# 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.

   **Solution:** categorical; discrete; ordinal

   (b) Type of house: split-level, ranch, colonial, other.

   **Solution:** categorical; discrete; nominal

(c) Income

> **Solution:** numeric; continuous; ratio

(d) SAT score

> **Solution:** numeric; discrete; interval

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.

> **Solution:** People who are listening to the radio during rush hour are disproportionately likely to be commuters driving home from work. People who are employed and drive to work are not representative of the population at large.

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.

> **Solution:** Observational

(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.

> **Solution:** Experimental

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

> **Solution:** Observational

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

> **Solution:** Observational

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.

   **Solution:** No. There could be many other reasons why death rates decreased, including improved medical technology in other areas. It could also be that the patients who came into the ER in the second year happened to be less sick, on average. In other words, there are many possible confounders.

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 $x$

   **Solution:** Dollars

   (b) Covariance between $x$ and $y$

   **Solution:** dollars × euros

   (c) Correlation between $x$ and $y$

   **Solution:** unitless

   (d) Skewness of $x$

   **Solution:** unitless

   (e) Variance of $y$

   **Solution:** dollars$^2$

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.

**Solution:** As we showed in class, the average deviation from the sample mean is zero regardless of the dataset. Taking the absolute value is similar to squaring the deviations: it makes sure that the positive ones don't cancel out the negative ones.

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

**Solution:** The variance is calculated from squared deviations. When $x$ is far from zero, $x^2$ is much larger than $|x|$ so large deviations "count more" when calculating the variance. Thus, the variance will be more sensitive to outliers.

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.

**Solution:**

$$\frac{1}{n}\sum_{i=1}^{n}(c + dx_i) = \frac{1}{n}\sum_{i=1}^{n} c + d\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = c + d\bar{x}$$

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

**Solution:**

$$\frac{1}{n-1}\sum_{i=1}^{n}[(c + dx_i) - (c + d\bar{x})]^2 = \frac{1}{n-1}\sum_{i=1}^{n}[d(x_i - \bar{x})]^2 = d^2 s_x^2$$

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

**Solution:** The new standard deviation is $|d|s_x$, the positive square root of the variance.

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

**Solution:** They are unchanged as long as $d$ is positive, but the sign will flip if $d$ is negative:

$$\frac{(c+dx_i)-(c+d\bar{x})}{ds_x} = \frac{d(x_i-\bar{x})}{ds_x} = \frac{x_i-\bar{x}}{s_x}$$

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. $Suicides = a + b \times spending$)

**Solution:**

$$b = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})r^2} \approx 0.29$$

$$a = \bar{Y} - b\bar{X} = 8325 - 0.29 \times 26750 \approx 663$$

$$Suicides = 663 + 0.29 \times Spending$$

(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)?

**Solution:** From the relationship derived above, we see that there is a positive relationship between spending and suicides. Hence, the lowest we can go would be to reduce spending to zero, which would imply that we could expect about 663 suicides.

(c) Compute the correlation between these two series

**Solution:** It's basically 1. $r = \frac{s_{xy}}{s_x s_y} = \frac{895000000}{53637*16688} \approx 0.993$

(d) What could explain this relationship?

**Solution:** Aside from conspiracy theories regarding NASA and mind control (I'm sure they exist somewhere), these two series, while correlated, are likely not to have any kind of causal relationship.

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

**Solution:** This is just like the regression problem from class, only with no slope. Differentiate with respect to $a$ and simplify as follows:

$$\begin{aligned}
-2\sum_{i=1}^{n}(y_i - a) &= 0 \\
\sum_{i=1}^{n}(y_i - a) &= 0 \\
\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} a &= 0 \\
\sum_{i=1}^{n} y_i &= na \\
a &= \frac{1}{n}\sum_{i=1}^{n} y_i \\
a &= \bar{y}
\end{aligned}$$

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.

**Solution:** All we need to do is replace $x_i$ with $z_{x_i}$ and $y_i$ with $z_{y_i}$ in the formulas we already derived for the regression slope and intercept:

$$a = \bar{y} - b\bar{x}, \quad b = \frac{s_{xy}}{s_x^2}$$

And use the properties of z-scores from class. Let $a*$ be the intercept for the regression with z-scores, and $b*$ be the corresponding slope. We have:

$$a* = \bar{z_y} - b^* \bar{z_x} = 0$$

since the mean of the z-scores is zero, as we showed in class. To find the slope, we need to know the covariance between the z-scores, and the variance of the z-scores for $x$:

$$b^* = \frac{s_{z_x z_y}}{s^2_{z_x}}$$

But since sample variance of z-scores is always one, $b^* = s_{z_x z_y}$. Now, by the definition of the sample covariance, the fact that the mean of z-scores is zero, and the definition of a z-score:

$$\begin{aligned} s_{z_x z_y} &= \frac{1}{n-1}\sum_{i=1}^{n}(z_{x_i} - \bar{z_x})(z_{y_i} - \bar{z_y}) \\ &= \frac{1}{n-1}\sum_{i=1}^{n} z_{x_i} z_{y_i} \\ &= \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right) \\ &= r_{xy} \end{aligned}$$

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 $s_x$.

    **Solution:**

    $$b = \frac{s_{xy}}{s_x^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

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

    **Solution:**

    $$a = \bar{y} - b\bar{x}$$

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

**Solution:**

$$r = \frac{s_{xy}}{s_x s_y}$$

(d) Show that

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

**Solution:**

$$\begin{aligned}
\hat{y} &= a + bx \\
\hat{y} &= (\bar{y} - b\bar{x}) + bx \\
\hat{y} - \bar{y} &= b(x - \bar{x}) \\
\hat{y} - \bar{y} &= \frac{s_{xy}}{s_x^2}(x - \bar{x}) \\
\hat{y} - \bar{y} &= \frac{s_{xy}}{s_x}\left(\frac{x - \bar{x}}{s_x}\right) \\
\frac{\hat{y} - \bar{y}}{s_y} &= \frac{s_{xy}}{s_x s_y}\left(\frac{x - \bar{x}}{s_x}\right) \\
\frac{\hat{y} - \bar{y}}{s_y} &= r\left(\frac{x - \bar{x}}{s_x}\right)
\end{aligned}$$

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

(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)

**Solution:** The key is picking a good time horizon to make your chart fit your objective. This time horizon makes it look like the stock has been tumbling throughout August and therefore, you should convince your client to sell. DISCLAIMER: This should not be construed as investment advice. I am not a financial adviser. You probably should not make investing decisions based on something like this.

```
install.packages("Quandl")
```

```
library(plotly)
library(data.table)
library(Quandl)
google <- Quandl("WIKI/GOOGL", type = "xts")
google <- data.table(date = index(google),
                     price = google$Close)

# Selling Google stock
sellChart <- google[date >= "2016-08-01"]
plot_ly(data = sellChart, x = date, y = price.Close,
        type = "scatter")
```

(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.

**Solution:** This chart makes it look like the stock has been climbing ever since Google went public and hence, the client should buy more because this one keeps going up, up, up!

```
# Buying Google stock
buyChart <- google[date >= "2006-01-01"]
plot_ly(data = buyChart, x = date, y = price.Close,
 type = "scatter")
```