Modelling Fjord Catchment Area with Fjord Valley Characteristics in British Columbia and New Zealand

By Nicholas Forbes-Smith

Aim

To analyse how fjord valley length and width can be used to model fjord catchment area in New Zealand and British Columbia.

Background

The dataset used contains data from 2 locations: fjords in the South Island of New Zealand (NZ), and fjords in British Columbia (BC), Canada.

The recorded variables are:

- *region* - location of the fjord, categorical variable (NZ or BC);
- *area* - catchment area of the fjord, in square km;
- *length, width* - characteristics of the fjord valley, in km.

Findings

1. British Columbian fjord area is best modelled by model BC13 (*Table 5*):

$$\mathbf{1.68\,\log(area) = 1.50\,\log(length) + 0.65\,\log(width)}$$

2. Fjord #62 is much wider than any other fjord in the New Zealand dataset (*Table 7*) and is an extreme outlier. If fjord #62 can be ignored, the area of fjords in NZ is best modelled by NZ7 model (*Table 9*):

$$\mathbf{area = -72.86 + 278.25\ width}$$

3. If Fjord #62 is valid, NZ fjord area is best modelled by NZ4 model (*Table 8*):

$$\mathbf{area = -144.61 + 431.27\ width - 71.51\ width^2}$$

Discussion

Unfortunately, all simpler models couldn't adequately model fjord area in British Columbia, so there are no alternative models to propose. If more data were available it might be possible to create a simpler model.

The model stated in *Finding 3* may produce unexpected results when predicting fjord catchment area if the fjord valley width is greater than 3.0km (*Figure 13*). More data is needed to verify the accuracy of the model for fjords with valley width greater than 3.0km.

Interestingly, the simplest model for NZ fjord catchment area (NZ2) could only be created in terms of valley width, not valley length (*Table 06*). Whereas the simplest model for predicting BC fjord catchment area (BC2) could only be created in terms of valley length, not valley width (*Table 02*). NZ fjords had a stronger relationship between valley length and width than BC fjords (*Table 01*).

The simple linear model was much more suited to modelling New Zealand fjord area, especially if fjord #62 can be ignored. *Figure 13* illustrates the dramatic difference between models which include fjord #62 and the model which ignores the data point. If fjord #62 cannot be discarded, more fjord data is required to confirm the proposed model (*NZ4*). Fjords which have a valley width of greater than 2km would be extremely valuable as there is a gap of 1km between the widest valley (fjord #62 – width of approximately 3km) and the second widest valley (fjord #66 – width of approximately 2km) in the New Zealand dataset.

Statistical Appendix

Computational Methods

- All statistical analyses were undertaken using Jupyter lab (alpha 0.26.5) with R (v 3.4.1 on x86_64-w64-mingw32 platform) and RStudio (v 1.0.153). All were installed as part of the Anaconda3 distribution (v 4.4.0)

- Libraries—MASS (functions used: rlm and boxcox), plot3D (functions used: scatter3D and text3D), sjPlot (function used: sjt.lm)

- Functions used during analyses:
    - read.csv() was used to import CSV files into R
    - lm() was used for non-robust regression
    - rlm() was used for robust regression
    - anova() was used to perform nested F-tests
    - summary() was used to print linear regression summary tables
    - sjt.lm() was used to format and export linear regression summary tables to HTML
    - influence.measures() was used to print influence measures summary table
    - seq() and predict() was used to create regression lines for models with non-linear transformations
    - plot() was used to display residual diagnostic plots
    - plot(), text(), abline(), lines() and legend() were used for all 2D plots
    - boxcox() was used to perform box cox transformations
    - scatter3D(), text3D(), expand.grid(), matrix() were used for all 3D plots

- All of the code used for statistical analyses is available at: *https://github.com/nf-s/Modelling-Fjord-Catchment-Area-with-Fjord-Valley-Characteristics-in-British-Columbia-and-New-Zealand*

Results

Correlation Matrix Tables for both regions

BC Fjords Correlation Table			
	Area	Length	Width
Area	1.00		
Length	0.84	1.00	
Width	0.58	0.56	1.00

NZ Fjords Correlation Table			
	Area	Length	Width
Area	1.00		
Length	0.61	1.00	
Width	0.88	0.64	1.00

Table 01: *The tables show that the correlation between valley width and height with fjord area is slightly higher with fjords in New Zealand than British Columbia. Interestingly, the valley characteristic with the highest correlation is flipped between both regions. In NZ valley width has a higher correlation with fjord area than valley length, but in BC valley length has a higher correlation with fjord area than valley width.*

There is also a significant correlation between length and width for both regions, this may cause collinearity issues when performing linear regression.

