Brooke Wheeler
1/28/22
Assignment #3

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.

____

I received no assistance on this assignment and/or did not discuss it with anyone other than Professor Miller.

Brooke Wheeler

Part 1: Working with Linear Regression Models

Instructions: For each question below, you must provide full explanations (**completely in your own words**) and show all work. You may type your work or include work done by hand.

Note: In class, and in the text, we wrote the least-squares estimator for slope as $b_1 = \frac{s_{xy}}{s_{xx}}$. An equivalent, and often more convenient, way to write it is $b_1 = r\frac{s_y}{s_x}$ where r is the sample correlation coefficient.

1. (***text* p. 33: 1.2**) The members of a health spa pay annual membership dues of $300 plus a charge of $2 for each visit to the spa. Let Y denote the dollar cost for the year for a member and X the number of visits by the member during the year. Express the relation between X and Y mathematically. Is it a functional relation or a statistical relation? Why?
   a) Y= 300 +2x. This is a functional relation since we know exactly how X relates to Y. We know that with each additional visit to the spa the charge goes up $2. This relationship can be described as Y= 300 +2x. All the membership dues will fall on this line perfectly.

2. **(based on *text* p. 37:1.29, 1.30)**

   a) Refer to regression model $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$. Assume that $X = 0$ is within the scope of the model. What is the implication for the regression function if β_0=0? How would the regression function plot on a graph?

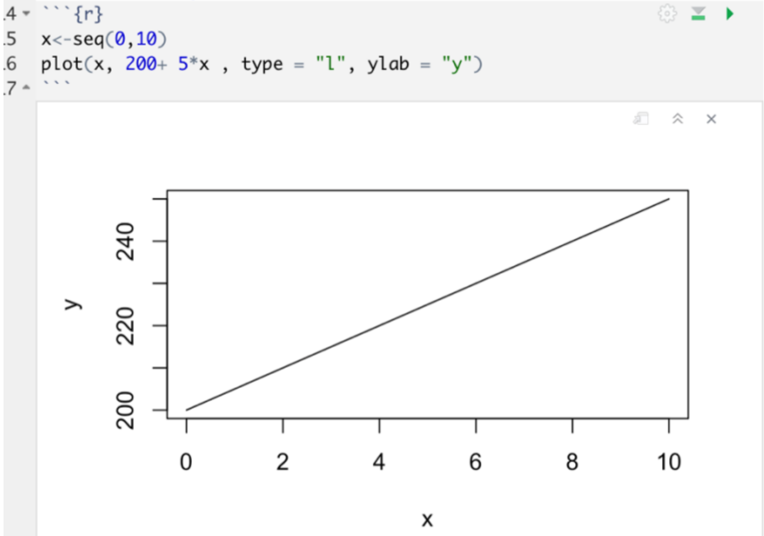
a. If the intercept is zero that means the regression line would go through the origin (0,0) and if X is zero that means it does not affect the dependent variable, Y. So the regression line would just be plotted as the given Y value.

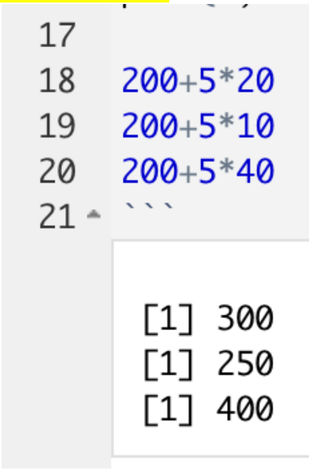
b) Refer to regression model $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$. What is the implication for the regression function if β_1=0? How would the regression function plot on a graph?

a. If the slope coefficient is zero that would make the slope*X value also equal to zero. This means that the regression line would be a straight line at the given intercept value.

3. (**Based on *text* p. 33: 1.6**) Suppose the regression parameters are $\beta_0 = 200$ and $\beta_1 = 5$.

a) Plot the regression equation. (You may do this with R or by hand.)



~~~
```{r}
x<-seq(0,10)
plot(x, 200+ 5*x , type = "l", ylab = "y")
~~~
```

y
200
220
240
0
2
4
6
8
10
x

i.

b) Predict the responses for X = 10, 20, and 40.

