Predicting Catchment Area of Fjords

This report discusses the relationship between catchment area of fjords and the characteristics of their respective valleys. A data set of fjords from two regions, New Zealand (NZ) and British Columbia (BC), was analysed to determine whether valley characteristics (length and width) could be used to predict the size of catchment area. Each region was analysed separately.

The data provided for the analysis is summarized visually in figure 1 and figure 2 below

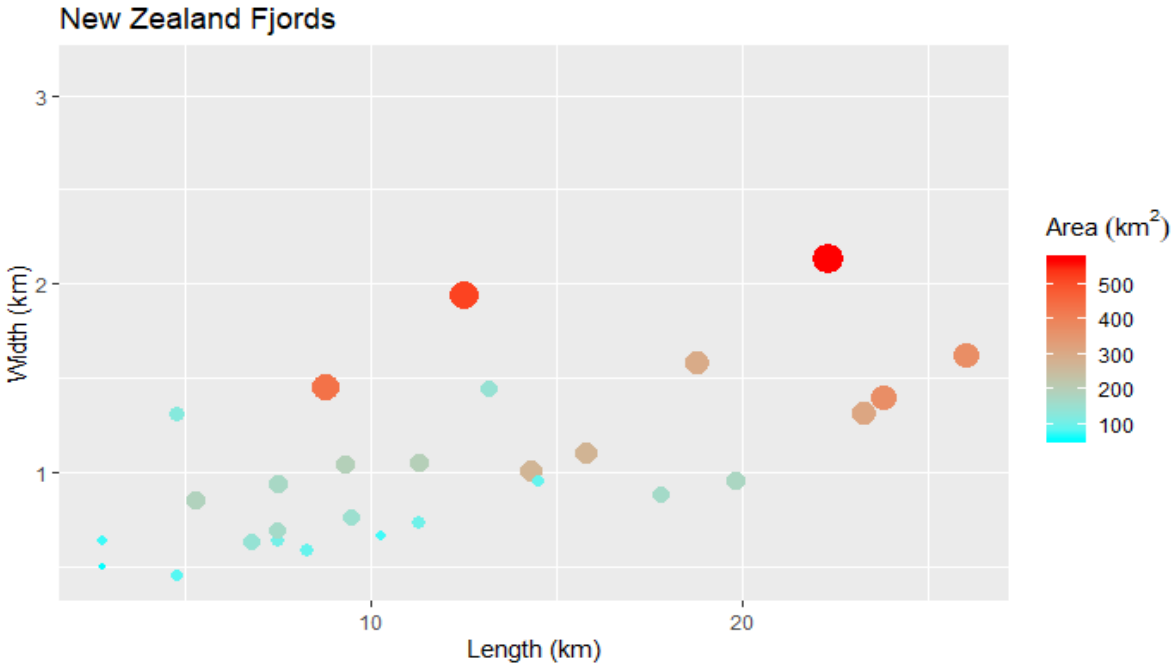

Figure 1 (above) shows an overview of the relationship that valley characteristics, length and width have with catchment area of fjords in New Zealand.

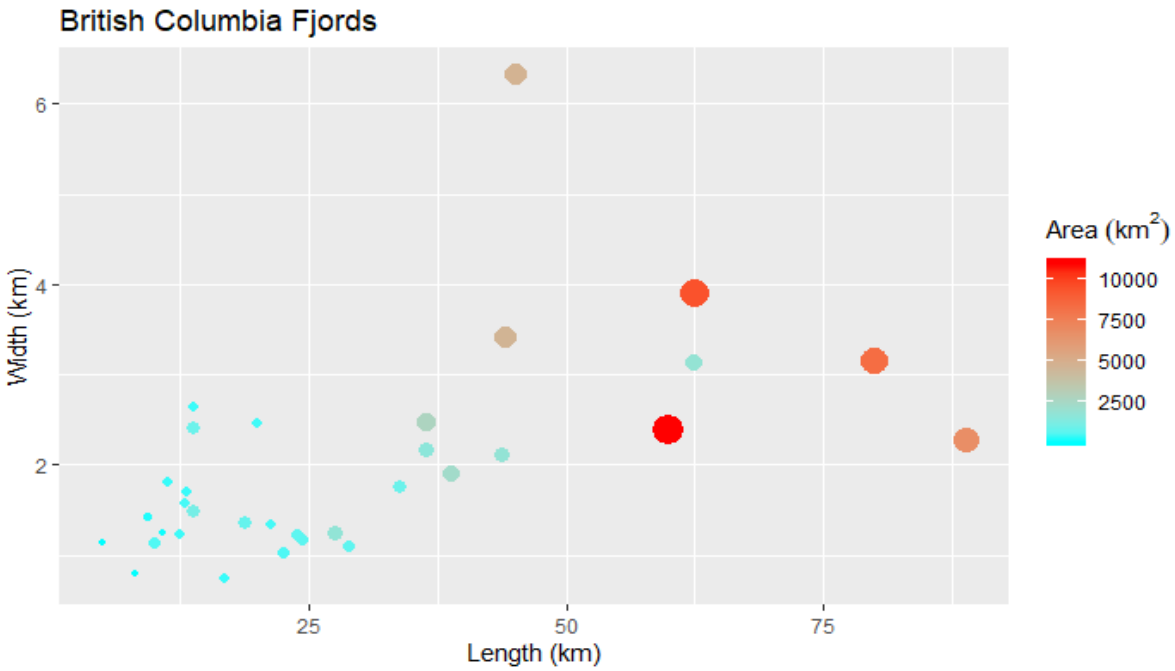

Figure 2 (above) shows an overview of the relationship that valley characteristics, length and width have with catchment area of fjords in British Columbia.

