Assignment 1 Submission

Himakar Gadham & Atikant Jain

April 19, 2024

Assignment 1 submitted by Himakar Gadham - 23783777, Atikant Jain - 24051868

Both of us contributed equally to all the questions. We both sat together in library and worked through all questions and both did final touches.

Question 1 Normal distribution. (8 marks)

(a) (1 mark) Find the probability that the user spends more than 15 minutes per month at the site.

```
mean <- 25
std_dev <- 4.0
prob_more_than_15 <- 1 - pnorm(15, mean, std_dev)
cat("Probability that the user spends more than 15 minutes per month:", round(prob_more_than_15, 4))
```

```
## Probability that the user spends more than 15 minutes per month: 0.9938
```

(b) (2 marks) Find the probability that the user spends between 20 and 35 minutes per month at the site.

```
prob_between_20_and_35 <- pnorm(35, mean, std_dev) - pnorm(20, mean, std_dev)
cat("Probability that the user spends between 20 and 35 minutes per month:",
    round(prob_between_20_and_35, 4))
```

```
## Probability that the user spends between 20 and 35 minutes per month: 0.8881
```

(c) (2 marks) What is the amount of time per month a user spends on Facebook, if only 1% of users spend this time or longer on Facebook?

```
time_for_1_percent <- qnorm(0.99, mean, std_dev)
cat("Time per month a user spends on Facebook if only 1% of users spend this time or longer:",
    round(time_for_1_percent, 4), "minutes")
```

```
## Time per month a user spends on Facebook if only 1% of users spend this time or longer: 34.3054 minu
```

(d) (3 marks) Between what values do the time spent of the middle 90% distribution of Facebook users fall?

```r
time_middle_90_lower <- qnorm(0.05, mean, std_dev)
time_middle_90_upper <- qnorm(0.95, mean, std_dev)
cat("Time spent by the middle 90% of distribution of Facebook users:",
    round(time_middle_90_lower, 4), "to", round(time_middle_90_upper, 4), "minutes")
```

```
## Time spent by the middle 90% of distribution of Facebook users: 18.4206 to 31.5794 minutes
```

Question 2 Blood fat concentration (11 marks)

(a) (6 marks) Conduct a two-independent sample t-test using R to determine whether the concentration of plasma cholesterol is significantly different between patients with no evidence of heart disease and those with narrowing of the arteries.

```r
# Given values
mean_no_disease <- 195.2745
var_no_disease <- 1303.9231
n_no_disease <- 51
mean_disease <- 216.1906
var_disease <- 1850.2488
n_disease <- 320

# test statistic
t_stat <- (mean_disease - mean_no_disease) / sqrt((var_no_disease / n_no_disease) +
                                                   (var_disease / n_disease))

# degrees of freedom using Welch's approximation
df <- ((var_no_disease / n_no_disease) + (var_disease / n_disease))^2 /
      ((var_no_disease^2 / (n_no_disease^2 * (n_no_disease - 1))) +
       (var_disease^2 / (n_disease^2 * (n_disease - 1))))

# two-tailed test
p_value <- 2 * pt(abs(t_stat), df = df, lower.tail = FALSE)

cat("Test Statistic:", t_stat, "\n")
```

```
## Test Statistic: 3.735664
```

```r
cat("Degrees of Freedom:", df, "\n")
```

```
## Degrees of Freedom: 74.57449
```

```r
cat("p-value:", p_value, "\n")
```

```
## p-value: 0.0003640548
```

1. **State the hypotheses**:
    - Null Hypothesis (H_0): There is no difference in the mean plasma cholesterol concentration ($\mu_1 = \mu_2$) between the two groups.
    - Alternative Hypothesis (H_1): There is a difference in the mean plasma cholesterol concentration ($\mu_1 \neq \mu_2$) between the two groups.

2. **Calculate the test statistic**: Given test statistic $t = 3.735664$
3. **Determine the sampling distribution**: Degrees of Freedom (df): 74.57449
4. **Determine the p-value**: p-value $= 0.0003640548$
5. **Interpretation of p-value**: p-value < 0.01: Reject the null hypothesis
6. **Conclusion**: Since the p-value ($= 0.0003640548$) is less than the significance level of 0.01, we reject the null hypothesis. There is evidence to suggest that there is a significant difference in mean plasma cholesterol concentration between patients with no evidence of heart disease and those with narrowing of the arteries.

