PSY30100-03 -- Assignment 3

Chapter 2: Look at Data -- Relationships

TA: Laura Lu

September 14, 2009

Question 1: 2.50 (page 108)

Each of the following statements contains a blunder (of correlation). Explain in each case what is wrong.

Review: Properties of Correlation (P.103)

- Correlation makes no distinction between explanatory and response variable.
- (In our textbook) Correlation requires that both variables be quantitative. And we cannot calculate a correlation between categorical variables.
- r is not changed when the unit of x, y, or both, changes.
- Positive (negative) r indicates positive (negative) association between x and y
- Always $-1 <= r <= 1$. Extreme values ±1 occur only when the points lie exactly along a straight line. $r = 0$ is called uncorrelated. The strength of linear relationship increases as r moves away from 0 toward either −1 or 1.
- Correlation measures only the strength of linear relationship between two variables. It does not describe curved relationship.
- Correlation is not resistant to outliers.

Question 1: 2.50 (page 108)

What's wrong?

(a) There is a high correlation between the gender of American workers and their income.

Ans:

The correlation doesn't work with nominal variables, and gender has a categorical (nominal) scale.

(we may say that there is a strong association between gender and income.)

Question 1: 2.50 (page 108)

What's wrong?

(b) We found a high correlation (r=1.09) students' ratings of faculty teaching and ratings made by other faculty members.

Ans:

Correlation couldn't be larger than 1!

$$-1 \le r \le 1$$

Question 1: 2.50 (page 108)

What's wrong?

(c) The correlation between planting rate and yield of corn was found to be r=0.23 bushel.

Ans:

Correlation has no unit!

Question 2: 2.62 (page 122)

Revenue and value of NBA teams.

$$\text{value} = 21.4 + (2.59 \times \text{revenue})$$

Review: Regression Line

$$\hat{y}_i = b_0 + b_1 x_i$$

$\hat{y}$ is the predicted *y* value (y hat)

b_1 is the **slope**

b_0 is the **intercept**

b_1: the amount by which y changes when x increases in one unit

b_0: the value of y when x =0

Question 2: 2.62 (page 122)

$$\text{value} = 21.4 + (2.59 \times \text{revenue})$$

- What is the slope of this line? And explain.

Ans: 2.59, meaning that (on the average) team value rises 2.59 units (dollars, $million, or whatever) from each one-unit increasing in revenue. (The ratio holds regardless of the unit, provided the same unit is used for both variables.)

Question 2: 2.62 (page 122)

(b) Use the line to predict the value of the Lakers from their revenue. What is the error in this prediction?

Ans:

value=21.4+(2.59*revenue)

=21.4+(2.59*149)

=407.31 (million dollars)

407.31-447=-39.69 (million dollars)

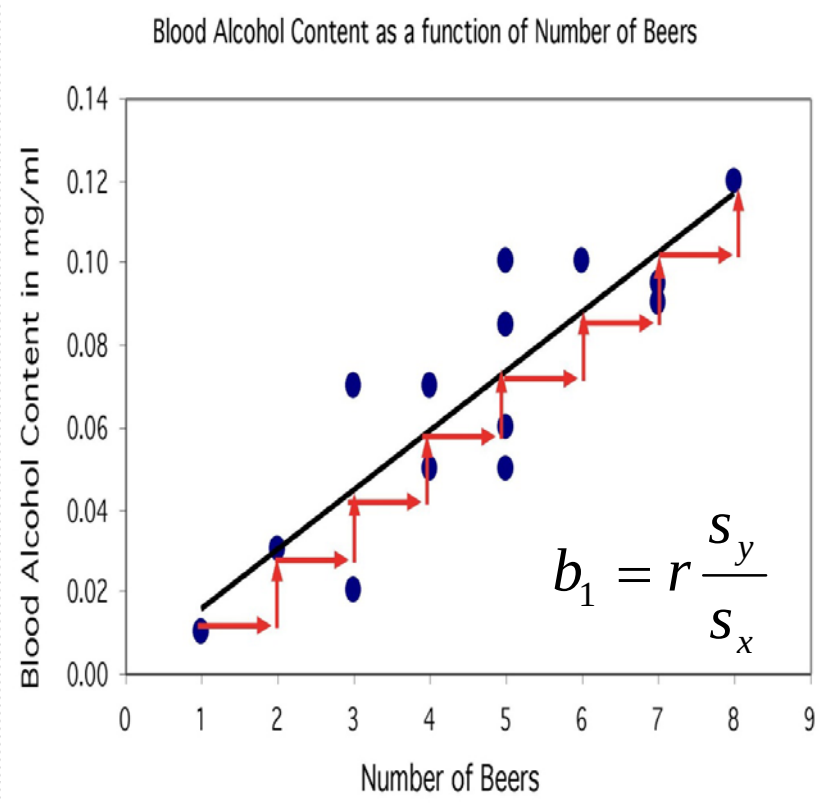
Question 2: 2.62 (page 122)