Findings

It was shown that catchment area could be predicted by at least one of the valley characteristics in each region. Although theses predictions were not 100% accurate, our models were able to explain 75% of variability of fjords in NZ and 83% of variability of fjords in BC.

Based on our data, the best equation for predicting catchment area of fjords in BC are;

$$\hat{Y}_{BC} = 1.3671 + 1.7216 \times \log X_{length}$$

where Y_{BC} is catchment area and X_{length} is valley length.

For NZ, the best equation was determined to be

$$\hat{Y}_{NZ} = -72.86 + 278.25 \times X_{width}$$

similarly, here Y_{NZ} is catchment area and X_{width} is valley length.

Exclusions

The equation relating to NZ is based on the exclusion of one data point from the analysis. Figure 3 shows the 'unusualness' of this data point.

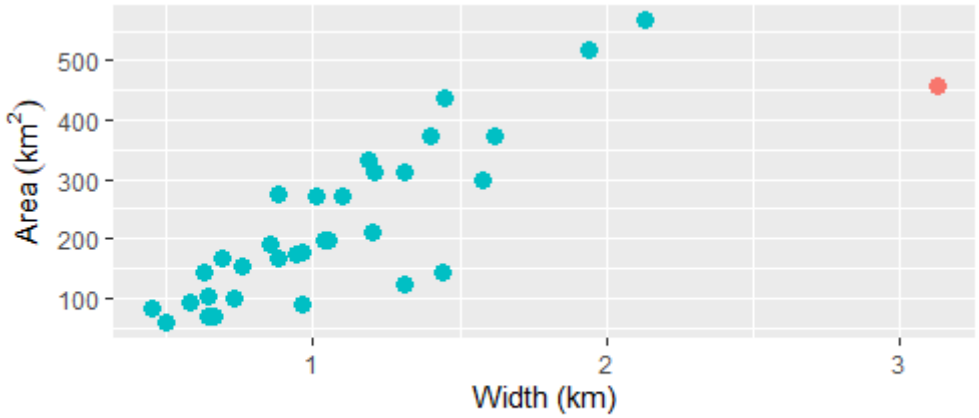

Figure 3 (above) shows how the outlier (shown in red) in the NZ Fjord data relates to other data points in the data set

Technical Details

The analysis starts with simple linear regression; one model for each predictor variable in each region. We now have 4 linear models. Where do we go from here?

For each region we choose the predictor variable, width or length, that gave us the highest R-Squared. From there, we run Residual Diagnostics and we start with NZ.

Fjords of New Zealand

The diagnostic plots showed that there was an influential point that we should consider removing. Further investigation indicates that the point is unusual in more than just its residual; its width is 3.7 standard deviations from the mean for NZ, making it an exceptional data point. This is illustrated in figure 3 above. Data point '62' is removed from further analysis.

New diagnostic plots are created now that the outlier has been removed. They show some negative skew in the residuals and some non-linearity (figure 4). Perhaps a transform of the response variable is a good idea?

