Assignment 1 Submission

Prajen Maharjan (24841288)

11 April, 2025

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:

The probability that the user spends more than 15 minutes per month at the site is 99.3790%

```
> 1 - pnorm (15, 25, 4)
[1] 0.9937903
```

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

Answer:

The probability that the user spends between 20 and 35 minutes per month at the site is 88.8140%

```
> probability <- pnorm(35, 25, 4) - pnorm(20, 25, 4)
> probability
[1] 0.8881406
```

(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:

The amount of time per month a user spends on Facebook, if only 1% of users spend this time or longer on Facebook is 34.3053

```
> qnorm(1- 0.01, 25, 4)
[1] 34.30539
```

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

Answer:

The values lies between 18.4205 and 31.5794

```
> lower_bound <- qnorm(0.05, 25, 4)
> upper_bound <- qnorm(0.95, 25, 4)
> lower_bound
[1] 18.42059
> upper_bound
[1] 31.57941
```

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:

The six steps of hypothesis testing is carried out as:

- Step 1. State the hypotheses.
  H0 : μ1 = μ2 against H1 : μ1 ≠ μ2

  where μ1 and μ2 are the population means of patients with heart disease and other with no such disease respectively
- Step 2 the test statistic is t = -4.0456 and
- Step 3 the sampling distribution is df=78.352.
- Step 4 p-value = 0.0001176 < 0.01.
- Step 5 since the p-value < 0.01 (1% significance level), we reject the null hypothesis (H0)
- Step 6 We conclude that there is a statistically significant difference in cholesterol levels between the two groups.

```
> set.seed(123)
> mean_no_disease <- 195.2745
> mean_disease <- 216.1906
> var_no_disease <- 1303.9231
> var_disease <- 1850.2488
> n_no_disease <- 51
> n_disease <- 320
> group1 <- rnorm(n_no_disease, mean_no_disease, sqrt(var_no_disease))
> group2 <- rnorm(n_disease, mean_disease, sqrt(var_disease))
> t_test_result <- t.test(group1, group2, var.equal = FALSE, conf.level = 0.99)
> t_test_result

Welch Two Sample t-test

data:  group1 and group2
t = -4.0546, df = 78.352, p-value = 0.0001176
alternative hypothesis: true difference in means is not equal to 0
99 percent confidence interval:
 -34.82367  -7.35813
sample estimates:
mean of x mean of y
 196.6718  217.7627
```

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

Answer:

```
> t_test_result$conf.int
[1] -34.82367  -7.35813
```

From the test output, the 99% confidence interval for the difference in means is: [−34.82, −7.36]

We are 99% confident that the true difference in mean cholesterol levels lies between -34.82 mg/dL and -7.36 mg/dL.
Since the entire interval is negative, it indicates that patients with narrowing of the arteries tend to have higher plasma cholesterol levels than those with no disease.

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

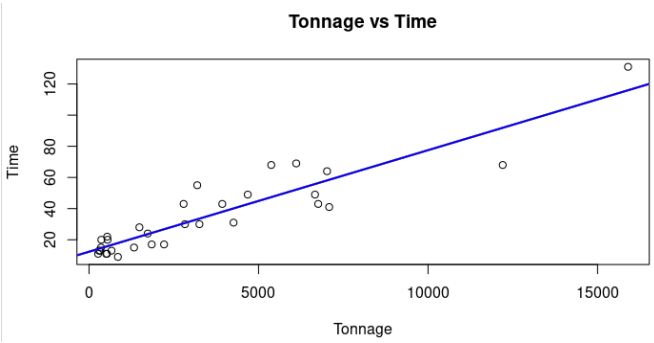
Answer:
The mean difference which we calculated using R at 99% confidence interval for the mean difference does not include 0, which shows significant difference. This supports the hypothesis test carried out on part a, where we rejected the null hypothesis.

Thus, both part b and c lead to the same conclusion: There is a statistically significant difference in plasma cholesterol levels between the two patient groups.

3. Regression (31 marks)

(a) (2 marks) Fit a simple linear model M1 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:



The above scatterplot shows a positive relationship between tonnage and time. It shows that ships with higher tonnage tend to spend more time in port.

