HW9

2024-11-11

Stat 5000 Homework #9
Fall 2024 due Fri, November 15th @ 11:59 pm Name: Sam Olson
Collaborators: Craig, Ethan, **The Hatman**

Q1

(a)

Suppose that six observations of the yield (Y) of a chemical process were taken at each of four temperature levels (X) for running the process, but you are only given information on the sample means and standard deviations for the observed yields at each temperature. The summary data are

Temperature (°C)	Sample Mean	Sample Variance	Sample Size
150	66	1.15	6
200	81	1.00	6
250	89	1.35	6
300	92	0.90	6

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$Var(X) = \sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Week 10, Slide 10

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

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})Y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

```r
temperature <- c(150, 200, 250, 300)
sampleMean <- c(66, 81, 89, 92)
sampleVariance <- c(1.15, 1.00, 1.35, 0.90)

temperatureMean <- mean(temperature)
tempVar <- var(temperature)
responseMean <- mean(sampleMean)

num <- sum(6 * (temperature - temperatureMean)*(sampleMean - responseMean))
denom <- sum(6 * (temperature - temperatureMean)^2)
b1 <- num/denom
b0 <- responseMean - (b1*temperatureMean)

b1
```

```
## [1] 0.172
```

```r
b0
```

```
## [1] 43.3
```

Week 10 Slide 15

$$Var(b_0) = \sigma^2 * \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right)$$

$$Var(b_1) = \sigma^2 * \frac{1}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Week 10 Slide 26

$$\hat{\sigma}^2 = MS_{error} = SS_{error}/(n-2)$$

See below (ANOVA table calculations for reasoning)

$$SS_{error} = \sum_{j=1}^{4}(n_j - 1)\hat{\sigma}_j^2$$

```r
n <- 24
temperature <- c(150, 200, 250, 300)
sampleMean <- c(66, 81, 89, 92)
sampleVariance <- c(1.15, 1.00, 1.35, 0.90)

temperatureMean <- mean(temperature)
tempVar <- var(temperature)
responseMean <- mean(sampleMean)

nRep <- rep(6, 4)
pooledVariance <- sum((nRep - 1) * sampleVariance) / sum(nRep - 1)
pooledVariance
```

```
## [1] 1.1
```

```
n <- 24
n_i <- 6

temperature <- c(150, 200, 250, 300)
sampleMean <- c(66, 81, 89, 92)
sampleVariance <- c(1.15, 1.00, 1.35, 0.90)

temperatureMean <- mean(temperature)
tempVar <- var(temperature)
responseMean <- mean(sampleMean)

hatY <- b0 + b1*temperature
hatYRep <- rep(hatY, each = 6)
sampleMeanRep <- rep(sampleMean, each = 6)
SSLack <- sum((sampleMeanRep - hatYRep)^2)
SSLack
```

```
## [1] 217.2
```

```
nRep <- rep(6, 4)
pooledVariance <- sum((nRep - 1) * sampleVariance) / sum(nRep - 1)
pooledVariance
```

```
## [1] 1.1
```

```
SSPure <- sum((n_i - 1) * sampleVariance)
SSPure
```

```
## [1] 22
```

```
SSPE <- SSPure
SSLOF <- SSLack
SSE <- SSLOF + SSPE
SS_error <- SSE
MSE <- SS_error / 22
```

```
temperature <- c(150, 200, 250, 300)
sampleMean <- c(66, 81, 89, 92)
sampleVariance <- c(1.15, 1.00, 1.35, 0.90)

temperatureMean <- mean(temperature)
tempVar <- var(temperature) * length(temperature)
tempRep <- rep(temperature, each = 6)

Varb1 <- MSE / sum((tempRep - temperatureMean)^2)
Varb0 <- MSE * ((1/n) + (temperatureMean^2 /sum((tempRep - temperatureMean)^2)))

SEb1 <- sqrt(Varb1)
SEb0 <- sqrt(Varb0)

SEb0
```

```
## [1] 2.791437
```

```
SEb1
```

```
## [1] 0.01204034
```

```
b1
```

```
## [1] 0.172
```

```
b0
```

```
## [1] 43.3
```

```
SEb0
```

```
## [1] 2.791437
```

```
SEb1
```

```
## [1] 0.01204034
```

Our estimates are: $b_0 = 43.3$ $SE(b_0) = 2.791437$ $b_1 = 0.172$ $SE(b_1) = 0.01204034$

(b)

Week 10 Slide 23

$$SS_{model} = b_1^2 \sum_{i=1}^{n} (x_i - \bar{x})^2$$

```
SSModel <- b1^2 * sum(6 * (temperature - temperatureMean)^2)
SSModel
```

```
## [1] 2218.8
```

Week 10 Slide 25

$$SS_{model} = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y}_i)^2$$

$$SS_{error} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

Week 11 Slide 40

$$SS_{error} = SS_{pureerror} + SS_{lack-of-fit}$$

$$SS_{lack-of-fit} = \sum_{i} \sum_{j} (\bar{Y}_{i.} - \hat{Y}_i)^2$$

$$\hat{Y}_i = b_0 + b_1 x_i$$

```
n <- 24
temperature <- c(150, 200, 250, 300)
sampleMean <- c(66, 81, 89, 92)
sampleVariance <- c(1.15, 1.00, 1.35, 0.90)

temperatureMean <- mean(temperature)
tempVar <- var(temperature)
responseMean <- mean(sampleMean)

varB0 <- 1/n + (temperatureMean^2 / sum(6 * (temperature - temperatureMean)^2) )
SEb0 <- sqrt(varB0)
varB1 <- 1 / sum(6 * (temperature - temperatureMean)^2)
SEB1 <- sqrt(varB1)

hatY <- b0 + b1*temperature
hatYRep <- rep(hatY, each = 6)
sampleMeanRep <- rep(sampleMean, each = 6)
SSLack <- sum((sampleMeanRep - hatYRep)^2)
SSLack
```

```
## [1] 217.2
```

```r
hatY <- b0 + b1*temperature
SSLack2 <- sum((sampleMean - hatY)^2)
SSLack2
```

```
## [1] 36.2
```

