STAT20 Homework #9

Seth Metcalf

Table of Contents

Introduction

This is Homework #8, which contains questions from Chapters 8, 9, & 10, and a ggplot question. *Due 16 April 2021.*

Chapter 8

1. Ch 8 A5:

Students named A, B, C, D, E, F, G, H, I, and J took a midterm and a final in a certain course. A scatter diagram for the scores is shown below.

a) Which students scored the same on the midterm as on the final?

Students A, F, and B.

b) Which students scored higher on the final?

Students G, C, and H

c) Was the average score on the final around 25, 50, or 75?

Average score on the final was about 50.

d) Was the SD of the scores on the final around 10, 25, or 50?

The standard deviation on the scores of the final was around 25.

e) For the students who scored over 50 on the midterm, was the average score on the final around 30, 50, or 70?

Average score on the final for students that scored over 50 on the midterm was 30.

f) True or false: on the whole, students who did well on the midterm also did well on the final.

False.

g) True or false: there is strong positive association between midterm scores and finalscores.

False.

2. Ch 8 Rev 3:

Suppose men always married women who were exactly 8% shorter. What would the correlation between their heights be?

Explanation

If they *always* married women that were 8% then there would be a 100% correlation, or a correlation of 1.

r = 1

3. Ch 8 Rev 4:

Is the correlation between the heights of husbands and wives in the U.S. around -0.9, -0.3, 0.3, or 0.9? Explain briefly.

a) Explanation

Since we know that taller men tend to marry taller women, the correlation coefficient would be positive. Furthermore, there most likely is a strong linear relationship between the two as height seems to be a strong factor in determining a spouse (however preferential it is it still seems to follow a general guideline that the taller you are, a wife is generally 6 inches shorter than her husband).

Overall this leads to the explanation that there is most likely a strong correlation between the two, estimating a 0.9 correlation coefficient.

Chapter 9

4. Based on Ch 9 A10:

Six data sets are shown on the next page. Calculate the correlation for (i) and (ii) in two ways. First, use the method in lecture code for 11/17: take the average of the products of the standard units. Second, check that you get the same answer if you use the cor() function. Find the correlations for the remaining data sets without doing any arithmetic, the other datasets are related to (i) and (ii) in ways that should make that easy.

a) Correlation

```
iiiSD = 2 # The SD of a data set containing numbers 1 through 7 is 2. This
occurs in both i and ii data on both axis.
iiiMean = 4 # The Mean of a data set containing numbers 1 through 7 is 4.
This occurs in both i and ii data on both axis.

# Correlation of data i calculating SUs
ix = c(1, 2, 3, 4, 5, 6, 7)
ixSU = (ix - iiiMean)/iiiSD
iy = c(2, 3, 1, 4, 6, 7, 5)
iySU = (iy - iiiMean)/iiiSD
iSU = (ixSU * iySU)
iCor = mean(iSU) # 0.7857143

# Correlation of data i using cor()
cor(ix, iy) # 0.7857143

## [1] 0.7857143

# Correlation of data ii
iix = c(1, 2, 3, 4, 5, 6, 7)
iixSU = (iix - iiiMean)/iiiSD
iiy = c(2, 3, 1, 4, 6, 5, 7)
iiySU = (iiy - iiiMean)/iiiSD
iiSU = (iixSU * iiySU)
iiCor = mean(iiSU) # 0.8571429

# Correlation of data ii using cor()
cor(iix, iiy) # 0.8571429

## [1] 0.8571429
```

Given this, the rest of the graphs are just variations of these two original graphs.

(iii) is just the reverse of (i) meaning that the correlation coefficient is the same: 0.7857143.

(iv) is just (i) but with 1 added to all numbers on the x axis, so the correlation coefficient is the same: 0.7857143.

(v) is just (i) but with 2 multiplied to all numbers on the x axis, so the correlation coefficient is the same: 0.7857143.

(vi) is just (ii) but with 3 multiplied to all numbers on the x axis, so the correlation coefficient is the same: 0.8571429.

5. Based on Ch 9 B2:

The National Health and Nutrition Examination Survey collects data on children. In this dataset, at each age from 6 to 11, the correlation between height and weight was just about 0.67. For all the children together, would the correlation between height and weight be just about 0.67, somewhat more than 0.67, or somewhat less than 0.67? Choose one option and explain. (Hint: consider drawing a scatter plot of what you think the dataset will look like.)

Explanation

I would think that the correlation would be more than 0.67. There seems to be positive correlation in the data of children that are collected and more data to back this up would most likely increase correlation to a certain extent.

6. Ch 9 C4:

For a certain data set, r = 0.57. Say whether each of the following statements is true or false, and explain briefly. If you need more information, say what you need and why.

a) There are no outliers.

It is impossible to say without looking at the graph itself. It could have perfect correlation except 1 outlier that drops the coeffecient all the way down to 0.57, or it could have 0 outliers but not perfectly correlated – more information needs to be given.

b) There is a non-linear association.

