Econometrics: Problem Set 4

1. Traffic crashes are the leading cause of death for Americans between the ages of 5 and 32. Through various spending policies, the federal government has encouraged states to institute mandatory seat belt laws to reduce the number of fatalities and serious injuries. In this exercise you will investigate how effective these laws are in increasing seat belt use and reducing fatalities. The data file 'seatbelt_data.csv' contains a panel of data from 50 U.S. states plus the District of Columbia for the years 1990 through 1997.

a. Estimate the effect of seat belt use on fatalities by regressing *FatalityRate* on *sb_useage*, *speed65*, *speed70*, *ba08*, *drinkage21*, *ln(income)*, and *age*. Does the estimated regression suggest that increased seat belt use reduces fatalities?
b. Do the results change when you add state fixed effects (you will need to load library(plm))? Provide an intuitive explanation for why the results changed.

Time and Entity Fixed Effects

We did not go over it in class, but we could include time and entity fixed effects:

$Y_{it} = \beta_1 X_{it} + \alpha_i + \gamma_t + u_{it}$

For example maybe driving cultures differ across states - this might lead to OVB - so we included state fixed effects in (b). Well also there are national laws and other ways the national government pushes for changes over time. In this sense we might also have omitted variables that are constant across states but vary over time - and if they are correlated to our Xs and our error term then they too can cause OVB. So to fix this we can also add time fixed effects.

This is simple to do in R. In your code calling your regression in (b), add the code: effect = "individual", now rerun. You should get the same answer, because you are telling it the 'fixed' effects are individual fixed effects, but this is the default anyway. Now to do time and entity fixed effects, instead include: effect = "twoways".

c. Do the results change when you add time fixed effects plus state fixed effects?
d. Which regression specification(a), (b), or (c)is most reliable? Explain why you think so.
e. Using the results in (c), discuss the size of the coefficient on sb_useage. Is it large? Small? How many lives would be saved if seat belt use increased 30% in the year 1997?
(*Hint*: Think how you might do this. *Fatalityrate* is death rate per million miles drive, *vmt* is millions of miles driven. So maybe use the millions of mile driven in all States in 1997 plus your estimate plus a 30% increase?)

2. Use the data set 'ins_data.csv' for this problem. It contains data on workers and their health and health insurance status (ie. whether or not they have it).

a. Are self-employed workers less likely to have heath insurance than 'regular' workers? If so is the difference large in a real-world sense? Is it statistically significant?
b. Self-employed people might be fundamentally different from normal workers in some ways. Control for some things that you may think are important and see if the self employed are less likely to have insurance.

c. How does health insurance status vary with age?
d. Is the effect of self employment on insurance status different for older workers than younger workers?
e. Some argue that the self employed are less likely to be insured, but still are just as healthy. Is this right? What if you only look at 'young' workers? What if you only look at 'old' workers? Are there any potential two-way causality problems that might undermine the interval validity of this kind of analysis?