$$SS_{pureerror} = \sum_i \sum_j (Y_{ij} - \bar{Y}_i) = \sum_{i=1}^{4} \sum_{j=1}^{6} (Y_{ij} - \bar{Y}_{i.})^2$$

$$s_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i.})^2$$

$$SS_{pureerror} = \sum_{i=1}^{4} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i.})^2 = \sum_{i=1}^{4} (n_i - 1) s_i^2$$

```r
n <- 24
n_i <- 6

temperature <- c(150, 200, 250, 300)
sampleMean <- c(66, 81, 89, 92)
sampleVariance <- c(1.15, 1.00, 1.35, 0.90)

temperatureMean <- mean(temperature)
tempVar <- var(temperature)
responseMean <- mean(sampleMean)

SSPure <- sum((n_i - 1) * sampleVariance)
SSPure
```

```
## [1] 22
```

```r
temperature <- c(150, 200, 250, 300)
sample_mean <- c(66, 81, 89, 92)
sample_variances <- c(1.15, 1.00, 1.35, 0.90)
n_i <- 6

SSPE <- SSPure
SSLOF <- SSLack
SSR <- SSModel

temperatureMean <- mean(temperature)
responseMean <- mean(sample_mean)

df_regression <- 1
df_residual <- length(temperature) * n_i - 2
df_lack_of_fit <- length(temperature) - 2
df_pure_error <- df_residual - df_lack_of_fit
df_total <- length(temperature) * n_i - 1
```

```r
SSE <- SSLOF + SSPE
SST <- SSR + SSE

MSR <- SSR / df_regression
MSE <- SSE / df_residual
MSLOF <- SSLOF / df_lack_of_fit
MSPE <- SSPE / df_pure_error

anova_table <- data.frame(
  "Source of Variation" = c("Regression on X", "Residuals", "- Lack-of-fit", "- Pure error", "Total"),
  "Degrees of Freedom" = c(df_regression, df_residual, df_lack_of_fit, df_pure_error, df_total),
  "Sum of Squares" = c(SSR, SSE, SSLOF, SSPE, SST),
  "Mean Square" = c(MSR, MSE, MSLOF, MSPE, NA)
)

# Display the table
print(anova_table)
```

```
##   Source.of.Variation Degrees.of.Freedom Sum.of.Squares Mean.Square
## 1     Regression on X                  1         2218.8  2218.80000
## 2           Residuals                 22          239.2    10.87273
## 3       - Lack-of-fit                  2          217.2   108.60000
## 4        - Pure error                 20           22.0     1.10000
## 5               Total                 23         2458.0          NA
```

Giving the following table

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square
Regression on X	1	2218.8	2218.8
Residuals	22	239.2	10.87273
- Lack-of-fit	2	217.2	108.60
- Pure error	20	22	1.1
Total	23	2458	

(c)

Compute the F-statistic for the lack-of-fit test and report the corresponding degrees freedom. Suppose the p-value is 0.0001, then interpret this result in the context of the study

Our General Approach:

$$F = \frac{MS_{\text{Lack of Fit}}}{MS_{\text{Pure Error}}} = \frac{108.60}{1.100000} = 98.72727$$

```
qf(p = 1 - .0001, df1 = 2, df2 = 20)
```

```
## [1] 15.11886
```

For comparison with the F Statistic, we have: Lack of Fit Df = 2 Pure Error Df = 20

So, the computed F-statistic for the lack-of-fit test is 98.72727 with degrees of freedom (2, 20). Given the provided p-value of of 0.0001 is very small (much less the typical $\alpha = 0.05$, and the F statistic much larger than the calculated F statistic of comparison, 15.11886), we have overwhelming evidence to reject the null hypothesis that the simple linear regression model is adequate (fits the relationship well) for the data provided.

Q2

The Berkeley Guidance Study enrolled children born in Berkeley, California, between January 1928 and June 1929, and then measure each child periodically until age 18. The data for all of the girls in the study who were measured at age 18 are posted in the file BGSgirls.dat in our course's shared folder on SAS Studio. There is one line for each girl in this data file, with the subject identification number, weight (in kilograms), and height (in centimeters), in that order from left to right

(a)

Compute least square estimates of the intercept (β_0) and slope (β_1) of a simple linear regression model for predicting weight (Y) from height (x). Report the parameter estimates and their standard errors. Is height a significant predictor of weight (yes or no)? Briefly justify your choice.

The REG Procedure
Model: MODEL1
Dependent Variable: Weight

Number of Observations Read	70
Number of Observations Used	70

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	1284.13346	1284.13346	22.42	<.0001
Error	68	3895.09926	57.28087		
Corrected Total	69	5179.23271			

Root MSE	7.56841	R-Square	0.2479
Dependent Mean	59.78429	Adj R-Sq	0.2369
Coeff Var	12.65954		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-58.48504	24.99519	-2.34	0.0222
Height	1	0.71014	0.14998	4.73	<.0001

Figure 1: CocoMelon

b0: -58.48504 b0 SE: 24.99519 b1: 0.71014 b1 SE: 0.14998

Is height a significant predictor of weight? Despite being a somewhat small estimate of the slope for b1 (magnitude smaller than 1), we do find overwhelming evidence to reject the null hypothesis that Height has

0 predictive power in estimating mean weight (have evidence to reject the null hypothesis at the $\alpha = 0.05$ level that the slope is 0).