i. X=10, Y=250

ii. X=20, Y=300

iii. X=40, Y= 400



```
200+5*20
200+5*10
200+5*40
```
```

```
[1] 300
[1] 250
[1] 400
```

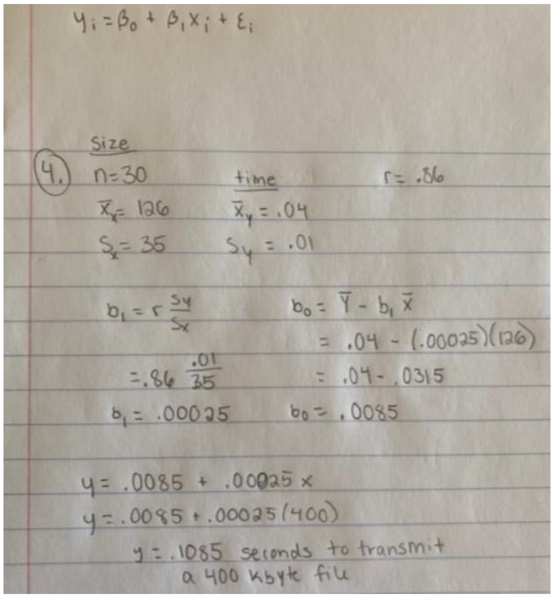
iv.

c) Explain the meaning of parameters β_0 and β_1.

i. β_0 represents the y-intercept, meaning that when x is zero the regression line crosses the y-axis at 200. β_1 represents the slope of the line so for every unit of change in X the value of y increases by 5.

4. 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.

Based on this data, fit a linear regression model and predict the time it will take to transmit a 400 Kbyte file.



$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

4.

Size	time	
$n = 30$		$r = .86$
$\bar{x}_x = 126$	$\bar{x}_y = .04$	
$S_x = 35$	$S_y = .01$	

$b_1 = r \frac{Sy}{Sx}$

$= .86 \frac{.01}{35}$

$b_1 = .00025$

$b_0 = \bar{Y} - b_1 \bar{x}$

$= .04 - (.00025)(126)$

$= .04 - .0315$

$b_0 = .0085$

$y = .0085 + .00025x$

$y = .0085 + .00025(400)$

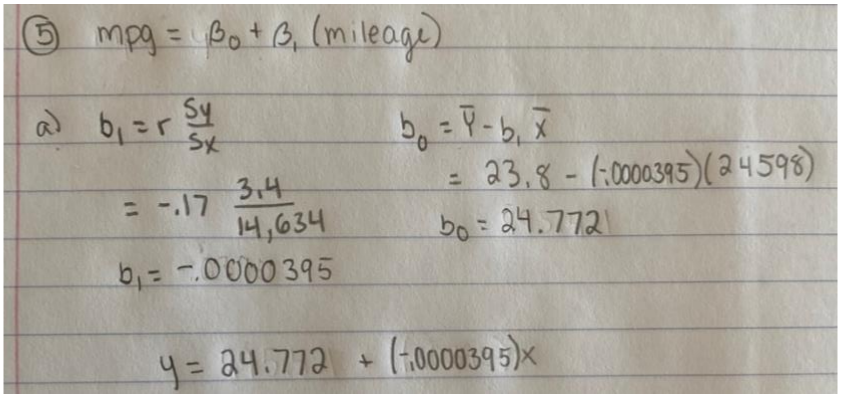
$y = .1085$ seconds to transmit a 400 kbyte file

5. 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 $r = -0.17$.

a) Compute the least squares regression line which describes how the number of miles per gallon depends on the mileage.



⑤ $mpg = \beta_0 + \beta_1 (\text{mileage})$

a) $b_1 = r \frac{s_y}{s_x}$

$= -.17 \frac{3.4}{14{,}634}$

$b_1 = -.0000395$

$b_0 = \bar{y} - b_1 \bar{x}$

