HW9

2024-11-11

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

Q1

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

(a)

Use this information to compute the least squares estimates of β_0 and β_1 for the simple linear regression model:

$$Y_{ij} = \beta_0 + \beta_1 x_i + \epsilon_{ij}$$

Report values for the estimated coefficients (b_0 and b_1) and their standard errors (S_{b_0} and S_{b_1}).

```
# Given data
temperature <- c(150, 200, 250, 300)
sample_mean <- c(66, 81, 89, 92)
sample_variance <- c(1.15, 1.00, 1.35, 0.90)
sample_size <- c(6, 6, 6, 6)

# Compute the weights based on sample variances and sample sizes
weights <- sample_size / sample_variance

# Weighted means of temperature and yield
weighted_mean_x <- sum(weights * temperature) / sum(weights)
weighted_mean_y <- sum(weights * sample_mean) / sum(weights)

# Weighted sums of squares
```

```r
SS_xy <- sum(weights * (temperature - weighted_mean_x) * (sample_mean - weighted_mean_y))
SS_xx <- sum(weights * (temperature - weighted_mean_x)^2)

# Compute b1 and b0
b1 <- SS_xy / SS_xx
b0 <- weighted_mean_y - b1 * weighted_mean_x

# Calculate standard errors for b0 and b1
s_squared <- sum(weights * (sample_mean - b0 - b1 * temperature)^2) / (sum(weights) - 2)
Sb1 <- sqrt(s_squared / SS_xx)
Sb0 <- sqrt(s_squared * (1 / sum(weights) + weighted_mean_x^2 / SS_xx))

# Display results
list(
  "Estimated Intercept (b0)" = b0,
  "Estimated Slope (b1)" = b1,
  "Standard Error of b0 (Sb0)" = Sb0,
  "Standard Error of b1 (Sb1)" = Sb1
)
```

```
## $`Estimated Intercept (b0)`
## [1] 44.44479
##
## $`Estimated Slope (b1)`
## [1] 0.1662539
##
## $`Standard Error of b0 (Sb0)`
## [1] 2.733743
##
## $`Standard Error of b1 (Sb1)`
## [1] 0.01162316
```

(b)

Complete the following ANOVA table:

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square
Regression on X	1	2022.0969	2022.096857
Residuals	22	200.9148	9.132492
- Lack-of-fit	2	178.9148	89.457413
- Pure error	20	22.0000	1.100000
Total	23	2223.0117	NA

```r
# Given data and results from previous calculations
sample_size <- c(6, 6, 6, 6)
sample_mean <- c(66, 81, 89, 92)
sample_variance <- c(1.15, 1.00, 1.35, 0.90)

# Total sample size and number of groups
N <- sum(sample_size)
r <- length(sample_size)
```

```r
# Weighted mean for sample means
weights <- sample_size / sample_variance
overall_weighted_mean <- sum(weights * sample_mean) / sum(weights)

# SST: Total Sum of Squares
SST <- sum(weights * (sample_mean - overall_weighted_mean)^2)

# Regression SS (SSR) using previously computed b1 and SS_xy
SSR <- b1 * SS_xy

# Residual SS (SSE)
SSE <- SST - SSR

# Pure Error SS (SSPE)
SSPE <- sum((sample_size - 1) * sample_variance)

# Lack-of-Fit SS (SSLF)
SSLF <- SSE - SSPE

# Degrees of Freedom
df_total <- N - 1
df_regression <- 1
df_residual <- N - 2
df_lack_of_fit <- r - 2
df_pure_error <- N - r

# Mean Squares
MSR <- SSR / df_regression
MSE <- SSE / df_residual
MSLF <- SSLF / df_lack_of_fit
MSPE <- SSPE / df_pure_error

# ANOVA Table
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, SSLF, SSPE, SST),
  "Mean Square" = c(MSR, MSE, MSLF, MSPE, NA)
)

anova_table
```

```
##   Source.of.Variation Degrees.of.Freedom Sum.of.Squares Mean.Square
## 1     Regression on X                  1      2022.0969 2022.096857
## 2           Residuals                 22       200.9148    9.132492
## 3       - Lack-of-fit                  2       178.9148   89.457413
## 4        - Pure error                 20        22.0000    1.100000
## 5               Total                 23      2223.0117          NA
```

(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

$$F = \frac{\text{MSLF}}{\text{MSPE}} == \frac{89.457413}{1.100000} \approx 81.32$$

MSLF (Mean Square for Lack of Fit) = 89.457413 **MSPE (Mean Square for Pure Error)** = 1.100000

Numerator degrees of freedom (for Lack of Fit) = 2 **Denominator degrees of freedom** (for Pure Error) = 20

So, the computed **F-statistic for the lack-of-fit test** is approximately **81.32** with degrees of freedom $(2, 20)$.

