THE CHARACTERISTICS OF THE FJORD VALLEY OF THE SOUTH ISLAND IN THE NEW ZEALAND AND THE BRITISH COLUMBIA CANADA.

Aims

- To ascertain whether valley features (width or length) can be used to get precise forecasts of catchment area.

Background

There is a vast difference of characteristics between each valley. This project would like to investigate whether valley features (width or length) can be used to get precise forecasts of catchment area. This report primarily based on a fjords dataset, which contains the data (width or length) on fjords from the South Island of New Zealand and British Columbia Canada.

Findings

1. The best model for the fjords in British Columbia Canada is the investigate models that use a log-log transformation for the length.
2. The best model for the fjords in New Zealand is the investigate models that use a log-log transformation for the width.

Discussion

A outlier in the fjords dataset of New Zealand is removed as it's a influential outlier, but a outlier in the fjords dataset of British Columbia is not removed as it's not a influential outlier. Log-log transformation is used for both British Columbia and New Zealand. Missing data for New Zealand are not removed as removing the missing data will affect the accuracy of the result.

Statistical methods and results

Variables

Separating the fjords in the South Island of New Zealand, and fjords in British Columbia Canada. “NZ_fjords_df” represent the fjords in the South Island of New Zealand, and “BC_fjords_df” represent the fjords in British Columbia Canada.

Removing the outliers from the fjords in the South Island of New Zealand.

“NZ_fjords_df_no” represent the fjords in the South Island of New Zealand without outliers.

Computational methods

R 4.0.2 and RStudio Desktop 1.3.1073 are used to do the analysis in this project

There following function in R are used:

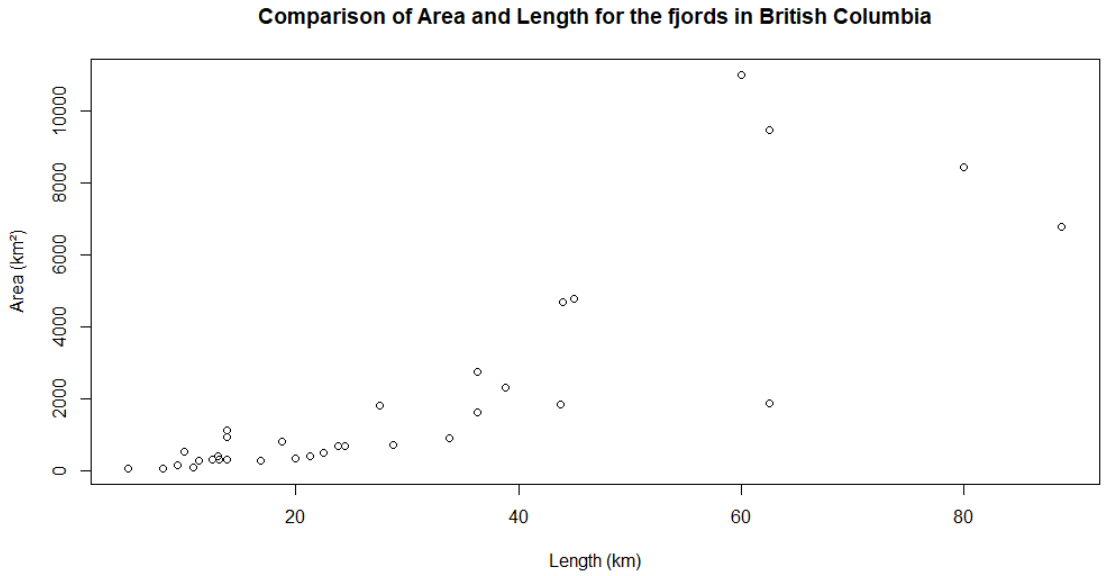
read.csv() for reading a file in a table format and creates a data frame from it.
subset() for getting a subset of vector, data frames or matrices which meet requirement.
plot() for creating a plot for a data.
lm() for fitting linear models.
layout() for showing data in a matrix format..
matrix() for creates a matrix format.
library() for loading a packages.
summary() for producing a summarized result of a data
boxcox() for calculating Box-Cox power transformations
abline() for adding straight lines to a plot

Package MASS is used in this project.