$= 23.8 - (-.0000395)(24598)$

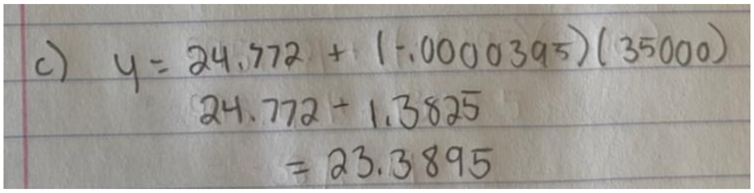
$b_0 = 24.772$

$y = 24.772 + (-.0000395)x$

b) What do the obtained slope and intercept mean in this situation?

i. The slope is equal to -.0000395 meaning that if the mileage increases by 1 mile then the average miles per gallon goes down by .0000395, or that if the mileage of a car goes up by 100,000 miles the average miles per gallon goes down by 3.95. The y intercept is equal to 24.772 which represents that an estimated average of 24.772 miles per gallon when a car has zero mileage.

c) You purchase a used car with 35,000 miles on it. Predict the number of miles per gallon.



c) $y = 24.772 + (-.0000395)(35000)$

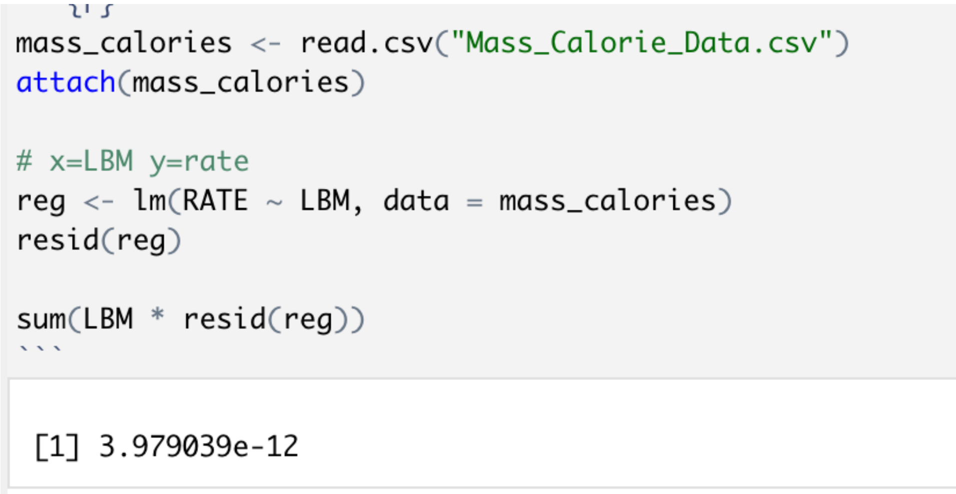
$24.772 + 1.3825$

$= 23.3895$

Part 2: More Properties of the LSRL

Instructions: In these problems, you will be working with the data set ***Mass_Calorie_Data.csv*** posted on Canvas. It is the data we used in some examples in class pertaining to twelve womens' lean body masses (in kg) and rates of burning calories (in calories per day). Load the data set into R and use the software to show that the properties below are satisfied. Include all relevant input, output, and explanations. [x= LBM and y=rate]

6. When the residuals are weighted by the level of the predictor variable in the i^{th} trial, the sum is zero. $\sum_{i=1}^{n} X_i e_i = 0$

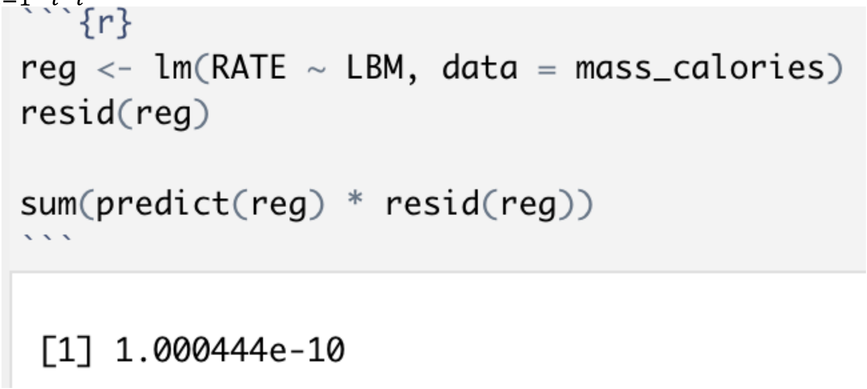