(b)

Plot weight versus height and insert the estimated regression line on the plot, and include the plot in your submission. What does this plot suggest?

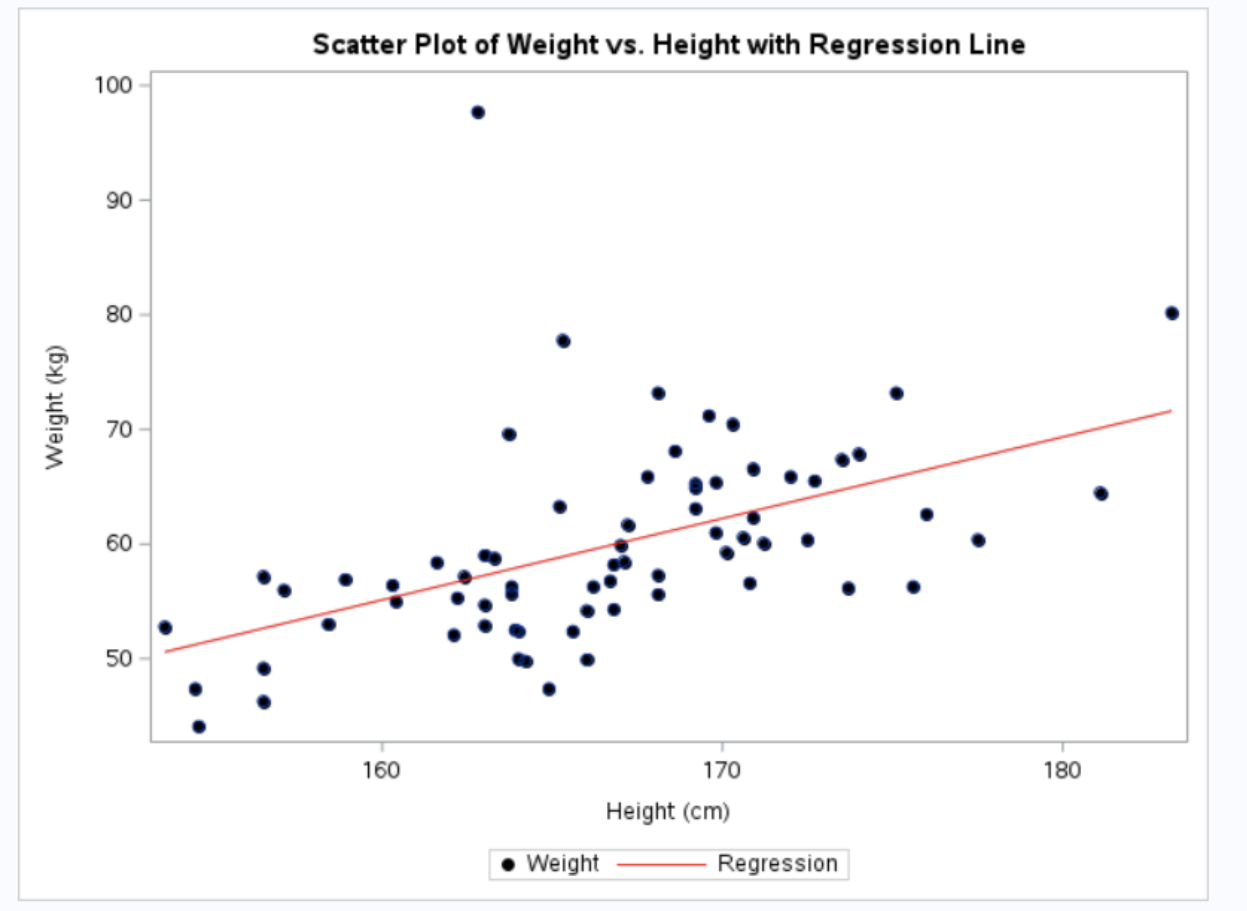

Figure 2: CocoMelon

The plot suggests that generally height and weight are postiively and linearly correlated. However, given the spread of values from the best fit line there is some evidence that there may be other factors or relationships to consider between height and weight.

(c)

Construct a plot of the studentized residuals versus $\hat{Y}_i$, where $\hat{Y}_i = b_0 + b_1 x_i$, and include the plot in your submission. What does this plot indicate?