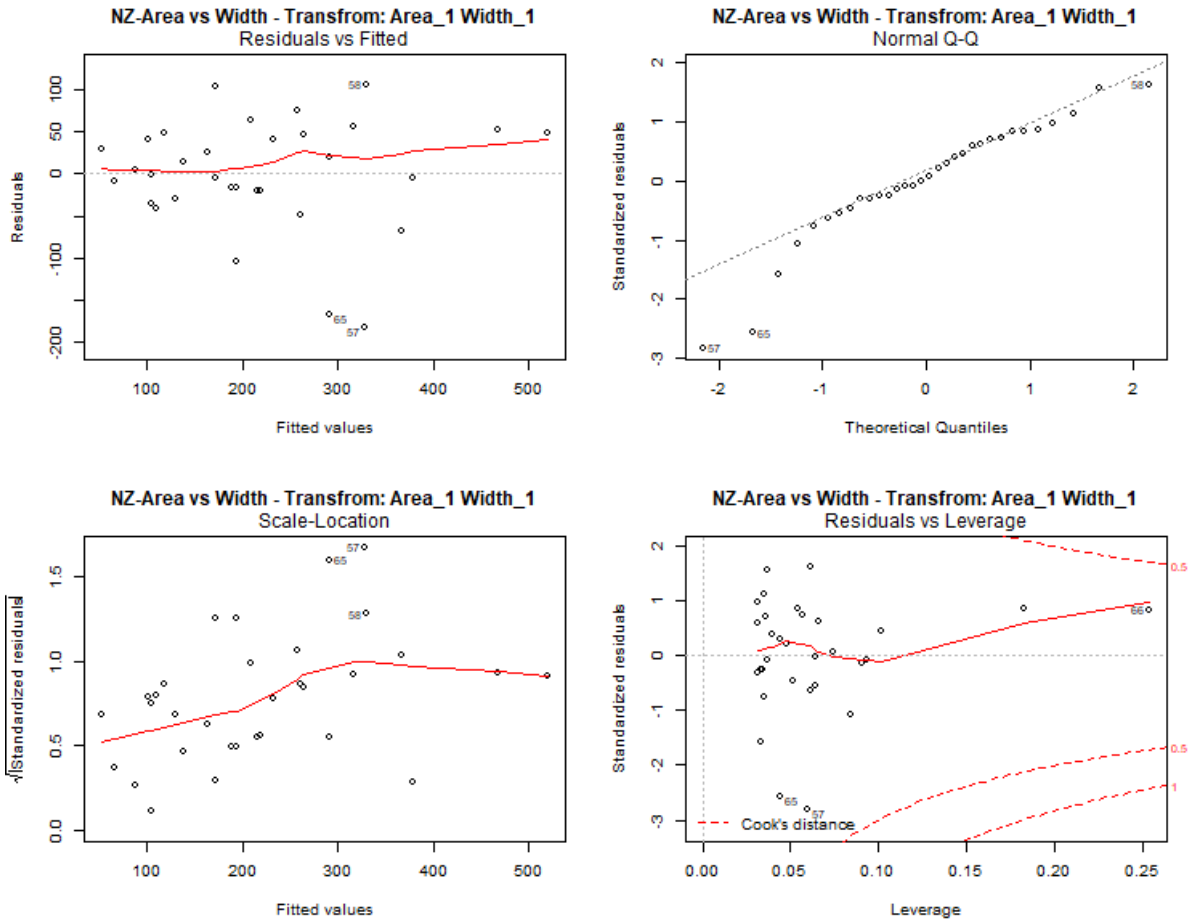

Figure 4: Diagnostic plots of regression performed on NZ fjord data. The two plots on the left-hand-side suggest some nonlinearity as well as non-constant variance. The data was not transformed. Despite the problematic diagnostics, this model achieved the highest R-Squared for the NZ fjord data.

A Boxcox showed that a square root transform of the response variable would be appropriate (figure 5). The transform was applied and then a new simple linear regression was constructed.

Diagnostic plots of the new regression did not any look better than they did before the transform. In fact, the Scale-Location plot looked worse than before. Furthermore, the R-Squared was lower. Based on these results it was decided that other transforms should be investigated, including a transform of the explanatory variable. Log transforms were tested because the researchers had suggested a log relationship between variables. Different permutations of log and square root transforms were applied. The results of these transforms can be seen in table 1 on the next page.

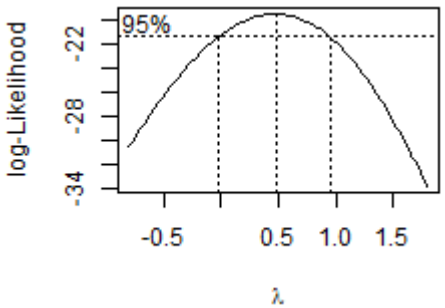

Figure 5. Boxcox of area ~ width for NZ fjord data.

In the end, it was found that the highest R-Squared came from a model with no transforms, despite the problematic diagnostic plots (figure 4). See figure 6 for a graphical view and see table 3 at the end of this report for numbers that define the model.
Out of all the models that were fit, the 'no-transform' diagnostic plots were the best looking. Only one other model had comparable diagnostics, but its R-Squared was lower and it involved a transform that reduced interpretability.

Finally, it is noted that because of the negative intercept, suspicious diagnostic plots, and removal of outliers, caution should be exercised when applying the suggested model outside the minimum and maximum values shown in table 3 at the end of this report.

NZ Width

Explanatory Transform	Response Transform	Intercept P-Value	Relationship P-value	R^2
1	1	0.968	7.06E-09	0.6662
1	0.5	2.40E-07	2.90E-08	0.6349
0	0.5	2.00E-16	4.91E-10	0.7183
1	0	2.00E-16	2.87E-07	0.4114
0	0	2.00E-16	1.27E-09	0.7007
Point 62 removed				
0	0.5	2.00E-16	1.29E-09	0.7123
1	1	0.0358	1.60E-10	0.7494
1	0.5	0.000197	6.11E-10	0.7262
1	0	2.00E-16	8.27E-09	0.6750
0.5	0.5	0.0371	5.78E-10	0.7272

Table 1. Here the numbers in the two transform columns on the left indicate the power to which the variable was raised when constructing a simple regression model. '0' indicates a log transform. The bottom half of the table shows results for models fit without the data point named '62'.

