

Psych 315, Winter 2021, Homework 4 Answer Key

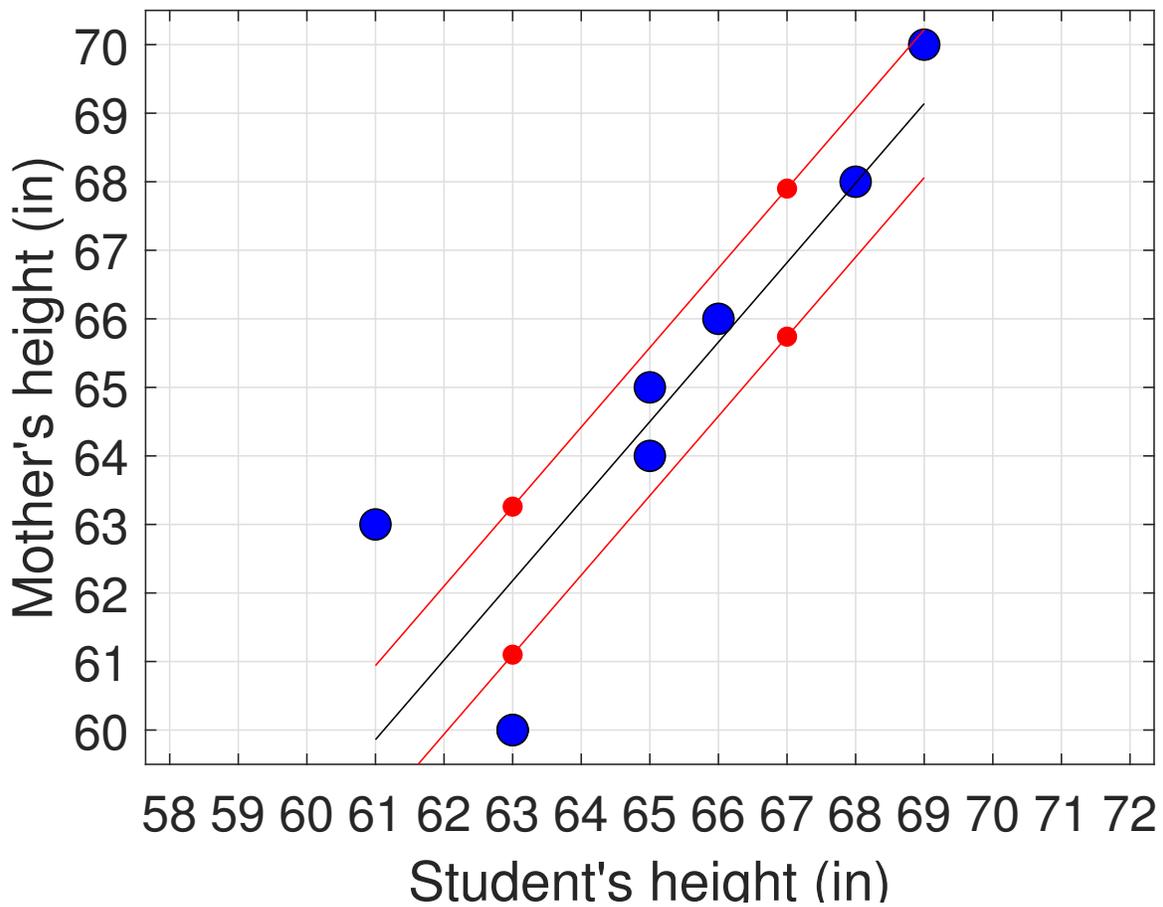
Due Friday, January 29th by 5pm.

Name \_\_\_\_\_ ID \_\_\_\_\_

Section [AA] (Natalie), [AB] (Natalie), [AC] (Ryan), [AD] (Ryan), [AE] (Kelly), [AE] (Kelly)

The scatterplot below plots Female students heights and their mother's heights for the 8 students who chose Yellow as their favorite color.

Round all answers to 2 decimal places.



1) Use R to load in the survey data, select the 8 students who chose Yellow as their favorite color, and calculate:

$\bar{x}$ : mean of student's heights  $\bar{y}$ : mean of mother's heights  $s_x$ : standard deviation of student's heights  $s_y$ : standard deviation of mother's heights  $r$ : correlation between students and mother's heights

Giant Hint: here's how to do this for the female students that chose 'Blue' as their favorite color:

```
# Load the survey data
survey <- read.csv("http://www.courses.washington.edu/psy315/datasets/Psych315W21survey.csv")
# Find female students that chose "Blue"
student.blue <- survey$gender == "Female" & survey$color=="Blue"
# Find the heights of these students (call it 'x'):
x <- survey$height[student.blue]
# Find their mother's heights (call it 'y')
y <- survey$mheight[student.blue]
# Find where there not NA's in both x and y:
goodId <- !is.na(x) & !is.na(y)
# Only include these pairs
x <- x[goodId]
y <- y[goodId]
# Means of x and y:
mx <- mean(x);
my <- mean(y);
# Standard deviations of x and y
sx <- sd(x);
sy <- sd(y);
# Correlation of x and y:
r <- cor(x,y)
mx
[1] 64.4375
my
[1] 63.625
sx
[1] 2.758418
sy
[1] 3.034745
r
[1] 0.6136708
```

