Assignment 1 Submission

Ian Chen, Marco Gunawan

11/04/2025

Statement of Contribution

Each member attempted each question, and we combined and discussed our personal findings to be set into this document.

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.

Answer: 0.9938

```
1 - pnorm(15, mean = 25, sd = 4)
```

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

Answer: 0.8881

```
pnorm(35, mean = 25, sd = 4) - pnorm(20, mean = 25, sd = 4)
```

(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?

Answer: 34.3054

```
qnorm(0.99, mean = 25, sd = 4)
```

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

Answer: (18.42059, 31.5794)

```
qnorm(0.05, mean = 25, sd = 4)
qnorm(0.95, mean = 25, sd = 4)
```

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.

1. Hypotheses:

   $H_0 : \mu_1 = \mu_2$ against $H_A : \mu_1 \neq \mu_2$

   where μ_1 and μ_2 are the population means of plasma cholesterol concentrations in patients with no evidence of heart disease and those with narrowing of the arteries respectively.

2. Assuming unequal variances, the test statistic t can be given as:

$$t = \frac{195.2745 - 216.1906}{\sqrt{\frac{1303.9231}{51} + \frac{1850.2488}{320}}} = -3.7357$$

   Similarly, the degrees of freedom df is given by:

$$df = \frac{\left(\frac{1303.9251}{51} + \frac{1850.2488}{320}\right)^2}{\frac{1}{51-1}\left(\frac{1303.9231}{51}\right)^2 + \frac{1}{320-1}\left(\frac{1850.2488}{320}\right)^2} = 74.5745$$

3. the sampling distribution thus is $t_{df=74.5745}$.

4. The calculated p-value $= 0.0003641 < 0.01$.

5. Decision. Given the p-value is less than the significance level (0.01), we reject the null hypothesis H_0 at the 1% significance level

6. Conclusion. Therefore, we conclude that the population means of plasma cholesterol concentrations in patients with no evidence of heart disease and those with narrowing of the arteries are significantly different.

```
#No Disease
mean1 <- 195.2745
var1 <- 1303.9231
n1 <- 51

#Disease
mean2 <- 216.1906
var2 <- 1850.2488
n2 <- 320
```

```
#calculate t and df
t_stat <- (mean1 - mean2)/sqrt(var1/n1 + var2/n2)
df <- (var1/n1 + var2/n2)^2 / ((var1/n1)^2/(n1-1) + (var2/n2)^2/(n2-1))

#calculate P_val
p_val <- 2 * pt(-abs(t_stat), df)
```

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

A 99% confidene interval for the mean difference in concentration of plasma cholesterol between the two groups of patients can be given as:

(-35.7164, -6.1158)

With the mean_difference $= -20.9161$, standard_error $= 5.5990$, critical t_value $= 2.6434$ and same degrees of freedom from (a), at df $= 74.5745 \sim 75$.

```
# Calculate the mean difference
mean_diff <- mean1 - mean2

# Calculate the standard error of the difference
```

```
se_diff <- sqrt(var1/n1 + var2/n2)

# Find critical t-value for 99% confidence interval (two sided)
t_crit <- qt(0.995, df)

# Calculate the margin of error
margin_error <- t_crit * se_diff

# Calculate the confidence interval
lower_ci <- mean_diff - margin_error
upper_ci <- mean_diff + margin_error
```

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

Since the confidence interval does not contain zero, this aligns with rejecting the null hypothesis at the 1% significance level in question (a).

The confidence interval is entirely negative, meaning that the mean plasma cholesterol concentration in the group with no heart disease is lower than the mean in the group with narrowing of arteries. The magnitude of the difference falling between 6.12 and 35.72 units with 99% confidence.

The fact that zero is not in the interval corresponds directly to the p-value (0.0003641) being less than 0.01, both leading to the same conclusion of rejecting H_0.

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.


