Problem Set 2

You can download the data file you need for question 4 here.

What do I submit?

- Your written up answers to exercise questions. If you work on a piece of paper, please scan using some sort of phone software (like Microsoft Lens or Adobe Scan) rather than just taking a picture.
- A do-file that runs your Stata analysis (for question 4).
- A log file that includes the output from running your do-file (for question 4).

Exercises

1. The following table shows, for eight vintages of delicious wine, purchases per buyer (y) and the wine buyer's rating (x) in a given year:

x	3.6	3.3	2.8	2.6	2.7	2.9	2.0	2.6
y	24	21	22	22	18	13	9	6

   (a) Estimate *by hand* the regression of purchases per buyer on the buyer's rating.

   (b) Interpret the slope of the estimated regression line.

   (c) Interpret the intercept of the estimated regression line .

2. Suppose that a random sample of 200 20-year-old men is selected from a population and that these men's height and weight are recorded. A regression of weight (measured in pounds) on height (measured in inches) yields
$\widehat{Weight} = -99.41 + 3.94 Height$
$R^2 = 0.81$; $SER = 10.2$

   (a) What is the predicted weight for someone who is 70 inches tall? 65 inches tall?
   (b) One 20-year-old man has a late growth spurt and grows 1.5 inches over the course of the year. What is the regression's prediction for the increase in his weight?
   (c) Suppose that you want to translate the results of this equation into centimeters and kilograms. What are the regression estimates from this new regression? Give all results, including estimated coefficients, R^2, and SER.
   (d) Interpret the R^2 value. Does it indicate anything about whether these estimates are likely to be biased? Explain.

3. Consider the savings function:

$$sav = \beta_0 + \beta_1 inc + u, u = e\sqrt{inc}$$

where e is a random variable with $E(e) = 0$ and $Var(e) = \sigma_e^2$. Assume that e is independent of inc.

   (a) Show that $E(u|inc) = 0$, so that the key zero conditional mean assumptionis satisfied. [Hint: If e is independent of inc, then $E(e|inc) = E(e)$]

   (b) Show that $Var(u|inc) = \sigma_e^2 inc$, so that the homoskedasticity assumption is violated. In particular, the variance of sav increases with inc. [Hint: $Var(e|inc) = Var(e)$ if inc and e are independent!]

   (c) Why might it be reasonable to assume that the variance of savings increases with family income?

4. The data file collegedistance.dta contains data from a random sample of high school seniors interviewed in 1980 and re-interviewed in 1986[1] Use these data to investigate the relationship between the number of completed years of education for young adults and the distance from each student's high school to the nearest four-year college. (Proximity to college lowers the cost of education, so that students who live closer to a four-year college should, on average, complete more years of higher education.)

   (a) Run a regression of years of completed education (ED) on distance to the nearest college ($Dist$), where $Dist$ is measured in tens of miles. (For example, $Dist = 2$ means that the distance is 20 miles.)[2]. Write the equation you estimated in the form $\widehat{ED} = \beta_0 + \beta_1 Dist$

   (b) How does the average value of years of completed schooling change when colleges are built close to where students go to high school?

   (c) Bob's high school was 20 miles from the nearest college. Predict Bob's years of completed education using the estimated regression. How would the prediction change if Bob lived 10 miles from the nearest college?

   (d) Does distance to college explain a large fraction of the variance in educational attainment across individuals? Explain.

   (e) Provide an example of a factor that might cause this model to violate the zero conditional mean assumption. Explain your reasoning.

   (f) What is the value of the standard error of the regression?[3] What are the units for the standard error (meters, grams, years, dollars, cents, or something else)?

   (g) Is the estimated regression slope coefficient statistically significant at the 5% level? What is the p-value associated with coefficient's t-statistic?

[1] These data were provided by Professor Cecilia Rouse of Princeton University and were used in her paper "Democratization or Diversion? The Effect of Community Colleges on Educational Attainment," Journal of Business and Economic Statistics, April 1995, 12(2): 217–224.

[2] Though we haven't covered regressions in our stata labs, I talk thorugh interpreting regression output in the Chapter 5 Video. You can regress a dependent variable `y` on an independent variable `x` with the command `regress y x`

[3] There are a few ways to find it in Stata's output. The easiest is to note that "root MSE" is the square root of the SER

(h) Construct a 95% confidence interval for the slope coefficient.

(i) Estimate a regression that restricts the sample to men, and calculate a 95% confidence interval for the slope. Do the same, restricting the sample to women. Does it look like the effect of distance on completed years of education is different?[4]

[4] Note that we cannot make claims about whether they are statistically different because the estimates come from two different samples! A hypothesis test here would be awesome, but we need to build a few more skills to do this.