A p-value of 0.0001 is very small (much less than a typical significance level, such as 0.05). This indicates strong evidence against the null hypothesis of the lack-of-fit test, which states that the simple linear regression model is adequate for the data.

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	69
Number of Observations with Missing Values	1

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	1505.24180	1505.24180	45.52	<.0001
Error	67	2215.55472	33.06798		
Corrected Total	68	3720.79652			

Root MSE	5.75048	R-Square	0.4045
Dependent Mean	59.23478	Adj R-Sq	0.3957
Coeff Var	9.70794		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-69.21339	19.05090	-3.63	0.0005
Height	1	0.77100	0.11428	6.75	<.0001

Figure 1: CocoMelon

(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?

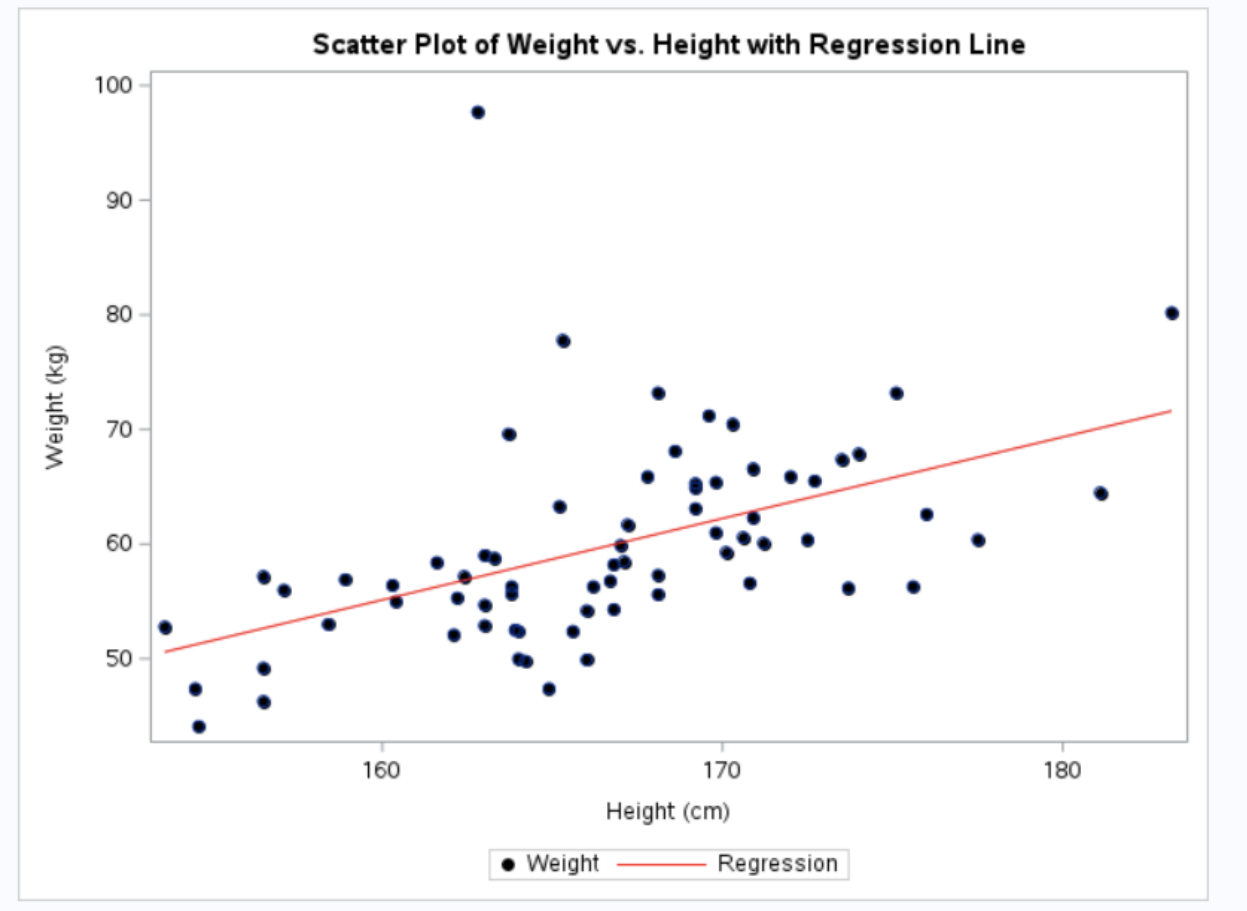

Figure 2: CocoMelon

(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?