```
{r}
mass_calories <- read.csv("Mass_Calorie_Data.csv")
attach(mass_calories)

# x=LBM y=rate
reg <- lm(RATE ~ LBM, data = mass_calories)
resid(reg)

sum(LBM * resid(reg))
```

```
[1] 3.979039e-12
```

a)

7. When the residuals are weighted by the predicted/ fitted value in the i^{th} trial, the sum is zero.$\sum_{i=1}^{n}\hat{Y}_i e_i = 0$



~~~
```{r}
reg <- lm(RATE ~ LBM, data = mass_calories)
resid(reg)

sum(predict(reg) * resid(reg))
~~~

```
[1] 1.000444e-10
```

a)

Part 3: Mini-Project

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

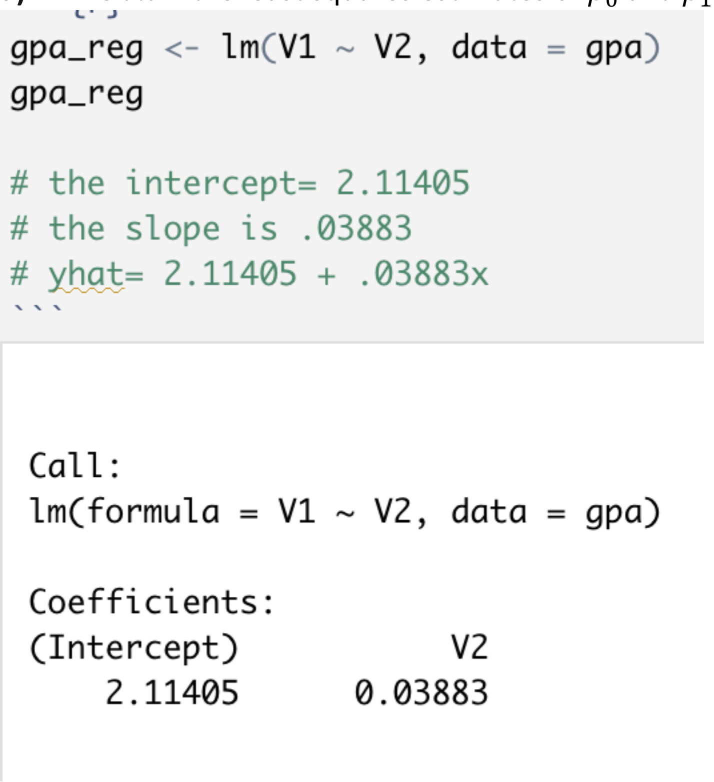
.

8. (based on *text* p. 35-36: **1.19, 1.23**). **Grade point average**. The director of admissions of a small college selected 120 students at random from the new freshman class in a study to determine whether a student's grade point average (GPA) at the end of the freshman year (Y) can be predicted from the ACT test score (X). The results of the study follow.

i	1	2	3	⋯	118	119	120
X_i	21	14	28	⋯	28	16	28
Y_i	3.897	3.885	3.778	⋯	3.914	1.860	2.948

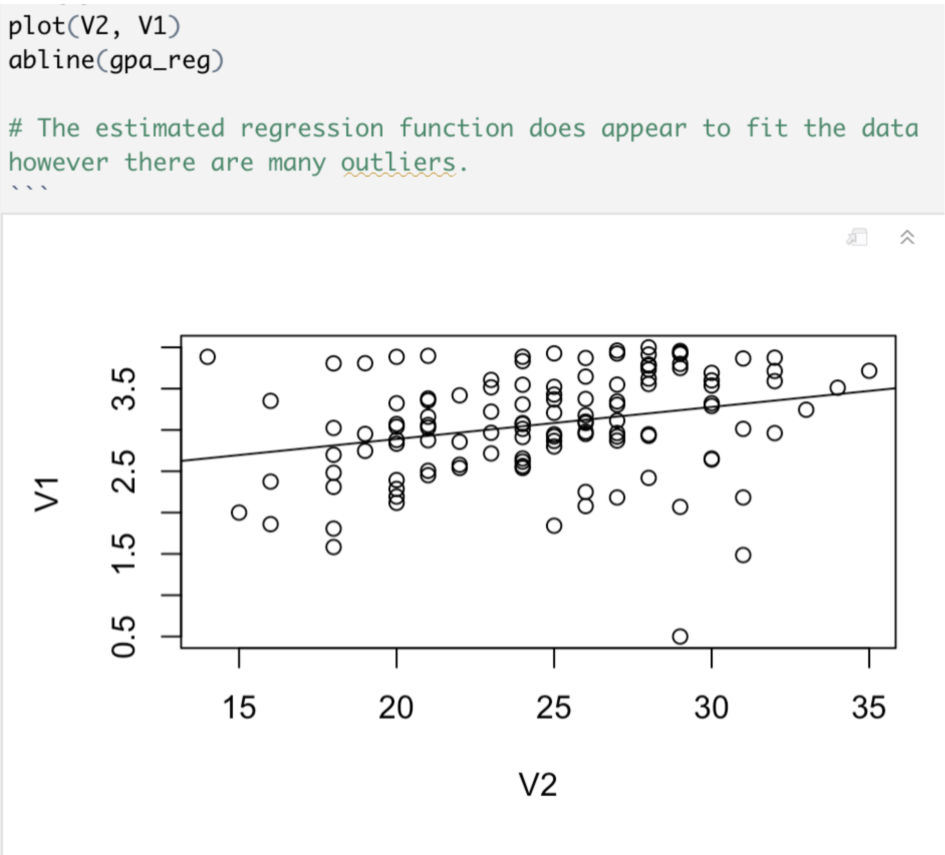
The full data set is available on Canvas. To read a text (ASCII) file, you can use an R command read.table("**CH01PR19.txt**")

