

Assignment 1 Submission

Ian Chen, Marco Gunawan

submission date

Statement of Contribution

Include this if pairing.

State what each team member has contributed to the assignment before answering the questions. Each student must contribute towards ALL questions. Your assignment will not be marked without this statement.

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

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

Answer: 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?

Answer: 34.3054

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

Answer: (18.42059, 31.5794)

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.

Answer:

1. $H_0 : \mu_1 = \mu_2$ 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. the test statistic is $t = -3.7357$

3. the sampling distribution is $t_{df=75}$.

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

5. Reject H_0 at the 1% significance level.
6. 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.

```
#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.

Answer:

A 99% confidence 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 ~ 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).

Answer:

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

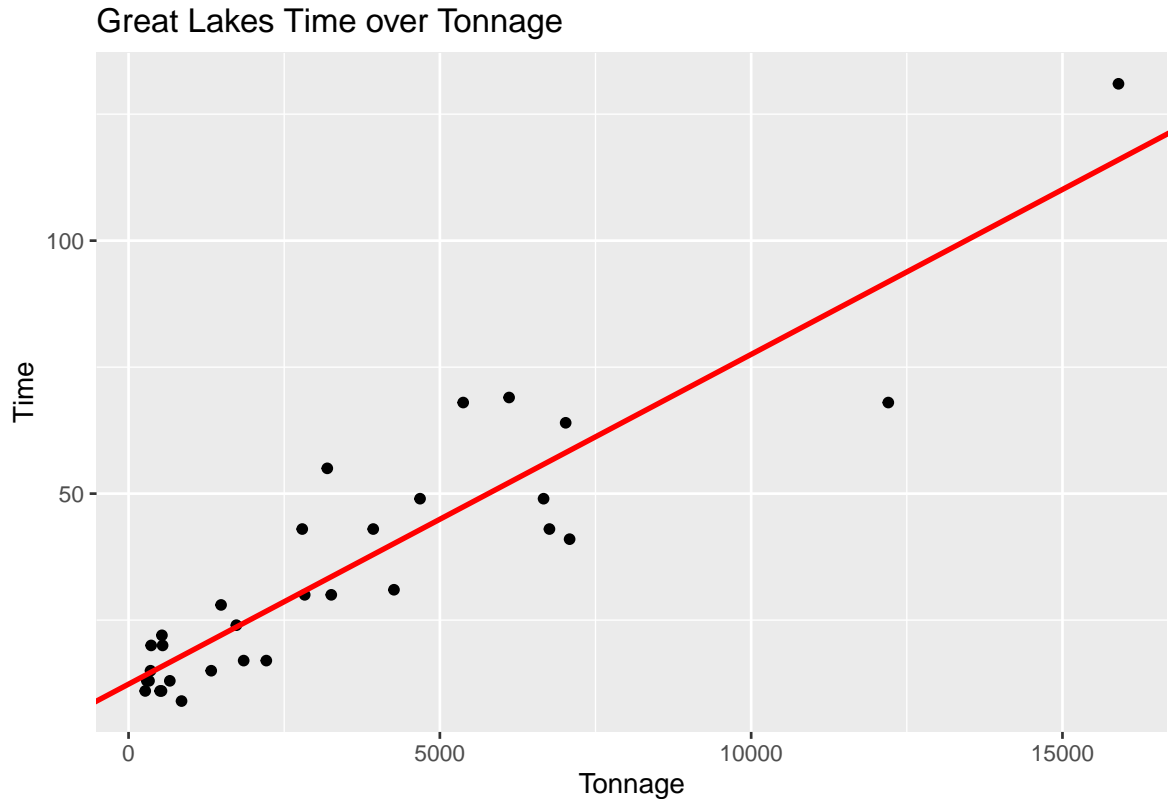
The confidence interval is entirely negative, meaning that the mean plasma cholesterol concentration in the group with no heart disease is lower than the mean in the group with narrowing of arteries. The magnitude of the difference falling between 6.12 and 35.72 units with 99% confidence.

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.

Answer:



- (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.

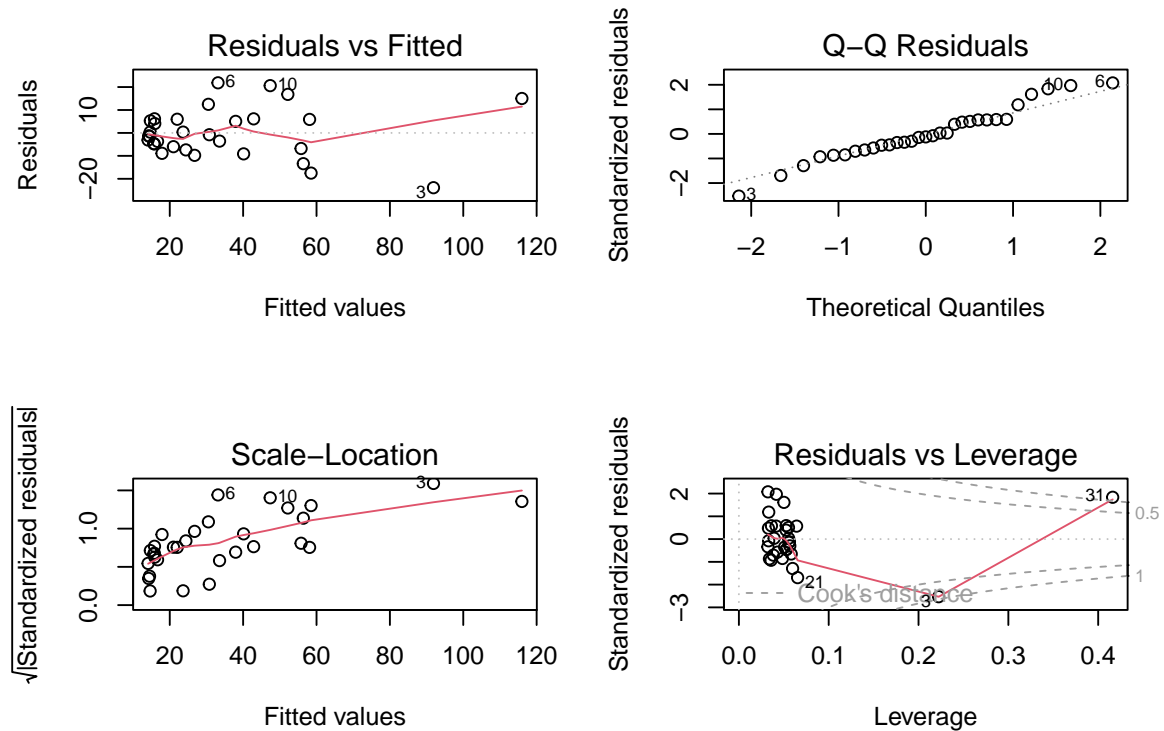
Answer:

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



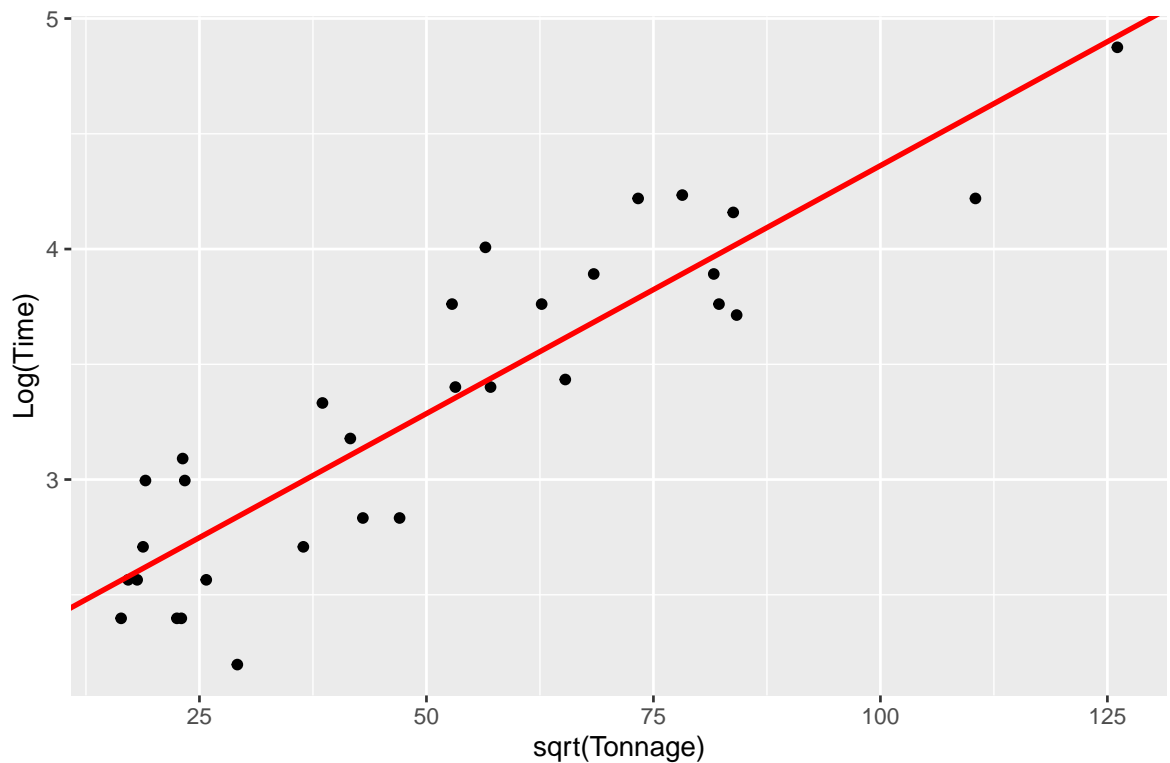