Both regions (British Columbia and New Zealand) are presented as separate analyses, mainly due to the findings from *Table 1* and because both regions reacted differently to various transformations.

British Columbia Results

British Columbian Fjord plots

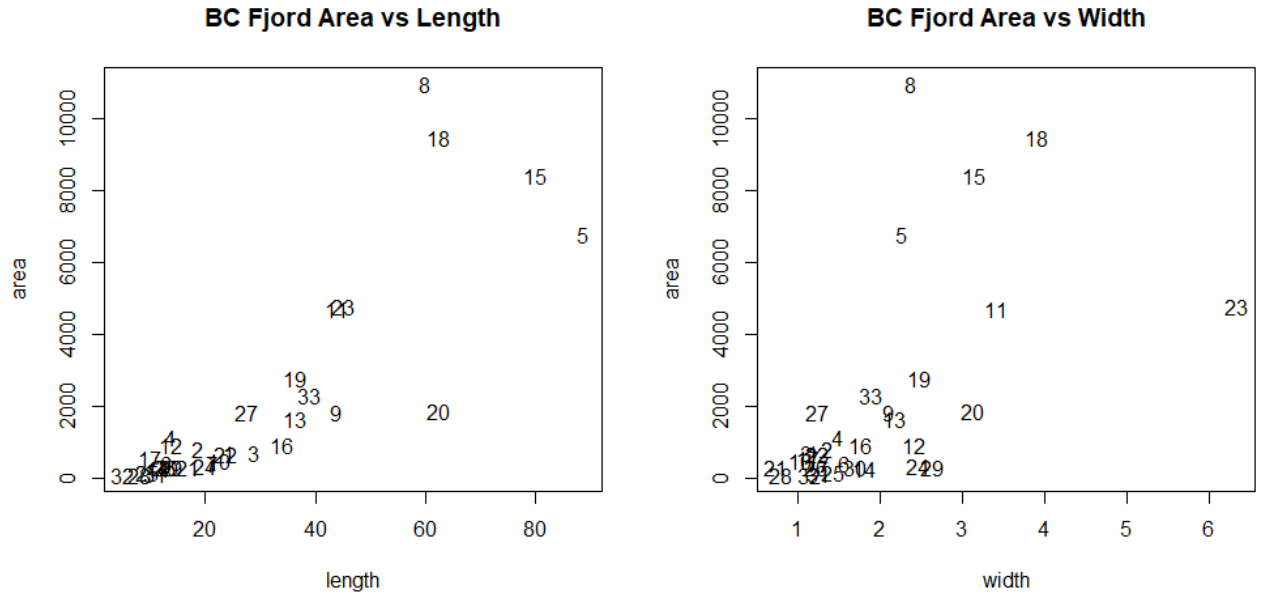

Figure 01: *This confirms the findings from the correlation matrix (Table 1) that there is a stronger correlation between length and area than width and area. Area vs length looks more homoscedastic than area vs width, but both plots show an increase in variance as area increases.*

The first regression models include combinations of width and length without any transformations:

British Columbia – Simple Linear Regression Summary

	BC1 – area=width*length			**BC2 –** area=width+length			**BC3 –** area= length		
	B	*std. Er*	*p*	*B*	*std. Er*	*p*	*B*	*std. Er*	*p*
(Intercept)	120.88	1446.44	0.934	-762.99	583.78	**0.005**	-274.82	481.46	**0.013**
length	39.87	45.24	0.385	100.15	15.83	**<.001**	112.88	13.31	**<.001**
width	-13.71	798.65	0.448	436.46	305.23	0.163			
length:width	27.63	19.47	0.167						
Observations	33			33			33		
R^2 / adj. R^2	.736 / .709			.718 / .699			.699 / .689		
F-statistics	27.008***			38.214***			71.959***		

Table 2: *While models BC1 and BC2 are statistically significant overall, the length and the width/length interaction coefficients are insignificant. All the BC3 model's coefficients are statistically significant and the adjusted R-squared value is also significant.*

The residual plots for the BC3 model show that the assumptions for regression aren't adequately met:

BC3 Regression Residual Plot – Area = Length

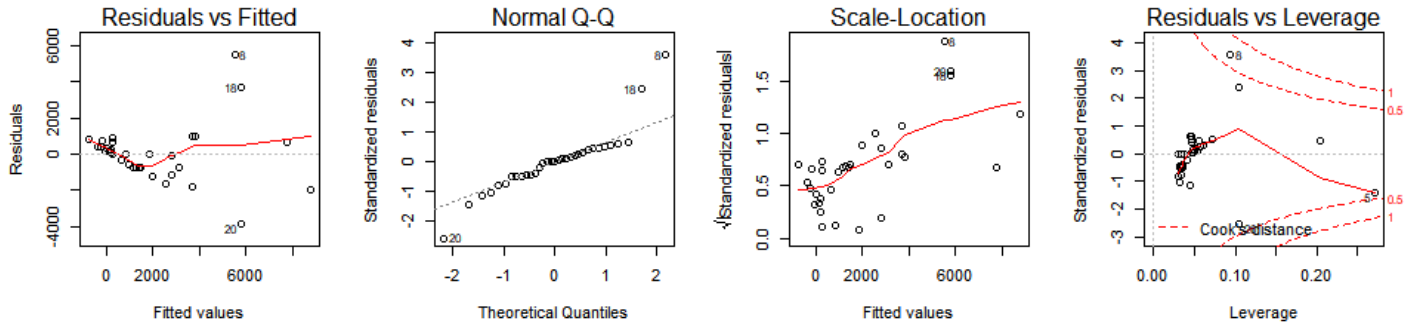

Figure 2: *BC3 model doesn't successfully meet some of the assumptions for linear regression. The variance is quite heteroscedastic, the data is not linear and there are a few points which break the normality of the data. There are also a few points (namely 5 and 8) which have large influence over the model.*

The plot of BC Fjord Area vs Length with the BC3 Regression Line further highlights the possible non-linear relationship between area and length:

BC3 (Area=Length) Plot – Area vs Length with Regression Line

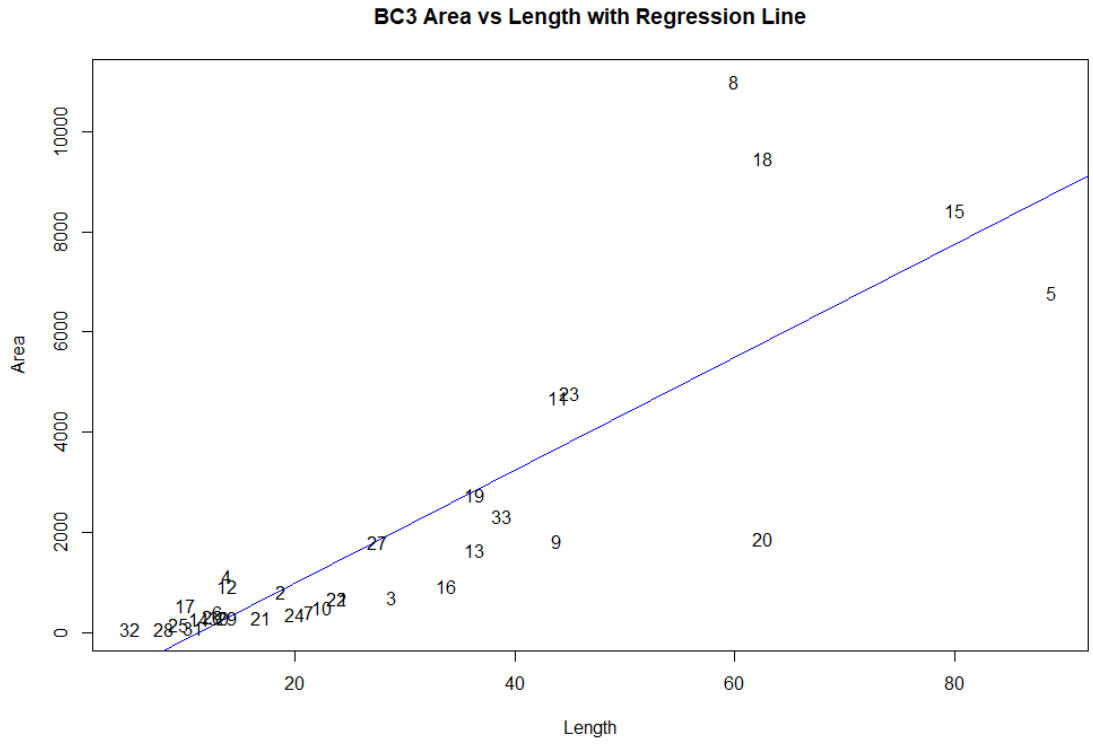

Figure 3: *This confirms the conclusions made from the previous table, the data doesn't seem to be linear and there isn't constant normally distributed variance from the regression line.*