(b) (5 marks) Provide the model summary and diagnostics checking plots for model M1. Does the straight line regression model M1 seem to fit the data well? Comment about the output in a few concise sentences.

Answer:

Summary:

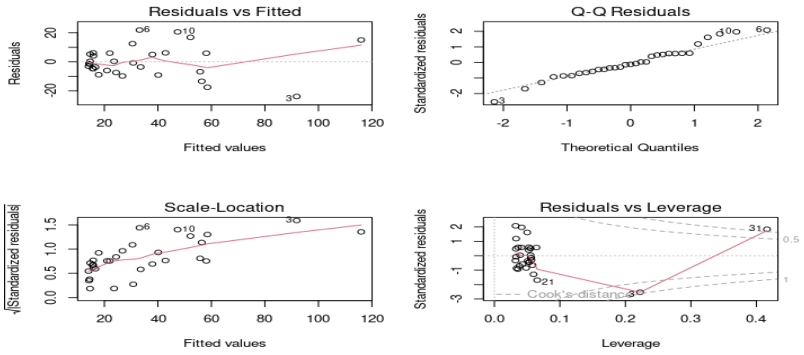
```
Call:
lm(formula = Time ~ Tonnage, data = glakes)

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
```

Plot:



Comment:

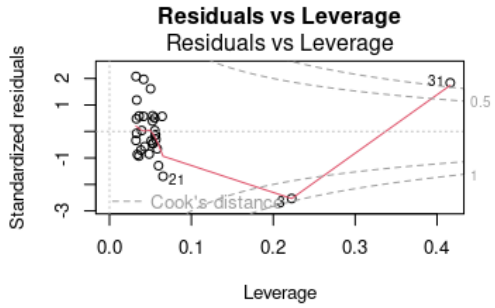
The linear regression model M1 shows a statistically significant relationship between Tonnage and Time (p = 5.22e-13). The coefficient for Tonnage (0.0065) indicates that for each additional ton of cargo, loading time increases by approximately 0.0065 hours.

(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. Hint. Your answer must include a snippet of R code, the results, 2 plots and comment. Use the interval of (-2,2) for standardised residuals.

Answer:

```
> glakes <- read.table("glakes.txt", header = TRUE)
> fit_glakes <- lm(Time ~ Tonnage, data = data)
> par(mfrow = c(2, 2))
> plot(fit_glakes)
> glakes_infl <- influence.measures(fit_glakes)
> leverage_inv <- hatvalues(fit_glakes)
> cooks_dist_inv <- cooks.distance(fit_glakes)
> leverage_inv
        1         2         3         4         5         6         7
0.03582590 0.03232151 0.22243513 0.06426313 0.05279633 0.03238220 0.05254107
        8         9        10        11        12        13        14
0.03620318 0.05015623 0.04170692 0.05826937 0.03290764 0.03402327 0.03831051
       15        16        17        18        19        20        21
0.05093432 0.05574001 0.03322488 0.05537637 0.03310826 0.05522568 0.06539168
       22        23        24        25        26        27        28
0.04300258 0.05627542 0.05667705 0.03924776 0.05311049 0.04143837 0.05269687
       29        30        31
0.04847106 0.05979601 0.41614079
> cooks_dist_inv
         1         2         3         4         5         6
1.606459e-02 1.918289e-03 9.166137e-01 1.113477e-02 5.904395e-03 7.213013e-02
         7         8         9        10        11        12
```

