for statistical significance, it seems that d83, d84, d85, lprbarr, lprbconv, lprbpris, lpolpc, and the constant are statistically significant, as their corresponding p-values would suggest. Furthermore, since the F-statistic is 0.0000, and is < 0.05, we can say that the regression model is statistically significant.

C2ii. Adding the logs of each wage variable in the dataset and estimating the model using fixed effects, we receive the following Stata output:



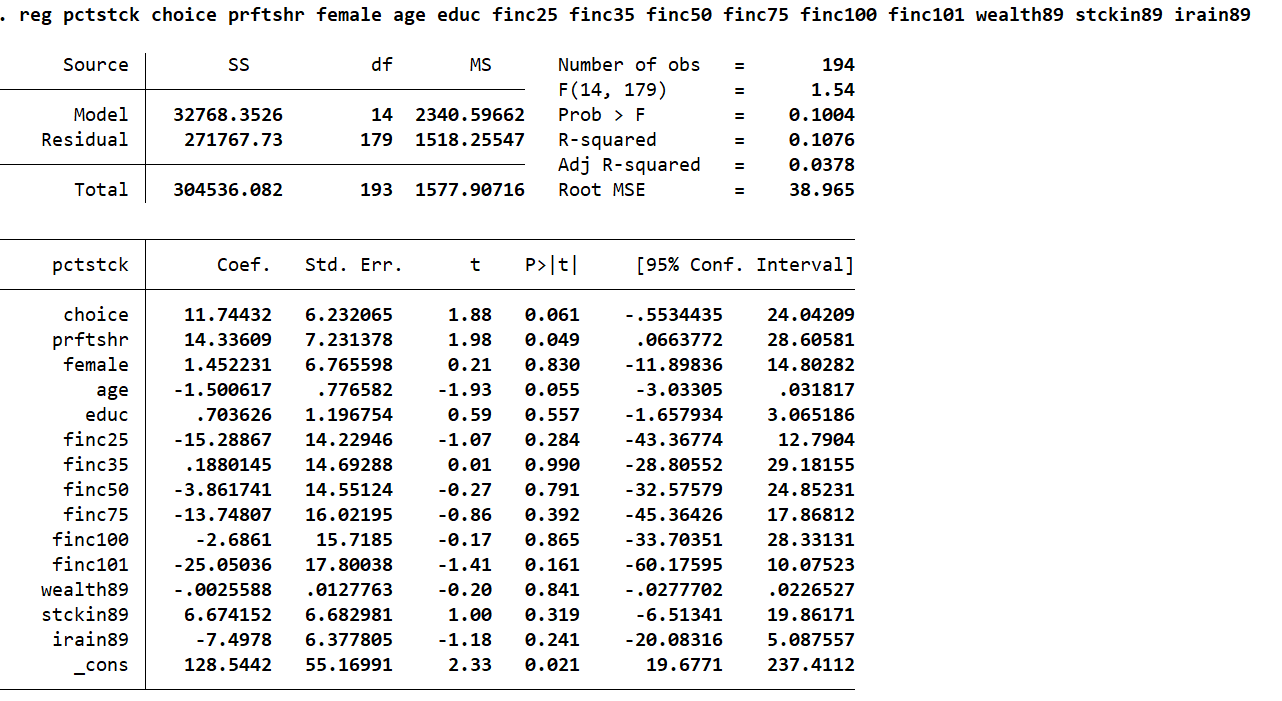


By including these wage variables in the fixed effects model, we can see that not only did the coefficients on d86 and d87 become positive, but also the coefficients on the criminal justice variables, such as lprbarr and lprbcov, have changed similarly. Thus, we can say that the log wage variables have little to no effect on the coefficients on the criminal justice variables.

Add the logs of each wage variable in the data set and estimate the model by fixed effects. How does including these variables affect the coefficients on the criminal justice variables in part (i)?

