Homework 2

Econ 103

Lecture Progress

We finished the Chapter 2 lecture and made it to slide 8 of the Chapter 3 lecture.

Homework Checklist

- ☐ **Book Problems (Chapter 2):** Do problems 1, 7, 8, 9bc, 13, 14, 16, 21, 23, 33, 35, 37, 41
- ☐ **Additional Problems:** See below.
- ☐ **R Tutorial:** Do R Tutorial 2 on the course website
- ☐ **Ask questions on Piazza**
- ☐ **Review Slides**

Additional Problems

1. For each variable indicate whether it is categorical or numeric; discrete or continuous; nominal, ordinal, interval, or ratio. Each one should have 3 classifications attached to it.
   (a) Grade of meat: prime, choice, good.
   (b) Type of house: split-level, ranch, colonial, other.
   (c) Income
   (d) SAT score
2. A drive-time radio show frequently holds call-in polls during the evening rush hour. Explain in no more than two sentences why such polls are likely to be biased.
3. Which of these studies are based on experimental data? Which are based on observational data?

(a) A biologist examines fish in a river to determine the proportion that show signs of disease due to pollutants poured into the river upstream.

(b) A Silicon Valley startup is trying to see what gathers more signups. They put a picture of a dog on their homepage which is shown to half of new customers and they put a picture of a cat on the homepage shown to the other half of customers.

(c) To understand how people respond to financial crises, an economist looks at banking data for individuals between 2000 and 2015.

(d) An industrial pump manufacturer monitors warranty claims and surveys customers to assess the failure rate of its pumps.

4. An emergency room institutes a new screening procedure to identify people suffering from life-threatening heart problems so that treatment can be initiated quickly. The procedure is credited with saving lives because in the first year after its initiation, there is a lower death rate due to heart failure compared to the previous year among patients seen in the emergency room. Do you agree? Explain.

5. Suppose that x_i is measured in US dollars and y_i is measured in euros. What are the units of the following quantities?

   (a) Interquartile Range of x

   (b) Covariance between x and y

   (c) Correlation between x and y

   (d) Skewness of x

   (e) Variance of y

6. The *mean deviation* is a measure of dispersion that we did not cover in class. It is defined as follows:

$$MD = \frac{1}{n}\sum_{i=1}^{n} |x_i - \bar{x}|$$

   (a) Explain why this formula averages the absolute value of deviations from the mean rather than the deviations themselves.

   (b) Which would you expect to be more sensitive to outliers: the mean deviation or the variance? Explain.

7. Consider a dataset $x_1, \ldots, x_n$. Suppose I multiply each observation by a constant d and then add another constant c, so that x_i is replaced by $c + dx_i$.

   (a) How does this change the sample mean? Prove your answer.

   (b) How does this change the sample variance? Prove your answer.

(c) How does this change the sample standard deviation? Prove your answer.

(d) How does this change the sample z-scores? Prove your answer.

8. You have the following data, which is real (barring some rounding to make the math nicer):

Year	**US Spending on Science, Space, and Tech ($millions)**	**Suicides by hanging, strangulation, and suffocation**
2006	24000	7500
2007	26000	8200
2008	28000	8600
2009	29000	9000

(a) Calculate the regression line of suicides on spending (i.e. $Suicides = a + b \times spending$)

(b) Assuming the relationship you derived in part (a) is true, how could we reduce the number of suicides? What is the lowest level of suicides we could attain (assuming you cannot have negative suicides)?

(c) Compute the correlation between these two series

(d) What could explain this relationship?

9. What value of a minimizes $\sum_{i=1}^{n}(y_i - a)^2$? Prove your answer.

10. Let

$$z_{x_i} = \frac{x_i - \bar{x}}{s_x}, \quad \text{and} \quad z_{y_i} = \frac{y_i - \bar{y}}{s_y}.$$

Show that if we carry out a regression with z_{y_i} in place of y and z_{x_i} in place of x, the intercept a will equal zero while the slope b will equal r, the sample correlation.

11. Let $\hat{y}$ denote our prediction of y from a linear regression model: $\hat{y} = a + bx$ and let r be the correlation coefficient between x and y.

(a) Express b in terms of s_{xy} and s_x.

(b) Express a in terms of b and the sample means of x and y.

(c) Express r in terms of the s_{xy}, s_x and s_y.

(d) Show that

$$\frac{\hat{y} - \bar{y}}{s_y} = r\left(\frac{x - \bar{x}}{s_x}\right)$$

12. You are a partner at Shady-Sleazy Consulting, LLC (motto: "Everything you want to hear and nothing you don't!"™).

    (a) You have been hired by a large investment bank to help them convince their clients that they should sell Google stock. Create a chart for their slide deck that supports this view. Making it in Excel is fine, though I encourage you to try in R (using the "Quandl" package makes it easy to get stock data)

    (b) You have been hired by the investment bank's rival as well and they want to convince their clients that they should invest more in a certain stock. As a partner of Shady-Sleazy Consulting, LLC, you have become adept at cutting corners, so you want to use the same data you've already collected. Create a chart that will convince their clients they should buy more Google stock.