Assignment 5: Regression

Akshat Valse

Due Saturday, July 29 at 8am

Please upload the `PDF` that you obtain by knitting the `Rmd` file that contains your `R` code and your text answering other questions. So this uploaded file will also show any output that `R` produces in addition to your code.

You should not touch the starter code as it will print out the necessary data frames and results for grading purposes.

```
set.seed(123)

# the code in this section loads libraries, installing them as necessary
# you may need to add more libraries
if (!require(pacman)) {install.packages(pacman)}
```

```
## Loading required package: pacman
```

```
pacman::p_load(ggplot2,readr,tidyr,dplyr)
```

Part 1: Multiple Linear Regression

This question involves the use of multiple linear regression on the Auto data set. You can get this from the ISLR package in R (install from CRAN)

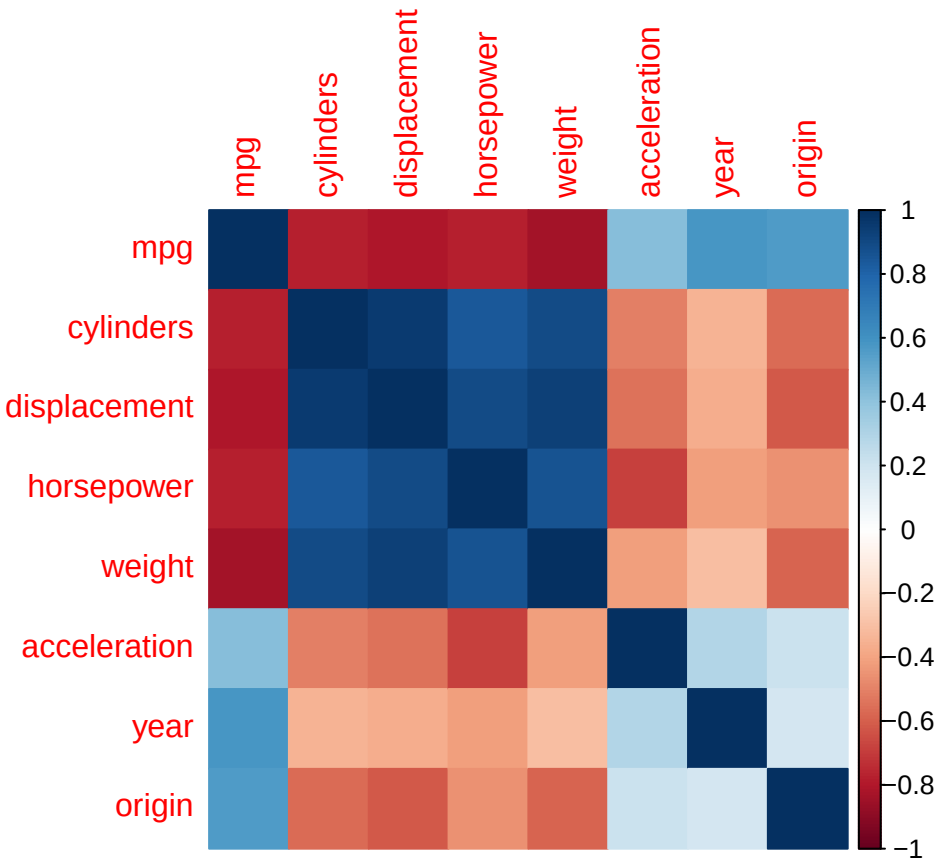
```
pacman::p_load(ISLR)
data(Auto)
data
```

Exercise 1.1

Produce a correlation plot between the variables. You will need to exclude the `name` variable, which is qualitative. You can calculate the matrix of correlations using the function `cor()`. You can use the `corrplot()` function to produce the correlation plot. To read more about `corrplot`, look up the function using `?corrplot`.