Results

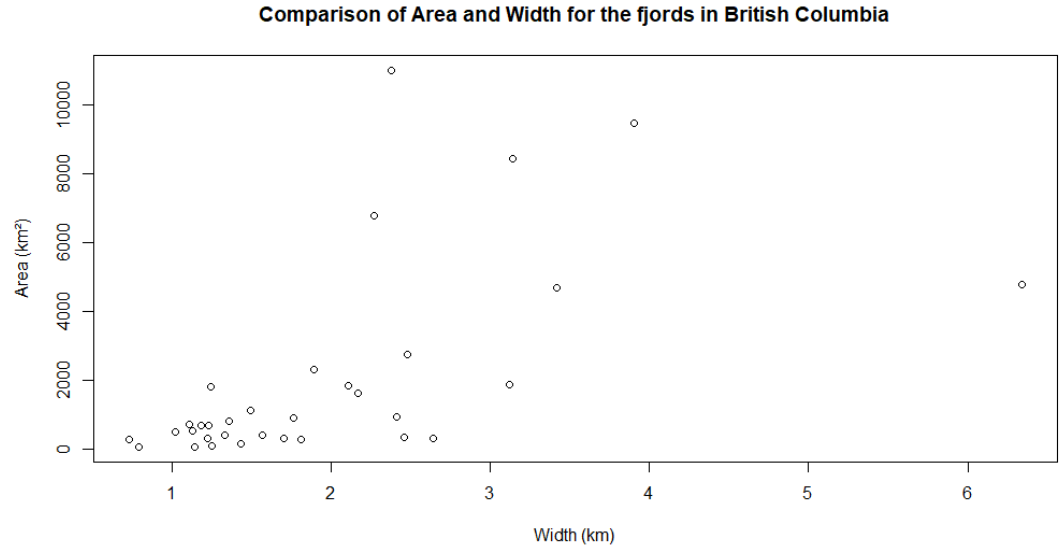
British Columbia

Figure 1: Comparison of Area and Length for the fjords in British Columbia



There is a great consistency between area and length

Figure 2: Comparison of Area and Width for the fjords in British Columbia



There is a good consistency between area and width

The scatter plot of area and length (Figure 1) has a greater consistency than the scatter plot of area and width (Figure 2)

Table 1: Summary of Simple Linear Regression on Comparison of Area and Length for the fjords in British Columbia

Simple Linear Regression on Comparison of Area and Length for the fjords in British Columbia

```
Call:
lm(formula = area ~ length, data = BC_fjords_df)

Residuals:
    Min      1Q  Median      3Q     Max
-3903.3  -763.0    20.1   655.1  5496.9

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1274.82     481.46  -2.648   0.0126 *
length        112.88      13.31   8.483  1.4e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1611 on 31 degrees of freedom
Multiple R-squared:  0.6989,    Adjusted R-squared:  0.6892
F-statistic: 71.96 on 1 and 31 DF,  p-value: 1.396e-09
```

The is a high multiple R-squared of length (0.6989) and a high adjusted R-squared of length (0.6892).

Table 2: Summary of Simple Linear Regression on Comparison of Area and Width for the fjords in British Columbia

Simple Linear Regression on Comparison of Area and Width for the fjords in British Columbia