C2iii. Some of log wage variables in C2ii. have an expected negative sign (e.g. lwfed), whereas some have an expected positive sign (e.g. lwtuc). However, since the model’s F-statistic, 0.0000, is less than 0.05, we can say that they are jointly significant.

C9i. Using OLS to estimate the model, we receive the following Stata output:

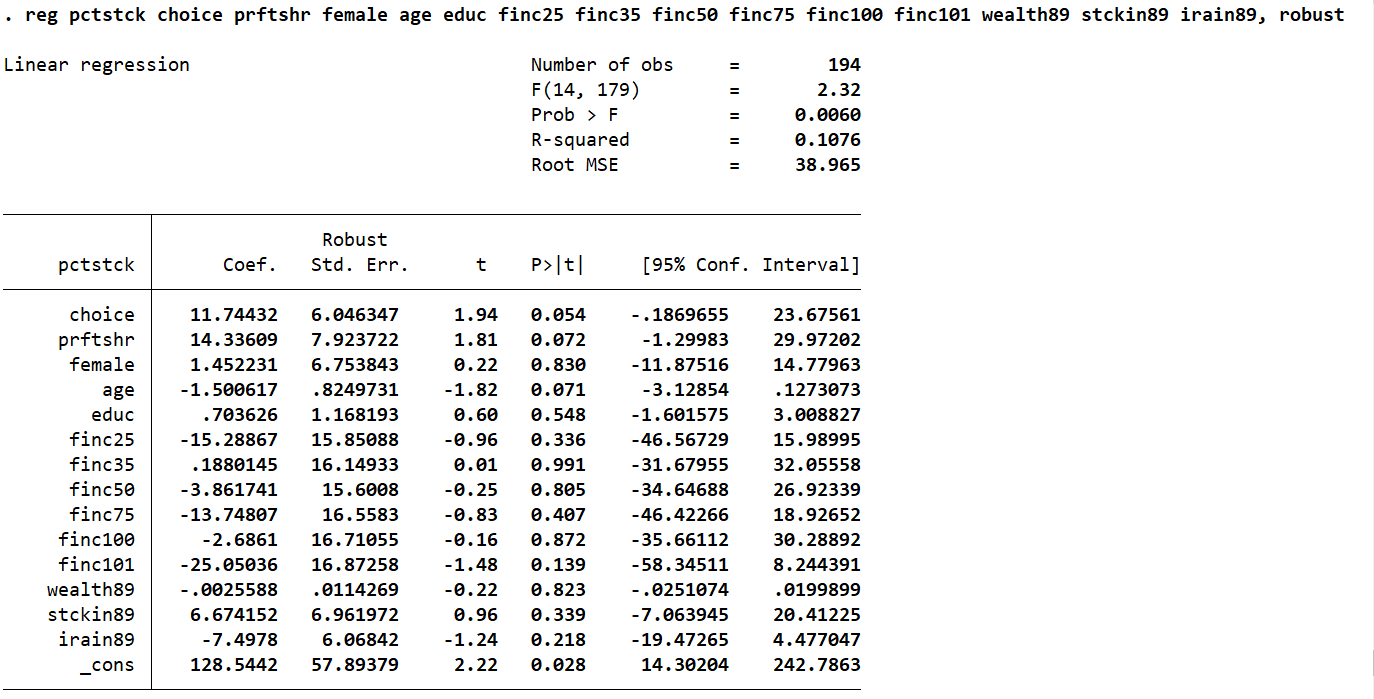


Interpreting the estimated coefficient on choice: A worker who has a choice in how to allocate pension funds among different investments is associated with a 11.74 increase in pctstck (Q: What is the variable label of the outcome variable? Could not find out its meaning in Stata), controlling for all other independent variables. Since its p-value is 0.061, and it is not < 0.05, we can say that the estimated coefficient on choice is not statistically significant.