You should use `method=color` for the correlation plot.

```
require(corrplot)

### YOUR CODE HERE
data = ISLR::Auto[, c(1,2,3,4,5,6,7,8)]
cor_matrix = cor(data, method = "pearson")
# Create the correlation plot
corrplot(cor_matrix, method = "color")
```




```
### END OF YOUR CODE
```

Exercise 1.2

Use the `lm()` function to perform a multiple linear regression with `mpg` as the response and all other variables except name as the predictors. Use the `summary()` function to print the results. Comment on the output. For instance:

- Is there a relationship between the predictors and the response?
- Which predictors appear to have a statistically significant relationship to the response?
- What does the coefficient for the `year` variable suggest?
- Interpret the coefficient on weight. What does the sign mean? What does the magnitude mean? (in this case you should be able to interpret it exactly – look up the units for the different variables here.)

```
### YOUR CODE HERE
require(pacman)
p_load(ggplot2, readr, tidyr, dplyr, ISLR)
data(Auto)

### Perform multiple linear regression
lm_model <- lm(mpg ~ . - name, data = Auto)

### Print the summary of the regression model
summary(lm_model)
```

```
##
## Call:
## lm(formula = mpg ~ . - name, data = Auto)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.5903 -2.1565 -0.1169  1.8690 13.0604
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -17.218435   4.644294  -3.707  0.00024 ***
## cylinders     -0.493376   0.323282  -1.526  0.12780
## displacement   0.019896   0.007515   2.647  0.00844 **
## horsepower    -0.016951   0.013787  -1.230  0.21963
## weight        -0.006474   0.000652  -9.929  < 2e-16 ***
## acceleration   0.080576   0.098845   0.815  0.41548
## year           0.750773   0.050973  14.729  < 2e-16 ***
## origin         1.426141   0.278136   5.127 4.67e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.328 on 384 degrees of freedom
## Multiple R-squared:  0.8215, Adjusted R-squared:  0.8182
## F-statistic: 252.4 on 7 and 384 DF,  p-value: < 2.2e-16
```

```
### END OF YOUR CODE
```

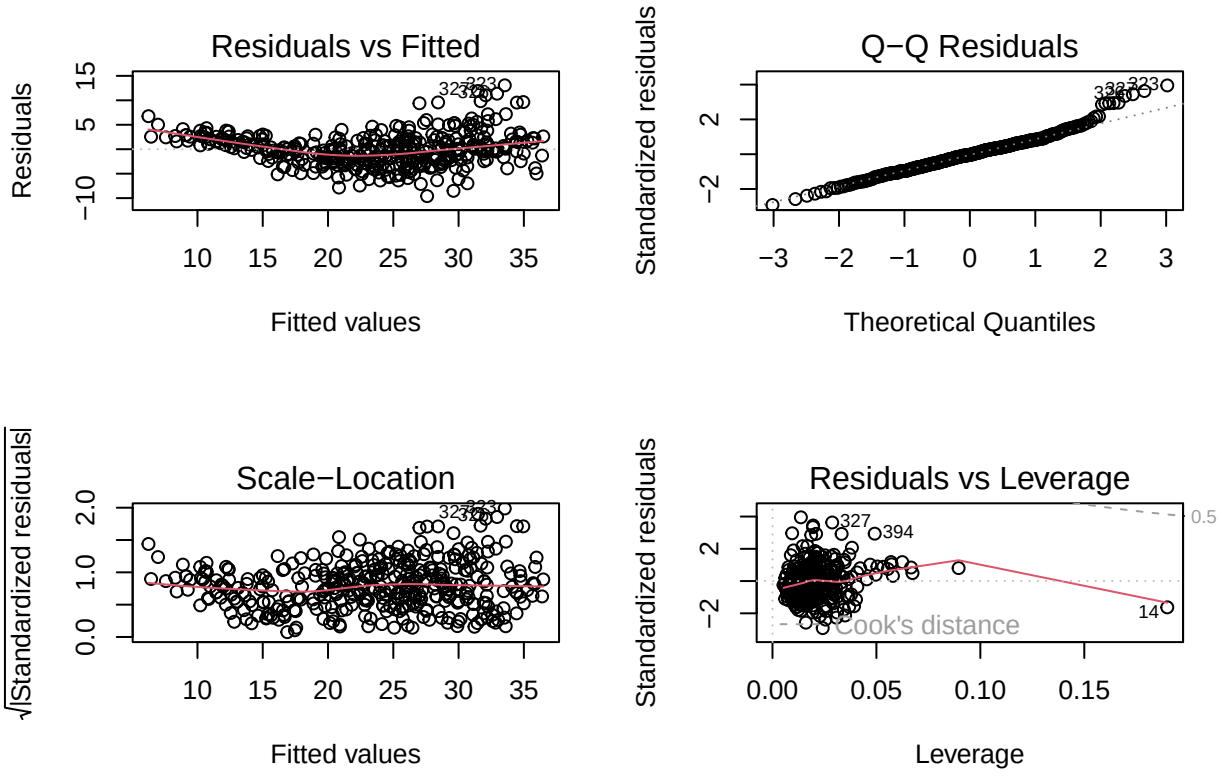
YOUR EXPLANATION HERE. The p-value for the F-statistic in the summary output will indicate whether there is a relationship between the predictors (independent variables) and the response variable (mpg). A small p-value (typically less than 0.05) indicates that at least one of the predictors is significantly related to the response, suggesting that there is a relationship between the predictors and the response.In the summary output, look for the "Pr(>|t|)" column. Predictors with a p-value less than 0.05 are considered statistically significant at the 5% significance level. These variables have a significant relationship with the response variable (mpg).The coefficient for the year variable represents the change in the response variable (mpg) for each one-unit increase in the year variable, while holding all other predictors constant. If the coefficient is positive, it suggesThe coefficient for the weight variable represents the change in the response variable (mpg) for each one-unit increase in weight, while holding all other predictors constant. Since weight is a negative value (-0.006), it suggests that as the weight of the car increases, the miles per gallon (mpg) decreases. The magnitude of the coefficient (-0.006) represents the amount of change in mpg for each one-unit increase in weight. For example, if weight increases by 1000 pounds, the mpg is expected to decrease by 6 (0.006 * 1000) units, while holding other predictors constant. The units for weight are typically in pounds, and the units for mpg are miles per gallon.ts that as the year increases, the mpg also increases. Conversely, if it is negative, it suggests that as the year increases, the mpg decreases.

