# Problem Set 2

1. To complete this exercise you need software that allows you to generate data from the uniform and normal distributions (you can use R).

   (i) Start by generating 500 observations $x_i$ – the explanatory variable – from the uniform distribution with range $[0, 10]$.

   (ii) Randomly generate 500 errors, $u_i$, from $Normal(0, 36)$ (i.e. mean zero with variance of 36. Note that you need to tell R the standard deviation rather than variance).

   (iii) Now generate the $y_i$ as

   $$y_i = 1 + 2x_i + u_i.$$

   Use the data to run the regression of $y_i$ on $x_i$. What are your estimates of the intercept and slope? Are they equal to the population values in the above equation? Explain.

   (iv) Obtain the OLS residuals, $\hat{u}_i$, and verify that equation $\sum_i \hat{u}_i = 0$ (subject to rounding error).

   (v) Compute $\sum_i \hat{u}_i x_i$. What do you conclude?

   (vi) Repeat parts (i), (ii), and (iii) with a new sample of data. Now what do you obtain for $\hat{\beta}_0$ and $\hat{\beta}_1$? Why are these different from what you obtained in part (iii)?

2. The data in 'wage2' on working men was used to estimate the following equation:

   $$\hat{educ} = 10.4 - 0.09 sibs + 0.13 meduc + 0.21 feduc$$

   where $n = 722$ $R^2 = 0.21$ and where educ is years of schooling, *sibs* is number of siblings, *meduc* is mother's years of schooling, and *feduc* is father's years of schooling.

   (i) Does *sibs* have the expected effect? Explain. Holding *meduc* and *feduc* fixed, by how much does *sibs* have to increase to reduce predicted years of education by one year? (A non-integer answer is acceptable here.)

(ii) Discuss the interpretation of the coefficient on *meduc*.

(iii) Suppose that Man A has no siblings, and his mother and father each have 12 years of education. Man B has no siblings, and his mother and father each have 16 years of education. What is the predicted difference in years of education between B and A?

3. In a study relating college grade point average to time spent in various activities, you distribute a survey to several students. The students are asked how many hours they spend each week in four activities: studying, sleeping, working, and leisure. Any activity is put into one of the four categories, so that for each student, the sum of hours in the four activities must be 168.

   (i) In the model

   $$GPA = \beta_0 + \beta_1 study + \beta_2 sleep + \beta_3 work + \beta_4 leisure + \epsilon,$$

   does it make sense to hold sleep, work, and leisure fixed, while changing study?

   (ii) Explain why this model violates Assumption MLR.3.

   (iii) How could you reformulate the model so that its parameters have a useful interpretation and it satisfies Assumption MLR.3?

4. Which of the following can cause OLS estimators to be biased?
   (i) Heteroskedasticity.
   (ii) Omitting an important variable.
   (iii) A sample correlation coefficient of 0.95 between two independent variables both included in the model.

5. Use the data in 'discrim' to answer this question. These are zip code–level data on prices for various items at fast-food restaurants, along with characteristics of the zip code population, in New Jersey and Pennsylvania. The idea is to see whether fast-food restaurants charge higher prices in areas with a larger concentration of black people.

   (i) Find the average values of *prpblck* and *income* in the sample, along with their standard deviations. What are the units of measurement of *prpblck* and *income*?

   (ii) Consider a model to explain the price of soda, *psoda*, in terms of the proportion of the population that is black and median income:

   $$psoda = \beta_0 + \beta_1 prpblck + \beta_2 income + \epsilon.$$

Estimate this model by OLS and report the results, including the sample size and R-squared. (Do not use scientific notation when reporting the estimates.) Interpret the coefficient on *prpblck*. Do you think it is economically large?

(iii) Compare the estimate from part (ii) with a simple regression estimate from *psoda* on *prpblck*. Is the discrimination effect larger or smaller when you control for income? Explain.

(iv) A model with a constant price elasticity with respect to income may be more appropriate. Report estimates of the model

$$ln(psoda) = \beta_0 + \beta_1 prpblck + \beta_2 ln(income) + \epsilon.$$

If *prpblck* increases by 0.20 (20 percentage points), what is the estimated percentage change in *psoda*?

(v) Now add the variable *prppov* to the regression in part (iv). What happens to $\hat{b}_{prpblck}$? Explain.

(vi) Find the correlation between $log(income)$ and *prppov*. Is it roughly what you expected?

(vii) Evaluate the following statement: "Because $log(income)$ and *prppov* are so highly correlated, they have no business being in the same regression."

6. The Bechdel test is a test (more of a minimum threshold) for whether a film portrays women in a 'normal/real life' way. A film passes the Bechdel test if there are at least 2 named female characters in the film who have a conversation together that doesn't pertain to a male character. (Whether this constitutes a good 'test' or not is definitely up for debate, but we will use it anyway as it highlights OVB nicely!).
   The dataset 'bechdel.csv' can be downloaded from blackboard (use the read.csv() function in R, as is used to load the flights_data.csv data on slide 47 in the multivariate regression slides).

   (i) Run a regression of the log of international grossing (adjusted for inflation in 2013 dollars) on whether the film passes the Bechdel test:

$$log(intgross\_13) = \beta_0 + \beta_1 pass\_test + \epsilon.$$

   What does this indicate about audiences preferences for films?

   (ii) Now include log of the budget of the film (adjusted for inflation in 2013 dollars), in particular take this to be the . How do you interpret the coefficient on the budget? Discuss what happens to the coefficient on *pass_test* in comparison to part (i). In particular, what does this mean for sexism in the film industry?