C9ii. The income, wealth, stock holding, and IRA holding control variables are important. One way of approaching this problem is by examining the p-values of the estimated coefficients on said variables. As we can see in the output above, they are all > 0.05, which means that such coefficients are not statistically significant, leading us to reject the null hypothesis for each one (e.g. there is no relationship between IRA and pctstck). Another way of approaching this is by examining the F-statistic. Since it is > 0.05, the independent variables in the model are not jointly significant.

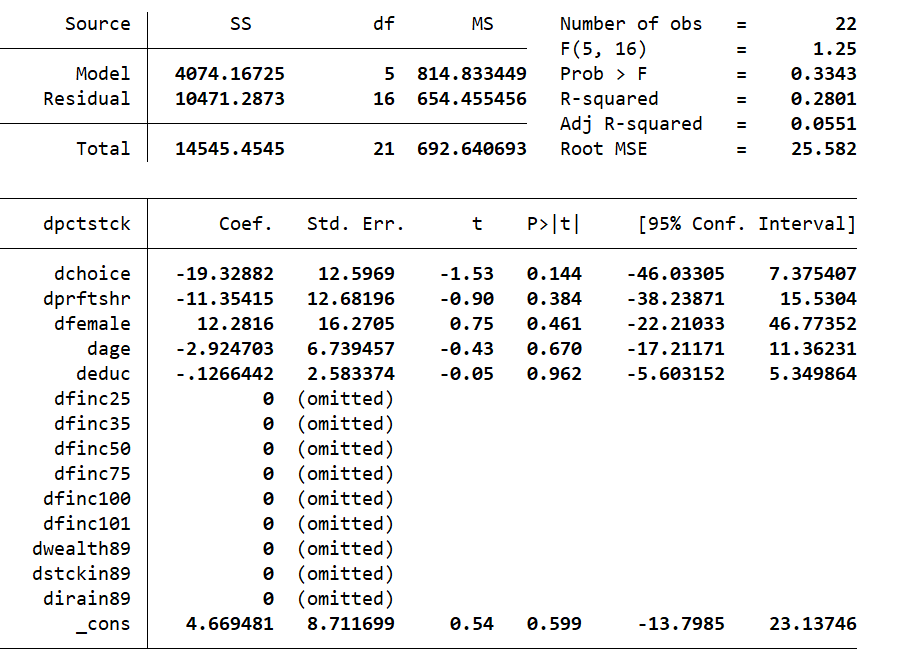
C9iii. By typing “tab id” in Stata, we can see that there are 194 observations for id, which stands for “family identifier”. Thus, there are 194 families, in total, in the dataset. However, there are 171 distinctive families in the dataset.

C9iv. Re-estimating the model in order to receive robust standard errors, we receive the following Stata output:



Comparing these robust standards errors to their standard counterparts in C9i., we can see that they differ from each other very slightly – with approximately 1-2-unit gaps. Given the small differences in these standard errors, I am not too surprised.

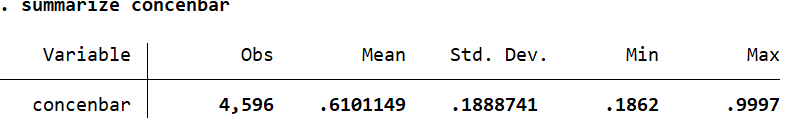
C9v. Estimate the equation by differencing across only the spouses within a family, we receive the follow Stata output:



The explanatory variables asked about in C9ii. are dropped in the first-differenced estimation because they are subtracted from each other (e.g. finc25 – finc25, wealth89 -wealth89, etc.), resulting in their coefficient values to be 0.

C9vi. The remaining variables in C9v. are not statistically significant, as their respective p-values are < 0.05. Given that the equation is estimated by a first difference of families with spouses, this is not surprising.

C14i. Computing the time averages of the variable concen and calling them concenbar, we receive the following Stata output:



There are approximately 1149 time averages (as typing “summarize id” would suggest). The smallest and highest values are 0.1862 and 0.9997, respectively.

C14ii. (Skip, as per Professor Piehl’s in-class instructions, since random effects is not covered).

