PSY30100-03 -- Assignment 2

Chapter 2: Look at Data -- Relationships

TA: Laura Lu

February 1, 2010

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, but gender has a categorical (nominal) scale.

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

Extension: another correlation

- Although in our textbook correlation requires that both variables be quantitative and we cannot calculate a correlation between categorical variables, there is one type of correlation for categorical variables:

  Point-biserial correlation

$$r_{pb} = \frac{M_1 - M_0}{s_n}\sqrt{\frac{n_1 n_0}{n^2}},$$

  where M1 is the mean value on the continuous variable X for all data points in group 1, M0 is the mean value on the continuous variable X for all data points in group 2. Further, n1 is the number of data points in group 1, n0 is the number of data points in group 2 and n is the total sample size, and

$$s_n = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2}.$$

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!

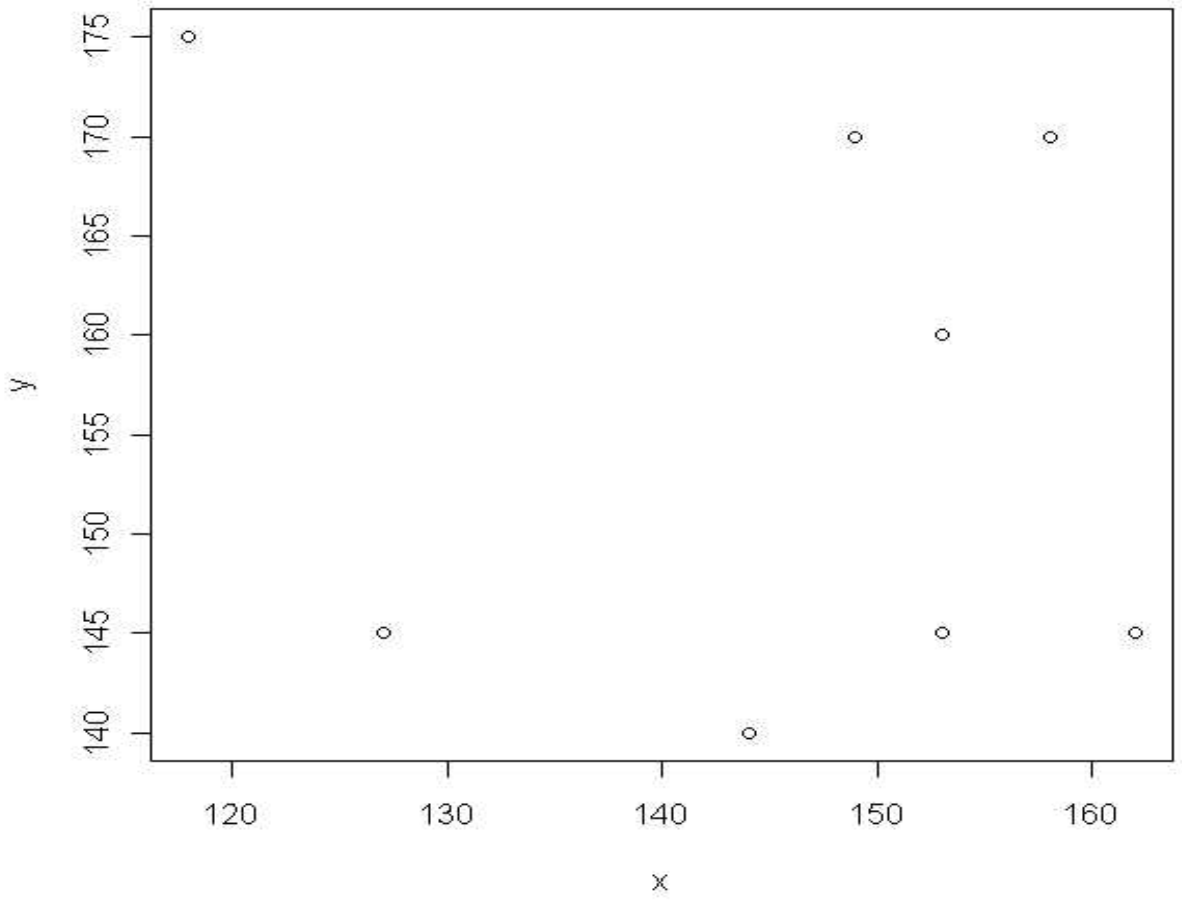
Question 2: 2.58 (page 122)

a) Plot data

b) Find the regression line

c) Draw the regression line

Question 2: 2.58 (page 122)