Exercise 1.3

Use the `plot` function to produce diagnostic plots of the linear regression fit. Comment on any problems you see with the fit. Do the residual plots suggest any unusually large outliers? Does the leverage plot identify any observations with unusually high leverage?

```
### YOUR CODE HERE
### Load necessary libraries and data (assuming it's already loaded)
require(pacman)
```

```r
p_load(ggplot2, readr, tidyr, dplyr, ISLR)
data(Auto)

### Perform multiple linear regression
lm_model <- lm(mpg ~ . - name, data = Auto)

### Produce diagnostic plots
par(mfrow = c(2, 2)) # Arrange the plots in a 2x2 grid
plot(lm_model)
```




```r
### END OF YOUR CODE
```

YOUR EXPLANATION HERE. Residual vs. Fitted plot: This plot helps to identify whether there are any patterns or trends in the residuals. Ideally, we want the residuals to be randomly scattered around the horizontal line at zero. If there are clear patterns, it suggests that the relationship between the predictors and the response is not adequately captured by the linear model.

Normal Q-Q plot: This plot checks whether the residuals are normally distributed. If the points closely follow the 45-degree reference line, it indicates that the residuals are approximately normally distributed. Deviations from this line suggest non-normality.

Scale-Location plot: Also known as the "Spread-Location" plot, it helps to assess the assumption of homoscedasticity. The plot shows the square root of the standardized residuals against the fitted values. Ideally, we want the points to be randomly scattered around a horizontal line, suggesting constant variance.