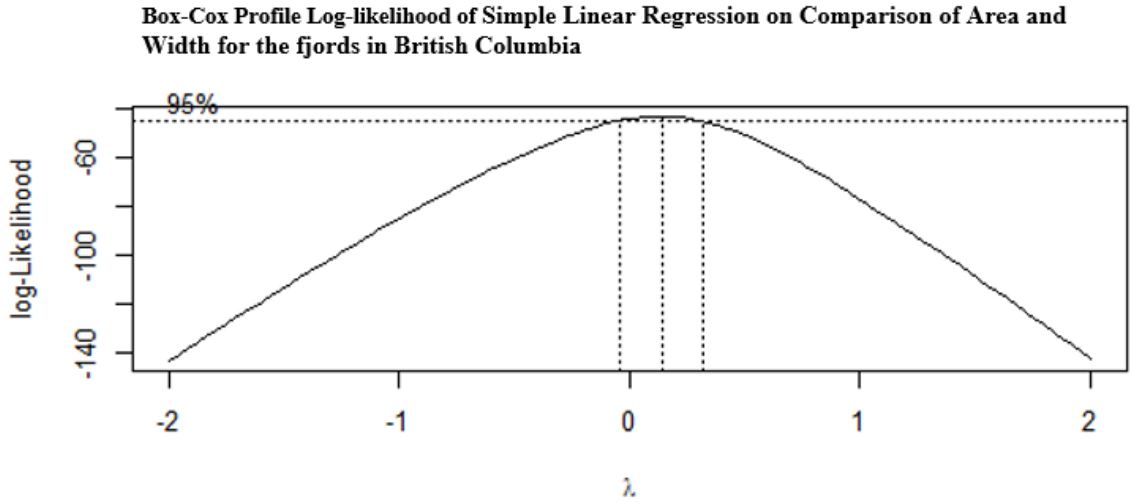
```
Call:
lm(formula = area ~ width, data = BC_fjords_df)

Residuals:
    Min      1Q  Median      3Q     Max
-3916.0 -1012.8  -281.0   -16.4  8335.8

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   -964.8      856.6  -1.126 0.268697
width         1522.7      379.3   4.015 0.000351 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2382 on 31 degrees of freedom
Multiple R-squared:  0.3421,    Adjusted R-squared:  0.3208
F-statistic: 16.12 on 1 and 31 DF,  p-value: 0.0003508
```

The is a low multiple R-squared of width (0.3421) and a low adjusted R-squared of width (0.3208).

The Multiple R-squared of length (0.6989) (Table 1) is 0.3568 higher than width (0.3421) (Table 2), the Adjusted R-squared of length (0.6892) (Table 1) is 0.3684 higher than width (0.3208) (Table 2).

Since the scatter plot of area and length (Figure 1) has a greater consistency than the scatter plot of area and width (Figure 2) and the Multiple R-squared of length (0.6989) (Table 1) is 0.3568 higher than width (0.3421) (Table 2), the Adjusted R-squared of length (0.6892) (Table 1) is 0.3684 higher than width (0.3208) (Table 2). The length is chosen for British Columbia.

Figure 3: Box-Cox Profile Log-likelihood of Simple Linear Regression on Comparison of Area and Width for the fjords in British Columbia



The lambda is close to 0.

Since the lambda is closer to 0 (Figure 3), a log transformation is used for the linear regression on comparison of area and width for the fjords in British Columbia.

Table 3: Summary of Log-Log Regression on Comparison of Area and Width for the fjords in British Columbia

Summary of Log Regression on Comparison of Area and Width for the fjords in British Columbia

```
Call:
lm(formula = log(area) ~ log(length), data = BC_fjords_df)

Residuals:
    Min      1Q  Median      3Q     Max
-0.9488 -0.5088 -0.0791  0.3736  1.1504

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.3671     0.4528   3.019  0.00504 **
log(length)   1.7216     0.1408  12.225 2.16e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5713 on 31 degrees of freedom
Multiple R-squared:  0.8282,    Adjusted R-squared:  0.8227
F-statistic: 149.4 on 1 and 31 DF,  p-value: 2.157e-13
```

There is a positive intercept, a very high multiple R-squared (0.8282), and a very high adjusted R-squared (0.8227). The log-log regression also has a very small residual standard error (0.5731), and the p-value is much less than 0.05.