Great Lakes Time over Tonnage
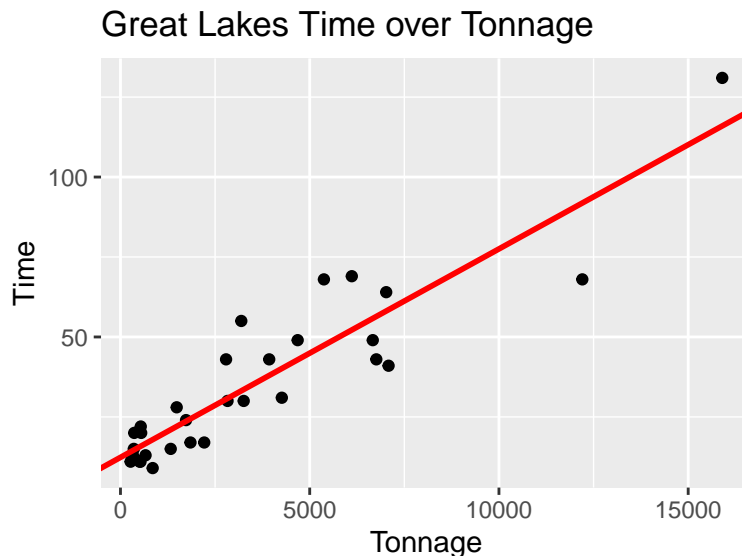
The output depicts a positive linear relationship between the volume of a ship's cargo (tonnage) and the time required to load and unload it. This suggests that the larger the tonnage, the longer it will take to load and unload it. However, the strength of the relationship is weak past the 7500 ton mark as there are insufficient data points in that range. Most observations fall below 7500 tons, with only two data points above this threshold. Hence this model may not be reliable in this higher tonnage range.

3

(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.

The model summary for model M_1:

```
##
## Call:
## lm(formula = Time ~ Tonnage, data = glakes)
##
## 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
```

Diagnostics checking plots for model M_1

Model Summary Analysis:

4
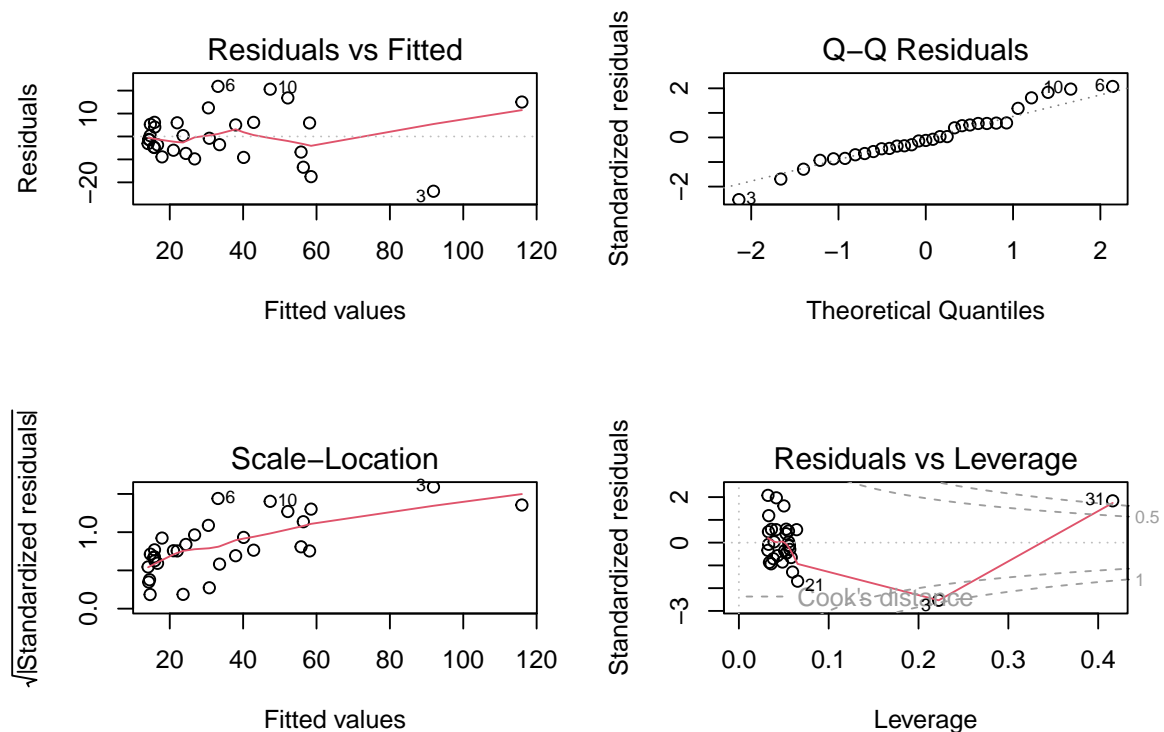
The fitted model equation is : Time= 12.3447+ 0.006518 × Tonnage

Slope interpretation ($\hat{\beta}_1 = 0.006518$): Time is expected to increase by 0.006518 units on average for every 1 unit increase in Tonnage.