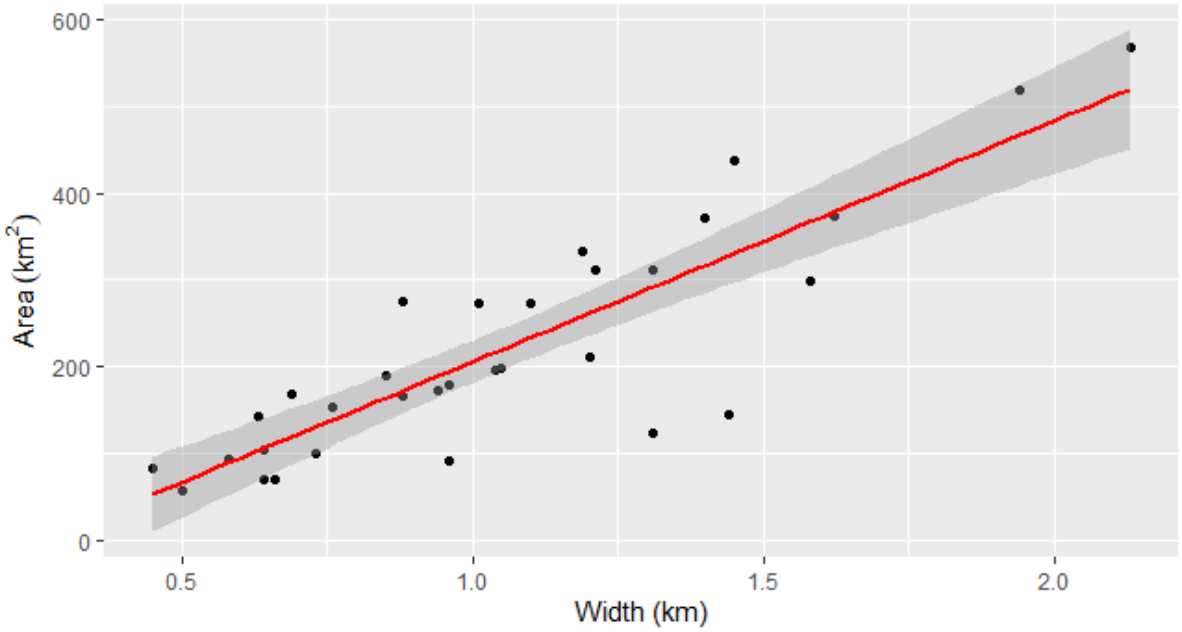

Figure 6. The line of fit (red) for what was determined to be the best linear regression model for the NZ Fjord data. The shaded area shows the 95% confidence interval for the line of fit. See table 3 for a numerical representation.

Fjords of British Columbia

For the BC data, using length in a simple regression gives a higher R-Squared than width does, so it seemed like it might be the better predictor. As with the NZ data, residual diagnostics suggest that the assumptions of least squares regression were not met perfectly (figure 7).