(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

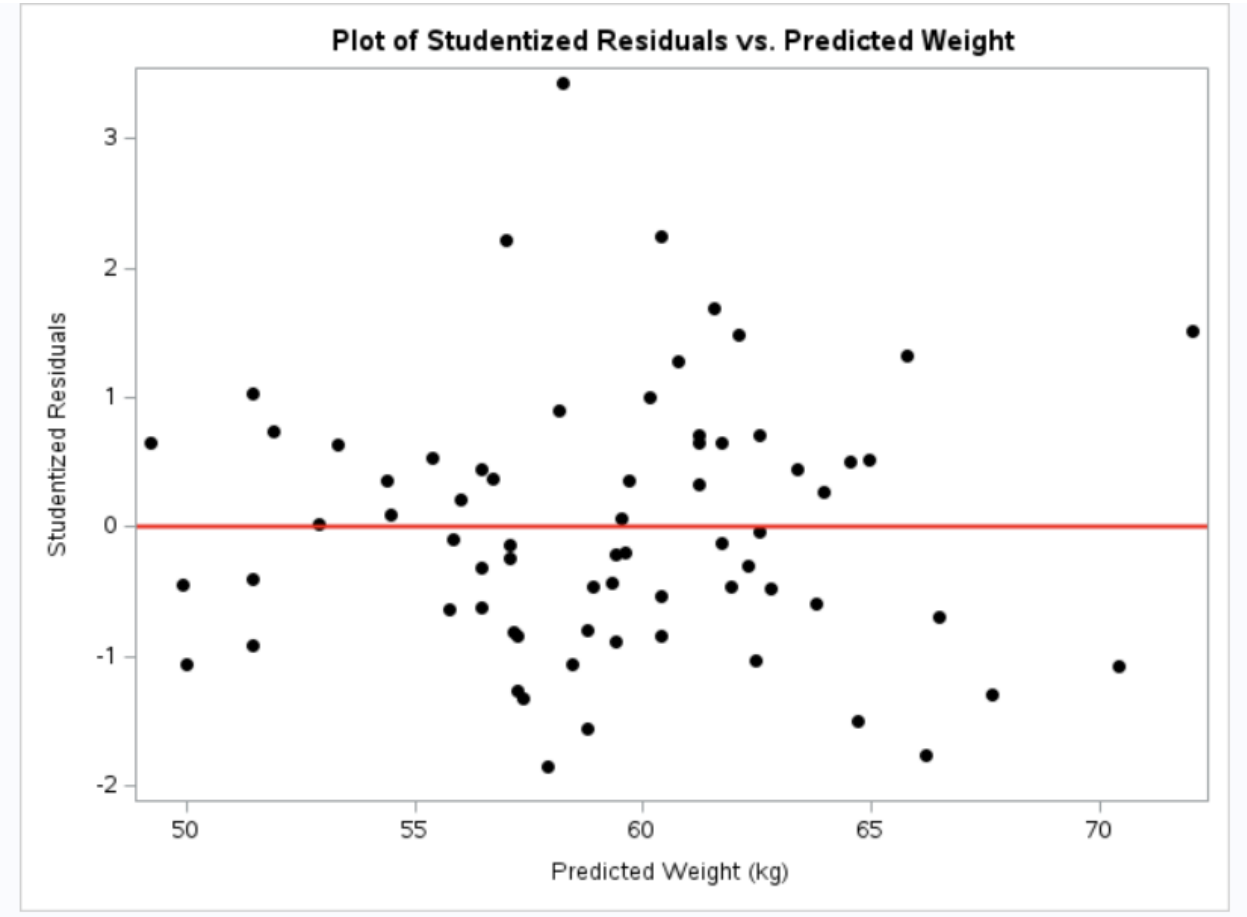

Figure 3: CocoMelon

```
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.

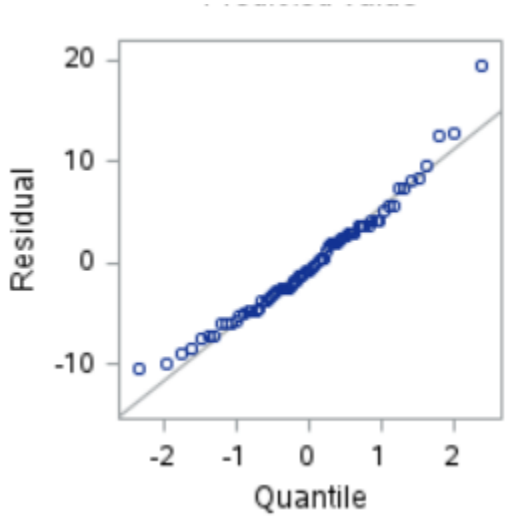

Figure 4: CocoMelon

(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.