(b) (3 marks) Determine a 99% confidence interval for the mean difference in concentration of plasma cholesterol between the two groups of patients.

```r
# Calculate the margin of error
margin_of_error <- qt(0.995, df) * sqrt(var_no_disease/n_no_disease + var_disease/n_disease)

# Calculate the confidence interval
lower_bound <- (mean_disease - mean_no_disease) - margin_of_error
upper_bound <- (mean_disease - mean_no_disease) + margin_of_error

# Print the results
cat("99% Confidence Interval for the Mean Difference in Concentration of Plasma Cholesterol:\n")
```

```
## 99% Confidence Interval for the Mean Difference in Concentration of Plasma Cholesterol:
```

```r
cat("(", lower_bound, ",", upper_bound, ")\n")
```

```
## ( 6.115758 , 35.71644 )
```

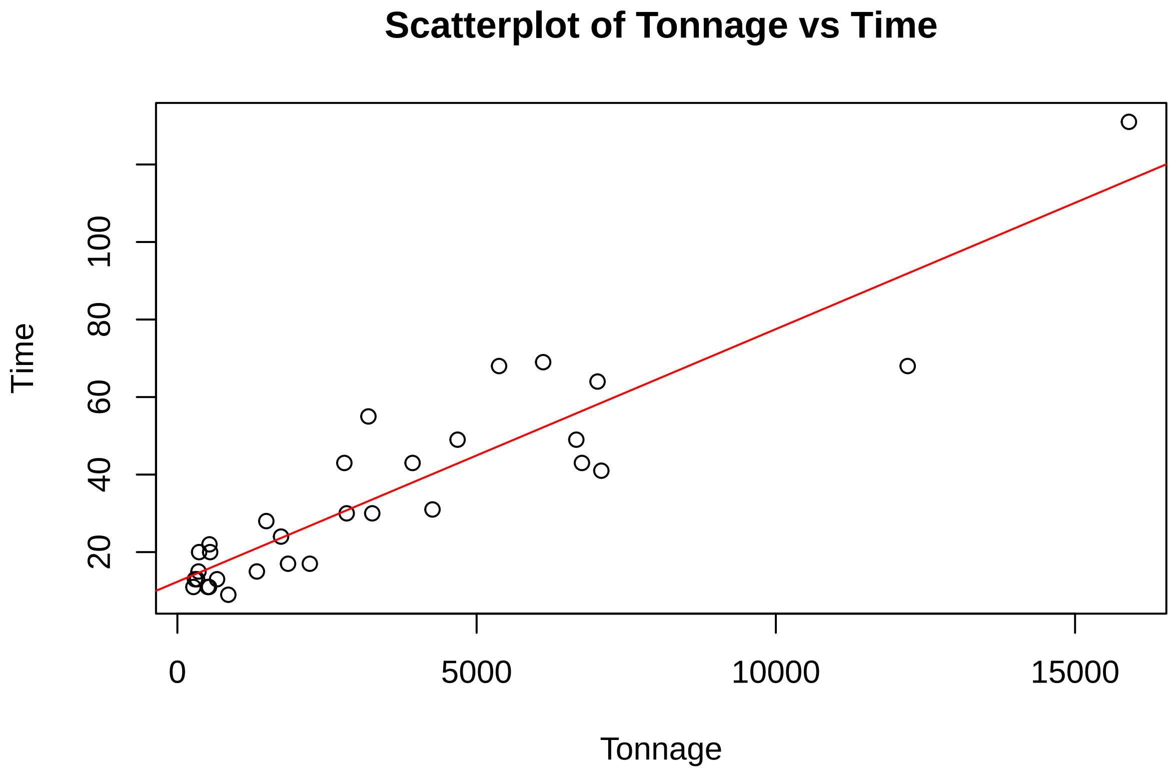
(c) (2 marks) Explain the correspondence between the confidence interval in (b) and a test of the hypotheses you listed in question (a).

The confidence interval (6.116, 35.716) indicates a 99% confidence that the true difference in mean plasma cholesterol concentration between the two groups lies within this range. In the hypothesis test, the null hypothesis (H0) states no difference, while the alternative (H1) suggests a difference. If the confidence interval excludes zero and the p-value is less than 0.01, we reject H0, indicating a significant difference in mean plasma cholesterol concentration between groups.

Question 3 Regression (31 marks)