Box cox transformation of the BC3 model suggested a log transformation on the dependent variable (area):

British Columbia –Linear Regression with Log Transformation on Dependent Variable Summary

	BC4 log(area)=width*length			**BC5** log(area)=width+length			**BC6** log(area)=length		
	B	*std. Error*	*p*	*B*	*std. Error*	*p*	*B*	*std. Error*	*p*
(Intercept)	3.89	0.58	**<.001**	4.85	0.24	**<.001**	5.15	0.2	**<.001**
length	0.08	0.02	**<.001**	0.05	0.01	**<.001**	0.06	0.01	**<.001**
width	0.8	0.32	**0.018**	0.27	0.12	**0.04**			
length:width	-0.01	0.01	0.083						
Observations	33			33			33		
R^2 / adj. R^2	.807 / .787			.786 / .772			.753 / .745		
F-statistics	40.520***			55.076***			94.620***		

Table 3: *Similarly to Table 2, the interaction between width and length is not significant, but the width coefficient (in BC5) is statistically significant. All models have a higher adjusted r-squared value than models in Table 2. The BC5 model has a lower F-statistic than BC6 and higher residual standard error, but has a higher adjusted R-squared value.*

The residual plot for BC5 shows that there is an obvious improvement in meeting the regression assumptions over the models with no log transformation:

BC5 Regression Residual Plot – (log(Area) = Width + Length)

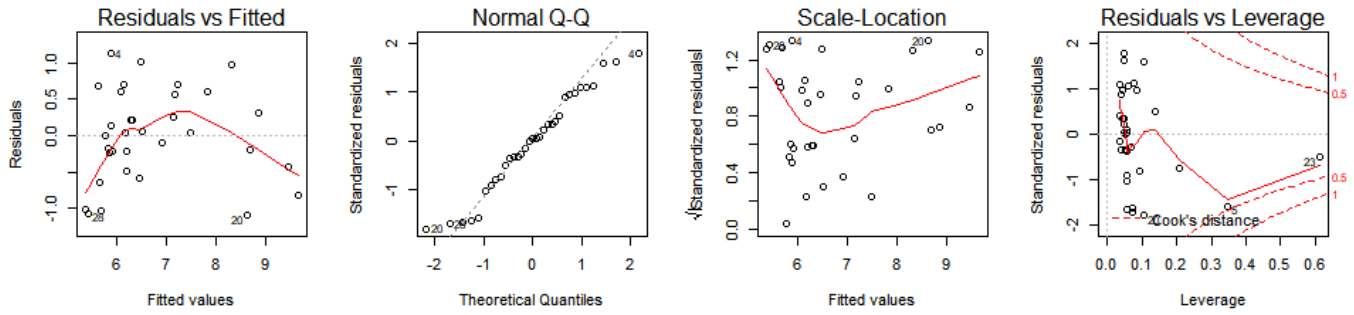

Figure 4: *This model shows an improvement in homoscedastic variance over the previous linear models (with no log transformation) and shows improvement in normality assumption.*

The residual plot for BC6 shows that the previous model (BC5) a more adequate model:

BC6 Regression Residual Plot – log(Area) = Length

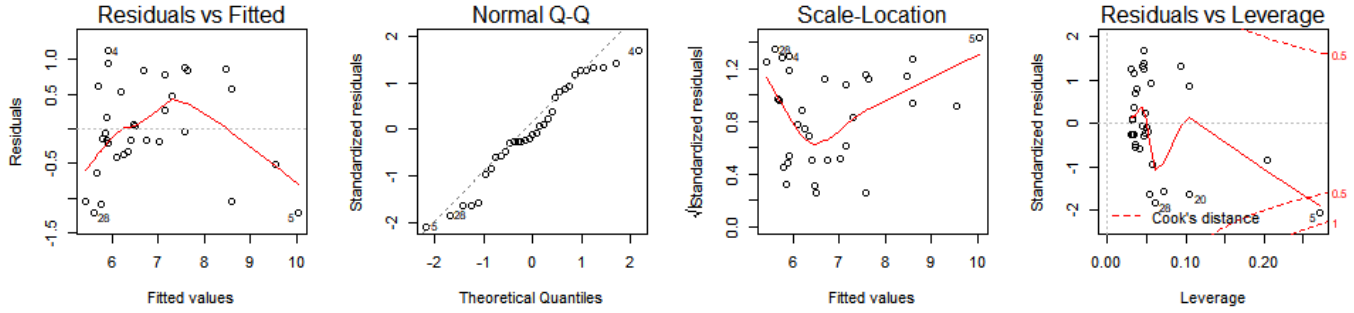

Figure 5: *This model is slightly worse than the previous in terms of homoscedasticity of variance and normality assumptions. The residuals vs leverage plot highlights that there is also a single point (#5) which has significant influence over the model*