2) Use R or your calculator to find the equation of the regression line and draw it by hand on the scatterplot.

$$m = r\left(\frac{s_y}{s_x}\right) \text{ and the y-intercept is: } b = \bar{Y} - (m)(\bar{X})$$

Here's how to do it in R:

```
# slope:
m <- r*sy/sx
# intercept:
b <- my - m*mx
mean of x: 65, mean of y: 64.5
sx = 2.5, sy = 3.32
Slope: m = (0.87)  $\frac{3.32}{2.5} = 1.16$ 
Intercept: b = 64.5 - (1.16)(65) = -10.9
Y = 1.16X + -10.9
```

3) Use R or your calculator to find the standard error of the estimate by calculating the sum of the squared residuals:

$$S_{yx} = \sqrt{\frac{\sum(Y - Y')^2}{n}}$$

in R:

```
# Find y on the regression line for every value of x:
yprime <- m*x+b
# Find the residuals:
residual <- y-yprime
# Use the residuals to calculate syx:
syx <- sqrt(sum( (y-yprime)^2)/length(x))
```

$$Y' = 64.5, 62.18, 62.18, 67.98, 65.66, 64.5, 59.86 \text{ and } 69.14$$

$$Y - Y' = 0.5, -2.18, -2.18, 0.02, 0.34, -0.5, 3.14 \text{ and } 0.86$$

$$\sum(Y - Y')^2 = 0.25 + 4.75 + 4.75 + 0 + 0.12 + 0.25 + 9.86 + 0.74 = 20.72$$

$$S_{yx} = \sqrt{\frac{20.72}{8}} = 1.61$$

4) Use the correlation as another way of calculating the standard error of the estimate. Your answer should be close, but not exactly the same due to rounding error.

$$S_{yx} = S_y \sqrt{1 - r^2}$$

$$(3.32) \sqrt{1 - 0.87^2} = 1.64 \text{ inches}$$

5) Use the regression line to predict the mother's height for a Female student that is 63 inches tall.

$$Y = mX + b = (1.16)(63) + -10.9 = 62.18 \text{ inches}$$

6) Assuming homoscedacity, find the range of mother's heights that covers the middle 50% of the heights of mothers of women that are 63 inches tall. Hint: The heights of the mothers of women that are 63 tall should be distributed normally with a mean determined by the regression line (problem 5) and a standard deviation equal to the standard error of the estimate (problem 4).

The Mother's heights should be distributed normally with a mean of 62.18 and a standard deviation of 1.61

Using table A, the z-scores covering the middle 50 percent of the normal distribution is  $z = +/- 0.67$

Converting to heights, the range is between  $62.18 - (0.67)(1.61)$  and  $62.18 + (0.67)(1.61)$  which is between 61.1 and 63.26 inches.

7) Repeat problems 5 and 6 but for students that are 67 inches tall. Note, because of homoscedasticity, the range above and below the predicted height should not change.

$$Y = mx + b = (1.16)(67) + -10.9 = 66.82 \text{ inches}$$

The Mother's heights should be distributed normally with a mean of 66.82 and a standard deviation of 1.61

Using table A, the z-scores covering the middle 50 percent of the normal distribution is  $z = +/- 0.67$

Converting to heights, the range is between  $66.82 - (0.67)(1.61)$  and  $66.82 + (0.67)(1.61)$  or between 65.74 and 67.9 inches

8) You should see that for any Female student's height, the middle 50% of the corresponding mothers heights should fall within the same range above and below the regression line.

Draw two parallel lines on the scatterplot, one above and one below the regression line that should cover the middle 50% of the mother's heights. Use the values from problems 6 and 7 as points on the lines.

9) Look at the scatterplot and calculate the actual percent of data points that fall between these two parallel lines. How close does it match to 50%?

5 of the 8 points fall between the parallel lines

This is  $100 \frac{5}{8} = 62.5$  percent of the points.

This is pretty close.

**10)** The correlation between SAT scores and IQ is around 0.5. Assume that SAT scores are normally distributed with a mean of 915 and a standard deviation of 88.24, and IQ scores are normally distributed with a mean of 100 and a standard deviation of 15.

**a)** Find the equation of the regression line that predicts IQs from SAT score. Hint: use the equations from problem 2. Give your answer in slope-intercept form.

