Extension Problems

Econ 103

Spring 2018

About This Document

Lecture #1

1. A long time ago, the graduate school at a famous university admitted 4000 of their 8000 male applicants versus 1500 of their 4500 female applicants.

   (a) Calculate the difference in admission rates between men and women. What does your calculation suggest?

   > **Solution:** The rate for men is 4000/8000 = 50% while that for women is 1500/4500 ≈ 33% so the difference is 17%. It appears that women are less likely to be accepted to the graduate school.

   (b) To get a better sense of the situation, some researchers broke these data down by area of study. Here is what they found:

	Men		Women	
	# Applicants	# Admitted	# Applicants	# Admitted
Arts	2000	400	3600	900
Sciences	6000	3600	900	600
Totals	8000	4000	4500	1500

   Calculate the difference in admissions rates for men and women studying Arts. Do the same for Sciences.

   > **Solution:** For Arts, the admission rate is 400/2000 = 20% for men versus 900/3600 = 25% for women. For Sciences 3600/6000 = 60% for men versus 600/900 ≈ 67% for women. In summary:
   >
   > | | Men | Women | Difference |
   > |---|---|---|---|
   > | Arts | 20% | 25% | -5% |
   > | Sciences | 60% | 67% | -7% |
   > | Overall | 50% | 33% | 17% |

(c) Compare your results from part (a) to part (b). Explain the discrepancy using what you know about observational studies.

> **Solution:** When we compare overall rates, women are less likely to be admitted than men. In each field of study, however, women are *more* likely to be admitted. In this example, field of study is a *confounder*: women are disproportionately applying to study Arts and Arts have much lower admissions rates than Sciences.

Lecture #2

2. 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.

3. Let m be a constant and $x_1, \ldots, x_n$ be an observed dataset.

(a) Show that $\sum_{i=1}^{n}(x_i - m)^2 = \sum_{i=1}^{n} x_i^2 - 2m\sum_{i=1}^{n} x_i + nm^2$.

Solution:

$$\begin{aligned}\sum_{i=1}^{n}(x_i - m)^2 &= \sum_{i=1}^{n}(x_i^2 - 2mx_i + m^2) \\ &= \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} 2mx_i + \sum_{i=1}^{n} m^2 \\ &= \sum_{i=1}^{n} x_i^2 - 2m\sum_{i=1}^{n} x_i + nm^2\end{aligned}$$

(b) Using the preceding part, show that $\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$.

Solution: Solving this requires two observations. First, note that $\bar{x}$ is a *constant*, i.e. that it does not have an index of summation. Second, note that $\sum_{i=1}^{n} x_i = n\bar{x}$. Hence, taking $m = \bar{x}$ in the formula from the preceding part,

$$\begin{aligned}\sum_{i=1}^{n}(x_i - \bar{x})^2 &= \sum_{i=1}^{n} x_i^2 - 2\bar{x}\sum_{i=1}^{n} x_i + n\bar{x}^2 \\ &= \sum_{i=1}^{n} x_i^2 - 2\bar{x}(n\bar{x}) + n\bar{x}^2 \\ &= \sum_{i=1}^{n} x_i^2 - 2n\bar{x}^2 + n\bar{x}^2 \\ &= \sum_{i=1}^{n} x_i^2 - n\bar{x}^2\end{aligned}$$

Lecture #3

4. 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:

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

Lecture #4

5. Define the z-scores

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

Show that if we carry out a regression with z_i in place of y_i and w_i in place of x_i, the intercept a^* will be zero while the slope b^* will be r_{xy}, the correlation between x and y.

Solution: All we need to do is replace x_i with w_i and y_i with z_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:

$$a^* = \bar{z} - b^*\bar{w} = 0$$

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

$$b^* = \frac{s_{wz}}{s_w^2}$$

But since sample variance of z-scores is always one, $b^* = s_{wz}$. 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:

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

6. This question concerns a phenomenon called *regression to the mean*. Before attempting this problem, read Chapter 17 of *Thinking Fast and Slow* by Kahneman.

   (a) 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.

   (b) 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. Show that

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

Solution:

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

(c) 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.

No extension problems for Lecture #5

Lecture #6