(c) What does the correlation say about the success of the regression line in predicting the values of the 29 teams? (hint: the coefficient of determination)

Review: Coefficient of determination, r^2

r^2, the coefficient of determination, is the square of the correlation coefficient.

r^2 represents **the percentage of the variance in *y* that can be explained by changes in *x*.**



Question 2: 2.62 (page 122)

(c) What does the correlation say about the success of the regression line in predicting the values of the 29 teams? (hint: the coefficient of determination)

Ans: The high correlation means that the line does a fairly good job of predicting value; specifically, the regression line explains about $r^2 \approx 86\%$ of the variability in team value.

Question 3: 2.126 (page 160)

Income <--> healthy

Ans: If a nation's population has high income, they have more money to spend on things that can help to keep them healthy: health care, medicine, better food, better sanitation, and so on.

On the other hand, if a nation's population is healthy, they can spend less on health care and instead put their money to more productive uses. Additional, they miss fewer work days, so they would typical earn more money.

Question 4: 2.154 (page 166)

$$\hat{y} = 46.6 + 0.41\,x$$

Q: Octavio scores 10 points above the class mean on the midterm. How many points above the class mean do you expect that he will score on the final?

Review: How to calculate the intercept and the slope

First we calculate the **slope of the line, b_1**;
from statistics we already know:

$$b_1 = r\frac{s_y}{s_x}$$

r is the correlation.

s_y is the standard deviation of the response variable y.

s_x is the the standard deviation of the explanatory variable x.

Once we know b_1, the slope, we can calculate **b_0, the intercept:**

$$b_0 = \bar{y} - b_1\bar{x}$$

where $\bar{x}$ and $\bar{y}$ are the sample means of the x and y variables

Question 4: 2.154 (page 166)

$$\hat{y} = 46.6 + 0.41\,x$$

Q: Octavio scores 10 points above the class mean on the midterm. How many points above the class mean do you expect that he will score on the final?

Ans: Note that $\overline{y} = 46.6 + 0.41\,\overline{x}$.

We predict that Octavio will score 4.1 points above the mean on the final exam because:

$$\hat{y} = 46.6 + 0.41(\overline{x} + 10)$$

$$= 46.6 + 0.41\,\overline{x} + 4.1$$

$$= \overline{y} + 4.1$$

Original definition of Regression

- Francis Galton (1908)
- "The children of tall parents are taller than average but not as tall as their parents."
- regression toward mediocrity

  This is for standardized regression only.

Standardized regression

- The original linear least-squares regression form

$$\hat{y} = b_0 + b_1 x$$

- The properties

$$b_1 = r\frac{s_y}{s_x} \qquad b_0 = \bar{y} - b_1\bar{x}$$

- Standardized regression (derived in the next slide)

$$z_y = r\, z_x$$

- When $|r| < 1$, then we say that the data points exhibit regression toward the mean.

Derivation of the standardized regression

$$z_{\hat{y}} = \frac{\hat{y} - \bar{y}}{s_y}$$

$$= \frac{(b_0 + b_1 x) - \bar{y}}{s_y}$$

$$= \frac{[(\bar{y} - b_1 \bar{x}) + b_1 x] - \bar{y}}{s_y} \quad = \frac{b_1 (x - \bar{x})}{s_y}$$

$$= \frac{b_1 s_x}{s_y} \frac{x - \bar{x}}{s_x} \quad = r\, z_x$$

Question 5:

- Determine whether each of the following statements regarding the correlation coefficient is true or false.

  A) The correlation coefficient equals the proportion of times that two variables lie on a straight line.

Ans: False.

Question 5:

B) The correlation coefficient will be +1.0 if all the data points lie on a perfectly horizontal straight line.

Ans: False.

The correlation coefficient will be 0 if all the data points lie on a perfectly horizontal straight line.

Question 5:

C) The correlation coefficient measures the strength of any relationship that may be present between two variables.

Ans: False.

The correlation coefficient measures the strength of linear relationship that may be present between two variables.

Question 5:

D) The correlation coefficient is a unitless number and must always lie between –1.0 and +1.0, inclusive.

Ans: True.