# Problem set 2

**Estimation and Post-Estimation (cross-section)**

The aim of this exercise is to help you practice the estimation and post-estimation commands reviewed in class 2. The first two parts will aim to estimate the returns to education in the US in 1976, first using OLS, then using 2SLS. The third part introduces Limited Dependent Variable models.

The dataset used is the same as last week:
http://personal.lse.ac.uk/lembcke/ecStata/card.zip

## I. OLS and regression diagnostic

1. From last week, you should remember how to open the dataset and the do-file editor. After doing so, estimate by standard OLS the coefficients of the Mincer equation for the individuals in your sample. In addition to education, experience and the square of experience, include controls for race and geography (ie, south and smsa).

2. Interpret your coefficients. Are they significant? Check that you understand all the figures in the output.

3. Check for multicollinearity. You should see strong collinearity between experience and its square. Explain why this is not a concern for our purposes. Check for influential points. Are there many "worrying" outliers? Also check for heteroskedasticity.

## II. Instrumental variables

From the previous exercise, you should be convinced that the Mincer equation is a strong and well specified relationship. However, the coefficient on education estimated by OLS has no causal interpretation (why?). In this section, we will attempt to estimate the returns to education by using "proximity to a 2/4 years college" as an instrument for education (remember question II.2. from last week).

1. Find the 2SLS estimator for the returns to education.
   Note: If you consider education to be endogenous, then you must also instrument for potential experience (defined as age-educ-6, and therefore strongly – negatively – correlated with the endogenous variable) and its square. Age and its square seem to be reasonable additional instruments (why?).

2. Interpret your results. Is the estimate what you expected?
   Also comment on the quality of first stage relationship and the size of the standard errors.

3. Test for endogeneity.

## III. Qualitative Response Model

Let us now investigate the determinants of the choice to go to college.

1. Define a dummy variable taking value 1 for individual who have been to college at least one year (corresponding to strictly more than 12 years of education in the US).

2. Fit a logistic regression model for this new variable, using as regressors the parents' education, geographic variable, race, and proximity to college. Interpret the output. (Be careful when you interpret your coefficients).

3. Check for outliers. Evaluate the quality of the fit.