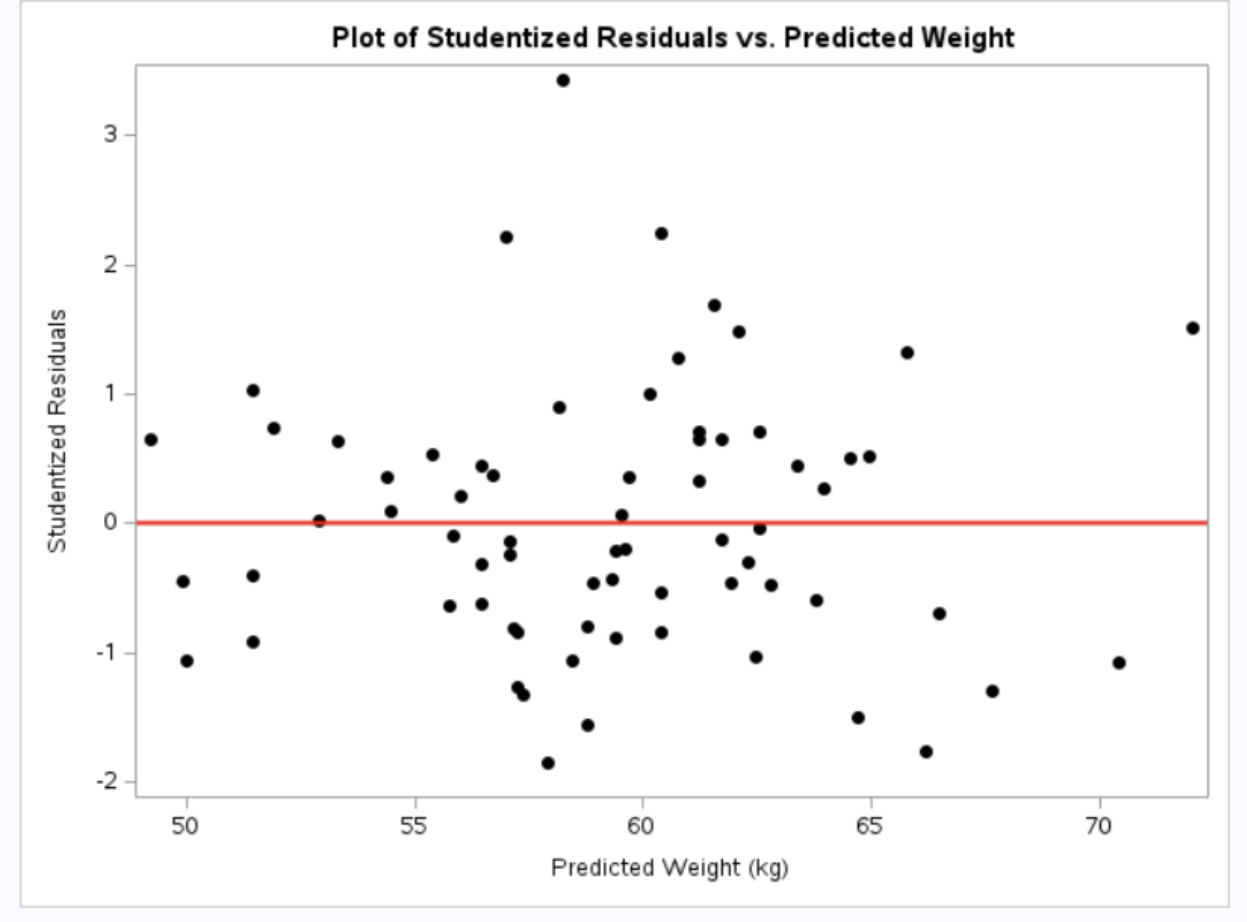

Figure 3: CocoMelon

This plot indicates that we generally see a random spread of (studentized) residuals across predicted values. However, we also observe a somewhat larger (wider) spread of residual values around predictions of 60kg, suggesting that there may be outliers to consider, considering this "spread" is primarily being caused by the large studentized residual at fitted value (weight in kg) of 60.

(d)

The diagnostic plots should indicate that there is one 18 year-old girl who is extremely heavy given her height. This observation may involve a value for either height or weight that was not properly recorded, or it may just correspond to an unusually heavy girl. You can delete this observation by replacing the value of the weight with a period. Because this is the only girl with weight exceeding 90 kg, you can delete this case in a data step by inserting the code:

```
if(weight > 90) then weight=. ;
```

Or you can use only the subset of data by

```
where weight le 90;
```

Re-fit the simple linear regression model. Do the diagnostic plots now appear to show that the data conform to the assumptions of the proposed regression model? If not, what problems remain? Include all relevant plots in your submission.

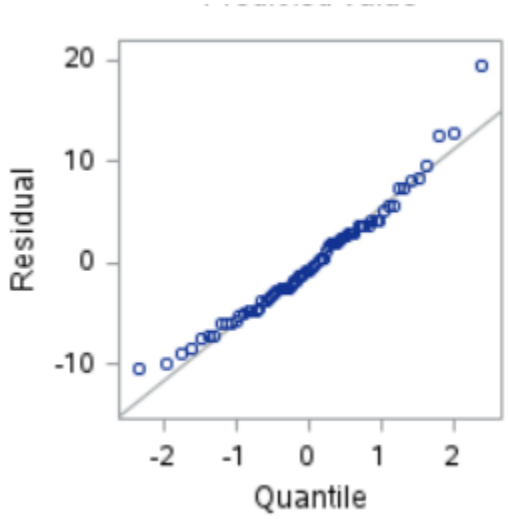

Figure 4: CocoMelon

The above is a snippet from the diagnostic plots provided by SAS, specifically the QQ (Quantile) plot to test whether the residuals of our model are normally distributed. We do generally observe the residuals fall in-line with the reference line, suggesting that normality is likely not violated.

To analyze whether Linearity or Equal Variance assumptions are being violated, we then turn to the residual plot by fitted values. We observe, similar to the interpretation given previously, that we tend to see a random spread of residuals across values. However, we do observe some outlying points near the middle. Overall, we have evidence to believe that both Linearity and Equal variance assumptions are not being violated. Furthermore, when reviewing the scatterplot of height and weight we do generally observe a positive linear relationship between the two variables, such that we have further reason to believe this assumption (linearity) is not being violated.