Box cox transformations of the previous models suggested a power transformation:

BC – Linear Regression with Log & Power Transformations on Dependent Variable Summary

	BC7 log(area)^2 = width+length			**BC8** log(area)^2 = length			**BC9** log(area)^3 = width+length			**BC10** log(area)^3 = length		
	B	*std. Er*	*p*	*B*	*std. Er*	*p*	*B*	*std. Er*	*p*	*B*	*std. Er*	*p*
(Intercept)	20.44	3.03	**<.001**	24.67	2.64	**<.001**	51.29	31.34	0.112	97.11	27.49	**0.001**
length	0.67	0.08	**<.001**	0.78	0.07	**<.001**	7.29	0.85	**<.001**	8.49	0.76	**<.001**
width	3.79	1.58	**0.023**				40.97	16.39	**0.018**			
Observations	33			33			33			33		
R^2 / adj. R^2	.821 / .809			.787 / .780			.835 / .824			.801 / .795		
F-statistics	68.760***			114.391***			76.043***			124.720***		

Table 4: *As the power of log(area) increases the model's adjusted r-squared and f-statistic increases.*

The following plot highlights the differences in Y transformation when only taking length into account:

BC3, BC6, BC8 & BC10 - Area vs Length with Regression lines

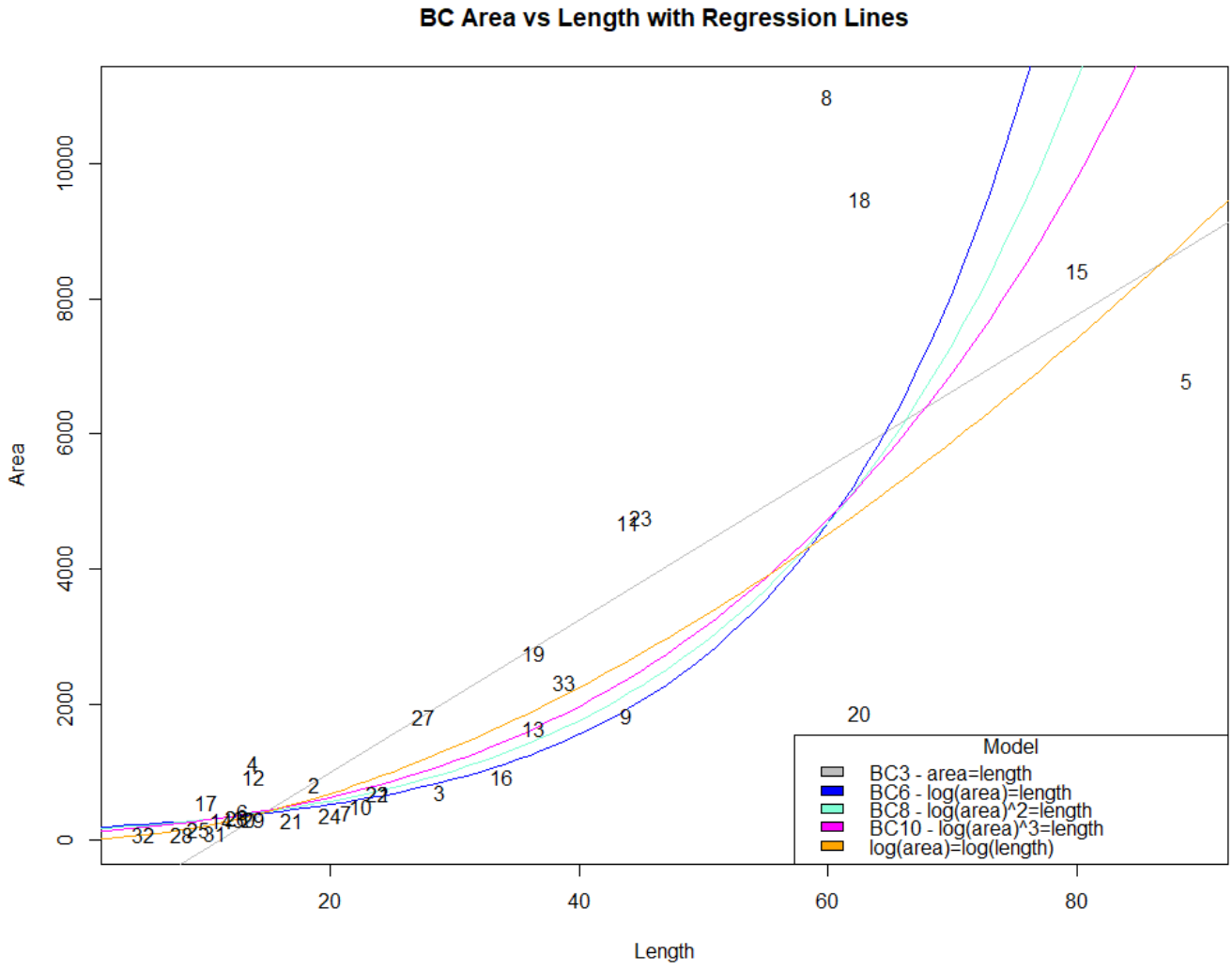

Figure 6: *The plot shows that as the power transformation of the dependent variable increases, the regression line represents the dataset more accurately. The log(area)=log(length) plot was added to show that a log transformation on the predictor variable has the effect of a higher power transformation on the dependent variable.*

The log(area)=log(length) model inspired the next set of models, which have a log transformation on dependent variable and both predictor variables (length and width):

BC – Linear Regression with Log Transformation on Dependent & Predictor Variables Summary