C14iii. (Skip, as per Professor Piehl’s in-class instructions, since random effects is not covered).

C14iv. (Skip, as well? Also deals with random effects).

C14v. (skip as well? again, also deals with re; based on C14iv.)

Standard Problems

4i. There are many possible measures of athletic success that can be included in a model that estimates the relationship between collegiate athletic performance and applicants. Examples of these measures include but are not limited to: number of matches won across games, such as basketball, baseball, and soccer, over a number of years; dummy variable for an athlete being male or female; and age required for athletes to participate in certain games, such as basketball, baseball, and soccer, over a number of years. Some timing issues that could occur with collecting the data necessary for estimating the relationship between collegiate athletic performance and applicants are: you need to consider when games tend to occur the most, as more games in a year could be associated with higher wins (and thus, skew results), and you also need to consider when do athletes – male or female – play the most, as some may have other responsibilities to attend to (e.g. do another job, attend college classes, etc.). Also, you have to consider the missing data between 1985-1990, 1990-1995, and 1995-2000, as it is only collected in 5-year intervals.

4ii. Other factors, or independent variables, that I might control for in such a relationship include but are not limited to: the amount of money spent on collegiate equipment (e.g. having bad uniforms may be associated with bad performance, and thus, less applications), the name of the college that athletes play for (e.g. athletes playing for Harvard may be associated with more applications), and the amount of male and female athletes that play (e.g. lots of male and little to no female athletes may be associated with high performance, and thus, more applications).

4iii. Using a fixed effects model would be best to estimate the effects of athletic success on the percentage change in applications. This is because using such a method would allow me to account for unobserved effects that vary across time and entities, despite it only having a limited number of independent variables. The model, then, based on my answers to 4i. and 4ii., would look something like:

*Applicationsit* = B0 δ *y90* + δ *y95* + B1 matchesit + B2 *money*it + B3 *name*it + B4 *female*it+ … + σi + ϵit

5i. The data can be classified as a cluster sample because the sample is drawn from college juniors and seniors for each class taken. They are each grouped, or clustered, by the classes that they have taken. Assuming standardized exam score is used as an outcome variable (which it is, indeed, as shown in the next problem), we can expect four observations for the typical junior/senior, as there are four independent variables based on the data collected about them, which are: SAT score, cumulate GPA, percentage of lectures attended, and whether a class taken is within a student’s major.

5ii. We develop a model that estimates the effects of one’s SAT score, one’s cumulate GPA, percentage of lectures that one has attended, and whether a class taken is within one’s major on one’s final exam performance, which is written as follows:

*finalexam*sc = B0 + B1 *SATs* + B2 *GPAs* + B3 %*lectures*sc + B4 *classmajor*sc + σi + ϵit

The variable that does not change within a subject is B0. While usually the constant is different for across entities, this particular one stays the same, despite students taking different classes (some may take the same classes together, but it is uncommon for juniors and seniors to be together). The classes that these students take, after all, may be correlated with each other.

5iii. If I pool all of the data and use OLS, I am assuming that the unobserved student characteristics that affect performance and attendance rate, ϵit, is uncorrelated with the independent variables in any past, current, and future time periods. Because of this assumption, we can also assume that σi is uncorrelated with said independent variables, as well. SAT score and prior cumulative GPA can help in this regard by adding them to the regression model, which would then eliminate the unobserved effects of said variables initially captured by σi (and thus, allow for accurate effects of said variables on final exam score).

5iv. If SAT score and prior cumulative GPA are not enough to capture student ability, then σi would be correlated with attendance (*%lectures*sc). This would then lead to the estimated coefficients on the independent variables created by the OLS regression model to be biased. Due to this, and that the data is a cross-section, we can re-estimate the model using fixed effects instead. This will not only the biasedness in the pooled OLS model, but also remove the SAT score and GPA variables from the model, since they would still not be enough to estimate the effect of lecture attendance on final exam performance.