Homework questions, week 1

Econ 103

1 Daily Homework questions

The questions in bold font are due on **Thursday 26th May**. You do not need to hand in the questions that are not in bold, though these will be useful to complete for your own understanding.

Lecture 1

Textbook questions:
Chapter 1: 3, 5, 7, **9, 13**
Chapter 2: 1

Solution: Solutions to textbook questions in back of textbook

Additional questions:

1. **For each variable indicate whether it is nominal, ordinal, or numeric.**
   (a) Grade of meat: prime, choice, good.

   Solution: ordinal

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

   Solution: nominal

   (c) Income

   Solution: numeric

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) In a pilot phase of a fund-raising campaign, a university randomly contacts half of a group of alumni by phone and the other half by a personal letter to determine which method results in higher contributions.

   Solution: Experimental

   (c) To analyze possible problems from the by-products of gas combustion, people with with respiratory problems are matched by age and sex to people without respiratory problems and then asked whether or not they cook on a gas stove.

   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.

Lecture 2

Textbook questions:
Chapter 2: 5, 7, **13, 15**, 17 [in part (b) skip MAD and MSD], **21** 23, 35.

Additional questions:

1. **Suppose that x_i is measured in centimeters and y_i is measured in feet. What are the units of the following quantities?**

   (a) Interquartile Range of x

   Solution: centimeters

   (b) Covariance between x and y

   Solution: centimeters $\times$ feet

   (c) Correlation between x and y

   Solution: unitless

   (d) Skewness of x

   Solution: unitless

   (e) Variance of y

   Solution: feet^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. 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:

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

Lecture 3 - Regression

Textbook questions:
Chapter 11: **1, 3**
Chapter 15: 1(a)

Solution: Solutions to textbook questions in back of textbook

Additional questions:

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.

2 R Tutorials

You should complete TryR levels 1-6 and R Tutorial #1 by **Monday 30th May**.

TryR can be accessed at `tryr.codeschool.com`.

R tutorials will be posted in Piazza, with solution code.