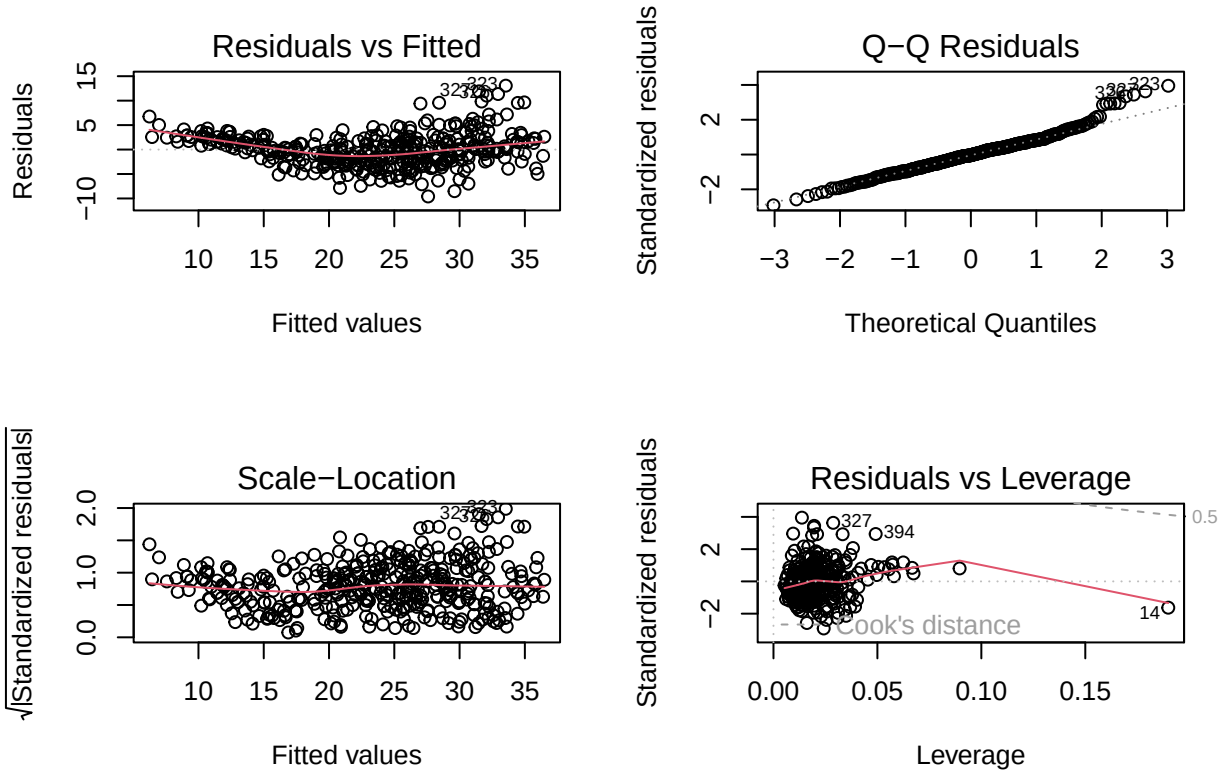
Leverage plot: This plot helps identify influential data points that have a high leverage on the regression fit.

Points with high leverage can significantly impact the regression line. Unusually high leverage points might be influential observations that are far away from the main cluster of data points.