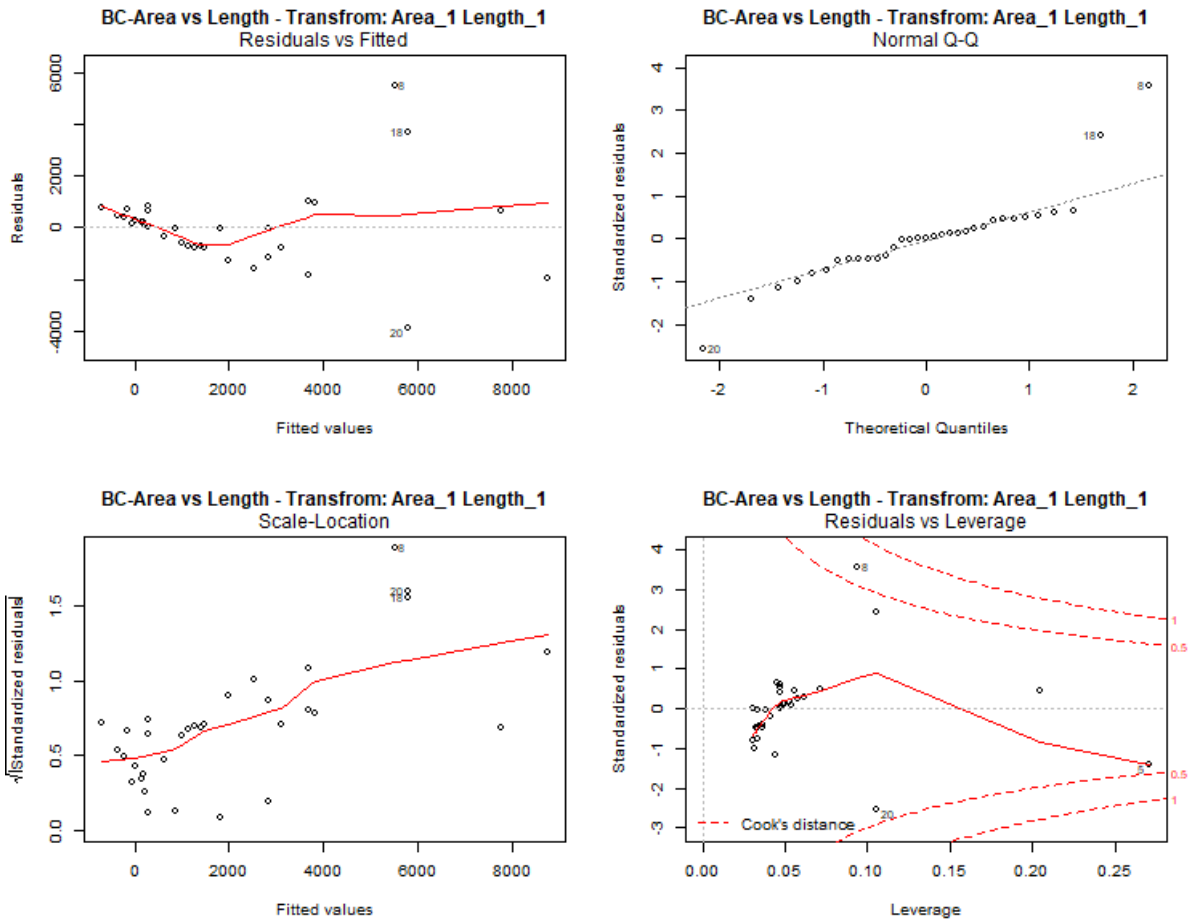

Figure 7. The initial diagnostic plots for simple linear regression using valley length to predict area catchment area in British Columbia. The plots suggest non-homogeneity of variance, and point to one possibly problematic data point, point '8'.

To address this, a transform of the response variable was considered. Boxcox suggested a log transform (figure 8). Applying this transform did not yield better diagnostics. In fact they looked much worse (figure 9).

Next, a boxcox was used on a fit without the two most influential points, points '5' and '8'. A log transform was still the suggestion according to boxcox. Applying it still did not improve the diagnostics and resulted in a lower R-Squared. It seemed like the best thing to do was to take some steps backwards.

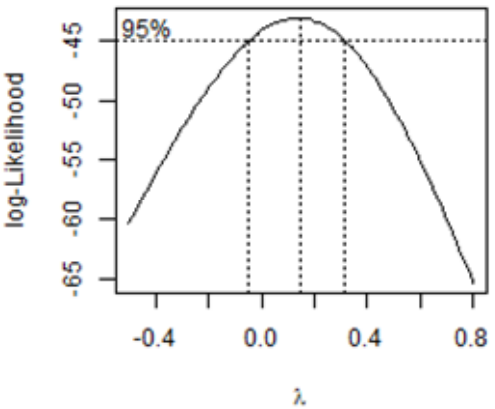

Figure 8. A boxcox suggests a log transform of the response variable using length as the predictor

