Intro to Econometrics

Due: Thursday, Mar 28
at the beginning of class

Prepare brief and precise answers to the following questions. You are encouraged to discuss the solutions in groups but should write up the solutions independently.

1. (SW Ex 7.11) A school district undertakes an experiment to estimate the effect of class size on test scores in second-grade classes. The district assigns 50% of its previous year's first grades to small second-grade classes (18 students per classroom) and 50% to regular-size classes (21 students per classroom). Students new to the district are handled differently: 20% are randomly assigned to small classes and 80% to regular-size classes. At the end of the second-grade school year, each student is given a standardized exam. Let Y_i denote the exam score for the ith student, X_{1i} denote a binary variable that equals 1 if the student is assigned to a small class, and X_{2i} denote a binary variable that equals 1 if the student is newly enrolled. Let β_1 denote the causal effect on test scores of reducing class size from regular to small.

   a) Consider the regression $Y_i = \beta_0 + \beta_1 X_{1i} + u_i$. Do you think that $\mathrm{E}[u_i|X_{1i}] = 0$? Is the OLS estimator of β_1 unbiased and consistent? Explain.

   b) Consider the regression $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$. Do you think that $\mathrm{E}[u_i|X_{1i}, X_{2i}]$ depends on X_1? Is the OLS estimator of β_1 unbiased and consistent? Explain. Do you think that $\mathrm{E}[u_i|X_{1i}, X_{2i}]$ depends on X_2? Will the OLS estimator of β_2 provide an unbiased and consistent estimate of the causal effect of transferring to a new school (that is, being a newly enrolled student)? Explain.

2. In class we showed that the formula for the OLS estimator in multiple regression is:

$$\hat{\beta} = (X'X)^{-1}X'Y\,.$$

   Suppose we have just one regressor, i.e.:

$$y_i = \beta_0 + \beta_1 x_i + u_i\,.$$

   We can rewrite the model as:

$$Y = X\beta + u$$

   where:

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \qquad X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \qquad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \qquad u = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix}$$

and our least squares estimator of β is:

$$\begin{aligned}\hat{\beta} &= (X'X)^{-1}(X'Y) \\ &= \Big(\frac{1}{n}X'X\Big)^{-1}\Big(\frac{1}{n}X'Y\Big)\end{aligned}$$

where $\frac{1}{n}X'X$ is a 2×2 matrix and $\frac{1}{n}X'Y$ is a 2×1 vector.

a) Write out $\frac{1}{n}X'X$ in terms of 1, $\bar{x}_n$ and $\frac{1}{n}\sum_{i=1}^n x_i^2$

b) Using the formula for the inverse of a 2×2 matrix, calculate

$$\Big(\frac{1}{n}X'X\Big)^{-1}$$

c) Using your answer to b), show that $\hat{\beta}$ recovers the usual formula for simple linear regression, i.e.:

$$\hat{\beta} = \begin{pmatrix} \bar{y}_n - \hat{\beta}_1 \bar{x}_n \\ \\ \frac{\frac{1}{n}\sum_{i=1}^n x_i y_i - \bar{x}_n \bar{y}_n}{\frac{1}{n}\sum_{i=1}^n x_i^2 - \bar{x}_n^2} \end{pmatrix}$$

3. (SW Ex 7.1) For this question, you will need to download the file `birthweight_smoking.csv` from the NYU classes page. The file `birthweight_smoking.R` shows you how to import the data and get started.

   The data are a random sample of 3000 babies born in Pennsylvania in 1989. The data include the baby's birthweight together with various characteristics of the mother, including whether she smoked during pregnancy. The variables are:

   - birthweight = baby's birthweight in grams
   - smoker = 1 if mother smoked during pregnancy, 0 otherwise
   - alcohol = 1 if mother drank alcohol during pregnancy, 0 otherwise
   - unmarried = 1 if mother is unmarried, 0 otherwise
   - nprevist = total number of prenatal visits

   a) Estimate the following three regression models:

   $$\begin{aligned} birthweight_i &= \beta_0 + \beta_1 smoker_i + u_i \\ birthweight_i &= \beta_0 + \beta_1 smoker_i + \beta_2 alcohol_i + \beta_3 nprevist_i + u_i \\ birthweight_i &= \beta_0 + \beta_1 smoker_i + \beta_2 alcohol_i + \beta_3 nprevist_i + \beta_4 unmarried_i + u_i \end{aligned}$$

   and construct 95% confidence intervals for the estimated effect of smoking on birth weight using each of the three regressions.

b) Does the coefficient on smoker in the first and second regressions suffer from omitted variables bias? Explain.

c) Consider the coefficient on *unmarried* in the third regression. A family advocacy group notes that the large coefficient suggests that public policies that encourage marriage will lead, on average, to healthier babies. Do you agree?
Hint: is unmarried a regressor or a control variable? Discuss some of the various factors that *unmarried* might be controlling for and how this affects the interpretation of its coefficient.