	BC11 log(area)=log(l)*log(w)			BC12 log(area)=log(l)+log(w)			BC13 (Robust) log(area)=log(l)+log(w)		
	B	*std. Err*	*p*	*B*	*std. Err*	*p*	*B*	*std. Err*	*p*
(Intercept)	1.94	0.74	**0.014**	1.81	0.46	**<.001**	1.68	0.44	**<.001**
log(length)	1.43	0.26	**<.001**	1.48	0.17	**<.001**	1.50	0.16	**<.001**
log(width)	0.33	1.15	0.778	0.59	0.25	**0.026**	0.65	0.24	**0.010**
log(length):log(width)	0.08	0.34	0.815						
Observations	33			33			33		
R^2 / adj. R^2	.855 / .840			.855 / .845			-		
F-statistics	57.022***			88.284***			-		
Res. std. err.	-			0.534			0.405		

Table 5: *The interaction between log(length) and log(width) is not significant, but the separated terms are significant for all other models. BC12 achieved a higher adjusted r-squared value over all previous models. BC12 also has the highest f-statistic for all models which include a width and length term. The robust regression model (BC13) has a slightly lower residual standard error than BC12 and all terms have a slightly lower p-value, except log(width) which has a significantly lower p-value.*

Unfortunately, all simpler models resulted in significantly worse adjusted R-squared and didn't satisfy all regression assumptions as well as BC12.

The residual plot for BC12 shows that the model meets all required regression assumptions more than any other model.

BC12 Regression Residual Plot – log(Area)=log(Length) + log(Width)

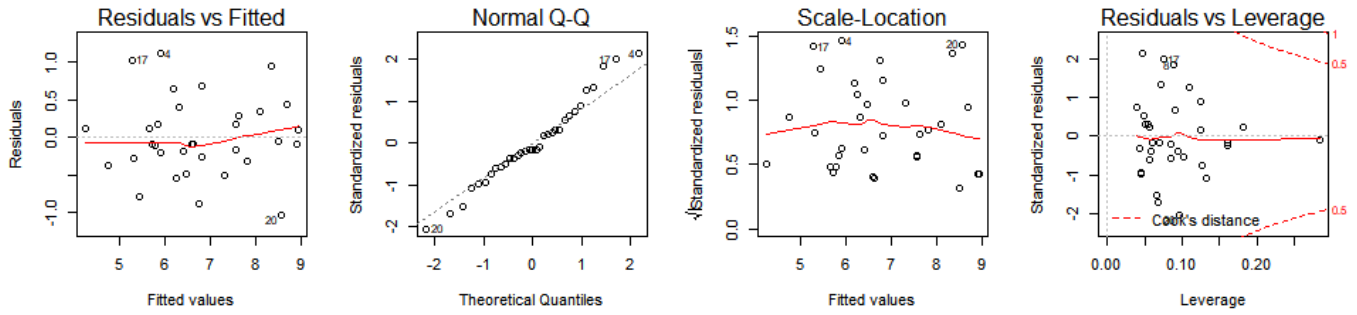

Figure 7: *BC12 successfully meets all assumptions for linear regression. The residuals vs leverage plot doesn't highlight any single points with significant influence over the model.*