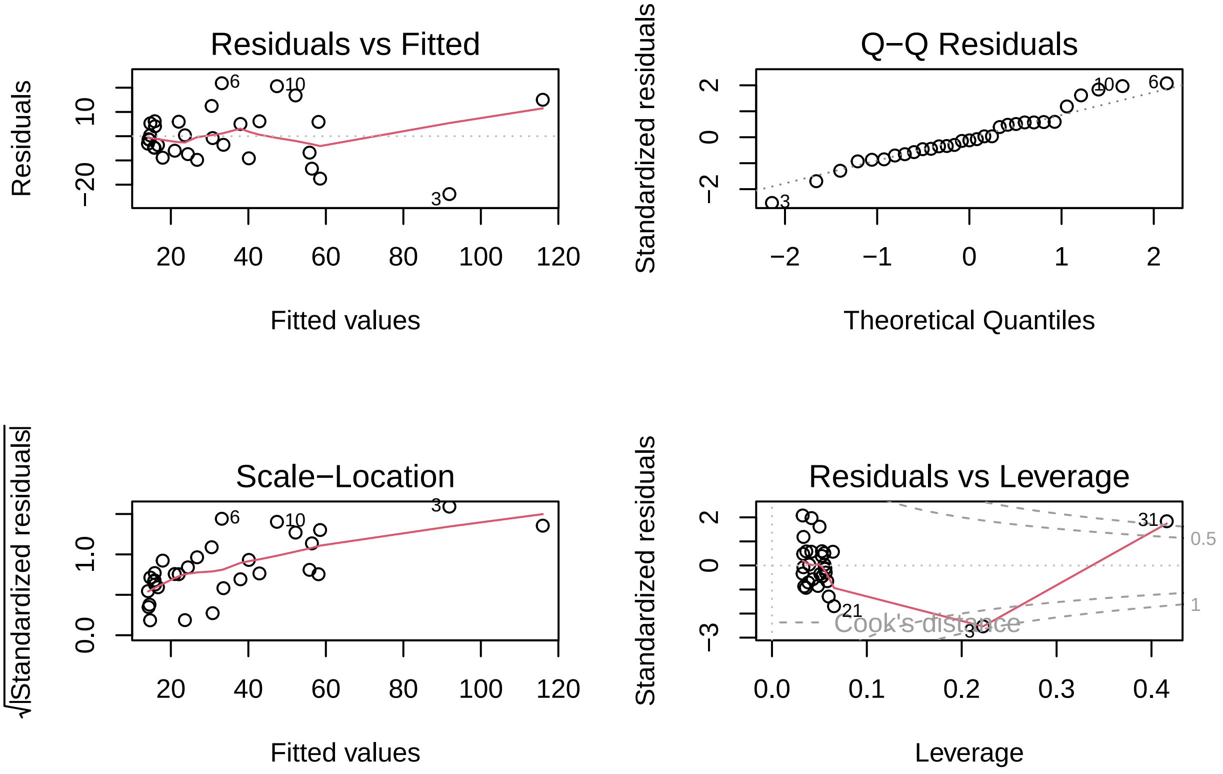
(a) (2 marks) Fit a simple linear model M_1 to these data. Present the appropriate scatterplot and plot the fitted line onto the scatterplot. Comment about the output in a few concise sentences.

The scatterplot depicts the association between the tonnage of liquid-carrying vessels and the duration spent in port. Overlaying the fitted line from the simple linear model $M1$, it becomes evident that while the points exhibit a general linear trend, they often diverge considerably from the line. Additionally, there are indications of potential curvature in the tonnage-time relationship, implying that a linear model may not adequately capture the underlying dynamics.



(b) (5 marks) Provide the model summary and diagnostics checking plots for model M_1. Does the straight line regression model M_1 seem to fit the data well? Comment about the output in a few concise sentences.

```
##
## Call:
## lm(formula = Time ~ Tonnage, data = great_lakes)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -23.882  -6.397  -1.261   5.931  21.850
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.344707   2.642633   4.671 6.32e-05 ***
## Tonnage      0.006518   0.000531  12.275 5.22e-13 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 10.7 on 29 degrees of freedom
## Multiple R-squared:  0.8386, Adjusted R-squared:  0.833
## F-statistic: 150.7 on 1 and 29 DF,  p-value: 5.218e-13
```



Comment: The model summary for M_1 indicates a linear regression model with the equation:

$$\text{Time} = 12.3447 + 0.0065 \times \text{Tonnage}$$

The Adjusted R-squared value is 0.833, suggesting that approximately 83.3% of the variance in the response variable (Time) is explained by the linear relationship with the predictor variable (Tonnage). The coefficient for Tonnage is statistically significant (p-value < 0.05), indicating a significant effect on the time spent in port.

However, diagnostics checking plots show some concerns. There appears to be curvature in the residuals vs. fitted plot, suggesting that the linear model M_1 might not be the best fit for the data. Additionally, the scale-location plot suggests heteroscedasticity in the residuals. Further model refinement or consideration of alternative models may be warranted.