- (c) (5 marks) Do you think there are outliers or influential points in the data? What influence do these points have on the model fit? Use leverage and Cook's distance for this investigation.

Answer:

Answer text here

- (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.

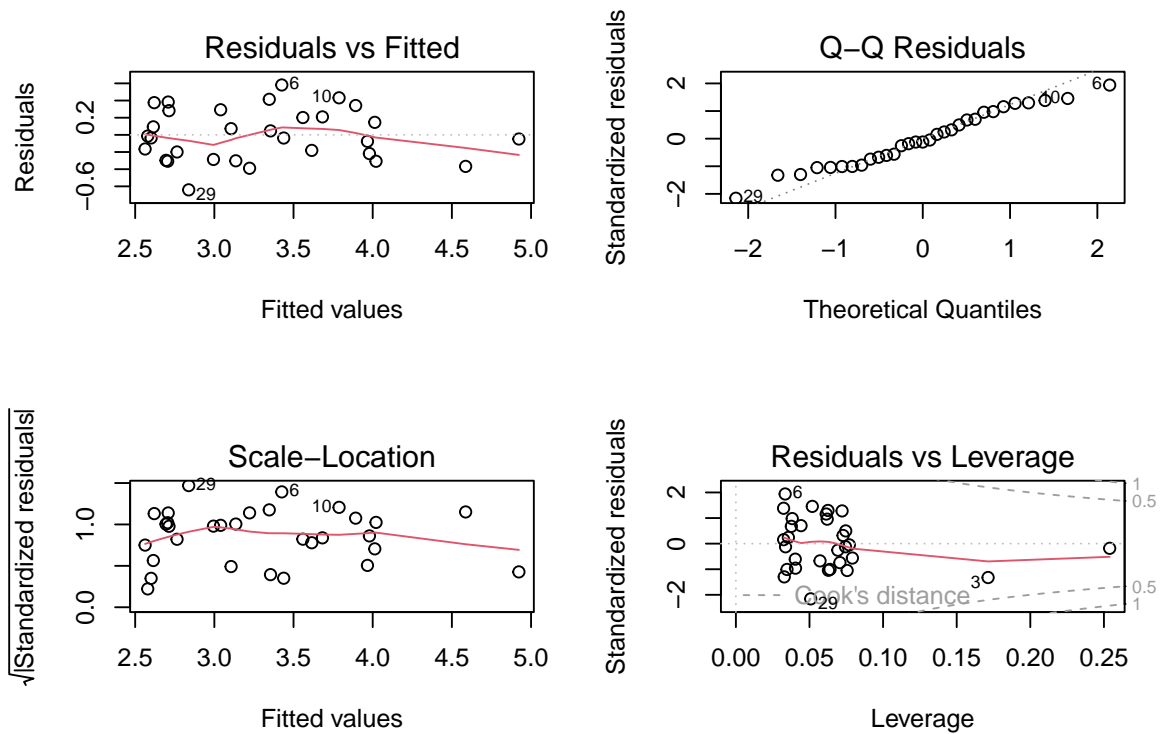
Answer:



- (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.

Answer:

```
##
## 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
```



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

Answer:

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 $t_{df=(n-2)}$ that is $t_{df=31-2=29}$.
4. The p-value = $P(t_{29} > 11.27) = 2.048434e-12$ (see below)

```
pt(11.27, 29, lower.tail=FALSE)
```

```
## [1] 2.048434e-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.

- (f) (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.

Answer:

Answer text here