Homework 3: Sections 2.5 - 2.6
KEY

STA209-04: Applied Statistics

February 8, 2019

Total Possible Points: 31

From the Book:

2.179) [4 pts]

a) [1 pts] The association is positive. Greater nurturing corresponds with larger hippocampus size.

b) [1 pts] This association is also positive. Larger hippocampus size corresponds with greater resiliency and ability to deal with stresses and strains of daily life.

c) [1 pts] The key facet of the experiment would be to randomize child-parent pairs to nurture and no-nurture groups. Parents in the nurture group would nurture their children while parents in the non-nurture group would not. A baseline measure of each child's hippocampus would also be obtained. The nurture/no-nuture treatment would persist throughout the development of the child and hippocampus size would be measured afterwards. There is no way this would be ethical.

d) [1 pts] Given that the data citing this association was not obtained via an unethical randomized experiment along the lines of what was described in (c), we cannot claim that maternal nurturing causes increased growth of the hippocampus in humans - we can only state the two are associated. On the other hand, we can make this claim for mice given that those data are derived from (randomized) experiments.

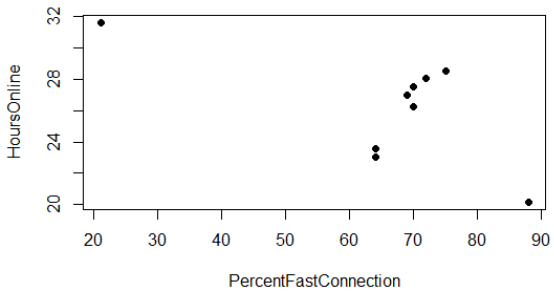
2.185) [3 pts] The word *correlation* is appropriately used when referring to the relationship between two *quantitative* variables. Since CBT is a categorical, dichotomous variable referring to whether or not an individual received that specific treatment, the word *correlation* should not be used. A more appropriate headline would be "Sleep improvements were strongly associated with CBT."

2.194) [6 pts]

a) [1 pts] A positive association would indicate that countries with higher percentage of fast connection users spend (on average) a greater amount of time online. Observing this positive relationship might make sense if one were to assume that greater proportions of high-speed connections are indicative of a reliance and heavy dependence on the Web. Thinking about the US for example, most of us spend hours accessing the web for course materials, entertainment, news, etc. All the while, we aim for higher connection speeds to accommodate this consumption of Internet resources.

b) [1 pts] A negative association would indicate that countries with higher percentage of *slow* connection users spend (on average) a greater amount of time online. This would make sense given that slower connections mean a user takes more time to access whatever material of interest on the web.

c) **[1 pts]** The scatterplot (shown below) exhibits a negative relationship when the outliers - Brazil and Switzerland - are included. Brazil is an outlier given its distance from other countries in terms of PercentFastConnection. Switzerland is an outlier in that it defies the clear positive linear trend observed when observing all other points (except Brazil).



d) **[1 pts]** Removing the outliers mentioned above yields a positive relationship.

e) **[1 pts]** With the outliers, the correlation is -0.65. Without, the correlation is 0.95. The correlation is clearly affected by outliers.

f) **[1 pts]** Correlation does not imply causation. We may not make this conclusion. We can only say that the two are associated.

2.212) **[3 pts]**

a) **[1 pts]** Using the provided regression equation:

$$102 - 3.34 * 8 = 75.28$$

$$102 - 3.34 * 14 = 55.24$$

b) **[1 pts]** For each year increase in the number of years playing football, there is an expected 3.34 percentage point decrease in cognition percentile.

c) **[1 pts]** Interpreting the intercept would not be appropriate in this context. Note that the intercept represents the predicted value for an individual who played football for zero years. The data only contain individuals who played between 7 and 18 years.

2.212) **[3 pts]**

a) **[1 pts]** Using the provided regression equation:

$$-170 + 4.82 * 60 = 119.20$$

$$-170 + 4.82 * 72 = 177.04$$

b) **[1 pts]** The slope of the line is 4.82. This describes the change in weight expected for each inch increase in height.

c) **[1 pts]** The intercept of the line is -170. It is not appropriate to interpret the slope in this context. There are no individuals contained in the data with a height of 0 inches.

d) **[1 pts]** Using the provided regression equation:

$$-170 + 4.82 * 20 = -73.60$$

It is inappropriate to use the regression line in this case since the data do not contain individuals (babies) whose heights are 20 inches. The data are collected from college-aged students. Using the regression line for predicting the weight of a 20 inch baby would be extrapolation.

Miscellaneous

S1) [3 pts]

a) [1 pts] Let X denote the insulin level and Y the BMI level. Since we are told the individual's insulin level is 0.4 standard deviations below average, we have $z_x = -0.4$. To get our predicted z-score for Y, z_y, we then multiply z_x by the correlation between X and Y: $-0.4 * 0.67 = -0.268$. Therefore, the predicted BMI for an individual whose insulin is 0.4 standard deviations below average is 0.268 standard deviations below average in BMI.

b) [1 pts] We first need to convert 103.1 into a z-score:

$$z_x = \frac{103.1 - 93.78}{9.36} = 0.996$$

Next, we multiply this z-score by the correlation between X and Y to get the predicted z-score for Y:

$$z_y = 0.996 * 0.67 = 0.667$$

Finally, we "unstandardize" z_y to get our answer in terms of the original units for Y:

$$0.667 * 4.68 + 25.01 = 28.13$$

c) [1 pts] Here, the roles of X and Y are reversed. We are being asked to use BMI to predict insulin. Therefore, using the formula discussed in class:

$$b = r_{xy}\frac{s_y}{s_x} = 0.67 * \frac{9.36}{4.68} = 1.34$$

S2) [3 pts]

a) [1 pts] Let X denote the fatty acid level and Y the uric acid level. Since we are told the individual's fatty acid level is 0.7 standard deviations above average, we have $z_x = 0.7$. To get our predicted z-score for Y, z_y, we then multiply z_x by the correlation between X and Y: $0.7 * 0.75 = 0.525$. Therefore, the predicted BMI for an individual whose fatty acid is 0.7 standard deviations above average is 0.525 standard deviations above average in uric acid.

b) [1 pts] We first need to convert 274.9 into a z-score:

$$z_x = \frac{274.9 - 271.99}{29.67} = 0.098$$

Next, we multiply this z-score by the correlation between X and Y to get the predicted z-score for Y:

$$z_y = 0.098 * 0.75 = 0.074$$

Finally, we "unstandardize" z_y to get our answer in terms of the original units for Y:

$$0.074 * 0.96 + 5.19 = 5.26$$

c) [1 pts] Using the formula discussed in class:

$$b = r_{xy}\frac{s_y}{s_x} = 0.75 * \frac{0.96}{28.67} = 0.025$$

S3) [6 pts] Several answers possible. Check whether they summarized the article and provided some appropriate rationale for why they agreed or disagreed with points raised in the article.