(c) (5 marks) Do you think there are outliers or influential points in the data? What influence do these points have on the model fit? Use leverage and Cook's distance for this investigation.

```r
m1.infl <- influence.measures(model_M1)
m1.infl
```

```
## Influence measures of
##   lm(formula = Time ~ Tonnage, data = great_lakes) :
##
##      dfb.1_ dfb.Tnng    dffit cov.r   cook.d    hat inf
## 1  -0.16212  0.05643 -0.17881 1.047 1.61e-02 0.0358
## 2  -0.04615  0.00270 -0.06098 1.100 1.92e-03 0.0323
## 3   0.53942 -1.39377 -1.50735 0.837 9.17e-01 0.2224   *
## 4   0.00453  0.10407  0.14746 1.121 1.11e-02 0.0643
## 5  -0.10680  0.06684 -0.10717 1.116 5.90e-03 0.0528
## 6   0.31074 -0.02504  0.40448 0.804 7.21e-02 0.0324
## 7   0.09078 -0.05662  0.09112 1.120 4.28e-03 0.0525
## 8   0.05148  0.03696  0.11196 1.087 6.42e-03 0.0362
## 9   0.06606  0.22803  0.38172 0.936 6.87e-02 0.0502
## 10  0.13562  0.20641  0.43365 0.840 8.44e-02 0.0417
## 11 -0.01327 -0.10764 -0.16111 1.106 1.32e-02 0.0583
## 12  0.05435  0.01225  0.08717 1.092 3.90e-03 0.0329
## 13 -0.08950 -0.03696 -0.16225 1.053 1.33e-02 0.0340
## 14 -0.13110  0.05543 -0.13946 1.078 9.90e-03 0.0383
## 15 -0.07979  0.04859 -0.08024 1.121 3.32e-03 0.0509
## 16 -0.03416  0.02220 -0.03421 1.134 6.06e-04 0.0557
## 17  0.18453 -0.03776  0.22137 1.005 2.41e-02 0.0332
## 18  0.00810 -0.00524  0.00811 1.135 3.41e-05 0.0554
## 19 -0.01121  0.00217 -0.01355 1.109 9.50e-05 0.0331
## 20  0.12119 -0.07827  0.12138 1.115 7.56e-03 0.0552
## 21 -0.01022 -0.33009 -0.46373 0.932 1.00e-01 0.0654
## 22 -0.11685  0.06003 -0.12010 1.096 7.38e-03 0.0430
## 23 -0.02909  0.01902 -0.02912 1.136 4.39e-04 0.0563
## 24 -0.07171  0.04711 -0.07177 1.130 2.66e-03 0.0567
## 25  0.00658 -0.00293  0.00694 1.116 2.49e-05 0.0392
## 26 -0.10396  0.06534 -0.10428 1.117 5.59e-03 0.0531
## 27  0.11296 -0.05511  0.11709 1.094 7.02e-03 0.0414
## 28  0.13748 -0.08593  0.13797 1.105 9.74e-03 0.0527
## 29 -0.18948  0.11067 -0.19136 1.072 1.85e-02 0.0485
## 30 -0.02250 -0.22385 -0.32985 1.013 5.31e-02 0.0598
## 31 -0.74091  1.55747  1.62159 1.434 1.20e+00 0.4161   *
```

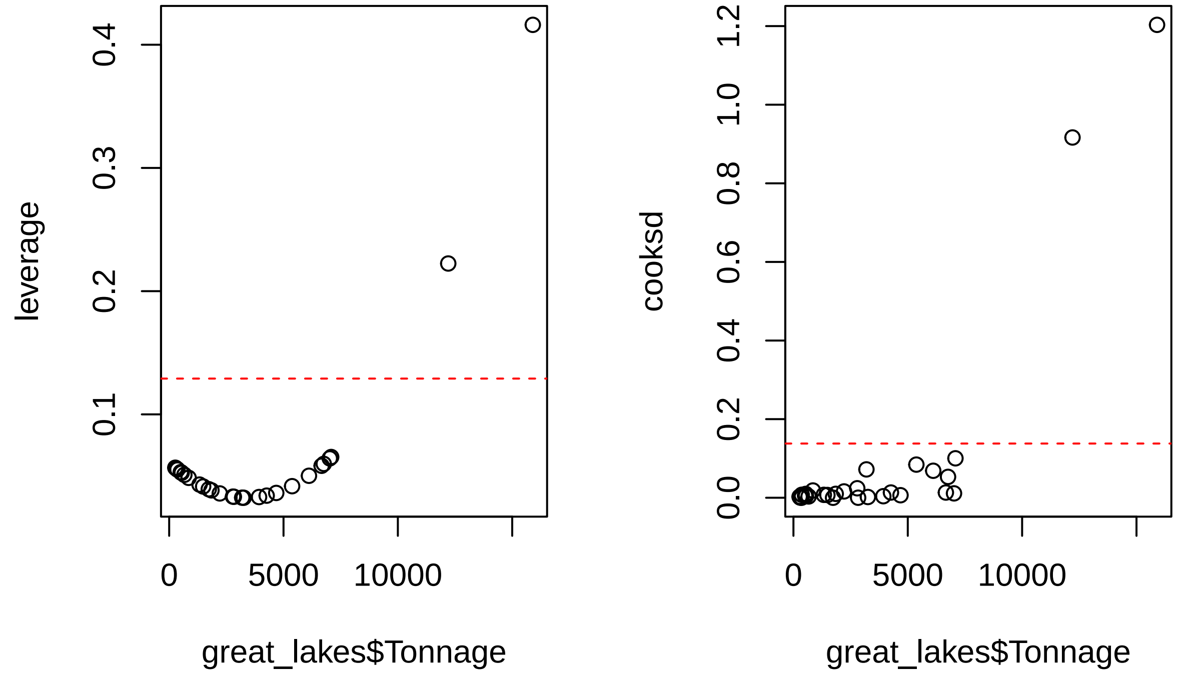
```
leverage <- hatvalues(model_M1) # Compute leverage Cook's distance
cooksd <- cooks.distance(model_M1) # Compute Cook's distance
n=31
```

Plots:

1. Leverage vs Cook's Distance: This plot helps identify influential points based on their leverage and Cook's distance. Points with high leverage and high Cook's distance are potentially influential.