Let X be SAT scores, and Y be IQ

The slope is  $r \frac{s_y}{s_x} = (0.5) \frac{15}{88.24} = 0.08$

The line goes through the means, so:

$$Y = (0.08)(X-915) + 100$$

$$Y = (0.08)X + 26.8$$

**b)** What is the expected IQ of a student with a SAT score of 1000?

$$\text{IQ} = (0.08)(1000) + 26.8 = 106.8$$

**c)** What is the proportion of variance of Y explained by X (the coefficient of determination)?

The coefficient of determination is  $r^2 = 0.5^2 = 0.25$

**d)** What is the total variance in the IQ scores?

The variance is the standard deviation squared:  $15^2 = 225$

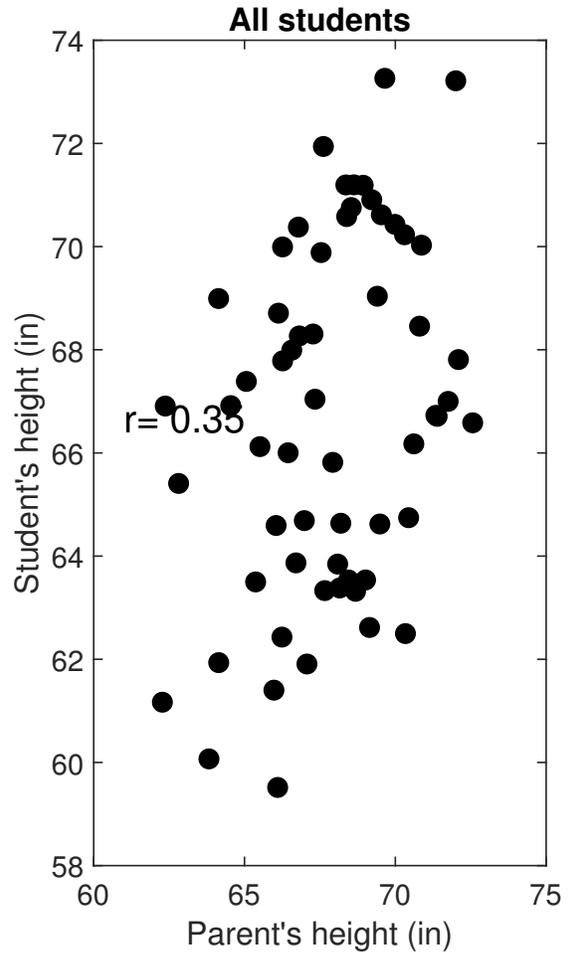
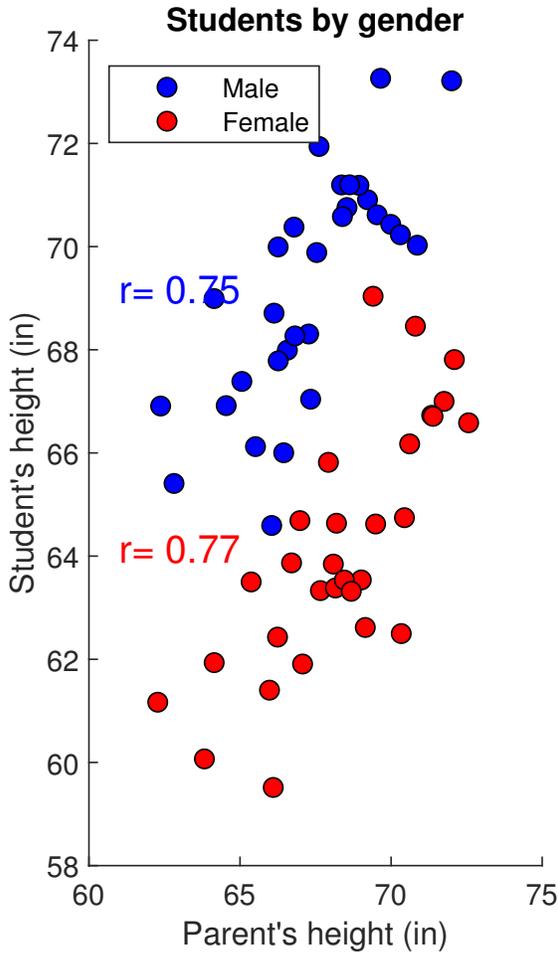
**e)** From parts c and d, calculate the amount of variance in IQ scores that is explained by SAT scores.

The amount of variance explained by SAT scores is equal to the total amount of variance in SAT scores multiplied by the proportion of variance accounted for, which is  $r^2$ .

$$(225)(0.25) = 56.25$$

11) Explain why the correlation between parent's heights and all student's heights might be lower than for the correlations you'd find for just the female or male students. Draw a picture if it helps.

While there may be a strong correlations within each gender combining students leads to added variance in the student's heights that is not explained by the parent's height. This leads to an overall lower correlation for the whole group than for the correlations within each gender.



12) Explain why the correlation between student's heights and video game playing time might be stronger for the whole group than for the correlations within male and female students. Again, draw a picture if it helps.

Suppose there is no correlation between height and video game playing within each gender. But since men play games more than women, and men are taller than women, the combined distribution is correlated.