Intercept interpretation ($\hat{\beta}_0 = 12.3447$): This means that when the tonnage is zero, the time taken to unload and load the ship is expected to be 12.3447 units. This measurement could be for the time the ship spends in port doing other activities such as refueling or stocking food for the crew.

The p-values for both coefficients are both extremely small, being under 0.001 which means we have strong evidence to reject the null hypothesis. Thus it is suggested that the tonnage has a significant impact on time.

Diagnostic Plots Analysis:

Residuals vs Fitted: The plot shows no clear pattern. This suggests that the relationship is approximately linear.

Q-Q Residuals: The majority of the data-points are on the diagonal line, suggesting the residuals are approximately normally distributed. However. there is some deviation towards the extremes indicating there may be some outliers within the data.

Scale-Location: The line trends slightly upwards which suggests that the variance may not be constant and is in fact increasing.

Residuals vs Leverage: The residuals vs leverage plot depicts some areas that require investigation. A large leverage means that the singular datapoint has a high influence on the output of the model. A high absolute value of the standardized residual means that they are poorly predicted by the model. Therefore the areas that require further investigation are those that have both a high leverage and a standard residual far from zero. Those values are datapoints 31 and 3.

(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.

Having a look at the Cook's distance plot:

```
par(mar = c(4, 4, 1.5, 1.5))
plot(glakes_lm_1, which = 5, cook.levels = 4 / (nrow(glakes) - 2))
abline(h = 2, col = "red", lty = 2)    # horizontal line at y = 2
abline(h = -2, col = "red", lty = 2)   # horizontal line at y = -2
```


5
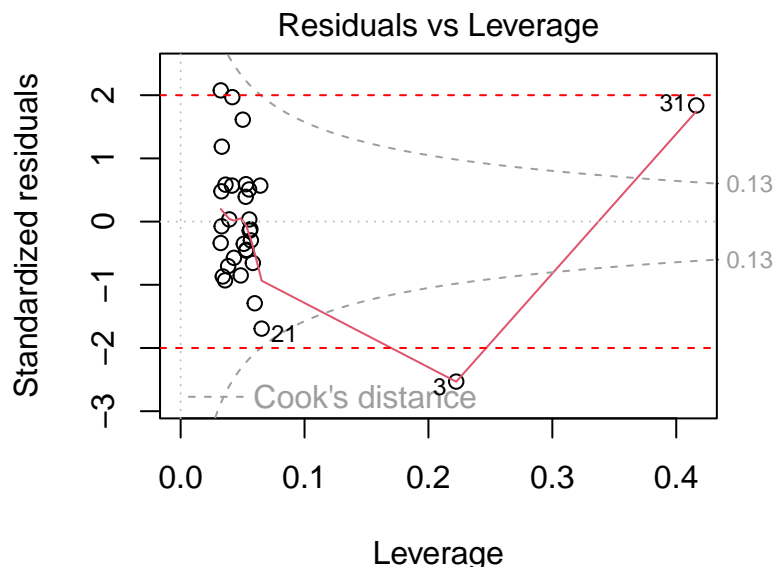
We say that a point is classified as a leverage point in simple linear regression, if it's leverage $h_{ii} > 4/n$. As $n = 31$, the rule is $h_{ii} > 0.13$. As leverage h_{ii} in the plot above has two points with $h_{ii} > 0.13$, there are two leverage points in the data.

We say that points are outliers if their standardized residuals have an absolute value of greater than 2. From the plot above, we can see that there are 2 points whose standardized residuals have an absolute value of greater than 2, such that there are 2 outliers.

