Problem Set 1

1. Suppose that you are asked to conduct a study to determine whether smaller class sizes lead to improved student performance of fourth graders.

   (i) If you could conduct any experiment you want, what would you do? Be specific.

   (ii) More realistically, suppose you can collect observational data on several thousand fourth graders in a given state. You can obtain the size of their fourth-grade class and a standardized test score taken at the end of fourth grade. Why might you expect a negative correlation between class size and test score?

   (iii) Would a negative correlation necessarily show that smaller class sizes cause better performance? Explain.

2. Let kids denote the number of children ever born to a woman, and let educ denote years of education for the woman. A simple model relating fertility to years of education is

   $$kids = \beta_0 + \beta_1 educ + u,$$

   where u is the unobserved error.

   (i) What kinds of factors are contained in u? Are these likely to be correlated with level of education?

   (ii) Will a simple regression analysis uncover the ceteris paribus effect of education on fertility? Explain.

3. The data set 'bwght' contains data on births to women in the United States. One choice for the dependent variable is infant birth weight in ounces (bwght), and an explanatory variable of interest is the average number of cigarettes the mother smoked per day during pregnancy (cigs). The following simple regression was estimated using data on $n = 1388$ births:

   $$\hat{bwght} = 119.8 - 0.51 cigs.$$

   (i) What is the predicted birth weight when $cigs = 0$? What about when

$cigs = 20$ (one pack per day)? Comment on the difference.

(ii) Does this simple regression necessarily capture a causal relationship between the child's birth weight and the mother's smoking habits? Explain.

(iii) To predict a birth weight of 125 ounces, what would cigs have to be? Comment.

(iv) It turns out that more than 500 births in the sample are above 125 ounces. Given your answer to (iii), does this surprise you? Notice that the proportion of women in the sample who do not smoke while pregnant is about 85%. Does this help explain things?

4. Consider the standard simple regression model $y = \beta_0 + \beta_1 x + u$ under the Gauss-Markov Assumptions SLR.1 through SLR.5. The usual OLS estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased for their respective population parameters. Let $\tilde{\beta}_1$ be the estimator of β_1 obtained by assuming the intercept is zero. (Hint, it is given by $\tilde{\beta}_1 = \frac{\sum_i x_i y_i}{\sum x_i^2}$)

   (i) Find $E\left(\tilde{\beta}_1\right)$ and verify that $\tilde{\beta}_1$ is unbiased for β_1 when the population intercept β_0 is zero. Are there other cases where $\tilde{\beta}_1$ is unbiased?

   (ii) Find the variance of $\tilde{\beta}_1$. (Hint: The variance does not depend on β_0.)

   (iii) Show that $Var\left(\tilde{\beta}_1 | X\right) \leq Var\left(\hat{\beta}_1 | X\right)$. (Hint: By Chauchy's inequality, for any sample of data, $\sum_{i=1}^n x_i^2 \geq \sum_{i=1}^n (x_i - \bar{x})^2$, with equality only when $\bar{x} = 0$.)

   (iv) Comment on the tradeoff between bias and variance when choosing between $\hat{\beta}_1$ and $\tilde{\beta}_1$.

5. Use the data in 'sleep75' from Biddle and Hamermesh (1990) to study whether there is a tradeoff between time spent sleeping (per week) and time spent in paid work. We could use either variable as the dependent variable. For concreteness, estimate the model

$$sleep = \beta_0 + \beta_1 totwrk + u,$$

   where *sleep* is minutes spent sleeping at night per week and *totwrk* is total minutes worked during the week.

   (i) Report your results in equation form along with the number of observations and R^2. What does the intercept in this equation mean?

(ii) If *totwrk* increases by 2 hours, by how much is sleep estimated to fall? Do you think this is a large effect?