2. Standardized Residuals vs Leverage: This plot helps identify outliers and influential points based on their standardized residuals and leverage. Points with standardized residuals outside the range (-2, 2) and/or high leverage are potentially outliers or influential points.

```
par(mfrow = c(1, 2)) #ploting the influtional measures against their x values
plot(leverage ~ great_lakes$Tonnage)
abline(h = 4 / n, col = "red", lty = 2)
plot(cooksd ~ great_lakes$Tonnage)
abline(h = 4 / (n-2), col = "red", lty = 2)
```

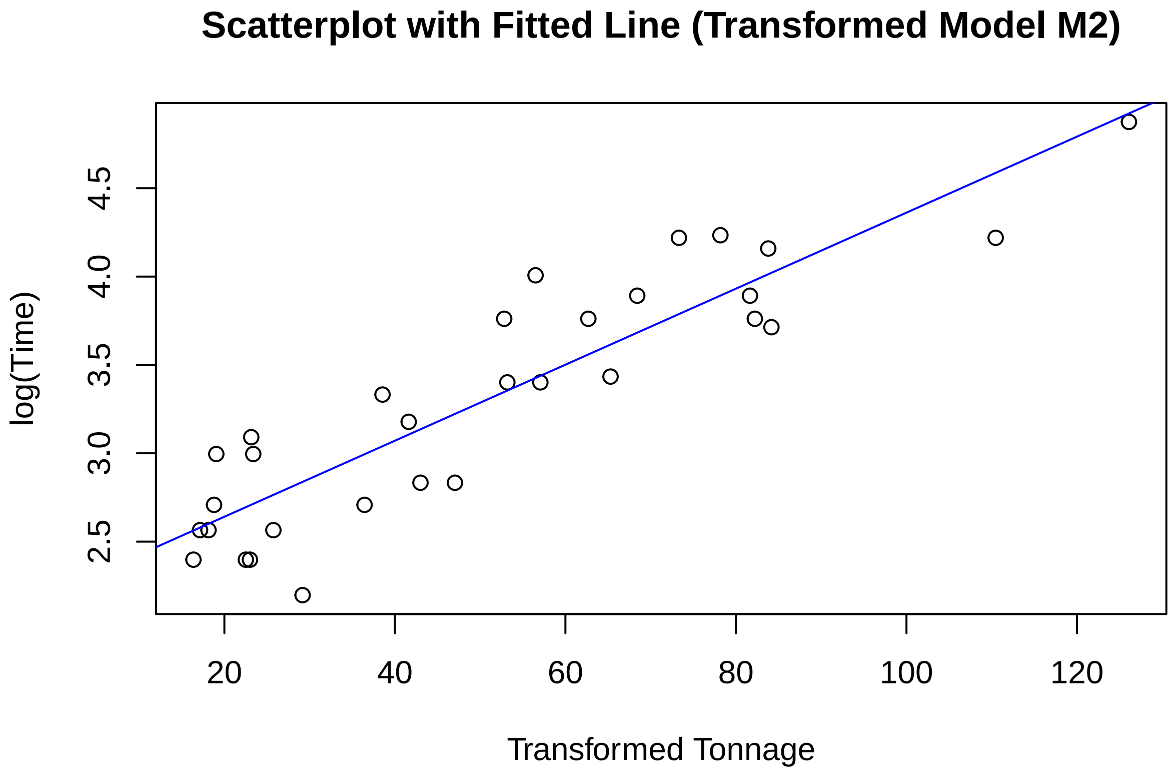


```r
par(mfrow = c(1, 1))
sort(leverage)
sort(cooksd)
```

Comment: We can identify 2 outliers (influential points) because they don not fall between (-2,2) in the standardised residuals vs leverage points plot, making them bad leverage points. These 2 points makes the least squares line inaccurate by dragging it away from it.

(d) (4 marks) Fit a regression model to the transformed M_2 model. Present the appropriate scatterplot and plot the fitted line onto the scatterplot. Does the transformed line regression model M_2 seem to fit the data well? Comment about the output in a few concise sentences.