Influential points are observations that have a big influence on the fitted line. This happens when the observations are both high leverage and is unusual (potentially an outlier). We note that the only point which meets both of these criteria is observation 3 and fitting a model after removing this point gives us:


Great Lakes Time over Tonnage
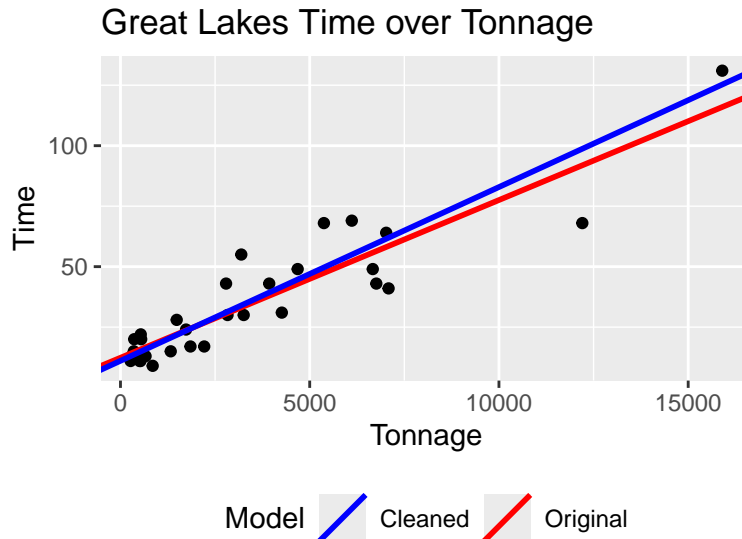
The new model better fits what appear to be a linear relationship

(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.

```
#model M_2 data transformations
log_Time = log(glakes$Time)
sqrt_Tonnage = sqrt(glakes$Tonnage)

#linear regression model M_2
glakes_lm_2 <- lm(log_Time~sqrt_Tonnage, data = glakes)
```
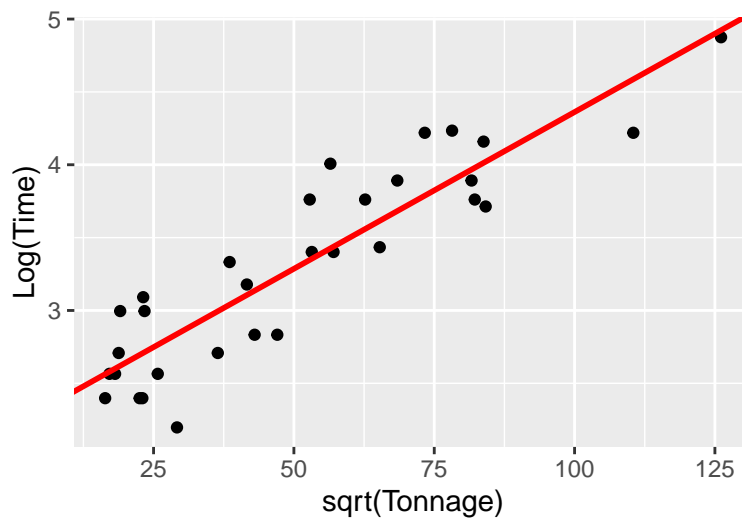
Just like M1, the transformed line (M2) depicts a positive linear relationship between the square root of the volume of a ship's cargo (tonnage) and the logarithm of the time required to load and unload it. This suggests that as the square root of the tonnage increases, the time required to load and unload the ship (as reflected by log(time)) also increases. The strength of the relationship also appears to have improved compared to M1. This is due to the transformations depicting the data-points being spread more evenly throughout the x-axis. However, past approximately the 85 unit mark of sqrt(tonnage), more data-points would be preferred. Overall, M2 fits the data better than M1, being more reliable in the high-tonnage ranges but could still use a few more data-points.