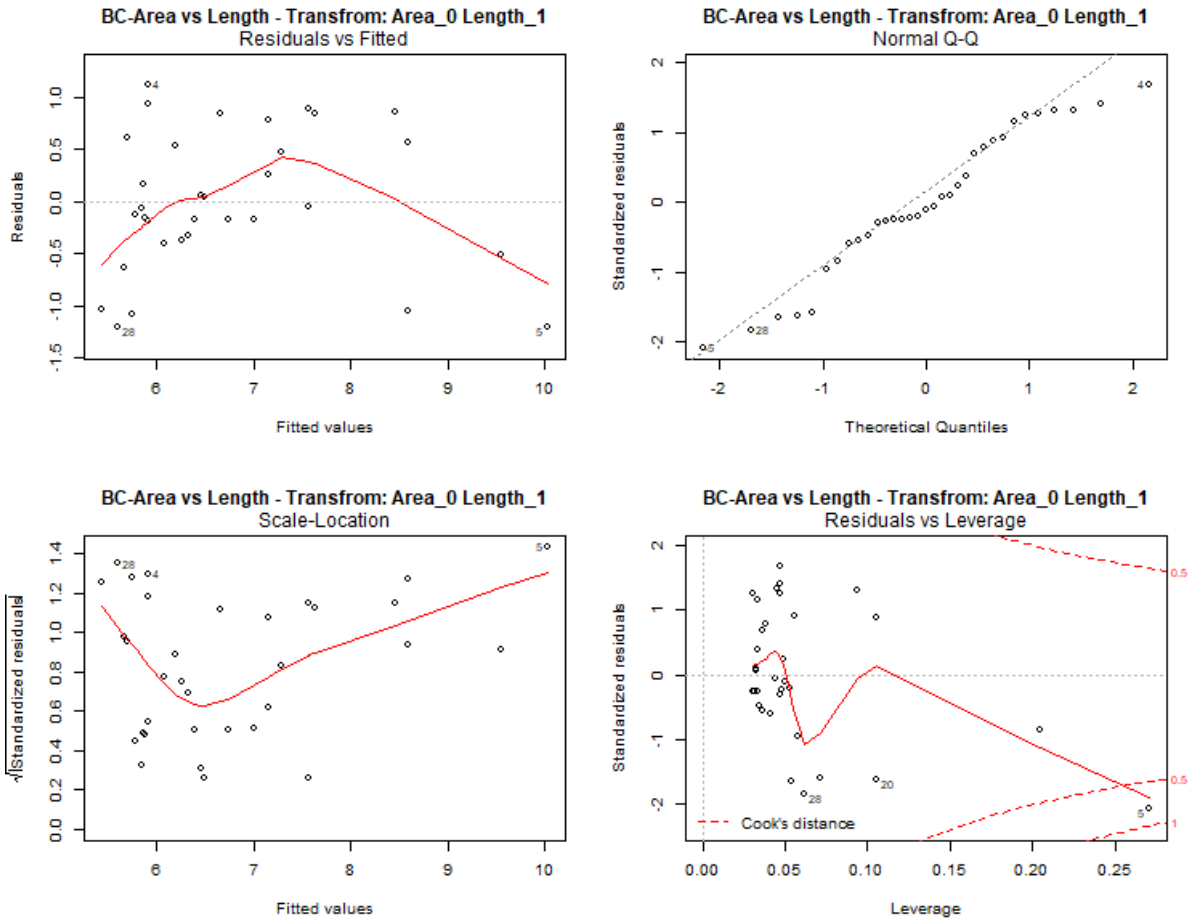

Figure 9

After putting all the data points back in, a transformation of the explanatory variable was considered. According to the researchers, log relationships have been used in similar situations. Therefore, a log transform was applied. The diagnostics looked much better at after this (figure 10).

At this point, it might have been sensible to accept this model. But with the drive of curiosity, and an intention to be thorough in the investigation, a new approach was taken. It was noted that models had been fit without points '5' and '8'. What if a fit was tried with only point '8' was removed? After all, the diagnostics had shown that '8' was the single most influential point. Perhaps things would go differently if we kept '5'. Without point '8', and without transforms, a new simple regression was fit.

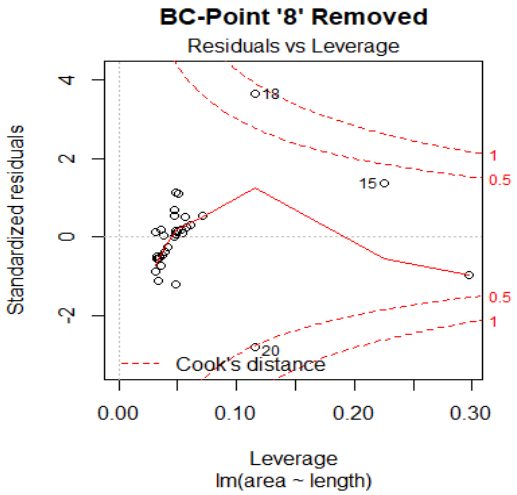

Figure 11. After removing '8', another point was deemed influential

The results showed that point '18' was now an influential point (figure 11). After removing '18', points '15' and '20' were shown to be influential. Finally, after removing points '15' and '18', there were no more influential points. Next a transform was tried without the influential points, '8', '15', 18' and '20'.