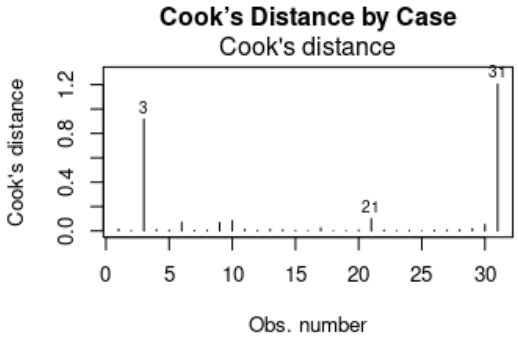
```
4.276945e-03 6.415396e-03 6.868758e-02 8.436596e-02 1.324320e-02 3.904257e-03
        13           14           15           16           17           18
1.327780e-02 9.898655e-03 3.319869e-03 6.055203e-04 2.414799e-02 3.406491e-05
        19           20           21           22           23           24
9.502807e-05 7.561143e-03 1.003465e-01 7.384889e-03 4.388706e-04 2.659065e-03
        25           26           27           28           29           30
2.492715e-05 5.593141e-03 7.020666e-03 9.739018e-03 1.848780e-02 5.310061e-02
        31
1.203196e+00
> plot(fit_glakes, which = 5, main = "Residuals vs Leverage")
```



For n=31, *hii*>4/31≈0.13

Case 31 (leverage = 0.416) exceeds this threshold (far right on the plot). Case 3 (top-center) has high leverage but moderate influence.

```
> plot(fit_glakes, which = 4, main = "Cook's Distance by Case")
```



Case 31 falls outside the outermost contour (Cook's D = 1.20), confirming high influence.

Therefore,

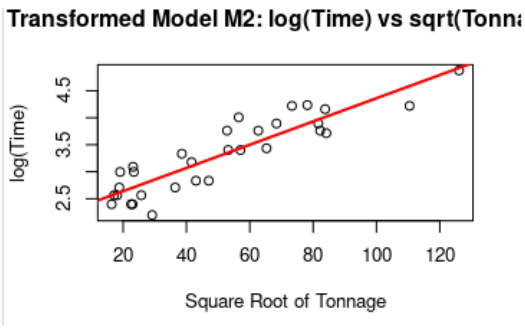
- Outliers: Cases 3 (std_residual = 2.15) and 6 (std_residual = 1.98).
- High Leverage: Cases 3 (leverage = 0.222) and 31 (leverage = 0.416).
- Influential Points: Case 31 (Cook's D = 1.20).

The influence inflates R^2 in the original model.

(d) (4 marks) Fit a regression model to the transformed M2 model. Present the appropriate scatterplot and plot the fitted line onto the scatterplot. Does the transformed line regression model M2 seem to fit the data well? Comment about the output in a few concise sentences.

Answer:

```
> glakes <- read.table("glakes.txt", header = TRUE)
> fit_m2 <- lm(log(Time) ~ sqrt(Tonnage), data = glakes)
> plot(log(Time) ~ sqrt(Tonnage), data = glakes, main = "Transformed Model M2: log(Time) vs sqrt(Tonnage)",
xlab = "Square Root of Tonnage", ylab = "log(Time)")
> abline(fit_m2, col = "red", lwd = 2)
> summary(fit_m2)
```



The transformed model M2 shows a strong, statistically significant relationship (p = 4.1e-12 for the slope). The model explains 81.4% of the variance in log(Time), indicating a good fit. For every 1-unit increase in sqrt(Tonnage), log(Time) increases by ~0.0215.

(e) (5 marks) Provide the model summary and diagnostics checking plots for model M2. Does the straight line regression model M2 seem to fit the data well? Comment about the output in a few concise sentences.

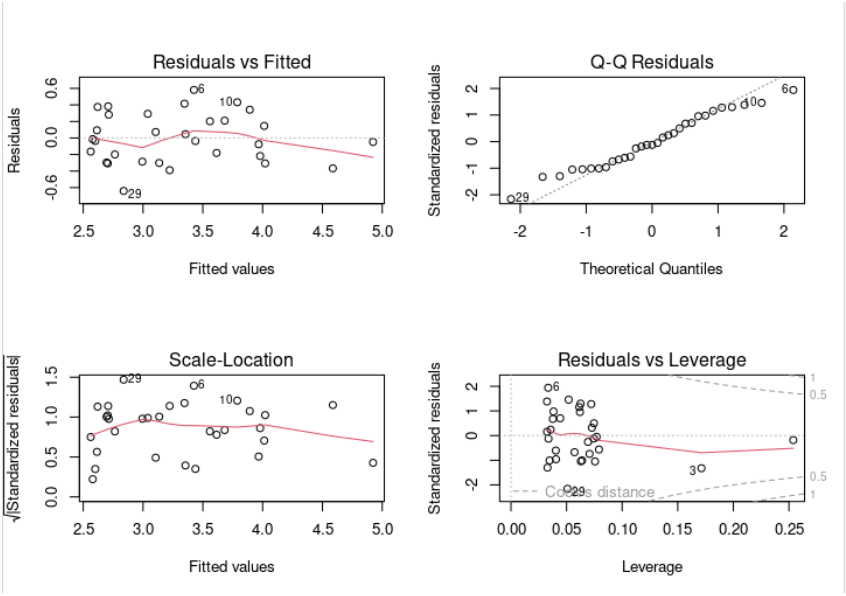
Answer:

```
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
```