Table 4: Summary of Log Regression on Comparison of Area and Length for the fjords in British Columbia

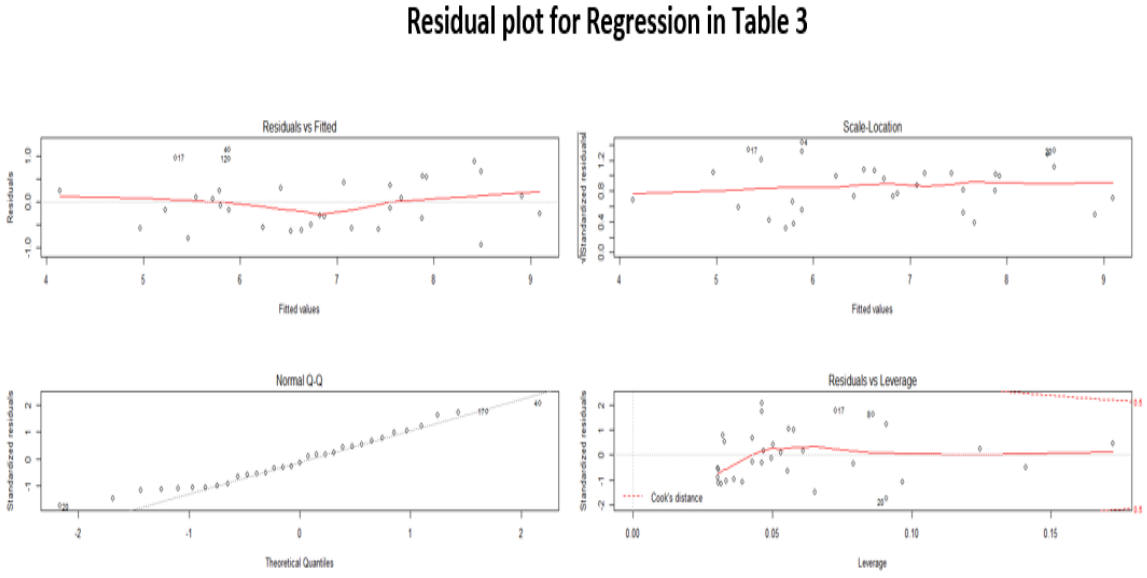
Summary of Log Regression on Comparison of Area and Length for the fjords in British Columbia

```
Call:
lm(formula = area ~ log(length), data = BC_fjords_df)

Residuals:
    Min      1Q  Median      3Q     Max
-3187.8 -1404.7   -33.0   697.1  6053.7

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -7447.1     1537.8  -4.843 3.37e-05 ***
log(length)   3025.7      478.2   6.327 4.86e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1940 on 31 degrees of freedom
Multiple R-squared:  0.5636,    Adjusted R-squared:  0.5495
F-statistic: 40.03 on 1 and 31 DF,  p-value: 4.859e-07
```

There is a negative intercept, a not high multiple R-squared (0.5636), and a not high adjusted R-squared (0.5495). The log regression also has a very high residual standard error (1940), and the p-value is much less than 0.05.

Since the log regression has a negative intercept, so the log-log transformation is used.