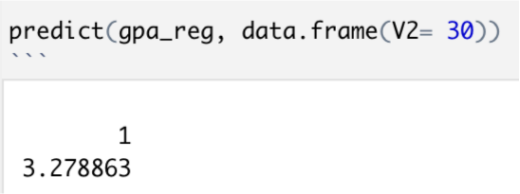
a) Obtain the least squares estimates of β_0 and β_1 and state the estimated regression function.



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

# the intercept= 2.11405
# the slope is .03883
# yhat= 2.11405 + .03883x
```
```

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

Coefficients:
(Intercept)           V2
    2.11405      0.03883
```

b) Plot the estimated regression function and the data. Does the estimated regression function appear to fit the data well?
```



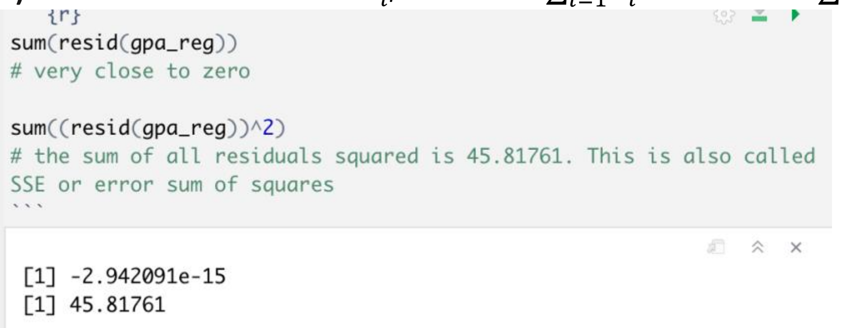
```
plot(V2, V1)
abline(gpa_reg)

# The estimated regression function does appear to fit the data
however there are many outliers.
```
```


V1
0.5
1.5
2.5
3.5
15
20
25
30
35
V2


**c)** Obtain a point estimate of the mean freshman GPA for students with ACT test score $X = 30$.



```
predict(gpa_reg, data.frame(V2= 30))
```
```

```
       1
3.278863
```

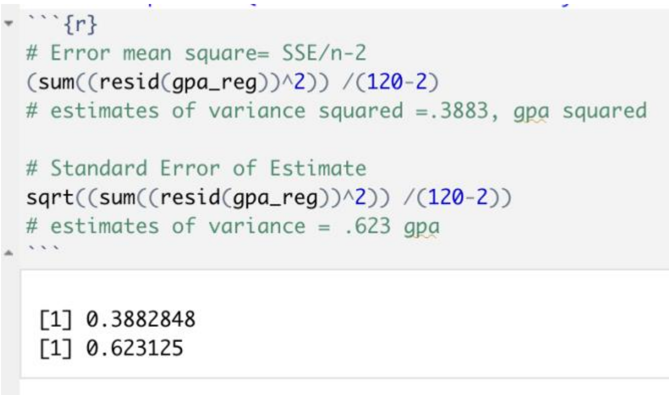
d) What is the point estimate of the change in the mean response when the entrance test score increases by one point?

a When the mean test score increases by one point the average gpa increases by .03883

e) Obtain the residuals $e_{i.}$ Show that $\sum_{i=1}^{n} e_i = 0$ and find $\sum_{i=1}^{n} {e_i}^2$.



```
{r}
sum(resid(gpa_reg))
# very close to zero

sum((resid(gpa_reg))^2)
# the sum of all residuals squared is 45.81761. This is also called
SSE or error sum of squares
```
```

```
[1] -2.942091e-15
[1] 45.81761
```

f) Obtain point estimates of σ^2 and σ. In what units is each of them expressed (include those units with your answers)?
```



~~~
```{r}
# Error mean square= SSE/n-2
(sum((resid(gpa_reg))^2)) /(120-2)
# estimates of variance squared =.3883, gpa squared

# Standard Error of Estimate
sqrt((sum((resid(gpa_reg))^2)) /(120-2))
# estimates of variance = .623 gpa
```

 [1] 0.3882848
 [1] 0.623125
~~~