Brooke Wheeler
2/9/22
Assignment #4

Due: Wednesday February 9, 2022 by **5 PM ET**

Please sign your name to the **appropriate** space below**.** Remember that you are permitted to receive (and provide) authorized assistance but must acknowledge it if you do.

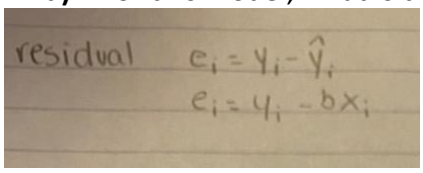
I received assistance on this assignment and/or discussed it with fellow classmates or a tutor.
Brooke Wheeler **recieved help from TA**_____
I received no assistance on this assignment and/or did not discuss it with anyone other than Professor Miller.
Brooke Wheeler

Part 1: The Regression Parameters

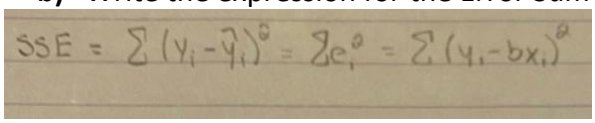
Instructions: In these problems we are going to focus on the regression parameters, β_0 and β_1, and their estimators, b_0 and b_1. Provide complete justification for each problem. You may type your work or include work done by hand.

1. You are given observations $X_1, X_2, X_3, \dots, X_n$ for the explanatory variable and observations $Y_1, Y_2, Y_3, \dots, Y_n$ for the response variable. The relationship between the response variable Y and the explanatory variable X can be described by the linear model $\hat{Y}_i = bX_i$.

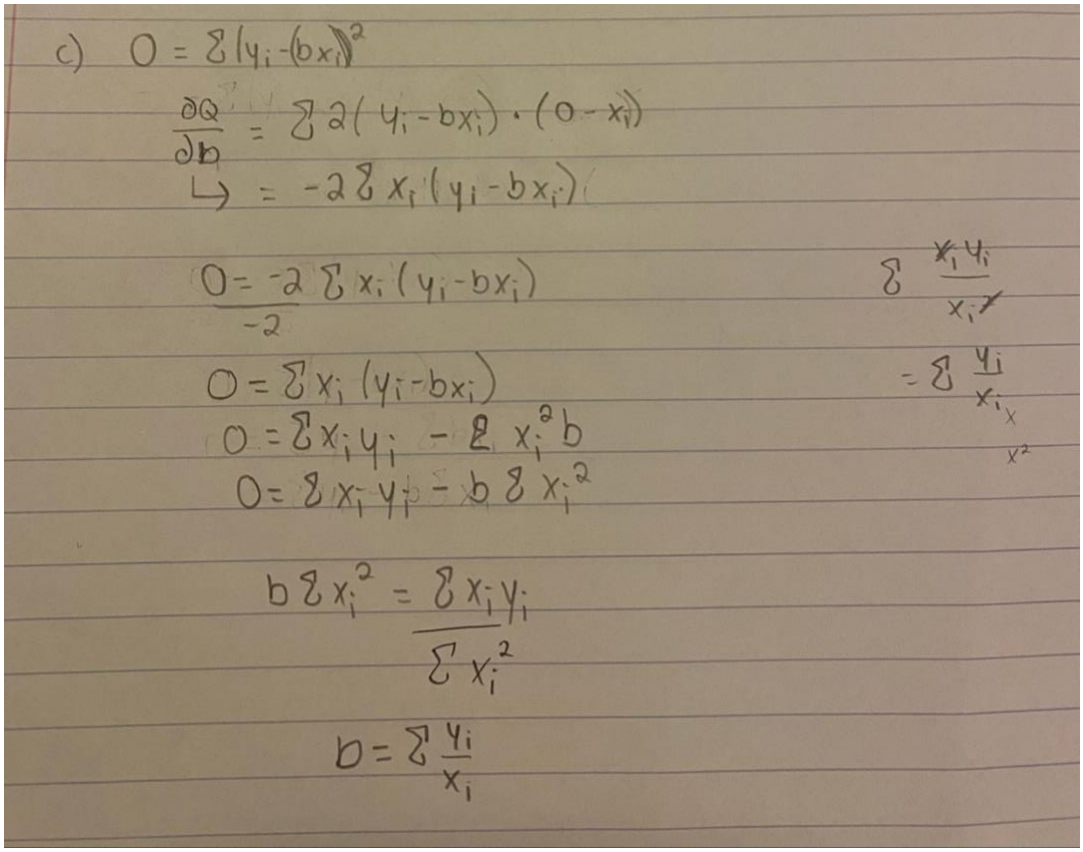
a) For this model, what is the formula for the residual, e_i?



b) Write the expression for the Error Sum of Squares, $SSE = Q(b_0, b_1)$, for this model.



c) Derive the formula for the least squares estimate of b. (Find the value of b which minimizes the SSE. This is like what we derived in the notes, except in this case we only have to minimize the SSE with respect to one variable, b.)