Similar to the last question, a look at the chart needs to occur. There could be a different (non-linear) association that the r coeffecient does not tell us, but we do not know until we look at the chart.

7. Ch 9 E3:

The correlation between height and weight among men age 18-74 in the U.S. is about 0.40. Say whether each conclusion below follows from the data; explain your answer.

a) Taller men tend to be heavier.

True, while there is not a strong correlation between the two, it still has a positive correlation and shows that the taller a person is, the heavier they are to some extent.

b) The correlation between weight and height for men age 18-74 is about 0.40.

This conclusion is correct, it is what is given to us and follows from the data.

c) Heavier men tend to be taller.

Also true, the correlation between height and weight is the same between weight and height and would therefore be about 0.40 as well. This means that, again while there might not be a strong correlation, there is a positive one that shows that the heavier a person is the taller they seem to be.

d) If someone eats more and puts on 10 pounds, he is likely to get somewhat taller.

Incorrect, it is said that there is a correlation between the two, not that there is causation between the two. They imply different things.

8. Ch 9 Rev 10:

In a study of 2005 Math SAT scores, the Educational Testing Service com-puted the average score for each of the 51 states, and the percentage of the high-school seniors in that state who took the test.14 (For these purposes, D.C. counts as a state.) The correlation between these two variables was equal to -0.84.

a) True or false: test scores tend to be lower in the states where a higher percentage of the students take the test. If true, how do you explain this? If false, what accounts for the negative correlation?

True. It could occur that the correlation between this is because in the states where only a few took it, they could have been specifically selected/privileged enough to be able to take the test. In these scenarios, higher scores would be expected between less students rather than having a large number (almost all) students take the test and have a more average test result.

b) In Connecticut, the average score was only 517. But in Iowa, the average was 608. True or false, and explain: the data show that on average, the schools in Iowa are doing a better job at teaching math than the schools in Connecticut.

False. It could be biased as was mentioned before. A comparison between the number/percentage of students that took the test between the two states would be needed and would allow for a better understanding.

Chapter 10

9. Ch 10 A2:

For the men age 18 and over in HANES: average height = 69 inches, SD =3 inches, average weight =190 pounds, SD = 42 pounds, r = 0.41. Estimate the average weight of the men whose heights were each of the following, and comment on the answers to c) and d).

a) 69 inches

Average height = 69 inches. This is 0 SDs away from the height, meaning that they were the average weight as well. Therefore:

Average weight of a 69 inch tall man = 190 pounds.

b) 66 inches

Since the height given to us was 66 inches, this is 1 standard deviation below the mean (69 inches - 1 SD = 66 inches), we can multiply the correlation coefficient (0.41) by the SD (42 inches): getting 17.22 pounds. 190 pounds is the average, we subtract this by one standard deviation of weight (- 17.22 pounds) and therefore get:

Average weight of a 66 inch tall man = 172.78 pounds.

c) 24 inches

Since the height given to us was 24 inches, this is 15 standard deviation below the mean (69 inches - 15 SDs = 66 inches), we can multiply the correlation coefficient (0.41) by the SD (42 inches): getting 17.22 pounds. Then, we multiply this number by 15 to signify 15 standard deviations below the average weight (17.22 * 15): getting 258.3 pounds. 190 pounds is the average, we subtract this by 258.3 pounds and therefore get:

Average weight of a 24 inch tall man = -68.3 pounds. Since this weight is negative, which is not possible, this answer does not make sense.

d) 0 inches

You can not have a height and weight of 0. Any answer to this question would not make sense.

10. Ch 10 C2:

For the first-year students at a certain university. the correlation between SAT scores and first-year GPA was 0.60. The scatter diagram is football-shaped. Predict the percentile rank on the first-year GPA for a student whose percentile rank on the SAT was:

a) 90%

Looking at a table of z-score values, we know that 90th percentile is approximately a z-score of 1.3. Knowing that the correlation is 0.60, multiply that by the z-score (0.60 * 1.3), and you get 0.78.

```
pnorm(0.78)

## [1] 0.7823046
```

They would be about the 78th percentile in GPA with a 90th percentile in SAT.

b) 30%

Looking at a table of z-score values, we know that 30th percentile is approximately a z-score of -0.5. Knowing that the correlation is 0.60, multiply that by the z-score (0.60 * -0.5), and you get -0.30.

```
pnorm(-0.30)

## [1] 0.3820886
```

They would be about the 38th percentile in GPA with a 30th percentile in SAT.

c) 50%

Looking at a table of z-score values, we know that 50th percentile is approximately a z-score of 0. Knowing that the correlation is 0.60, multiply that by the z-score (0.60 * 0), and you get 0.

```
pnorm(0)

## [1] 0.5
```

They would be about the 50th percentile in GPA with a 50th percentile in SAT.

d) unknown

If the score is unknown then it would be the best guess to be about average, meaning that a person with an unknown percentile in SAT would be about 50th percentile in GPA.