BC12 & BC13 – Area vs Length vs Width with Regression Planes

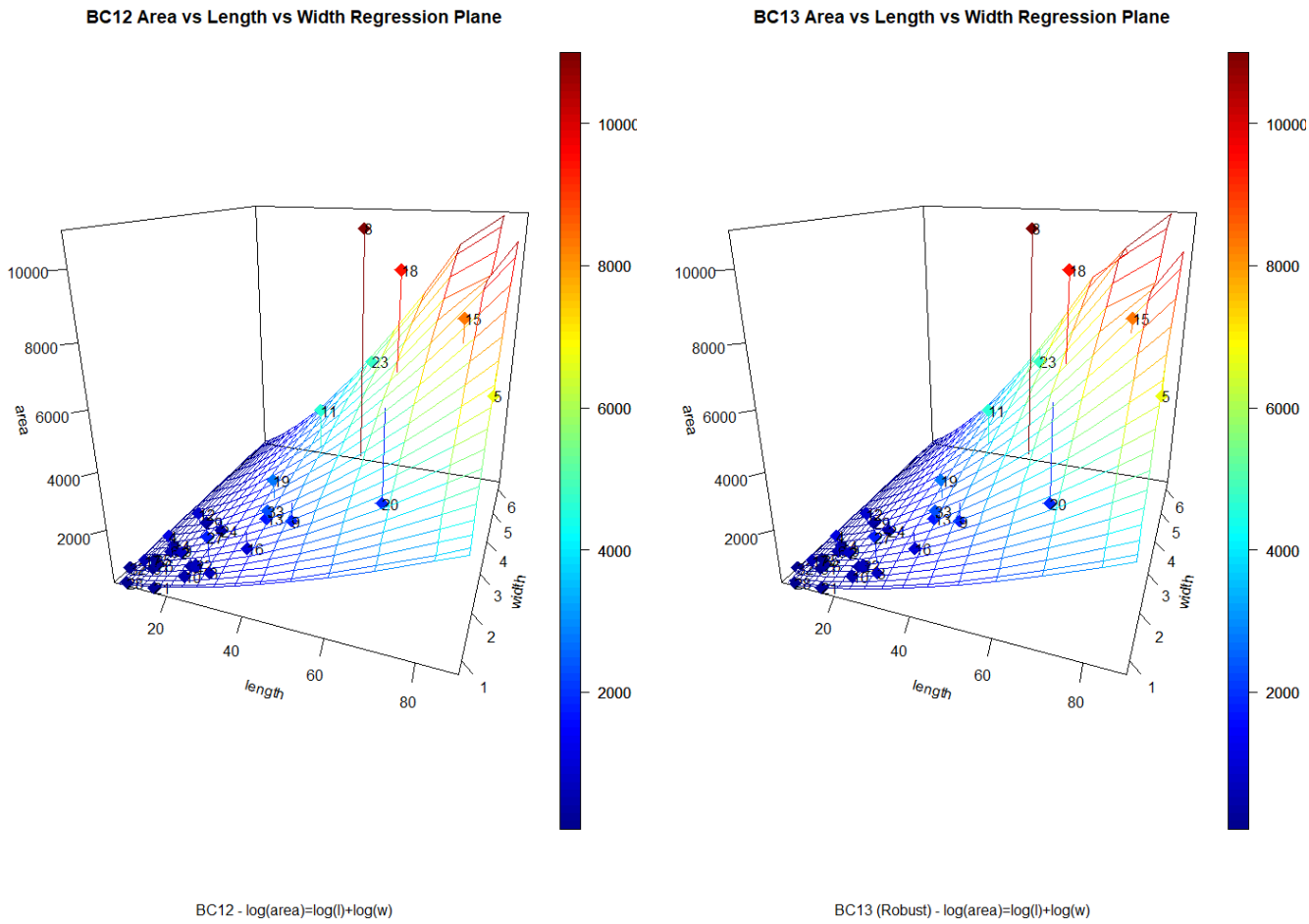

Figure 8: *Both models are quite similar, the robust model has slightly smaller residuals for points 18, 15 and 5. Unfortunately, for both models there is quite a large residual for points 8 and 20. As both models are to a log scale (rather than the linear scale displayed in the plot), the residuals for these points aren't that large relative to other points. This is shown in Figure 7, in the residuals vs fitted and the residuals vs leverage, which shows that no points have significant influence or a significant cook's distance.*

More fjord data, especially with higher values for width and length, would allow to further refine the proposed models and may help to explain some of the outlying points.

As stated in the *Findings*, I recommend model BC13.

