STAT20 Homework #10

Seth Metcalf

Table of Contents

Introduction

This is Homework #10, which contains questions from Chapters 10, 11, and 12. *Due 23 April 2021.*

Chapter 10

1. Ch 10 D2:

An instructor standardizes her midterm and final each semester so the class average is 50 and the SD is 10 on both tests. The correlation between the tests is around 0.50. One semester, she took all the students who scored below 30 at the midterm, and gave them special tutoring. They all scored above 50 on the final. Can this be explained by the regression effect?

Answer yes or no, and explain briefly.

No, the regression effect states that it will be closer to the average (of 50) because they all scored above the average (of 50) on the final then it can not be explained by this effect.

2. Ch 10 E2:

In Pearson's study, the sons of the 72-inch fathers only averaged 71 inches in height. True or false: if you take the 71-inch sons, their fathers will average about 72 inches in height.

Explain briefly.

False. There could be the case where many different inches of father's heights averaged to the son height of 71 inches and because of this it could be the case that the average son's height of 71 inches has a different average height of the father.

3. Ch 10 Rev 4:

In one study, the correlation between the educational level of husbands and wives in a certain town was about 0.50; both averaged 12 years of schooling completed, with an SD of 3 years.

a) Predict the educational level of a woman whose husband has completed 18 years of schooling.

Since 18 is 2 SDs (SD of 3) above the average of 12 years. Multiply the amount of standard deviations by the correlation coefficient (2 * 0.5) and you get the amount of SDs you expect the significant other to be (1 SD). Add this to the average (12 + 1 SD) and you get:

15 years is the average education that you expect a women to have.

b) Predict the educational level of a man whose wife has completed 15 years of schooling.

Since 15 is 1 SD (SD of 3) above the average of 12 years. Multiply the amount of standard deviations by the correlation coefficient (1 * 0.5) and you get the amount of SDs you expect the significant other to be (0.5 SDs). Add this to the average (12 + 0.5 SD) and you get:

13.5 years is the average education that you expect a women to have.

c) Apparently, well-educated men marry women who are less well educated than themselves. But the women marry men with even less education. How is this possible?

Most likely this is due to very high deviations on the LSRL line which causes there to be conflicting results especially with a high SD that this study carries.

Chapter 11

4. Ch 11 B3:

At a certain college, first-year GPAs average about 3.0, with an SD of about 0.5; the correlation with high-school GPA is about 0.6. Person A predicts first-year GPAs just using the average. Person B predicts first-year GPAs by regression, using the high-school GPAs.

Which person makes the smaller r.m.s. error? Smaller by what factor?

Since person A is just using the average then they are just using the SD to predict as well, meaning that they have the SD of y (0.5) as their eorr.

Person B is using regression, meaning that they are using the equation sqrt(1 - r^2) * SD of Y. Since sqrt(1 - r^2) = 0.8, multiplying this by the SD of y will yield 0.4 error, meaning that it is smaller using regression over just the average.

5. Ch 11 E1:

The following results were obtained for about 1000 families:

avg height of husband 68 inches, SD 2.7 inches;

avg height of wife 63 inches, SD 2.5 inches, r 0.25

a) What percentage of the women were over 5 feet 8 inches?

```
1 - pnorm(2.0)

## [1] 0.02275013
```

5’8" inches is exactly 2 SDs over the mean for women, we can use the pnorm function in order to calculate the Z-score and find that rougly 2.3% of women are over 5’8" tall.

b) Of the women who were married to men of height 6 feet, what percentage were over 5 feet 8 inches?

```
husbandavg = 68 # inches
wifeavg = 63 # inches
husbandSD = 2.7 # inches
wifeSD = 2.5 # inches
spousecorrelation = 0.25 # correlation coefficient
husbanddiff = 72 - husbandavg # 72 inches is 6 foot
husbandSDsAboveAvg = husbanddiff/husbandSD # 1.481481 SDs
wifeAdd = husbandSDsAboveAvg * spousecorrelation * wifeSD
wifeavg + wifeAdd

## [1] 63.92593
```

63.92593 is the predicted height of the wife

6. Ch 11 Rev 7:

The freshmen at a large university are required to take a battery of aptitude tests. Students who score high on the mathematics test also tend to score high on the physics test. On both tests, the average score is 60; the SDs are the same too. The scatter diagram is football-shaped. Of the students who scored about 75 on the mathematics test:

i) Just about half scored over 75 on the physics test.

ii) More than half scored over 75 on the physics test.

iii) Less than half scored over 75 on the physics test.

Choose one option and explain.

Because the graph is football shaped we know that it is not perfect (1.0) correlation coefficient. Knowing this, since both averages and SDs are the same on the tests, we know that even if they got a 75 on the math test, the average student will still score slightly lower than 75 on the physics test even with strong correlation (0.9).

Chapter 12

7. Ch 12 A3:

Here are summary statistics for heights of fathers and sons:

avg height of fathers 68 inches, SD 2.7 inches;

avg height of sons 69 inches, SD 2.7 inches, r 0.5

a) Find the regression equation for predicting the height of a son from height of father.

Slope of the regression line is given by r * SD_y/SD_x. This is 0.5 * 2.7/2.7 = 0.5.

Y-intercept is given by the mean of y - (slope * mean of x) = 69 * (0.5 * 68) = 69 - 34 = 35

This regression line is given by: y = 35 + 0.5x

b) Find the regression equation for predicting the height of a father from height of son.