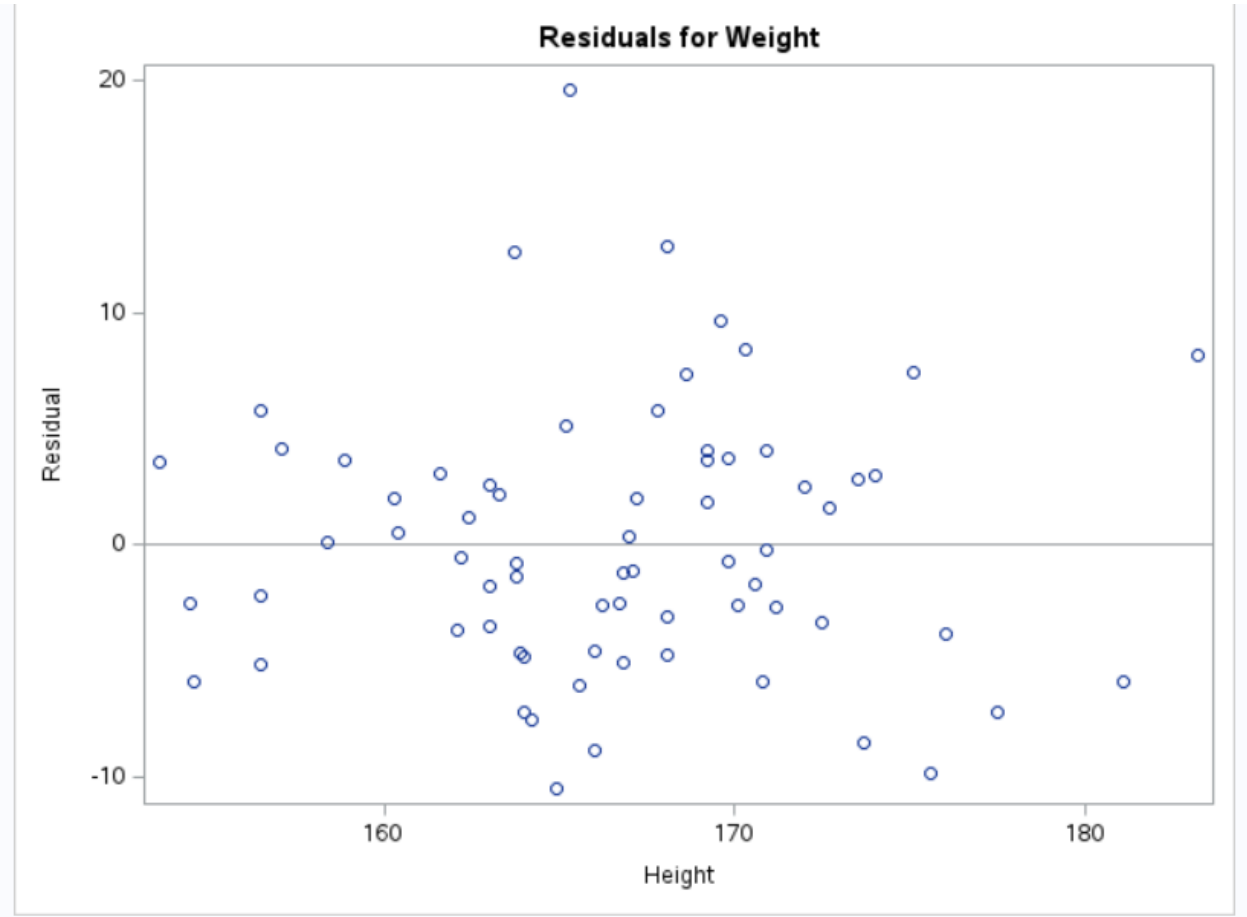

Figure 5: CocoMelon

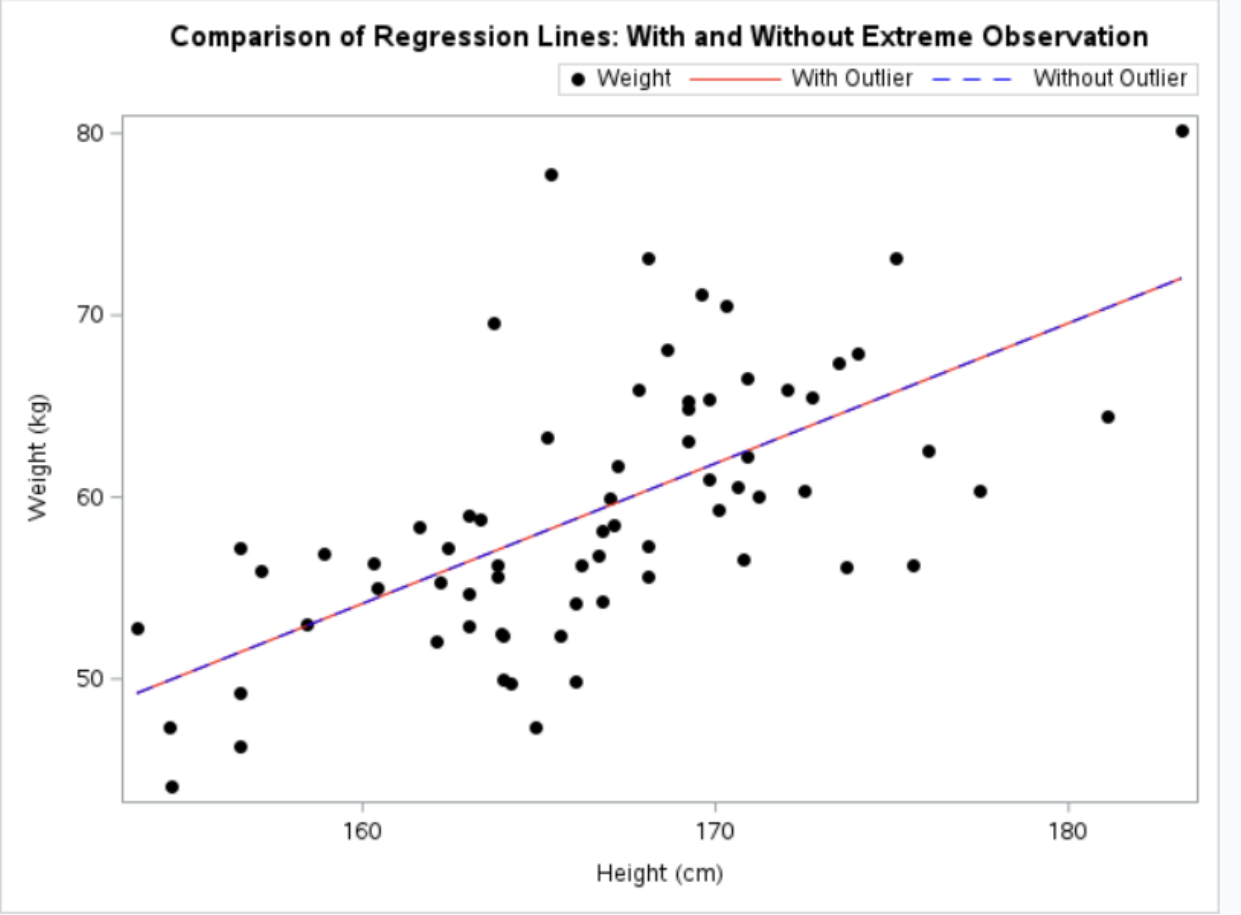

Figure 6: CocoMelon

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.

(b)

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

(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?