Review: Regression Line

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

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

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

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

Review: correlation r

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

In order to obtain r, we have to obtain

$$\bar{x},\ \ s_x,\ \ \bar{y},\ \ s_y$$

Question 2: 2.58 (page 122)

Step 1:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = 145.5 \qquad s_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} = 15.37159$$

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = 156.25 \qquad s_y = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2} = 14.07886$$

Step 2:

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

$$= \frac{1}{8-1}\left[\left(\frac{153-145.5}{15.37}\right)\left(\frac{145-156.25}{14.1}\right) + \ldots + \left(\frac{153-145.5}{15.37}\right)\left(\frac{160-156.25}{14.1}\right)\right]$$

$$\approx -0.203$$

Question 2: 2.58 (page 122)

Step 3:

$$b_1 = r\frac{s_y}{s_x} \approx -0.203 \times \frac{14.1}{15.37} \approx -0.1844$$

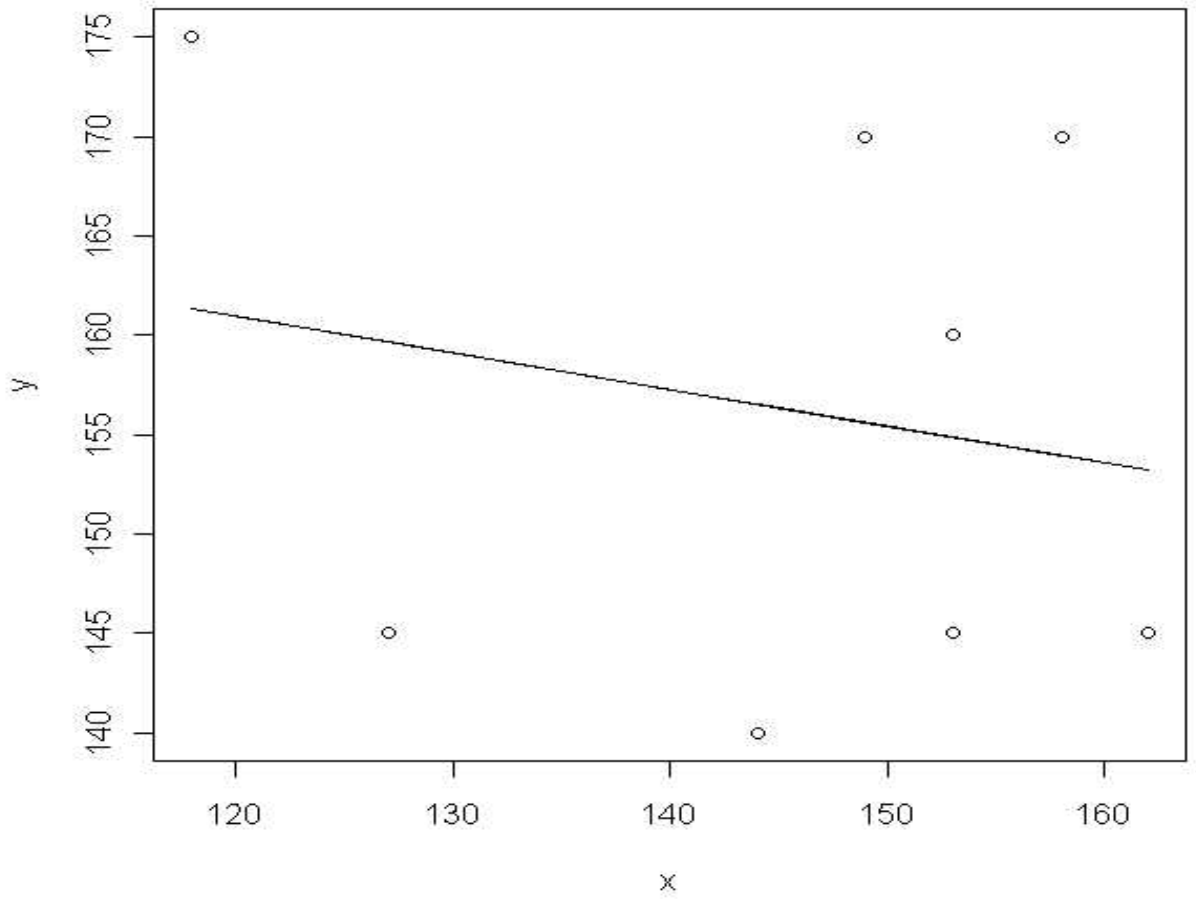
Step 4:

$$b_0 = \bar{y} - b_1\,\bar{x} \approx 156.25 - (-0.1844) \times 145.5 \approx 183$$

Step 5:

$$\hat{y}_i = b_0 + b_1 x_i = 183 - 0.1844\,x_i$$

Question 2: 2.58 (page 122)



Question 3: 2.62 (page 122)

- Revenue and value of NBA teams.

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

(a) What is the slope of this line? And explain.

Ans: 2.59.

(you may have your own explanation) It means 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 3: 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 3: 2.62 (page 122)

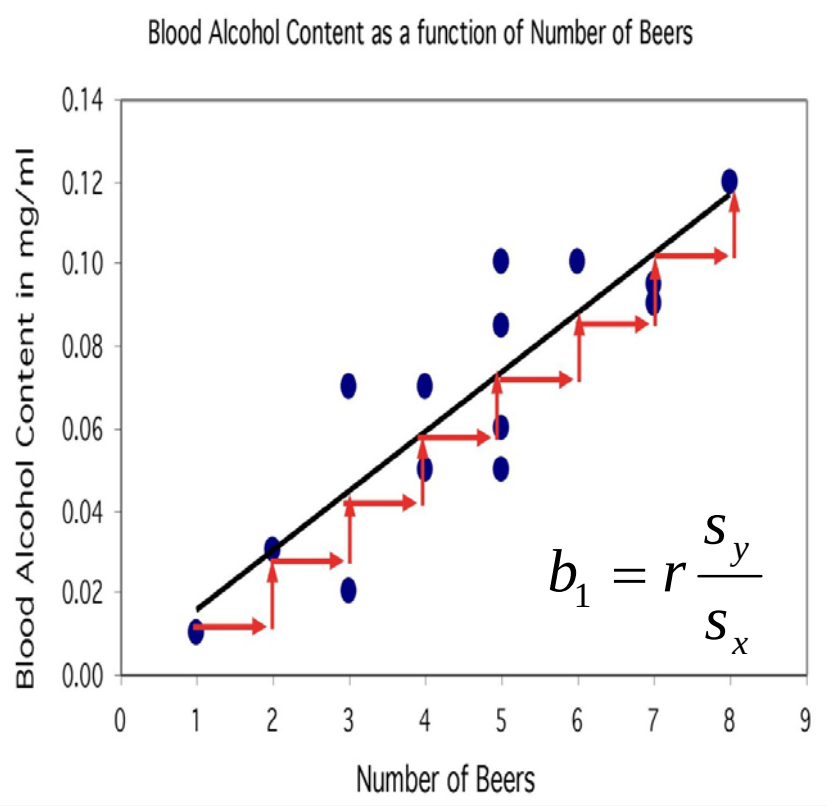
(c) What does the correlation say about the success of the regression line in predicting the values of the 29 teams?

(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 3: 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 4: 2.123,2.124 (p.166)

Q: 2.123

Ans: There are a lot of answers. You may have your own answer.

Mine: Age is one of lurking variables. Married men would generally be older than single men, so they would have been in the work force longer and therefore had more time to advance in their careers.

Question 4: 2.123,2.124 (p.166)

Q: 2.124

Ans: A large company has more workers who might be laid off and often pays its CEO a higher salary because, presumably, there is more work involved in running a large company than a small one. Smaller companies typically pay less and have fewer workers to lay off.

Question 5: 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:

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

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

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

Caution!

- In our text book, “This is an example of regression toward mean” which is not appropriate.

- regression toward mean is for **standardized** regression only!

Standardized regression

- The general 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

$$z_y = r\, z_x$$

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

Extension: Derivation of standardized regression

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

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

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

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