Problem Set #3

Econ 103

Part I – Problems from the Textbook

Chapter 11: 1, 3
Chapter 15: 1(a), 5(a)

Part II – Additional Problems

1. 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:

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

2. 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:

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

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:

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

3. 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:

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

(e) (3 points) Using the equation derived in (d), briefly explain "regression to the mean."

Solution: The formula shows that unless r is one or negative one, perfect positive or negative correlation, our best linear prediction of y based on knowledge given x is closer to the mean of the y-observations (relative to the standard deviation of the y-observations) than x is to mean of the x-observations (relative to the standard deviation of the x-observations). If x is very large, for example, we would predict that y will be large too, but not as large.

4. Lothario, an unscrupulous economics major, runs the following scam. After the first midterm of Econ 103 he seeks out the students who did extremely poorly and offers to sell them "statistics pills." He promises that if they take the pills before the second midterm, their scores will improve. The pills are, in fact, M&Ms and don't actually improve one's performance on statistics exams. The overwhelming majority of Lothario's former customers, however, swear that the pills really work: their scores improved on the second midterm. What's your explanation?

Solution: This is an example of regression to the mean. The students Lothario seeks out were both unprepared for the midterm *and* got unlucky: the correlation between exam scores is less than one. It is very unlikely that they will be unlucky twice in a row, so their performance on the second exam will almost certainly be higher. Our best guess of their second score is closer to the mean than their first score.