Problem Set 7

If you would like a small amount of feedback on your work, questions 1 and/or 2 can be handed in to your class teacher. You can do this through blackboard (under the heading 'Exercise Hand-in'). No marks will be given for this and this will have no impact on your final exam grade. The deadline to hand the work in is Monday 26th 09:00am.

1. In 1985, neither Florida nor Georgia had laws banning open alcohol containers in vehicle passenger compartments. By 1990, Florida had passed such a law, but Georgia had not. (You have no variables pertaining to the law; only the knowledge of what is given above).

   (i) Suppose you can collect a random sample of the population in both states, for 1985 and 1990. Let arrest be a binary variable equal to 1 if a person was arrested for drunk driving during the year. Without controlling for any other factors, write down a linear probability model that allows you to test whether the open container law reduced the probability of being arrested for drunk driving. Which coefficient in your model measures the effect of the law?

   (ii) Why might you want to control for other factors in the model? What might some of these factors be?

   (iii) Now, suppose that you can only collect data for 1985 and 1990 at the county level for the two states. The dependent variable would be the fraction of licensed drivers arrested for drunk driving during the year. How does this data structure differ from the individual-level data described in part (i)? What econometric method would you use?

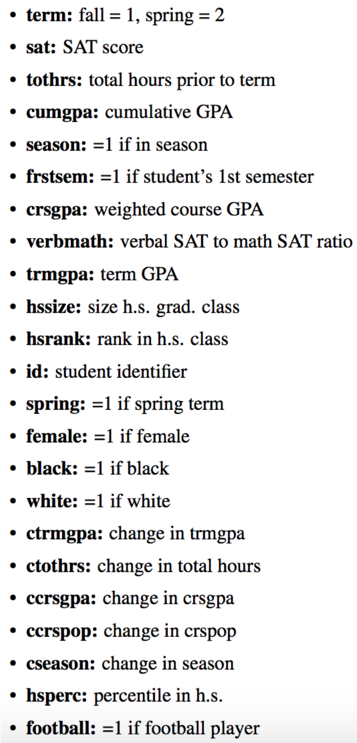
2. Suppose that, for one semester, you can collect the following data on a random sample of college juniors and seniors for each class taken: a standardized final exam score, percentage of lectures attended, a dummy variable indicating whether the class is within the student's major, cumulative grade point average prior to the start of the semester, and SAT score.

   (i) 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? i.e. which variables require only an s subscript? If you estimate this using a fixed effects regression, which coefficients will you not be able to identify?

(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?

(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?

3. Use 'GPA3' for this exercise. The data set is for 366 student-athletes from a large university for fall and spring semesters. (A similar analysis is in Maloney and McCormick (1993), but here we use a true panel data set.) Because you have two terms of data for each student, an unobserved effects model (fixed effects or panel data model) is appropriate. The primary question of interest is this: Do athletes perform more poorly in school during the semester their sport is in season? A description of each of the variables is given below:



- **term:** fall = 1, spring = 2
- **sat:** SAT score
- **tothrs:** total hours prior to term
- **cumgpa:** cumulative GPA
- **season:** =1 if in season
- **frstsem:** =1 if student's 1st semester
- **crsgpa:** weighted course GPA
- **verbmath:** verbal SAT to math SAT ratio
- **trmgpa:** term GPA
- **hssize:** size h.s. grad. class
- **hsrank:** rank in h.s. class
- **id:** student identifier
- **spring:** =1 if spring term
- **female:** =1 if female
- **black:** =1 if black
- **white:** =1 if white
- **ctrmgpa:** change in trmgpa
- **ctothrs:** change in total hours
- **ccrsgpa:** change in crsgpa
- **ccrspop:** change in crspop
- **cseason:** change in season
- **hsperc:** percentile in h.s.
- **football:** =1 if football player

(i) Use pooled OLS to estimate a model with term GPA (*trmgpa*) as the dependent variable. The explanatory variables are *spring*, *sat*, *hsperc*,

$female$, $black$, $white$, $frstsem$, $tothrs$, $crsgpa$, and $season$. Interpret the coefficient on $season$. Is it statistically significant?

(ii) Most of the athletes who play their sport only in the fall are football players. Suppose the academic ability of football players is systematically lower than that of other athletes. If ability is not adequately captured by SAT score and high school percentile, explain why the pooled OLS estimators will be biased.

(iii) In the lectures, we subtracted the group means in order to remove the individual (or time) effects. An alternative way to remove these effects is to difference the data over time, that is, instead of using $y_{it} - \bar{y}_i$, we can use $y_{it} - y_{i(t-1)}$. Use the data differenced across the two terms (they have the same variable names but with an additional 'c' at the start.). Which variables will drop out? Does it make sense to include *spring*? What is the effect of being 'in your sporting season' now?

(iv) Can you think of one or more potentially important, time-varying variables that have been omitted from the analysis?

4. Use the state-level data on murder rates and executions in 'murder' for the following exercise.

(i) Consider the panel data model

$$mrdrte_{it} = \lambda_t + \mu_i + \beta_1 exec_{it} + \beta_2 unem_{it} + u_{it},$$

where λ_t simply denotes time effects and μ_i is the unobserved state effect. 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.

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

(iii) Now, using 1990 and 1993, estimate the equation by fixed effects. Include both time and individual fixed effects. (You can decide yourself whether to include the time dummy manually yourself or using the two-way model). Is there evidence of a deterrent effect? How strong is it?

(iv) 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?

(vi) Estimate the equation using two-way fixed effects, dropping Texas from the analysis. Now, what do you find? What is going on?

(vii) 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. What can we conclude about the robustness of our previous results?

5. Use the panel data in 'crime4' for this question.

   Run a regression of the change in the log of crime rate per capita, *clcrmrte*, on the lag of the change in the log of police per capita. Include fixed effects for the year and the county. Have the dynamics in this model caused an endogeneity issue? Explain why we have used the lag of the change in police per capita and the change in crime rate, rather than just the current levels. Compare the coefficient from this regression with the coefficient from a regression of *crmrte* on *polpc*.