Part 2: Inferences in Regression

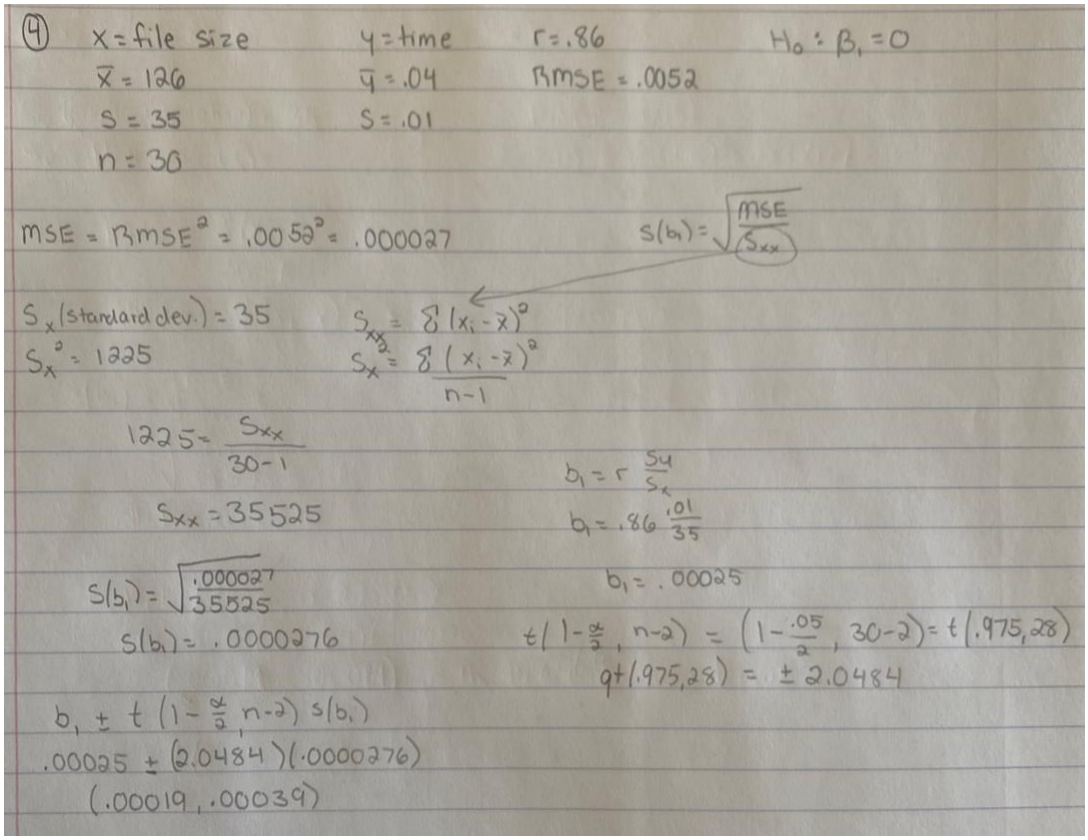
Instructions: For each question below, you must show all work. Use R to find p-values and critical values. You may type your work or include work done by hand.

2. (Continued from Assignment 3) The time it takes to transmit a file always depends on the file size. Suppose you transmitted 30 files, with the average size of 126 Kbytes and the standard deviation of 35 Kbytes. The average transmittance time was 0.04 seconds with the standard deviation of 0.01 seconds. The correlation coefficient between the time and the size was 0.86.

In the previous homework, we fit a regression model that predicted the time it will take to transmit a 400 Kbyte file. According to this model, the standard deviation of responses is estimated by

$$s = \sqrt{MSE} = s_Y \sqrt{\frac{n-1}{n-2}(1-r^2)} = 0.0052.$$

a) Construct a 95% confidence interval for the regression slope.

