17.874, Spring 2004
Problem Set 5

1. Dummy Variables. Dummy or indicator variables in regression are analogous to differences of means. Here you will do some algebraic manipulations and simulations to explore this idea.

a. Consider the regression:

$$y_{\mathrm{i}} = \alpha + \beta X_{\mathrm{i}} + \epsilon_{\mathrm{i}},$$

where X_{i} is a dummy variable. Suppose there are n observations total, and n_1 and n_0 in the groups $X = 1$ and $X = 0$, respectively. Denote the mean of Y when $X = 1$ as $\bar{Y}_1$ and the mean of Y when $X = 0$ as $\bar{Y}_0$. Show that the least squares estimator of b is identical to the difference between the group means. Show that the least squares estimator a is identical to the group mean $\bar{Y}_0$.

b. Show that the variance of b equals the variance of the differences of means.

Now let us introduce another variable, Z. The regression then is:

$$y_{\mathrm{i}} = \alpha + \beta X_{\mathrm{i}} + \gamma Z_{\mathrm{i}} + \epsilon_{\mathrm{i}},$$

Assume ϵ is normal with mean 0 and variance 1. Let $\alpha = .5$, $\beta = .1$ and $\gamma = 1$. In STATA, set the number of observations equal 435. Generate random values of Z from the normal distribution with mean 0 and variance 1. Let $X = 1$ for the first 326 percent of cases and 0 for the remaining cases. (Use the commands **set x = 0; set x = 1 if _n¡327**).

c. Regress y on X and regress y on X and Z. How do your estimates differ? Is there a bias when Z is not included? Which is more efficient?

d. Now, we construct a slightly different dataset where X and Z are correlated. Let $X = -1$ if $Z + u < -.25$ and let $X = +1$ if $Z + u > .25$, where u is a normally distributed random variable with standard deviation .2. (In STATA, use, for instance, **gen X = -1 if z + (.2*invnorm(uniform()) < -.25**). Now, regress y on X and regress y on X and Z. How do your estimates differ from part (c)? Is there a bias when Z is not included? Is there a gain in efficiency?

2. Simulation of Instrumental Variables.

Generate data that fit the following structure. There are 500 observations. Z and X_1 are independently and normally distributed (mean 0, variance 1 is fine). X_2 depends on X_1 and Z as follows:

$$X_2 = .5X_1 - .5Z + u,$$

where u is normal with mean 0 and variance 1. Y depends on X_1 and X_2 plus an error.

a. Regress Y on X_1 and X_2.

b. Regress Y on X_1, X_2, and Z.

c. Regress Y on X_2 ommitting X_1.

d. Regress X_2 on Z. Generate the predicted values, $\hat{X}_2$ (use the command **predict**). Regress Y on $\hat{X}_2$. Report the coefficients and standard errors.

e. The last regression is an implementation of instrumental variables, but gives the wrong standard errors. Implement the right IV estimates in STATA using the command **reg y X2 (Z)**.)

f. Using the formulas from class and the variances and covariances, calculate the correct $V(b_{\text{IV}})$.

g. Comment on the analyses you have done. How does IV work? How does it correct for bias? How much loss of efficiency is there?