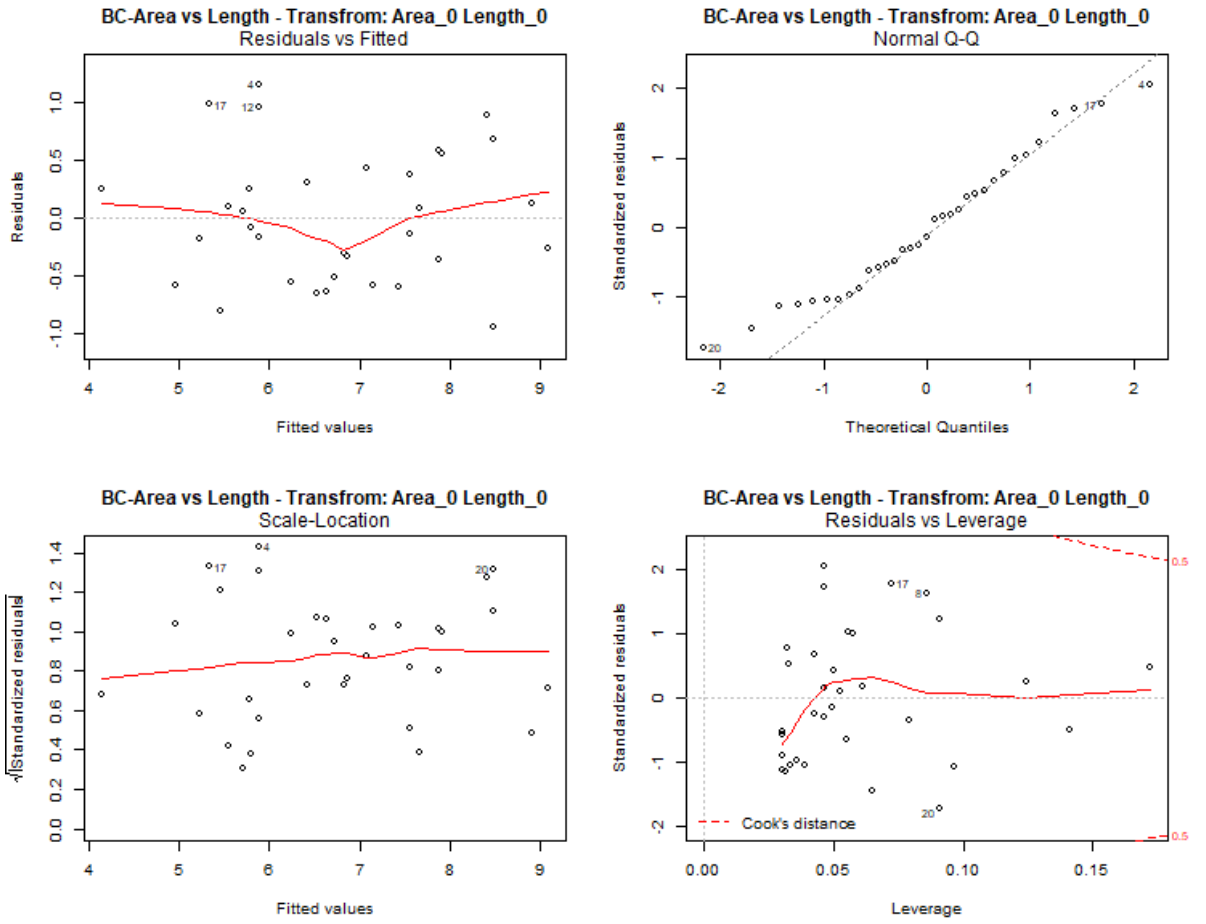

Figure 10. Residual Diagnostics after log transforms of both variables with *all data points.*

Diagnostics of the new regression with log transforms of both variables, and four influential points removed, showed that point '5' was now an influential point. It was removed and a new model was fit with a log transform of the response variable. The diagnostics did not look too bad (figure 12). However the R-Squared was not as good as what was achieved in previous steps.

For each iteration of the process so far, R-Squared and the P-Values of the coefficients were recorded. Additionally, transforms were tried with various data points removed. The results of these tests are shown in table 2.

From table 2 we can see that a log transform of both variables, with all data points, yielded the highest R-Squared. See figure 13 for a

BC Length

Explanatory Transform	Response Transform	Intercept P-Value	Relationship P-value	R^2
1	1	0.0126	1.40E-09	0.6989
0	1	3.37E-05	4.86E-07	0.5636
0	0	0.00504	2.16E-13	0.8282
1	0	2.00E-16	6.18E-11	0.7532
Point 23 removed				
log	log	0.00399	7.14E-13	0.8247
Points 5 & 8 Removed				
1	0	2.00E-16	4.12E-10	0.7453
Point 8 Removed				
1	1	7.21E-03	2.10E-10	0.7448
0	0	9.55E-04	4.64E-12	0.8125
Points 18 & 8 Removed				
1	1	4.29E-03	2.94E-11	0.7873
0	0	9.55E-04	4.64E-12	0.8125
Points 8,15,18&20 Removed				
1	1	0.00153	1.94E-11	0.8163
1	0	2.00E-16	2.60E-08	0.6887
0	0	3.57E-03	7.60E-11	0.7969
Points 5,8,15,18&20 Removed				
1	0	2.00E-16	2.11E-09	0.7542
0	0	0.7632	1.29E-09	0.7632

Table 2. Shows R-Squared values for transforms with and without various data points removed. Transforms are expressed as power to which a variable was raised.

graphical view of the model, and see table 3 at the end of this report for numbers that define the model.