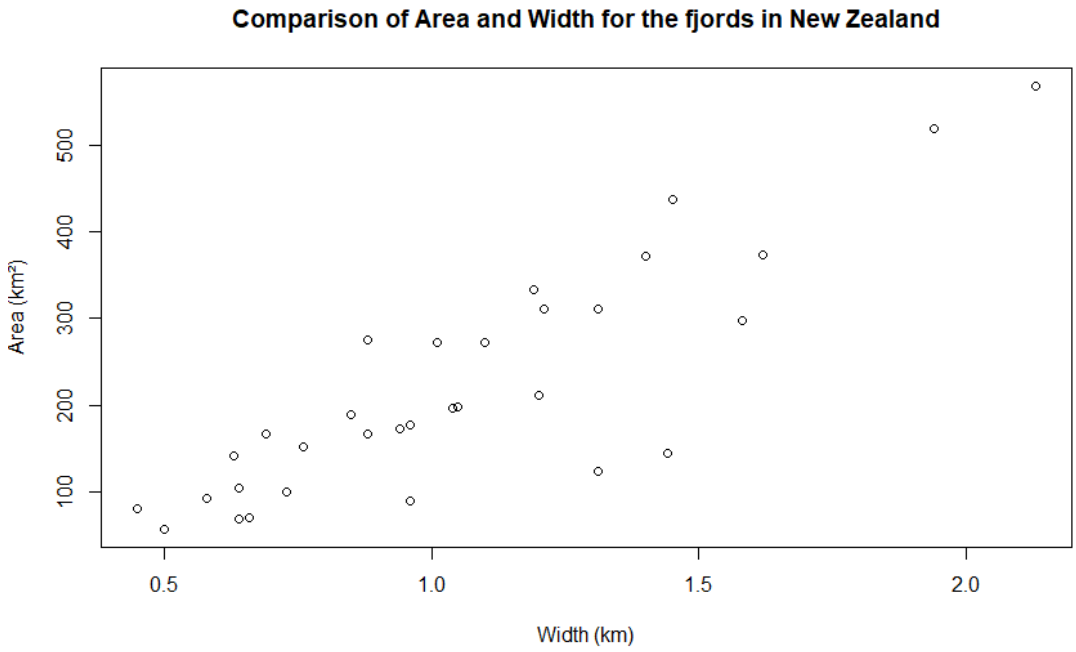
Table 4: Residual plot for Regression in Table 3



The top-left figure (Residuals vs Fitted) and top-right figure (Scale-Location) show there is no particular pattern (homoscedasticity). The bottom-left figure (Normal Q-Q) show a straight line, which mean the residuals are distributed normally. The bottom-right figure (Residuals vs Leverage) show there is no influential outlier.

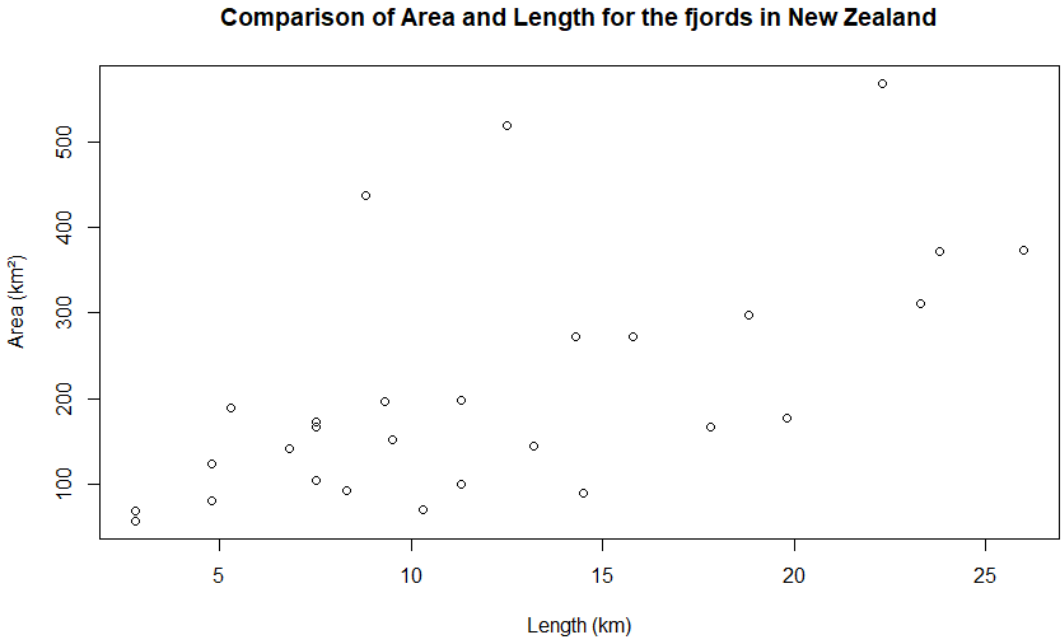
New Zealand

Figure 4: Comparison of Area and Width for the fjords in New Zealand



There is a great consistency between area and width

Figure 5: Comparison of Area and Length for the fjords in New Zealand



There is a good consistency between area and length

The scatter plot of area and width (Figure 4) has a greater consistency than the scatter plot of area and length (Figure 5)