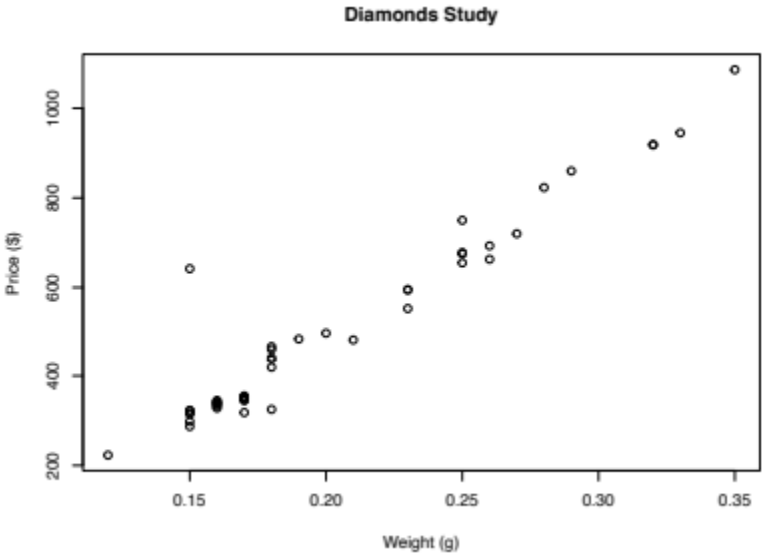

Figure 7: CocoMelon

(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 8: CocoMelon

(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 9: CocoMelon

(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 10: CocoMelon