(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 = glakes)
##
## 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
```
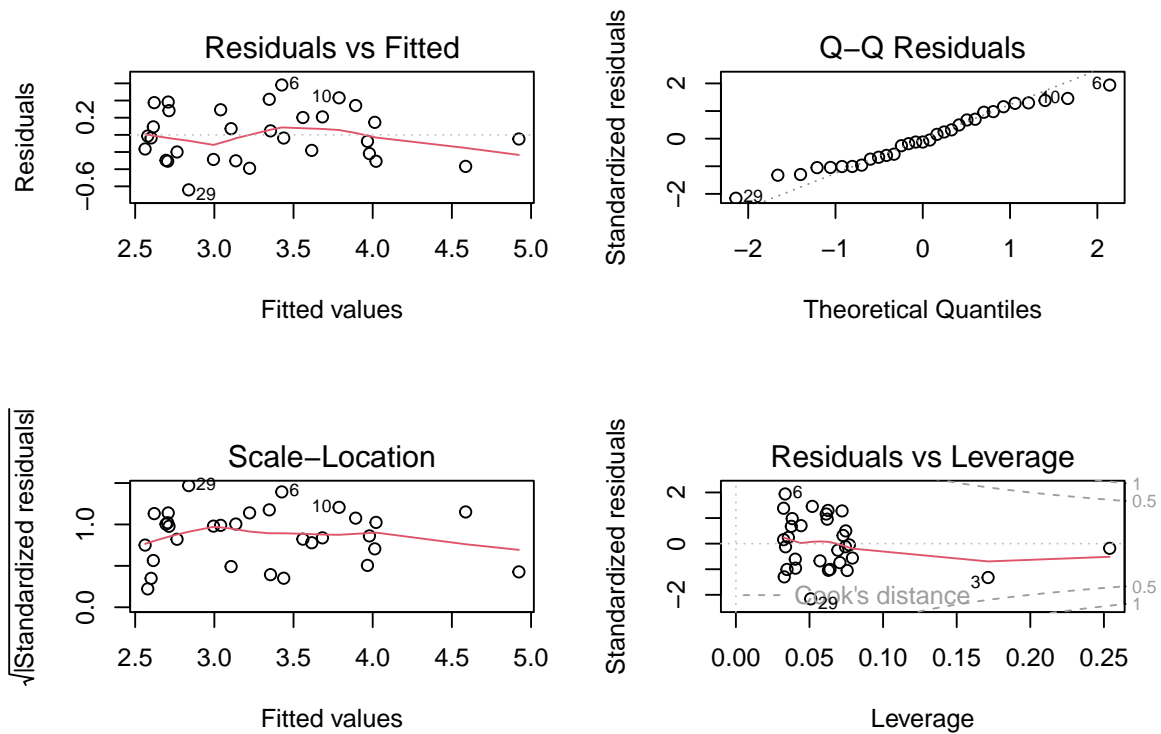
Model Summary Analysis:

The fitted model equation is: $\log(\text{Time}) = 2.2104 + 0.02151 \times \text{sqrt(Tonnage)}$

Slope interpretation ($\hat{\beta}_1 = 0.02151$): For every unit increase in the square root of tonnage, the log of time is expected to increase by an average of 0.0215 units.

Intercept interpretation ($\hat{\beta}_0 = 2.2104$) When the tonnage is zero, the logarithm of time is expected to be 2.2104. This means the time is approximately 9.12 units. This measurement could be for the time the ship spends in port doing other activities such as refueling or stocking food for the crew.

The p-values for both coefficients are both less than 0.001. This provides strong evidence that rejects the null hypothesis. Therefore, it is suggested that there is a significant relationship between the square root of tonnage and logarithm of time.

Diagnostic Plots Analysis:

Residuals vs Fitted: The plot depicts no clear patterns. Therefore, the relationship is suggested to be reasonably linear.

Q-Q Residuals: Most of the observations lie along the diagonal, except a few data points at the extremities of the graph. This suggests the residuals are generally linear.

Scale-Location: There is no clear pattern visible but that may be due to the limited amounts of data past the 4.0 range of fitted values. Based off the data given, it is suggested that the variance is constant.

Residuals vs Leverage: Again we are looking for data points in the graph with both a high leverage and a standardized residual far from zero. For the most part all data points follow that trend except data point 3, which has a slightly low standardized residual value for it being larger in leverage. Hence further investigation of data point 3 is recommended.

8

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

1. $H_0 : \beta_1 \leq 0$ against $H_A : \beta_1 > 0$, where β_1 is the slope of the model M_2. Note that the alternative hypothesis is one sided.

2. The test statistic (from R output as shown in question (e)):

$$t = \frac{\hat{\beta} - 0}{se(\hat{\beta})} = \frac{0.021514}{0.001909} = 11.27$$

3. The sampling distribution for the test statistic t is $t_{\text{df}=(n-2)}$, that is $t_{df=31-2=29}$.

4. The p-value $= P(t_{29} > 11.27) = 2.048434 \times 10^{-12}$

5. **Decision**: Given the p-value is less than the significance level (0.05), we reject the null hypothesis H_0 at the 5% significance level.

6. **Conclusion**: Therefore, we conclude that the slope is statistically significantly 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.

When Tonnage $= 10,000$, for model M_1, the confidence interval of the mean response is given as (69.3647, 85.6821) and the prediction interval is given as (54.1705, 100.8763).

