STAT 5000 HOMEWORK #9
FALL 2024 DUE FRI, NOV 15TH @ 11:59 PM NAME: _____

Directions: Type or clearly handwrite your solutions to each of the following exercises. Partial credit cannot be given unless all work is shown. You may work in groups provided that each person takes responsibility for understanding and writing out the solutions. Additionally, you must give proper credit to your collaborators by providing their names on the line below (if you worked alone, write "No Collaborators"):

_____

1. **[+25]:** 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)	150	200	250	300
Sample Mean	66	81	89	92
Sample Variance	1.15	1.00	1.35	0.90
Sample Size	6	6	6	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}).

(b) Complete the following ANOVA table:

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square
Regression on X			
Residuals			
Lack-of-fit			
Pure error			
Total			

(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.

2. **[+30]:** 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.

Use SAS to complete the following exercises:

(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.

(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?

(c) Construct a plot of the studentized residuals versus $\hat{Y}_i$, where $\hat{Y}_i = b_o + b_1 x_i$, 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

```
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.

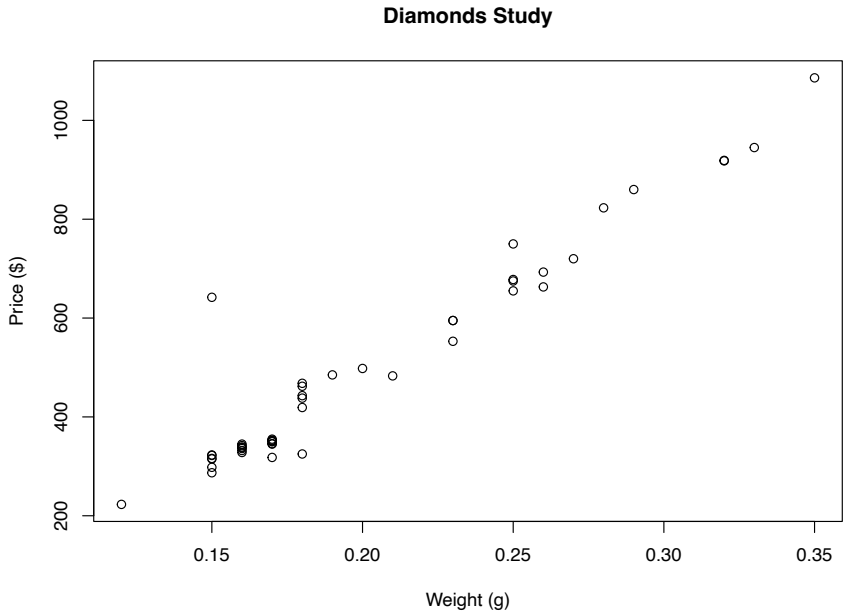
(e) Plot the estimated regression lines with the extreme observation included and the extreme observation removed on the same plot. Include the plot in your submission. Did deleting the observation in part (d) have a large effect on any of the parameter estimates? Briefly justify your response.

3. **[+45]:** 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?



(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
```

(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
```

(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
```

(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
```

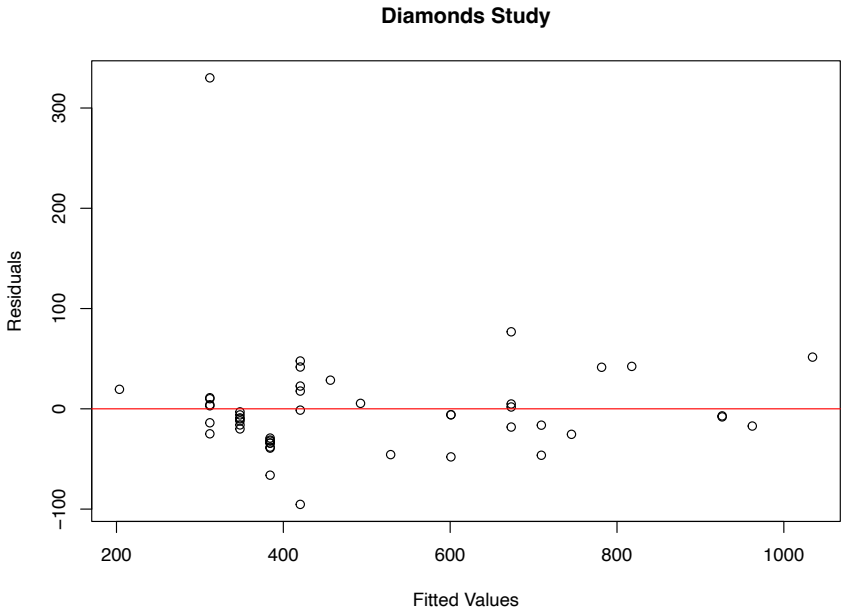
(h) A 95% confidence interval for the conditional mean price of all diamonds in the population with a weight of 0.2 grams is shown below. Give the interpretation of this interval.

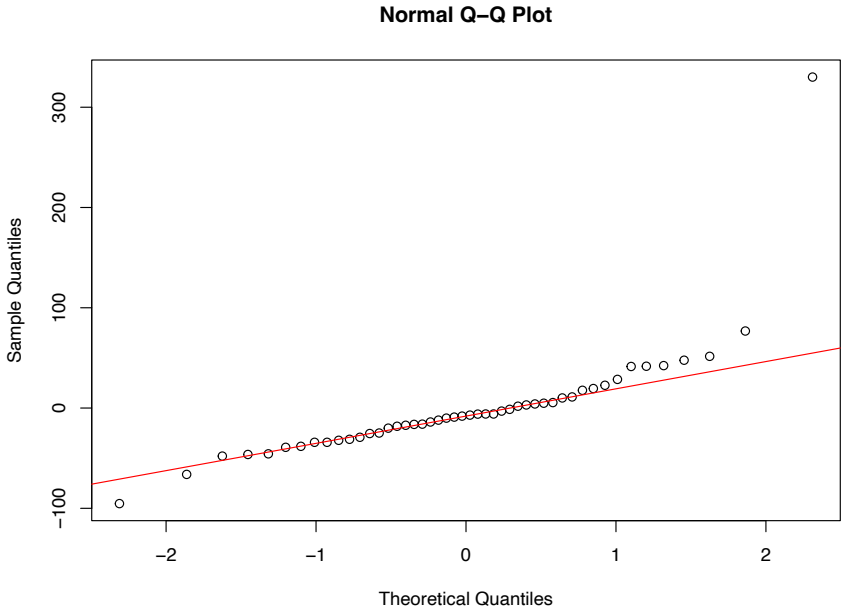
```
     fit      lwr      upr
492.5573 475.4583 509.6563
```

(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
```

(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?





Total: 100 points **# correct:** _____ **%:** _____