**Work Problems Chapter 13**

Suppose I want to know something about the study habits of undergraduate college students. I collect a random sample of 200 students and find that they spend 12 hours per week studying, on average, with a standard deviation of 5 hours. I am curious how their social lives might be associated with their studying behavior, so I ask the students in my sample how many other students at their university they consider "close friends." The sample produces an average of 6 close friends with a standard deviation of 2. Please use this information to answer the following questions. The correlation between these two variables is -.40.

1. Assume that "Hours spent studying" is the *Y* variable and "Close friends" is the *X* variable. Calculate the regression coefficient (i.e., the slope) and wrap words around your results. What, exactly, does this regression coefficient tell you?

   $b = r \times \frac{s_y}{s_x}$ **so b = (-.40)** $\frac{5}{2}$ **→ (-.40)(2.50) = -1.00 So for every increase of one friend there is a corresponding decreases of one hour in study time, on average.**

2. What would the value of the standardized regression coefficient be in this problem? How do you know?

   **The standardized regression coefficient would be -.40. We know this because in a regression model with a single predictor variable, the correlation coefficient is the same as the standardized regression coefficient.**

3. Calculate the intercept and wrap words around your result.

   $a = \bar{Y} - b\bar{X}$**, so 12 - (-1.00)(6) → 12 + 6 = 18. The expected value of *Y* (hours spent studying) when *X* (number of friends) equals zero is 18.**

4. If you know that somebody studied had 10 close friends, how many hours per week would you expect her to study?

   $Y_{\text{-predicted}}$ **= -1(10) + 18. So the predicted value of *Y* is -10 + 18 = 8. For a person with 10 friends, we would predict 8 hours spent studying per week. (Notice that the hours spent studying score is well below the average. This is because the number of friends is well above the average, and the association between number of friends and hours spent studying is negative.)**

5. What, exactly, is a residual (when talking about regression)?

   **A residual is the difference between the predicted value of *Y* and the observed value of *Y* for a given value of *X*. So on the last question, we found a predicted value of *Y***

**of 8 at a given value of *X*, 10. But in a real sample, a person with 10 friends may have study for 20 hours per week, or 16, or 4, or whatever. The difference between the predicted score and the actual score is a residual.**

6. Regression is essentially a matter of drawing a straight line through a set of data, and the line has a slope and an intercept. In regression, how is it decided where the line should be drawn? In other words, explain the concept of least squares to me.

   **In any scatterplot of data, there are an infinite number of straight lines that can be drawn through the data. But there is only one straight line that will produce the smallest sum of squared residuals between the data points and the straight line. This line that produces the smallest sum of squared residuals is the regression line in an ordinary least squares regression, and the line will have a slope (the regression coefficient) and a point at which it intersections they *Y* axis (the intercept).**

7. Now suppose that I add a second predictor variable to the regression model: Hours per week spent working for money. And suppose that the correlation between the hours spent working and hours spent studying is -.50. The correlation between the two predictor variables (number of close friends and hours spent working for money) is -.30.

   a. What effect do you think the addition of this second predictor variable will have on the overall amount of variance explained ($R^2$) in the dependent variable? Why?

      **The addition of this second predictor variable should add to the overall amount of variance explained in the dependent variable. The correlation between the second predictor variable (hours spent working) and the dependent variable (hours spent studying) is moderate (*r* = -.50), so this predictor variable will explain some of the variance in the dependent variable. In addition, the correlation between the two predictor variables is not too strong (*r* = -.30), so each should be able to explain unique portions of variance in the dependent variable. Together, these two predictor variables should be able to explain more of the variance in the dependent variable than either would alone.**

   b. What effect do you think the addition of this second predictor variable will have on the strength of the regression coefficient for the first predictor variable, compared to when only the first predictor variable was in the regression model? Why?

      **Because the two predictor variables are correlated with each other, the addition of the second predictor variable will most likely reduce the size of the regression coefficient for the first predictor variable.**