Slope of the regression line is given by r * SD_y/SD_x. This is 0.5 * 2.7/2.7 = 0.5.

Y-intercept is given by the mean of y - (slope * mean of x) = 68 * (0.5 * 69) = 68 - 34.5 = 33.5

This regression line is given by: y = 33.5 + 0.5x

8. Ch 12 A3 (Continued):

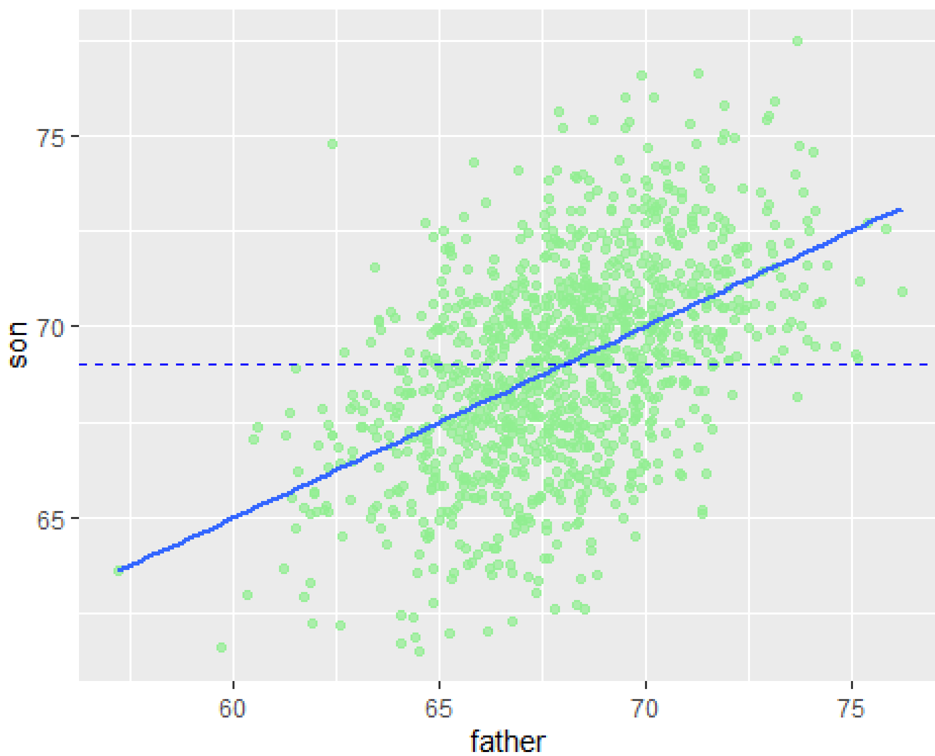
Continuing the previous problem, load the data HW10.fatherson.csv (in the Data folder under Files) and plot the points using ggplot(), including the regression line for predicting son’s height from father’s height, using geom_smooth() without the confidence interval (as in lecture code for 4/15 or 4/20). Also add a horizontal line through the average son's height. Verify that you get the same equation for the regression line as you did using formulas. The data is simulated, but has the same summary statistics as above.

Graph:

```
library(ggplot2)
fatherson = read.csv("C:\\Users\\sethc\\Documents\\STAT20\\Homework 10\\HW10.fatherson.csv")
sonmean = mean(fatherson$son)

ggplot(fatherson,aes(x=father,y=son)) +
  geom_point(alpha=0.7,color="light green")  +
  geom_smooth(method="lm", se=F)  +
  geom_hline(yintercept = sonmean, linetype="dashed", color = "blue")

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



9. Ch 12 B1:

For a large sample of men, the regression equation for predicting height from education is

predicted height = (0.25 inches per year) x (education) + 66.75 inches

Predict the height of a man with 12 years of education; with 16 years of education.

With 12 years of education:

(0.25 * 12) + 66.75 inches = 69.75 inches.

With 16 years of education:

(0.25 * 16) + 66.75 = 70.75 inches.

Does going to college increase a man's height? Explain.

Even though there seems to be a correlation between a man's height and college years, it does not mean that going to college increases a man's height. It could be an external factor, like the more attractive a man is the more likely he is to be successful (which has been shown many times), and this could explain the fact that the more a man has a better education the higher his height is.

10. Ch 12 Rev 7:

A statistician is doing a study on a group of undergraduates. On average, these students drink 4 beers a month, with an SD of 8. They eat 4 pizzas a month, with an SD of 4. There is some positive association between beer and pizza, and the regression equation is:

predicted number of beers = ___ x number of pizzas + 2.

However, the statistician lost the data and forgot the slope of the equation. (Perhaps he had too much beer and pizza.) Can you help him remember the slope?

Explain.

We know the equation y-intercept = y_mean - (slope * x_mean). The y-intercept is given to us through the equation (2). And the means for both of them (4 beers amd 4 pizzas) are also given to us in the original problem.

If we rearrange the equation a bit, (subtract he y_mean from both sides and divide by the x_mean), we can get the equation that the -slope = (y-intercept - y_mean)/x_mean. Plugging in the numbers that were mentioned before and we can get the equation -slope = (2 - 4)/4, meaning that the -slope = -0.5, multiply both sides by -1 and we get:

the slope = 0.5.