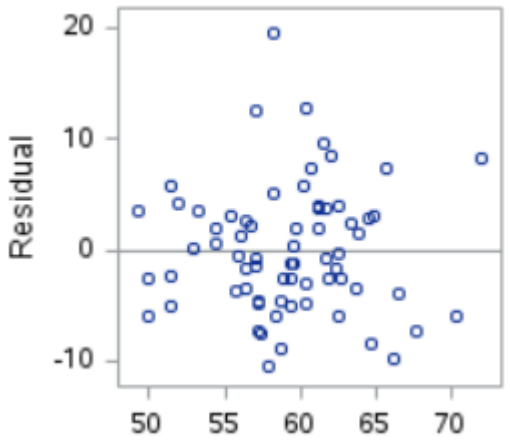

Figure 5: CocoMelon

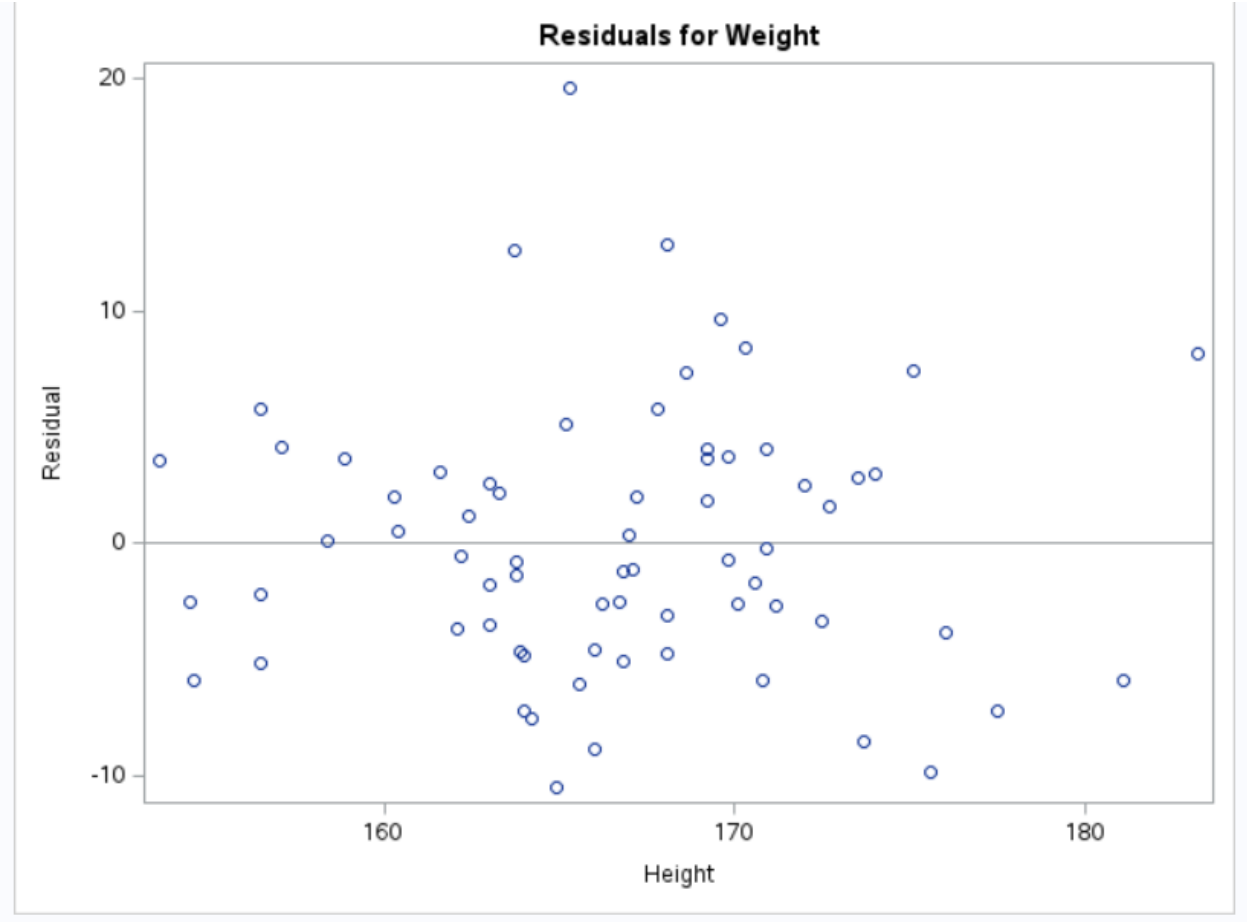

Figure 6: CocoMelon

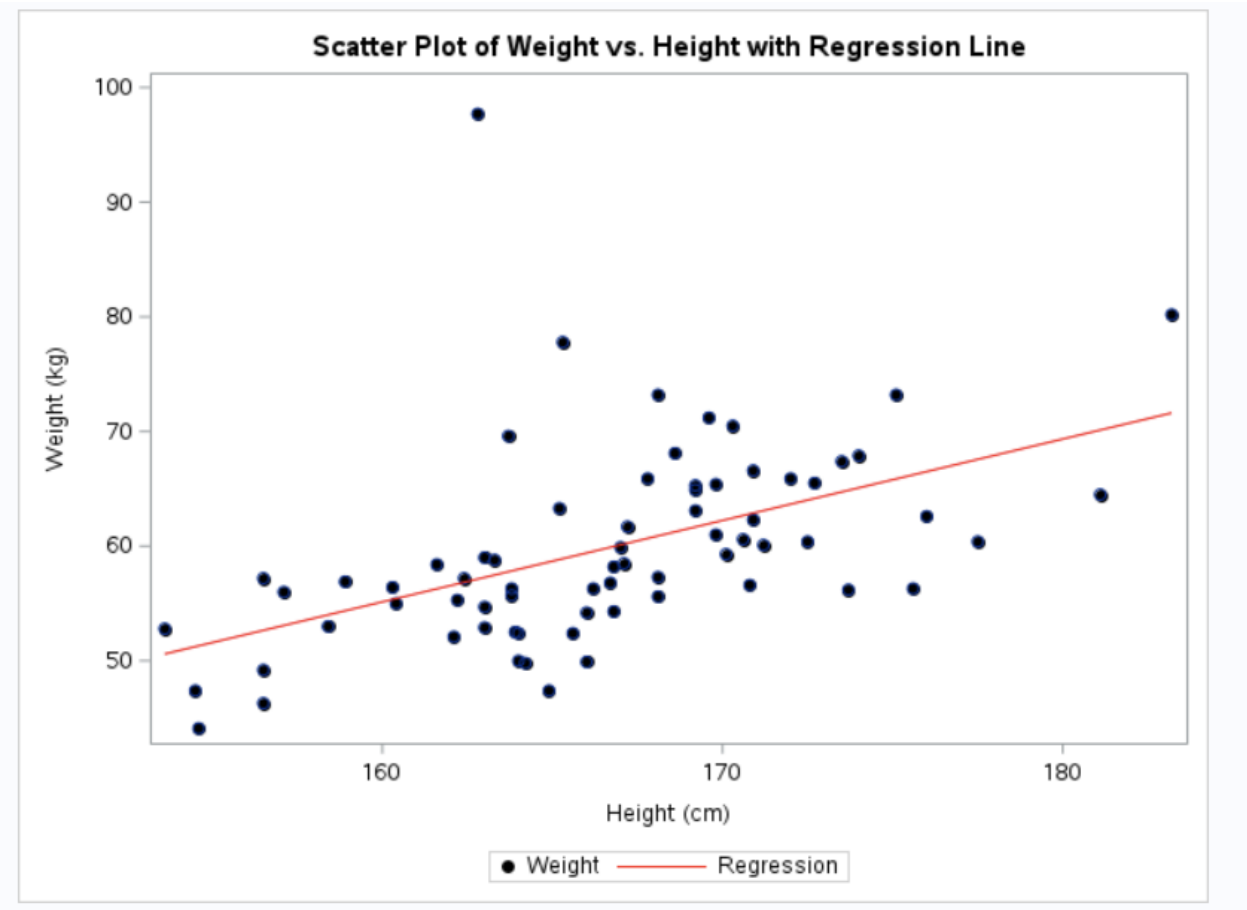

Figure 7: CocoMelon

(e)

Plot the estimated regression lines with the extreme observation included and the extreme observation removed on the same plot. Include the plot in your submission. Did deleting the observation in part (d) have a large effect on any of the parameter estimates? Briefly justify your response.