```
#95% confidence and prediction interval for model 1
newdata_1= data.frame(Tonnage = 10000)
conf_int_1 <- predict(glakes_lm_1, newdata = newdata_1,
                      interval = "confidence", level = 0.95)
pred_int_1 <- predict(glakes_lm_1, newdata = newdata_1,
                      interval = "predict", level = 0.95)
```

For model M_2, as it is a transformed model, when Tonnage $= 10,000$, $\sqrt{Tonnage} = 100$. The confidence interval of the mean response is given as (4.1400, 4.5837) and the prediction interval is given as (3.7002, 5.0234).

```
#95% confidence and prediction interval for model 2
newdata_2= data.frame(sqrt_Tonnage = 100)
conf_int_2 <- predict(glakes_lm_2, newdata = newdata_2,
                      interval = "confidence", level = 0.95)
pred_int_2 <- predict(glakes_lm_2, newdata = newdata_2,
                      interval = "predict", level = 0.95)
```

For both of these models, the prediction intervals are wider than the confidence intervals because they include both:

1. Uncertainty about the mean response (as in confidence intervals)
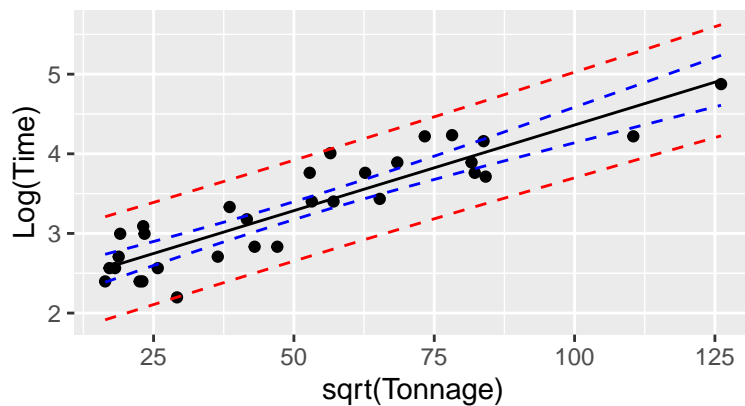2. Random variation of individual observations around the mean

Taking a look at the plots for the fitted model, the confidence and prediction intervals for each of M_1 and M_2 respectively:

Confidence and Prediction Intervals for mod
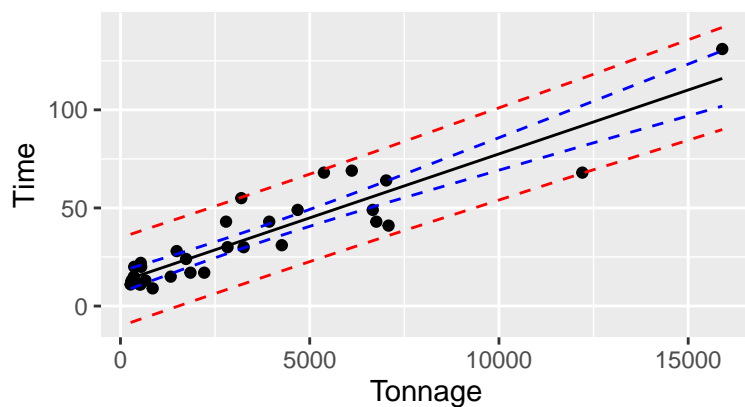
We see that the prediction intervals is slimmer than the confidence interval, and are tightest in the middle of the data, widening at the extremes.

Most data points also falls within the prediction interval, indicating the model captures the overall variability well.