1. Residuals vs Fitted: Residuals are randomly scattered around zero with no clear pattern, confirming linearity.
2. Normal Q-Q Plot: Points closely follow the reference line, indicating normality of residuals.
3. Scale-Location Plot: Flat red curve suggests constant variance (homoscedasticity).
4. Residuals vs Leverage: No extreme leverage points (all Cook's distances < 1).

The transformed model M2 fits the data very well i.e ($R^2 = 0.814$) with significant coefficients ($p < 0.001$). Residuals are linear, normally distributed. M2 also does not show any severe leverage or residuals.

(f) (4 marks) Perform a hypothesis testing for a positive slope at a significance level of 5% based on model M2.

Answer:

Step 1. State the hypotheses.

- H_0: $\beta_1 \leq 0$ (The slope of sqrt(Tonnage) is zero or negative).
- H_1: $\beta_1 > 0$ (The slope is positive, implying sqrt(Tonnage) increases log(Time)).

Step 2. The test statistic is: $t = 11.27$, derived from the estimated slope ($\beta_1 = 0.0215$) and its standard error ($SE = 0.0019$).

Step 3. The sampling distribution is: t-distribution with df = 29 (degrees of freedom = $n - 2 = 31 - 2$).

Step 4. p-value = 4.1×10^{-12} (from the model summary), which is < 0.05.

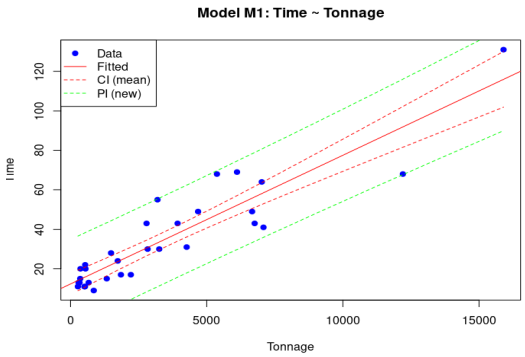
Step 5. Decision: Since the p-value < 0.05 (5% significance level), we reject H_0.

Step 6. Conclusion:
There is statistically significant evidence (at α = 0.05) to conclude that the slope is positive. This confirms that as ship tonnage increases, loading/unloading time (log(Time)) significantly increases.

(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 M1 and transformed model M2 respectively. Provide two scatterplots that consist the fitted model, the confidence and prediction 3 intervals for each of M1 and M2 respectively. Comment about the output in a few concise sentences.