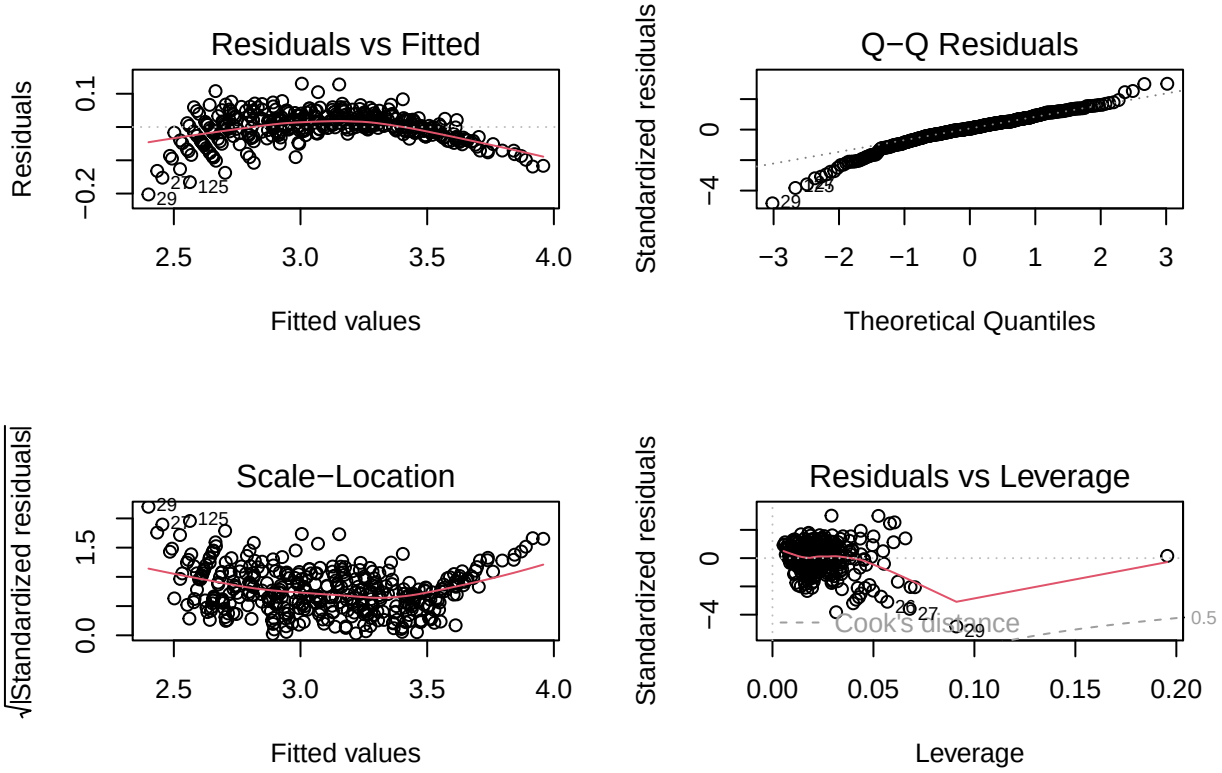
Exericse 1.4

Try a few different transformations of the response, such as $\log(Y)$, $\sqrt{Y}$, Y^2. Comment on your findings.

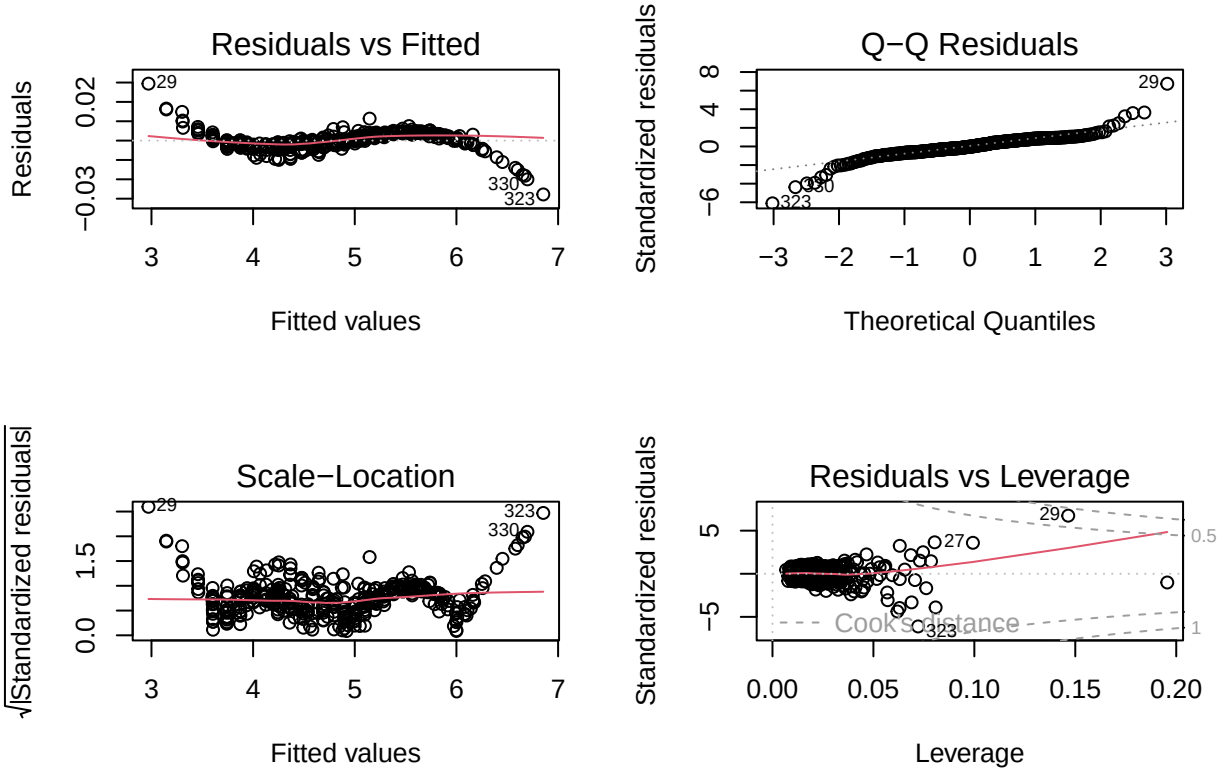
```r
### YOUR CODE HERE
### Load necessary libraries and data (assuming it's already loaded)
require(pacman)
p_load(ggplot2, readr, tidyr, dplyr, ISLR)
data(Auto)

### Original linear regression model
lm_model_original <- lm(mpg ~ . - name, data = Auto)

### Log transformation of the response variable
Auto$log_mpg <- log(Auto$mpg)
lm_model_log <- lm(log_mpg ~ . - name, data = Auto)

### Square root transformation of the response variable
Auto$sqrt_mpg <- sqrt(Auto$mpg)
lm_model_sqrt <- lm(sqrt_mpg ~ . - name, data = Auto)

### Squared transformation of the response variable
Auto$squared_mpg <- Auto$mpg^2
lm_model_squared <- lm(squared_mpg ~ . - name, data = Auto)

### Produce diagnostic plots for each model
par(mfrow = c(2, 2)) # Arrange the plots in a 2x2 grid
plot(lm_model_original)
```