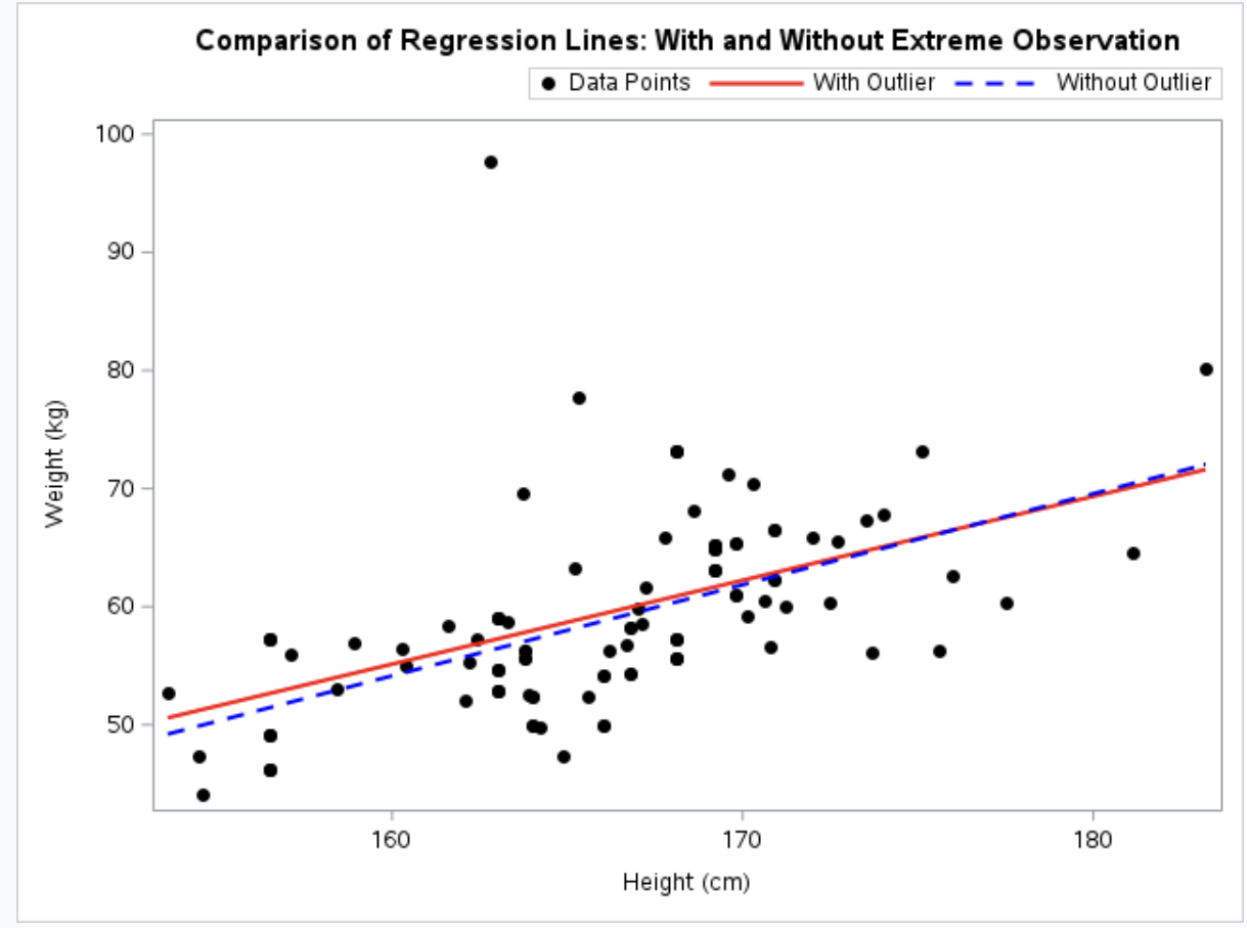

Figure 8: CocoMelon

We observe the two fit lines to be remarkably close to one another, with a slight divergence for heights less than 170cm. This suggests that despite removing an obvious outlying point, the overall impact that point had on our predicted best fit line was negligible, or doesn't dramatically change the magnitude nor does it change the signage of the slope of the estimated regression line. The intercept term of the fit line is also different, but again, not especially so in terms of the value corresponding to the intercept term (though not explicitly shown in the above graph).

Q3

One factor that may explain the price of a diamond is the weight of the diamond. Data were collected for a sample of 48 diamonds, including the weight in grams (g) and the price (in Singapore dollars) of each diamond. These data are located in the file diamonds.csv posted in Canvas. The R code that generated the output below is included in Canvas in the diamonds Hmwk9.R file for your reference.

(a)

Write the simple linear regression model for this problem (including assumptions). Give the definition of the parameter values β_0, β_1, and σ^2 in the context of the response and explanatory variables.

Assumptions: We have the independence and fixed-values-for-x assumptions, as well as linearity, constant variance, and normality (particularly and specifically for the error terms/residuals).

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

Where

$$\epsilon \sim N(0, \sigma^2)$$

β_0: the conditional mean price of the diamond(s) when weight of the diamond is 0g

β_1: the increase (or change) in the conditional mean price of the diamond(s) (change in conditional mean of the response) when weight of the diamond is increased by 1g

σ^2: the expected value of the mean-squared error of the residuals, or the variance of the residuals, where the residual is the vertical distance between observed value of diamond price and the predicted value of diamond price

(b)

Write the simple linear regression model for this problem in vector-matrix notation. Give the first 4 rows of the design matrix $\mathbf{X}$.

The overall simple linear regression model is:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

Where

$$\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$$

Or:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ 1 & X_3 \\ 1 & X_4 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

Where we have observed values and estimates:

$$\begin{bmatrix} 355 \\ 328 \\ 350 \\ 325 \\ \vdots \\ 316 \end{bmatrix} = \begin{bmatrix} 1 & 0.17 \\ 1 & 0.16 \\ 1 & 0.17 \\ 1 & 0.18 \\ \vdots & \vdots \\ 1 & 0.15 \end{bmatrix} \begin{bmatrix} -229.94 \\ 3612.5 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \vdots \\ \epsilon_{48} \end{bmatrix}$$

The first 4 rows of the design matrix $\mathbf{X}$ are:

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ 1 & x_4 \end{bmatrix} = \begin{bmatrix} 1 & 0.17 \\ 1 & 0.16 \\ 1 & 0.17 \\ 1 & 0.18 \end{bmatrix}$$

(c)

Describe the scatterplot, shown below, of the weight and price of the 48 diamonds in this sample. What do you notice about the relationship between these two values?

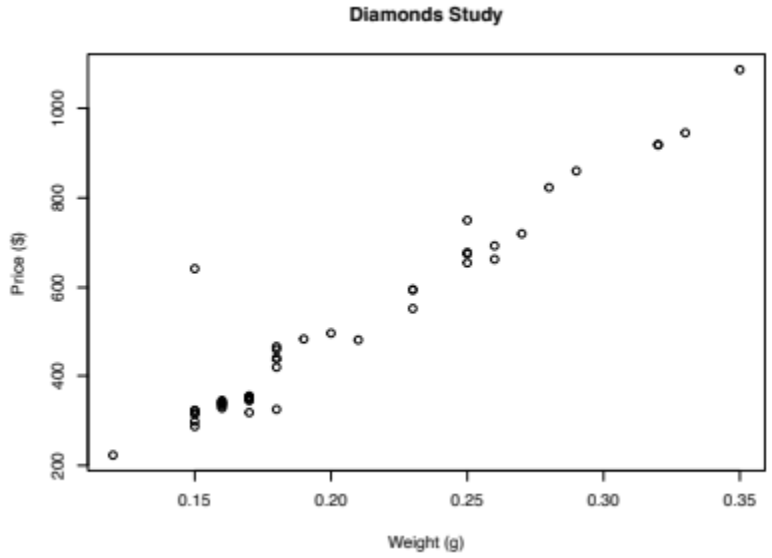

Figure 9: CocoMelon

We see that generally weight and price have a positive linear relationship with one another.

(d)

The output below includes the sample correlation coefficient between the weight and price of the diamonds. How does the value of the correlation reinforce your description from part (c).

```
> cor(diamonds$weight, diamonds$price)
[1] 0.9622006
```

Figure 10: CocoMelon

The above correlation reinforces both the direction of the relationship (positive and linear) as well as the strength of their linear fit (value close to 1 being a strong linear correlation).