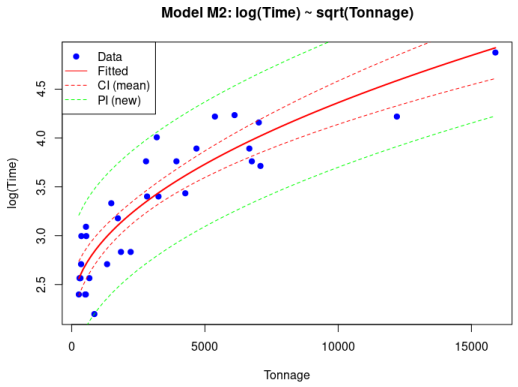
Answer: **For M1**

```
> glakes <- read.table("glakes.txt", header = TRUE)
> m1 <- lm(Time ~ Tonnage, data = glakes)
> newdata <- data.frame(Tonnage = 10000)
> m1_pred <- predict(m1, newdata, interval = "confidence", level = 0.95)
> m1_pred_pred <- predict(m1, newdata, interval = "prediction", level = 0.95)
> m1_pred
     fit      lwr      upr
1 77.5234 69.36467 85.68213
> m1_pred_pred
     fit      lwr      upr
1 77.5234 54.17047 100.8763
> plot(glakes$Tonnage, glakes$Time, main = "Model M1: Time ~ Tonnage", xlab = "Tonnage", ylab = "Time",
pch = 19, col = "blue")
> abline(m1, col = "red")
> tonnage_seq <- seq(min(glakes$Tonnage), max(glakes$Tonnage), length.out = 100)
> m1_conf <- predict(m1, newdata = data.frame(Tonnage = tonnage_seq), interval = "confidence")
> m1_pred_band <- predict(m1, newdata = data.frame(Tonnage = tonnage_seq), interval = "prediction")
> lines(tonnage_seq, m1_conf[, "lwr"], col = "red", lty = 2)
> lines(tonnage_seq, m1_conf[, "upr"], col = "red", lty = 2)
> lines(tonnage_seq, m1_pred_band[, "lwr"], col = "green", lty = 2)
> lines(tonnage_seq, m1_pred_band[, "upr"], col = "green", lty = 2)
> legend("topleft", legend = c("Data", "Fitted", "CI (mean)", "PI (new)"), col = c("blue", "red", "red", "green"),
lty = c(NA, 1, 2, 2), pch = c(19, NA, NA, NA))
```



For M2:

```
> glakes <- read.table("glakes.txt", header = TRUE)
> m2 <- lm(log(Time) ~` I(Tonnage^0.5), data = glakes)
> newdata <- data.frame(Tonnage = 10000)
> m2_pred_log <- predict(m2, newdata, interval = "confidence", level = 0.95)
> m2_pred <- exp(m2_pred_log)
> m2_pred
      fit      lwr      upr
1 78.39767 62.79776 97.87284
> m2_pred_pred_log <- predict(m2, newdata, interval = "prediction", level = 0.95)
> m2_pred_pred <- exp(m2_pred_pred_log)
> m2_pred_pred
      fit    lwr      upr
1 78.39767 40.455 151.9267
> plot(glakes$Tonnage, log(glakes$Time), main = "Model M2: log(Time) ~ sqrt(Tonnage)",  xlab = "Tonnage",
ylab = "log(Time)", pch = 19, col = "blue")
> tonnage_seq <- seq(min(glakes$Tonnage), max(glakes$Tonnage), length.out = 100)
> pred_log <- predict(m2, newdata = data.frame(Tonnage = tonnage_seq))
> lines(tonnage_seq, pred_log, col = "red", lwd = 2)
> conf_log <- predict(m2, newdata = data.frame(Tonnage = tonnage_seq), interval = "confidence")
> pred_band_log <- predict(m2, newdata = data.frame(Tonnage = tonnage_seq), interval = "prediction")
> lines(tonnage_seq, conf_log[, "lwr"], col = "red", lty = 2)
> lines(tonnage_seq, conf_log[, "upr"], col = "red", lty = 2)
> lines(tonnage_seq, pred_band_log[, "lwr"], col = "green", lty = 2)
> lines(tonnage_seq, pred_band_log[, "upr"], col = "green", lty = 2)
> legend("topleft",  legend = c("Data", "Fitted", "CI (mean)", "PI (new)"),  col = c("blue", "red", "red", "green"),
lty = c(NA, 1, 2, 2), pch = c(19, NA, NA, NA))
```



1. M2's intervals (CI: 62.8–97.9, PI: 40.5–151.9) are tighter than M1's (CI: 69.4–85.7, PI: 54.2–100.9), indicating better predictive accuracy with transformations.
2. M2's curve fits the nonlinear trend better than M1's straight line.
3. M2 is preferable for port planning due to its robust handling of scale and variance.