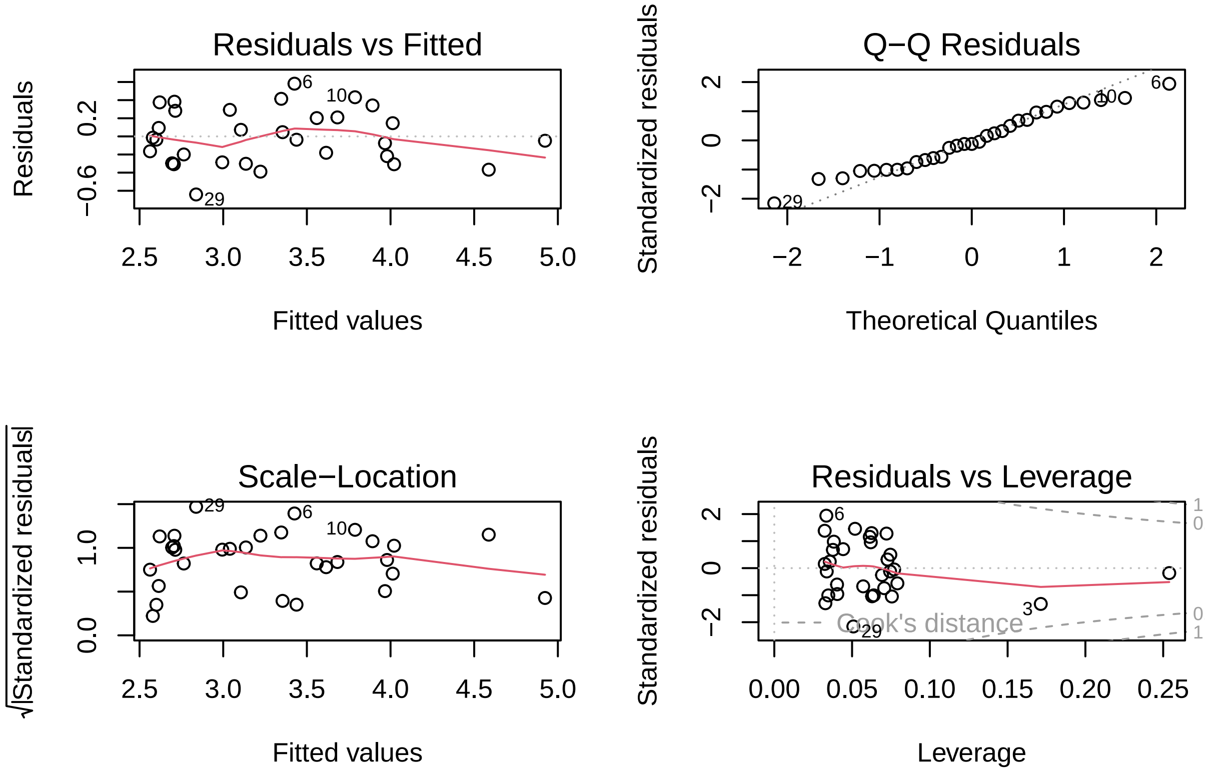
```r
# Fit the regression model to the transformed predictor variable
model_M2 <- lm(log(Time) ~ sqrt(Tonnage), data = great_lakes)

# Scatterplot with the transformed predictor variable and fitted line
plot(log(Time) ~ sqrt(Tonnage), main = "Scatterplot with Fitted Line (Transformed Model M2)",
     xlab = "Transformed Tonnage", ylab = "log(Time)", data = great_lakes)
abline(model_M2, col = "blue")
```



Comment: The transformed regression model M_2 fits the data reasonably well. The scatterplot with the fitted line indicates a positive linear relationship between the transformed predictor variable (sqrt(Tonnage)) and the log of Time. However, there seems to be some curvature in the relationship, suggesting that a linear model might not capture all the nuances of the data. Further exploration or consideration of alternative models may be warranted to improve the fit.

(e) (5 marks) Provide the model summary and diagnostics checking plots for model M_2. Does the straight line regression model M_2 seem to fit the data well? Comment about the output in a few concise sentences.

```
##
## Call:
## lm(formula = log(Time) ~ sqrt(Tonnage), data = great_lakes)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.6408 -0.2522 -0.0357  0.2457  0.5814
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.210424   0.111580   19.81  < 2e-16 ***
## sqrt(Tonnage) 0.021514   0.001909   11.27  4.1e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3048 on 29 degrees of freedom
## Multiple R-squared:  0.8141, Adjusted R-squared:  0.8077
## F-statistic:   127 on 1 and 29 DF,  p-value: 4.098e-12
```



Comment: Based on the provided output, the model summary for M_2 shows that it is a linear regression model with the equation:

$$\log(\text{Time}) = 2.2104 + 0.0215 \times \text{Tonnage\_transformed}$$

The Adjusted R^2 value of 0.8077, indicating that approximately 80.77% of the variance in the log-transformed response variable (Time) is explained by the linear relationship with the transformed predictor variable (Tonnage_transformed). The coefficient for Tonnage_transformed is statistically significant (p-value $<$ 0.001), suggesting that the transformed tonnage has a significant effect on the log of time. However, the diagnostics checking plots, such as residuals vs fitted plot, indicate some curvature in the residuals, which implies that the linear model M_2 might not be the best fit for the data. Additionally, the scatterplot and fitted line might not capture all the nuances in the relationship between the transformed tonnage and the log of time. Further model refinement or consideration of alternative models may be necessary to improve the fit and capture the underlying patterns in the data.