(e)

Using the output shown below, give the equation for the least squares regression line to predict the price of a diamond from its weight.

```
Call:
lm(formula = price ~ weight, data = diamonds)

Residuals:
   Min     1Q Median     3Q    Max
-95.31 -26.37  -7.56  10.32 330.07

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -229.94      31.63  -7.271 3.58e-09 ***
weight       3612.50     150.76  23.962  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 58.81 on 46 degrees of freedom
Multiple R-squared:  0.9258,    Adjusted R-squared:  0.9242
F-statistic: 574.2 on 1 and 46 DF,  p-value: < 2.2e-16
```

Figure 11: CocoMelon

$$\hat{Y_i} = b_0 + b_1 x_i = -229.94 + 3612.50 x_i$$

(f)

Use the ANOVA Table shown below to conduct a test of significance for the linear regression model

```
            Df  Sum Sq Mean Sq F value Pr(>F)
weight       1 1986120 1986120   574.2 <2e-16 ***
Residuals   46  159112    3459
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 12: CocoMelon

$H_0 : \beta_1 = 0$ $H_a : \beta_1 \neq 0$

With an F statistic of 574.2 and a p-value of <2e-16, we have overwhelming evidence to reject the null hypothesis in favor of the alternative hypothesis that $\beta_1 \neq 0$, such that we have evidence that there is a significant linear relationship between weight of diamonds (g) and price ($).

(g)

A 95% confidence interval for the slope parameter in the simple linear regression model is shown below. Give an interpretation of this interval.

```
                2.5 %    97.5 %
(Intercept) -293.6016 -166.2819
weight      3309.0375 3915.9530
```

Figure 13: CocoMelon

We are 95% confident, that the population slope is between 3309.0375 to 3915.953.

Also, we are 95% confident that the true expected increase (or change) in mean price (conditional mean price) when diamond weight increases by 1g is between $3309.04 to $3915.95.

(h)

A 95% confidence interval for the conditional mean price of all diamonds in the population with a weight of 0.2 grams is shown below. Give the interpretation of this interval.

```
     fit      lwr      upr
492.5573 475.4583 509.6563
```

Figure 14: CocoMelon

We are 95% confident, that the true mean price (conditional mean price) for diamonds weighting 0.2g is between $475.46 to $509.66.

(i)

A 95% prediction interval for the price of a diamond in the population with a weight of 0.3 grams is shown below. Give the interpretation of this interval.

```
       fit      lwr      upr
853.8068 730.5604 977.0533
```

Figure 15: CocoMelon

We are 95% confident, that the price for a diamond weighting 0.3g is between $730.56 to $977.05.

(j)

Examine the residual plots shown below. Is there any reason to suspect the model assumptions do not hold or that there are influence points?

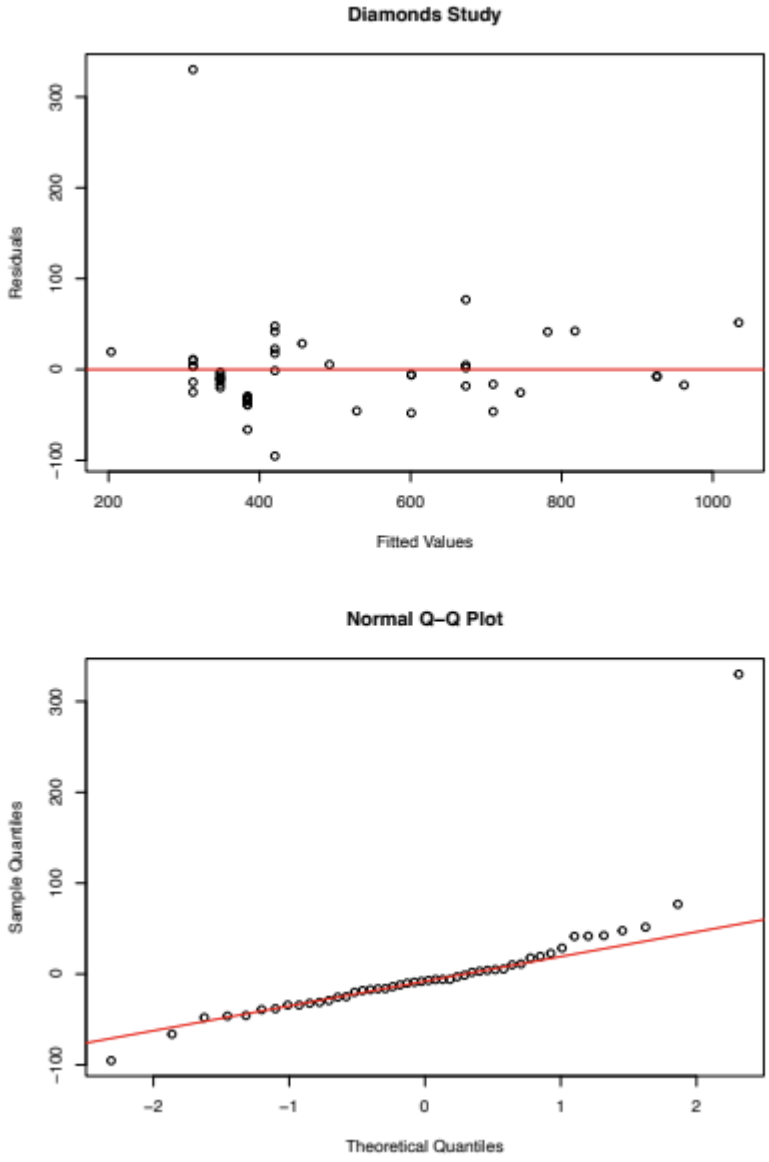

Figure 16: CocoMelon

Linearity: We would anticipate the spread of residuals to appear rather random across the range of fitted values, however, we tend to see clusters of either positive or negative residual values (in addition to one particular outlying value). Because of this, I have some concerns that our linearity assumption is being violated.

Equal Variance: We generally observe the spread of residuals to be consistent across the range of fitted values, such that we would have reason to believe that our equal variance assumption is not being violated.

Normality: Given the above QQ (Quantile) Plot, we observe the residuals track rather nicely against the reference line, with some deviations from the reference line for higher/larger values of the tail. Generally,

this is good and would indicate that normality is likely not being violated in our model.