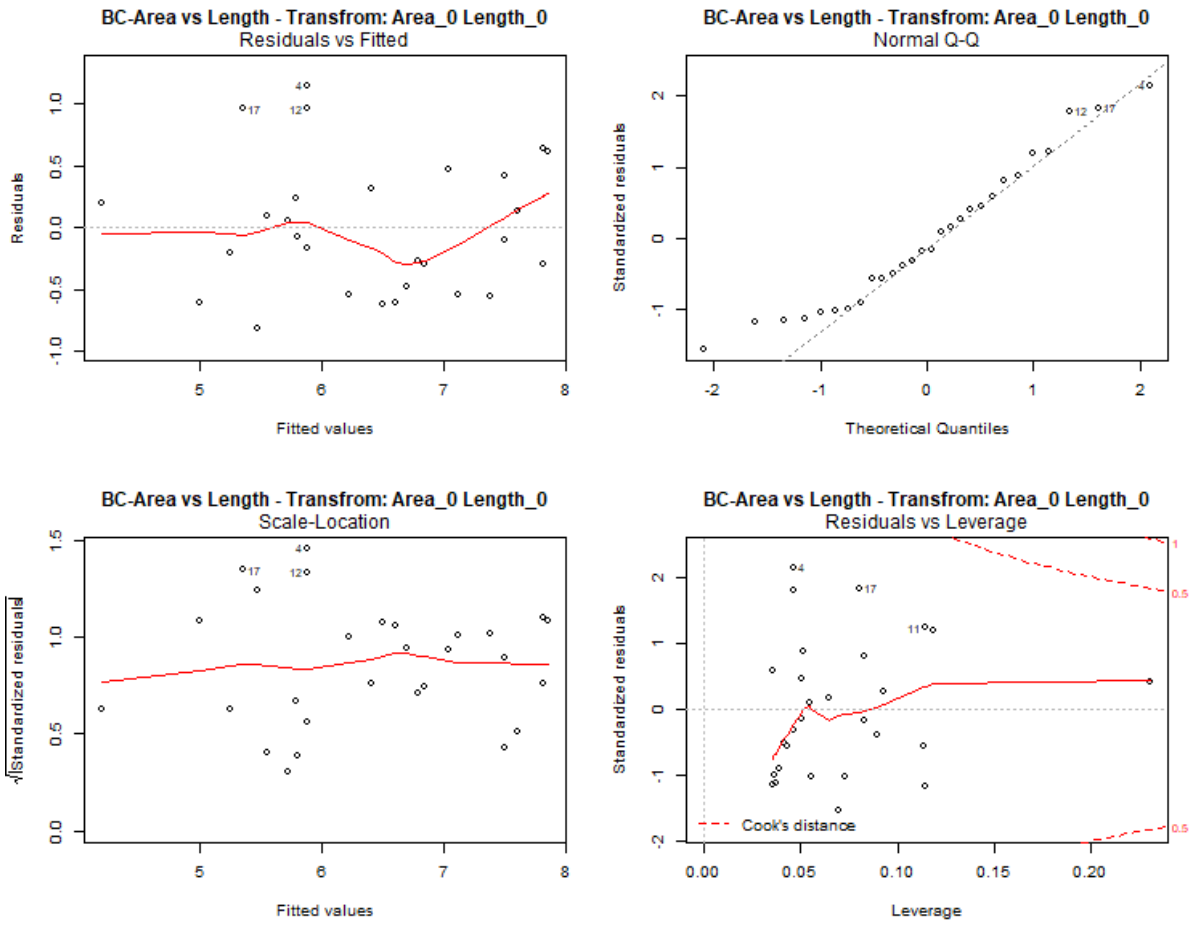

Figure 12. Residual Diagnostics after log transforms of both variables and removing points '5', '8', '15', '18' and '20'. R-Squared was lower than it is was before removing any data points.

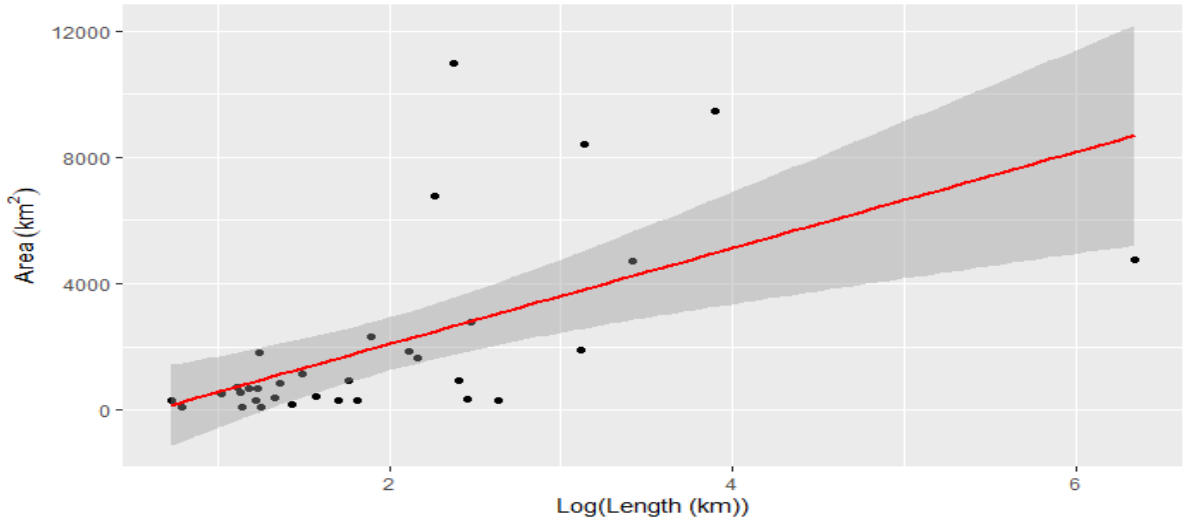

Figure 13. Line of fit (red) for the best model of the BC data. The shaded area shows the 95% confidence interval for the line of fit. See table 3 for the numerical representation.

British Columbia Fjord Prediction Model			
		Confidence Interval	
	Estimate	2.50%	97.50%
Intercept	1.3671	0.4435	2.2906
Width	1.7216	1.4343	2.0088
Note the intecept confidence interval does not include 0. Model was fit using length data with the bounds below.			
Minimum safe width (km)		5	
Maximum safe width (km)		88.8	

New Zealand Fjord Prediction Model			
		Confidence Interval	
	Estimate	2.50%	97.50%
Intercept	-72.86	-140.6	-5.160224
Width	278.25	218.25	338.242
When making prediction, proceed with caution when doing so out side the values below			
Minimum safe width (km)		0.45	
Maximum safe width (km)		2.13	

Table 3. Shows the final models for fjord data from each region.

Appendix – R Funcitons and Packages

All analysis was done using R version 3.5.2

The following table contains information on the functions and packages used.

Aspect of Report	Function and Package used to Produce aspect of Report
Figure 1, 2, 3, 6 and 13	ggplot() from library(ggplot2)
All diagnostic plots	plot() – built in R function
Results	lm() and summary() – built in R functions
Boxcox – to determine transformations	Boxcox – built in R function
All tables	Not from R, but from Excel