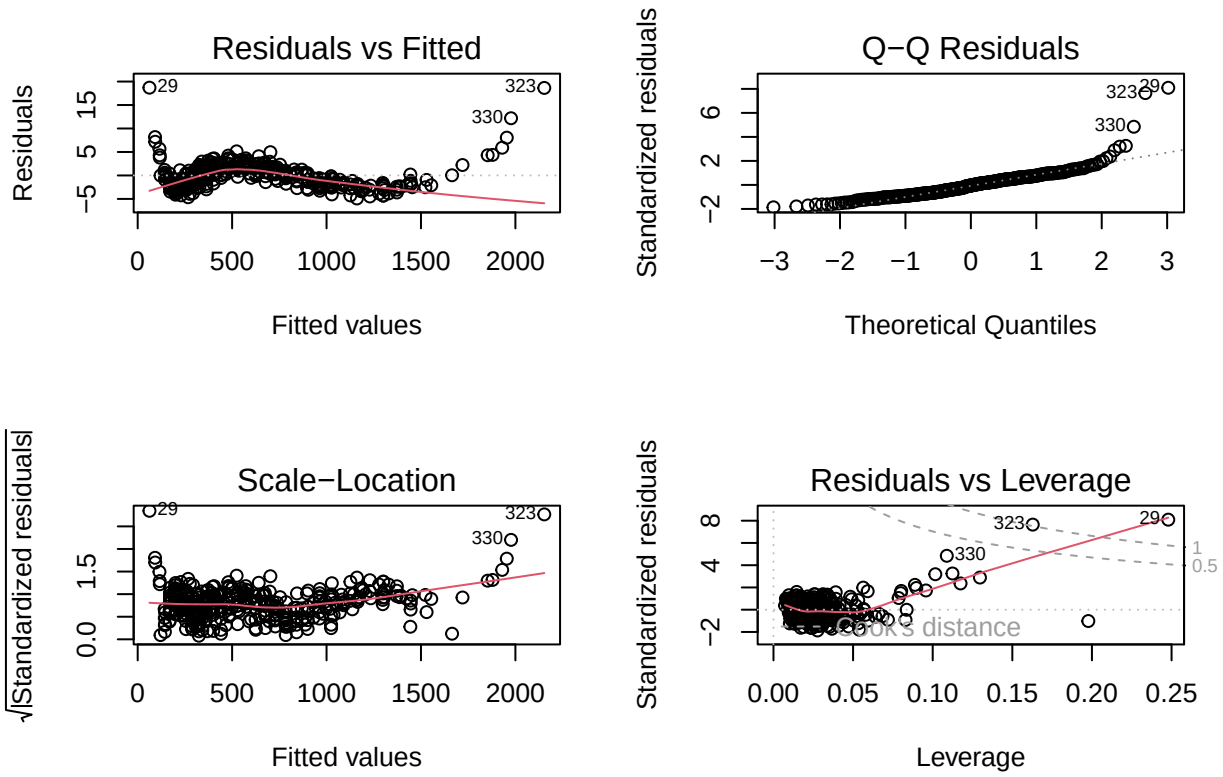

```
plot(lm_model_log)
```




```
plot(lm_model_sqrt)
```




```
plot(lm_model_squared)
```




```
### END OF YOUR CODE
```

YOUR EXPLANATION HERE. Original Linear Model: This is the initial linear regression model without any transformation. We can evaluate the assumption of linearity and check for the presence of influential outliers.

$\log(Y)$ Transformation: The log transformation of the response variable may be useful when the relationship between the predictors and the response is multiplicative rather than additive. The residual vs. fitted plot may show an improved linear relationship, especially if there was exponential growth or decay in the original data. The normal Q-Q plot should also show less deviation from normality.

$\sqrt{Y}$ Transformation: The square root transformation can help when the variance of the response variable is not constant. It often stabilizes the variance and can make the model more robust to outliers.

Y^2 Transformation: Squaring the response variable can be helpful when the relationship between the predictors and the response is curvilinear. The model may capture quadratic relationships better.