(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 11: CocoMelon

(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 12: CocoMelon

(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 13: CocoMelon

(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?

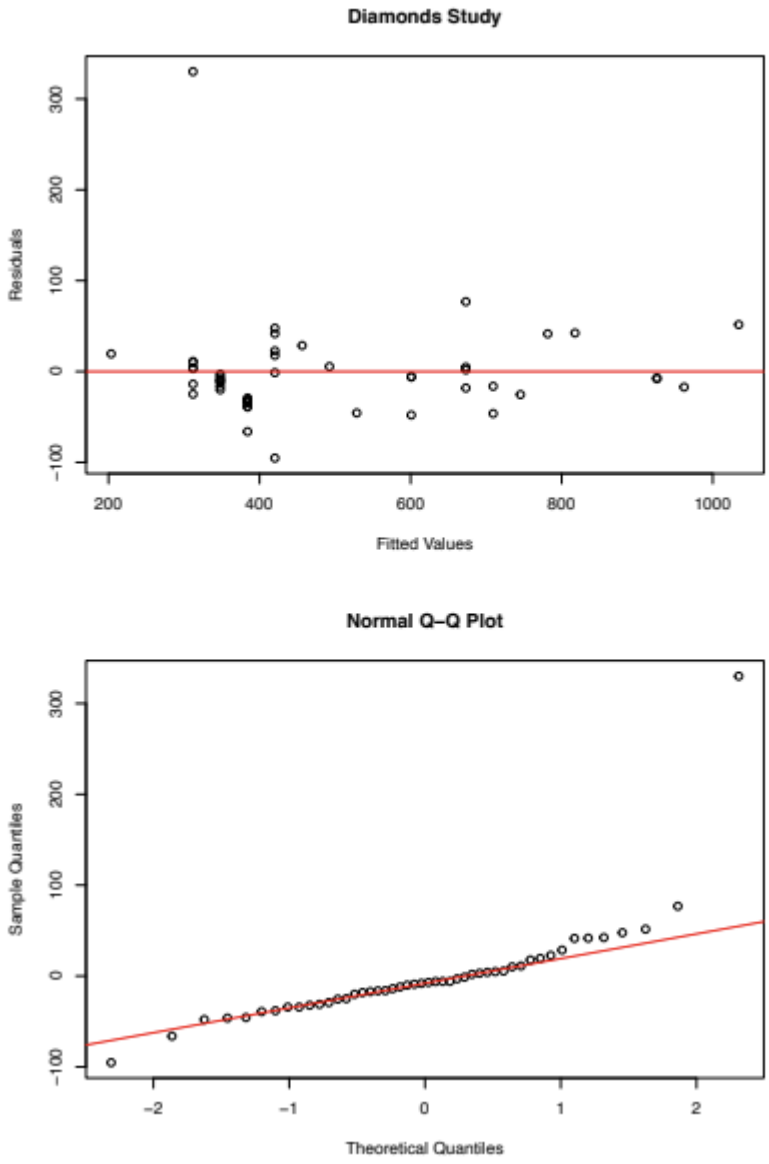

Figure 14: CocoMelon