Table 5: Summary of Simple Linear Regression on Comparison of Area and Width for the fjords in New Zealand

Summary of Simple Linear Regression on Comparison of Area and Width for the fjords in New Zealand

```
Call:
lm(formula = area ~ width, data = NZ_fjords_df)

Residuals:
     Min       1Q   Median       3Q      Max
-183.010  -23.135    1.784   47.440  106.207

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   -72.86      33.15  -2.198   0.0358 *
width         278.25      29.38   9.471  1.6e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 66.96 on 30 degrees of freedom
Multiple R-squared:  0.7494,    Adjusted R-squared:  0.741
F-statistic: 89.71 on 1 and 30 DF,  p-value: 1.595e-10
```

The is a high multiple R-squared of length (0.7494) and a high adjusted R-squared of length (0.741).

Table 6: Summary of Simple Linear Regression on Comparison of Area and Length for the fjords in New Zealand

Summary of Simple Linear Regression on Comparison of Area and Length for the fjords in New Zealand

```
Call:
lm(formula = area ~ length, data = NZ_fjords_df)

Residuals:
     Min       1Q   Median       3Q      Max
-150.900  -53.181   -8.186   16.683  303.440

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   56.765     44.430   1.278 0.212674
length        12.720      3.221   3.950 0.000533 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 110.8 on 26 degrees of freedom
  (4 observations deleted due to missingness)
Multiple R-squared:  0.375,     Adjusted R-squared:  0.3509
F-statistic:  15.6 on 1 and 26 DF,  p-value: 0.0005332
```

The is a low multiple R-squared of length (0.375) and a low adjusted R-squared of length (0.3509).

The Multiple R-squared of length (0.7494) (Table 5) is 0.3744 higher than width (0.375) (Table 6), the Adjusted R-squared of length (0.741) (Table 5) is 0.3901 higher than width (0.3509) (Table 6)

Since the scatter plot of area and width (Figure 4) has a greater consistency than the scatter plot of area and length (Figure 5) and the Multiple R-squared of length (0.7494) (Table 5) is 0.3744 higher than width (0.375) (Table 6), the Adjusted R-squared of length (0.741) (Table 5) is 0.3901 higher than width (0.3509) (Table 6). The length is chosen for New Zealand.

Table 7: Summary of Log-Log Regression on Comparison of Area and Width for the fjords in New Zealand

Summary of Log-Log Regression on Comparison of Area and Width for the fjords in New Zealand

```
Call:
lm(formula = log(area) ~ log(width), data = NZ_fjords_df)

Residuals:
     Min       1Q   Median       3Q      Max
-0.78707 -0.13996  0.03605  0.22771  0.54072

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.24698    0.05967  87.934  < 2e-16 ***
log(width)   1.32859    0.15614   8.509  1.7e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3371 on 30 degrees of freedom
Multiple R-squared:  0.707,     Adjusted R-squared:  0.6973
F-statistic:  72.4 on 1 and 30 DF,  p-value: 1.704e-09
```

There is a positive intercept, a very high multiple R-squared (0.707), and a very high adjusted R-squared (0.6973). The log-log regression also has a very small residual standard error (0.3371), and the p-value is much less than 0.05.