11. Ch 10 Rev 3:

Pearson and Lee obtained the following results in a study of about 1,000 families: average height of husband = 68 inches, SD = 2.7 inches, average height of wife = 63 inches, SD = 2.5 inches, r = 0.25. Predict the height of a wife when the height of her husband is:

a) 72 inches

The height of a wife when the husband is 72 inches. 72 - 68 (average height of a husband) = 4. 4 (difference in height given - average height)/2.7 = about 1.48 SDs. Multiply the correlation coefficient (0.25) by the SD of a wife (2.5). You get 0.625 inches. Multiply this by the number of SDs away (1.48) and you get 0.925. Add this number to the average of a wife (63 inches) and get you get the estimate that:

63.93 inches is the height of the wife.

b) 64 inches

The height of a wife when the husband is 64 inches. 64 - 68 (average height of a husband) = -4. -4 (difference in height given - average height)/2.7 = about -1.48 SDs. Multiply the correlation coefficient (0.25) by the SD of a wife (2.5). You get 0.625 inches. Multiply this by the number of SDs away (-1.48) and you get -0.925. Add this number to the average of a wife (63 inches) and get you get the estimate that:

62.07 inches is the height of the wife.

c) 68 inches

The height of a wife when the husband is 68 inches. 68 - 68 (average height of a husband) = 0. 0 (difference in height given - average height)/2.7 = about 0 SDs. Multiply the correlation coefficient (0.25) by the SD of a wife (2.5). You get 0.625 inches. Multiply this by the number of SDs away (0) and you get 0. Add this number to the average of a wife (63 inches) and get you get the estimate that:

63 inches is the height of the wife.

d) unknown

Since this is unknown, the best guess that we could have is to assume the average meaning that 63 inches is the height of the wife.

ggplot Question

family.csv ggplot

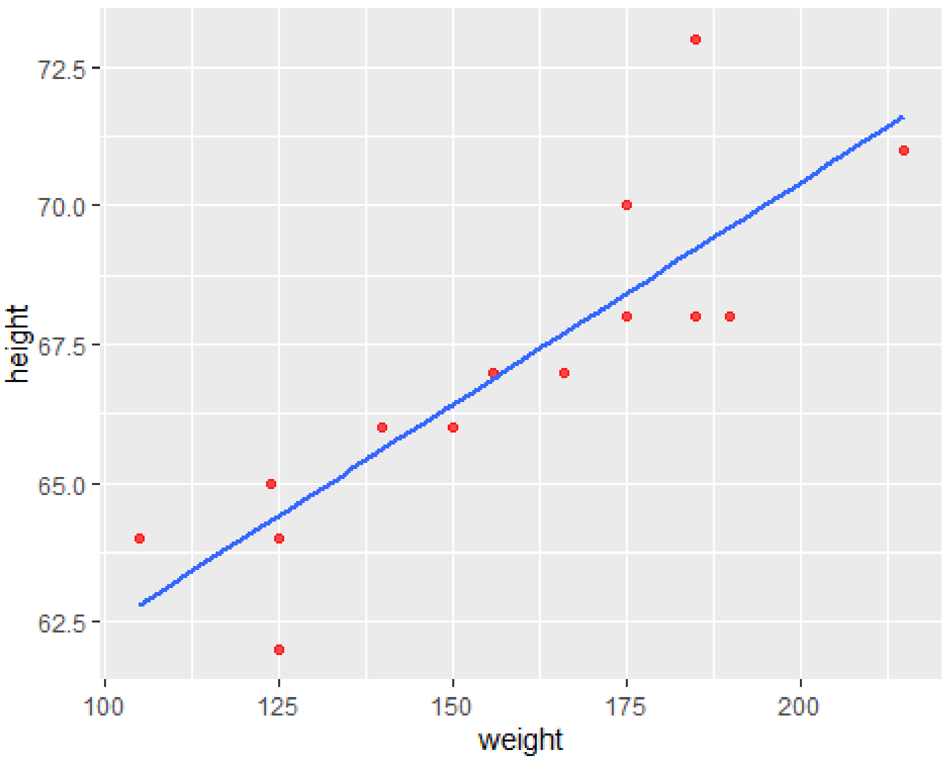
For the data in the file family.csv (in the Data folder under Files) make a plot using ggplot() with height on the horizontal axis and weight on the vertical axis. Look at lecture code from 4/8 for examples. Include this in a code chunk as in the .rmd file to make it appear in the file you submit.

Graph

```
library(ggplot2)
family = read.csv("C:\\Users\\sethc\\Documents\\STAT20\\Homework
9\\family.csv")

ggplot(family,aes(x=weight,y=height)) +
  geom_point(alpha=0.7,color="red")  +
  geom_smooth(method="lm", se=F)

## `geom_smooth()` using formula 'y ~ x'
```


72.5
70.0
67.5
65.0
62.5
height
100
125
150
175
200
weight