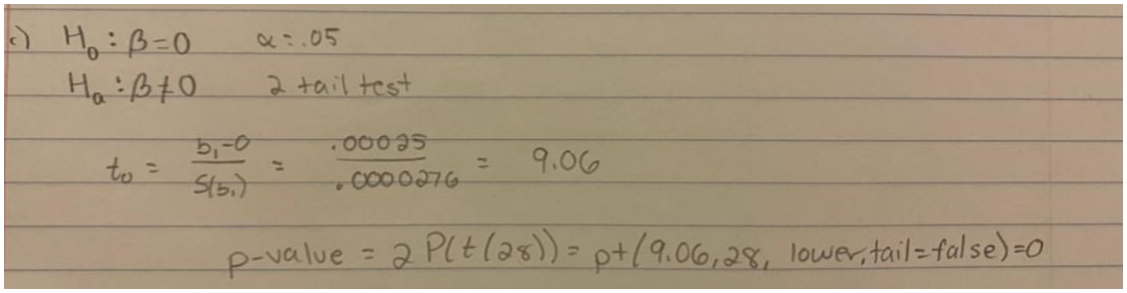


The 95% confidence interval is from .00019 to .00039.

b) Based on this interval, is the slope significant at the 5% level?

i. Based on the interval the slope is significant at the 5% level. Since zero is not included in the confidence interval we can reject the null hypothesis at the level of .05 that there is not linear association between file size and time it takes to send the file. Meaning we can support that there is a linear association.

c) State the null and alternative hypotheses that would be used in b). Calculate the test statistic and the p-value.



d) When you answered questions b) and c), it was correct to conduct a two-sided test. However, in this given example, why does it make more sense to consider an upper-tail alternative?

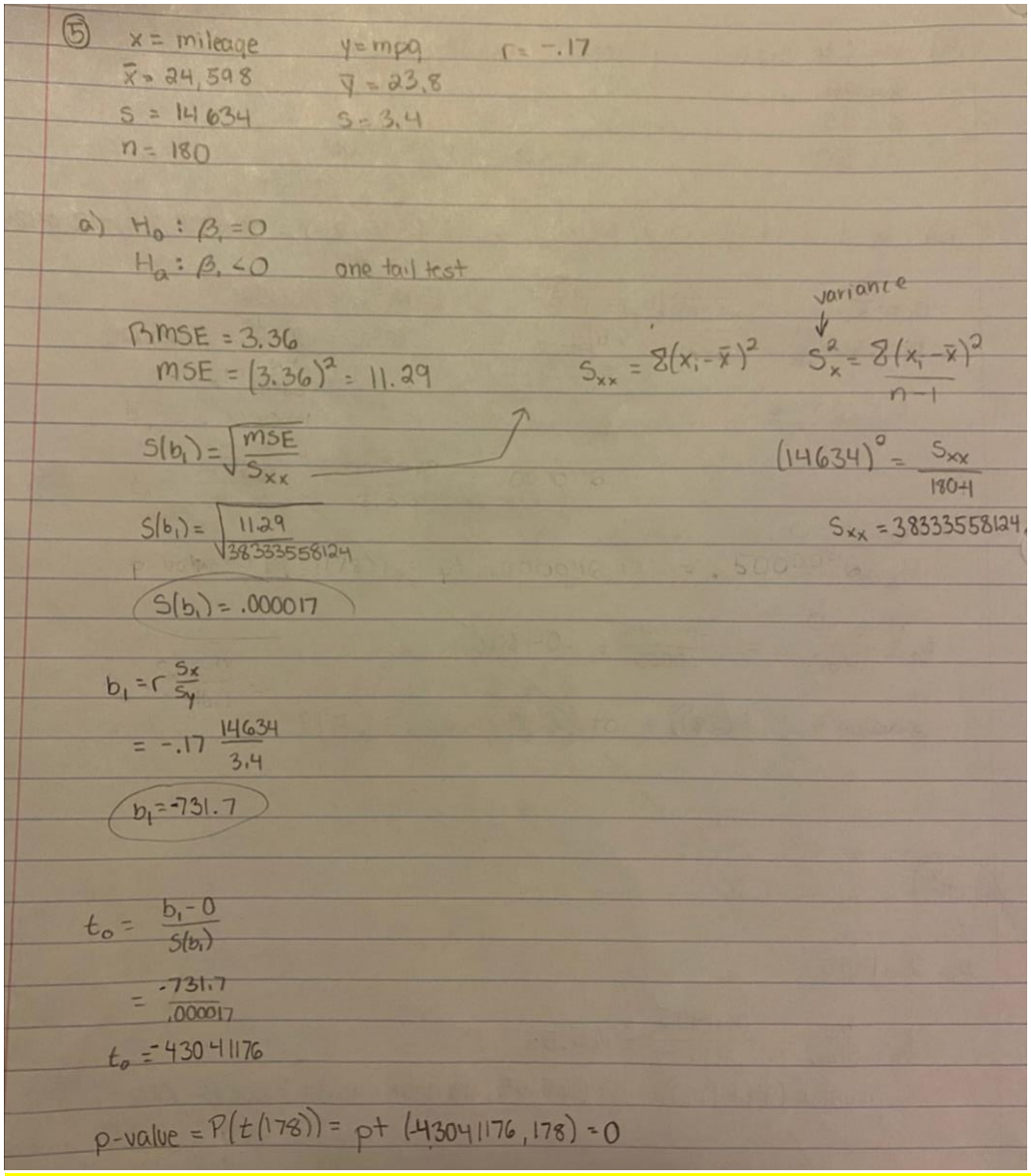
i. In this example it would make sense that file size effects the time it takes to send the file and it would make sense that the bigger the file size the longer it takes. Above we only found evidence to support the hypothesis that there is a linear association between file size and time it takes to send the file. If we did a upper-tail test our alternative hypothesis would be that the slope is greater than 0, meaning that there is a positive slope, that the larger the file size the longer it takes to send the file.