Table 8: Summary of Log Regression on Comparison of Area and Width for the fjords in

New Zealand

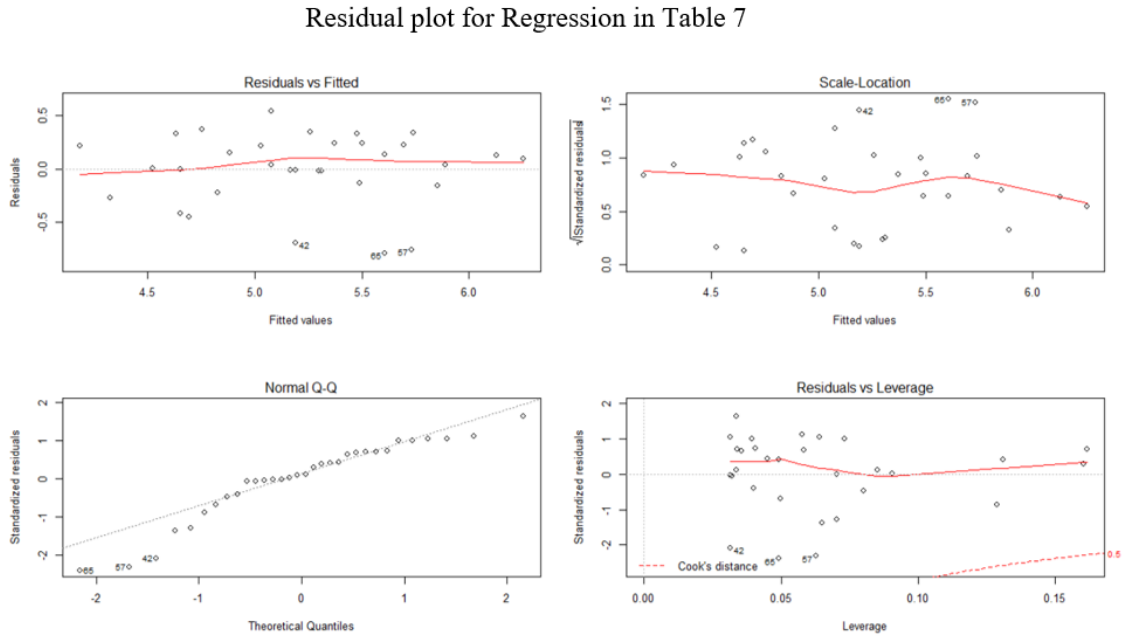
Summary of Log Regression on Comparison of Area and Width for the fjords in New Zealand

```
Call:
lm(formula = area ~ log(width), data = NZ_fjords_df)

Residuals:
    Min      1Q  Median      3Q     Max
-183.769  -37.757    9.403   46.274  129.157

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   225.86      13.20  17.107  < 2e-16 ***
log(width)    281.68      34.55   8.154 4.22e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 74.59 on 30 degrees of freedom
Multiple R-squared:  0.6891,    Adjusted R-squared:  0.6787
F-statistic: 66.48 on 1 and 30 DF,  p-value: 4.216e-09
```

There is a positive intercept, a high multiple R-squared (0.6891), and a high adjusted R-squared (0.6787). The log regression also has a not big residual standard error (74.59), and the p-value is much less than 0.05.

Since the Multiple R-squared of the Log-Log Regression (0.707) (Table 7) is 0.3744 higher than the Log Regression (0.0179) (Table 8), the Adjusted R-squared of the Log-Log Regression (0.6973) (Table 7) is 0.0186 higher than the Log Regression (0.6787) (Table 8). The residual standard error the Log-Log Regression (0.3371) (Table 7) is 74.2529 less than the Log Regression (74.59) (Table 8), so the log-log transformation is used.

Table 9: Residual plot for Regression in Table 7



The top-left figure (Residuals vs Fitted) and top-right figure (Scale-Location) show there is no particular pattern (homoscedasticity). The bottom-left figure (Normal Q-Q) show a straight line, which mean the residuals are distributed normally. The bottom-right figure (Residuals vs Leverage) show there is no influential outlier.