(f) (4 marks) Perform a hypothesis testing for a positive slope at a significance level of 5% based on model M_2.

```
## critical_value:  -2.04523
```

```
## t_stat:  11.27
## p_value:  4.097893e-12
```

Since the absolute value of the t-statistic (11.279) exceeds the critical value (2.04523) and the p-value (4.097893×10^{-12}) is less than 0.05, we reject the null hypothesis. Therefore, there is sufficient evidence to conclude that the slope of the regression line for model M_2 is significantly different from zero and is positive.

(g) (6 marks) Compare a 95% confidence interval of the mean response and a 95% prediction interval for a new value when Tonnage = 10,000 using the untransformed model M_1 and transformed model M_2 respectively. Provide two scatterplots that consist the fitted model, the confidence and prediction intervals for each of M_1 and M_2 respectively. Comment about the output in a few concise sentences.

```r
new_data <- data.frame(Tonnage = 10000)
# 95% confidence interval of the mean response for a new value when Tonnage = 10,000
round(predict(model_M1, newdata = new_data, interval = "confidence", level = 0.95), 4)
```

```
##       fit     lwr     upr
## 1 77.5234 69.3647 85.6821
```

```r
# 95% prediction interval of the mean response for a new value when Tonnage = 10,000
round(predict(model_M1, newdata = new_data, interval = "prediction", level = 0.95), 4)
```

```
##       fit     lwr      upr
## 1 77.5234 54.1705 100.8763
```

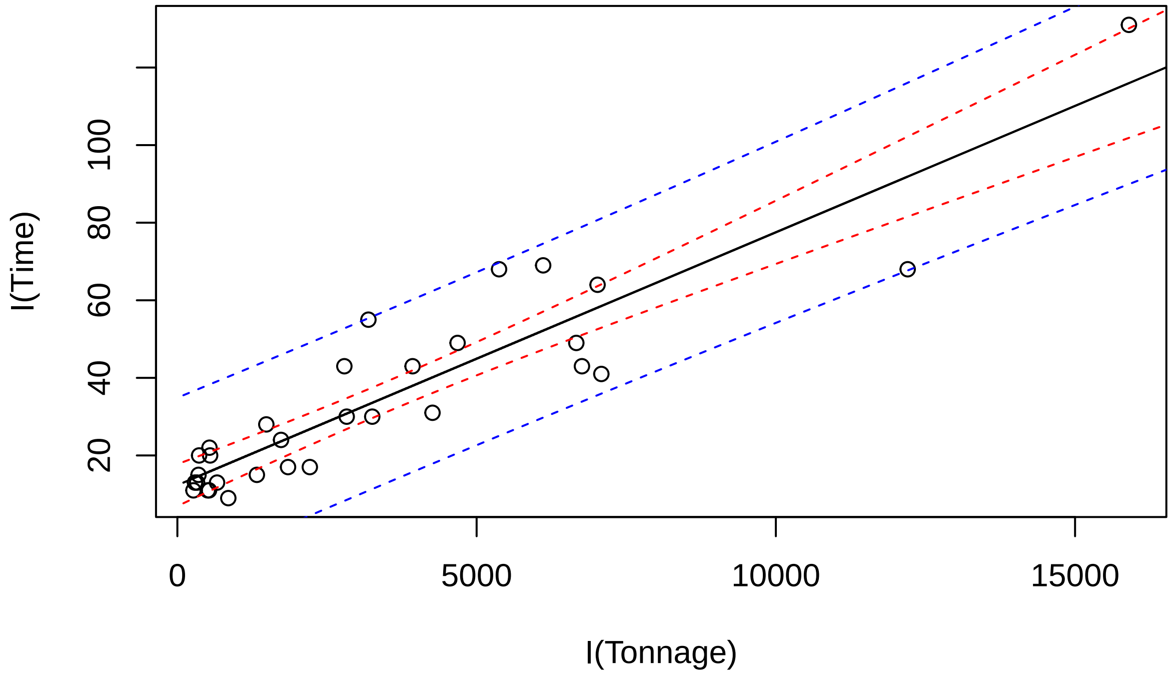
```r
# Model M2
# 95% confidence interval of the mean response for a new value when Tonnage = 10,000
round(predict(model_M2, newdata = new_data, interval = "confidence", level = 0.95), 4)
```

```
##      fit    lwr    upr
## 1 4.3618 4.1399 4.5837
```