3. (Continued from Assignment 3) At a gas station, 180 drivers were asked to record the mileage of their cars and the number of miles per gallon. The results are summarized in the table.

	Sample mean	Standard deviation
Mileage	24,598	14,634
Miles per gallon	23.8	3.4

The sample correlation coefficient is $r = -0.17$. In the previous assignment, we fit a regression model that described how the number of miles per gallon depends on the mileage. According to this model, the standard deviation of responses is estimated by $s = \sqrt{MSE} = s_Y\sqrt{\frac{n-1}{n-2}(1-r^2)} = 3.36$.

a) Do the given data present a significant evidence that cars with higher mileage are less economic? Formulate appropriate null hypothesis and alternative and conduct the test.



Since the p-value is less than alpha=.05 we reject the null hypothesis and support that there is a negative association between mileage and mpg. Meaning that as cars have higher mileage they get less mpg.

Part 3: Mini-Project

Instructions: You will need R to answer the following questions. For each, provide any relevant R output and clearly state a conclusion with full support for your answer.

4. (*based on text **p. 90-91: 2.4***)

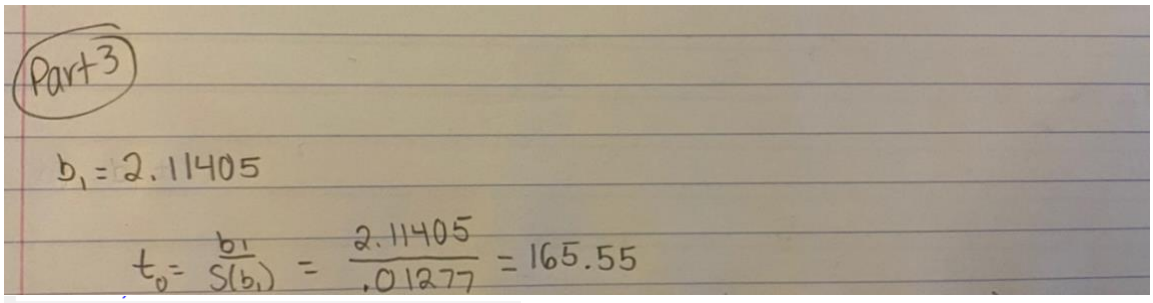
Grade point average (This data set was already used in Assignment 3 and is posted on Canvas).

a) Obtain a 99% confidence interval for β_1. Does it include zero? Why might the director of admissions be interested in whether the confidence interval includes zero?

```
gpa_reg <- lm(V1 ~ V2, data = gpa)
gpa_reg

# the intercept= 2.11405
# the slope is .03883

confint(gpa_reg, level = .99)
# The 99% confidence interval for B_1 is .00538 to .07227. The confidence
interval does not include zero meaning that we reject the null hypothesis
that there is no linear association between ACT scores and gpa at the end
of freshman year. The director of admission would want to see if students
gpa's are effected by their performance on the ACT, they then could use
those scores to predict gpa of students.

```

```
Call:
lm(formula = V1 ~ V2, data = gpa)

Coefficients:
(Intercept)           V2
    2.11405      0.03883

                  0.5 %     99.5 %
(Intercept) 1.273902675 2.95419590
V2          0.005385614 0.07226864
```

b) Test whether or not a linear association exists between student's ACT score (X) and GPA at the end of the freshman year (Y). Use a level of significance of 0.01. State the hypotheses, p-value, decision, and conclusion.




```
# null hypothesis: slope=0
# alt. hypothesis: slope does not = 0
# two tail
summary(gpa_reg)

2*(pt(165.55, 118, lower.tail = FALSE))
```

```
[1] 1.456486e-141
```

Since the p-value is less than alpha = .01 we reject the null hypothesis that there is no association between gpa and ACT scores and support that there is an association between ACT scores and GPA at the end of freshman year.