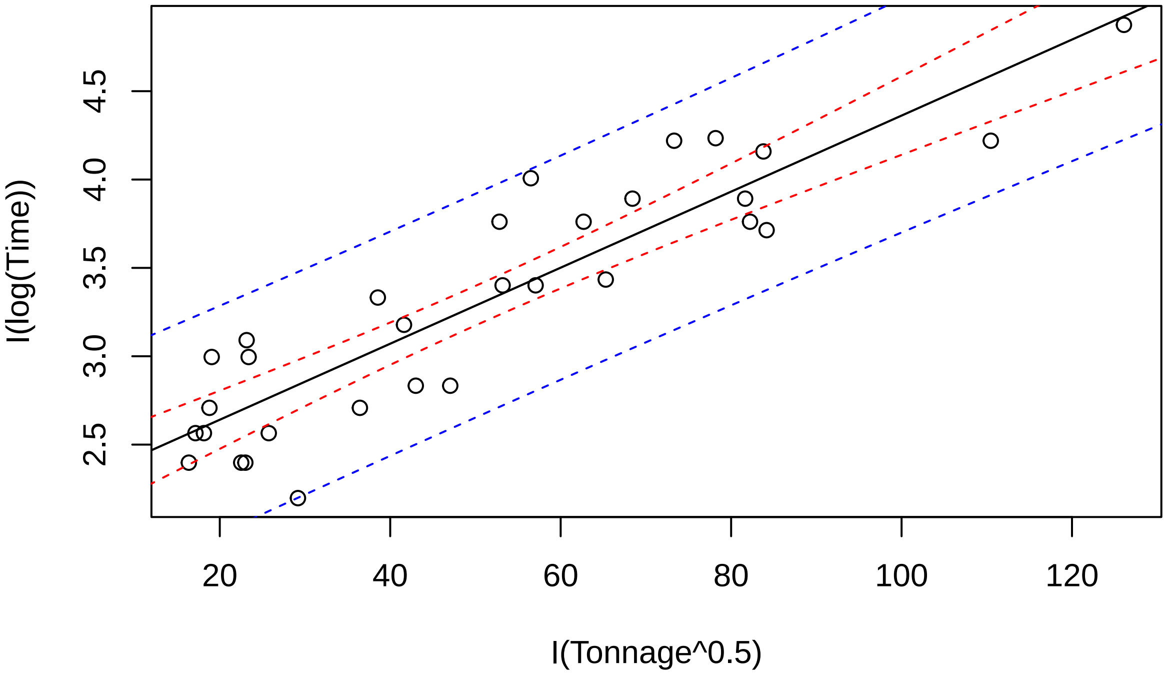
```r
# 95% prediction interval of the mean response for a new value when Tonnage = 10,000
round(predict(model_M2, newdata = new_data, interval = "prediction", level = 0.95), 4)
```

```
##      fit    lwr    upr
## 1 4.3618 3.7002 5.0234
```

```r
# Model M1
with(great_lakes, plot(I(Time)~ I(Tonnage), type = "p"))
with(great_lakes, lines(fitted(model_M1) ~ Tonnage))
new1 = data.frame(Tonnage = seq(100,20000,100))
CIs_M1 = predict(model_M1, new1, interval = "confidence")
PIs_M1 = predict(model_M1, new1, interval = "predict")
matpoints(new1$Tonnage, CIs_M1, lty = c(1, 2, 2), col = c("black", "red", "red"), type = "l")
matpoints(I(new1$Tonnage), PIs_M1, lty = c(1, 2, 2), col = c("black", "blue", "blue"), type = "l")
```



```
# Model M2
with(great_lakes, plot(I(log(Time))~ I(Tonnage^0.5), type = "p"))
with(great_lakes, lines(fitted(model_M2) ~ Tonnage))
new2 = data.frame(Tonnage = seq(100,20000,100))

CIs = predict(model_M2, new2, interval = "confidence")
PIs = predict(model_M2, new2, interval = "predict")
matpoints(I(new2$Tonnage)^0.5, CIs, lty = c(1, 2, 2), col = c("black", "red", "red"), type = "l")
matpoints(I(new2$Tonnage)^0.5, PIs, lty = c(1, 2, 2), col = c("black", "blue", "blue"), type = "l")
```



The 95% confidence interval and prediction interval for a new value when Tonnage = 10,000 are notably different between the untransformed model M_1 and the transformed model M_2.

For Model M_1: - Confidence interval for the mean response: Lower Bound: 69.3647, Upper Bound: 85.6821 - Prediction interval for a new value: Lower Bound: 54.1705, Upper Bound: 100.8763

For Model M_2: - Confidence interval for the mean response: Lower Bound: 4.1399, Upper Bound: 4.5837 - Prediction interval for a new value: Lower Bound: 3.7002, Upper Bound: 5.0234

The scatterplots visualize the fitted model along with the confidence and prediction intervals for both models. In Model M_1, the intervals are wider, reflecting greater variability in predictions due to the untransformed nature of the data. Conversely, Model M_2 exhibits narrower intervals, indicating reduced variability in predictions attributed to the transformation. These plots underscore the impact of transformation on the predictive uncertainty of regression models.