New Zealand Results

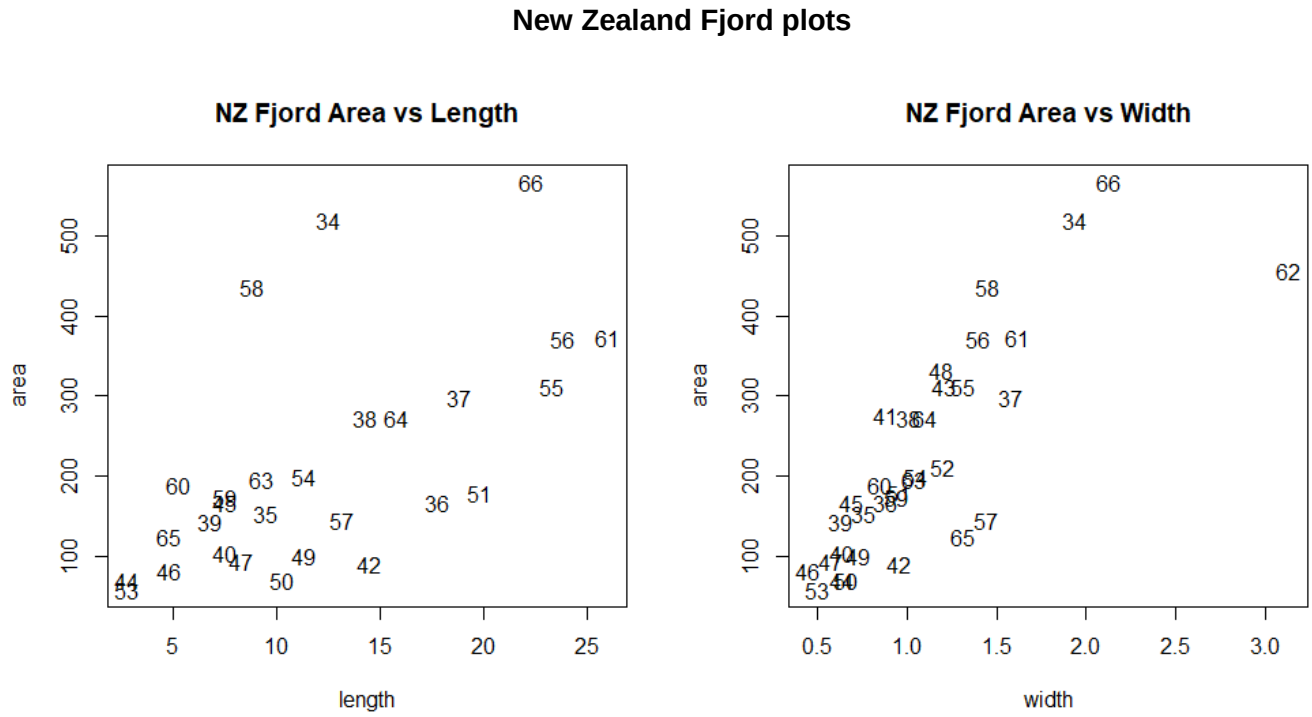

Figure 10: *This confirms observations made from correlation matrix in Table 1, width has as stronger correlation with area than length. The left plot shows that there is probably a linear relationship between length and area, but the variance is not constant or normally distributed. The right plot highlights a potential outlier – point #62. If this point is removed there is a strong linear relationship between width and area.*

Initially point 62 is left in the model, but later models will show that the point causes overfitting and results which are questionable (intuitively they don't make complete sense in the physical world).

The first regression models include combinations of width and length without any transformations:

New Zealand – Simple Linear Regression Summary

	NZ1 – area=width*length			**NZ2 –** area=width+length			**NZ3 –** area=width		
	B	*std. Er*	*p*	*B*	*std. Er*	*p*	*B*	*std. Er*	*p*
(Intercept)	-16.38	70.82	0.819	-83.62	33.61	**0.02**	-1.29	32.24	0.968
width	196.95	72.52	**0.012**	263.36	38.36	**<.001**	204.91	26.05	**<.001**
length	-4.3	6.07	0.485	1.65	2.52	0.519			
width:length	5.15	4.77	0.292						
Observations	28			28			33		
R^2 / adj. R^2	.793 / .768			.783 / .766			.666 / .655		
F-statistics	30.714***			45.198***			61.871***		

Table 6: *Length is insignificant for the models NZ1 and NZ2, width is significant for NZ3 but the intercept coefficient is insignificant.*

The residual plot for NZ3 highlights the massive influence data point #62 has over the model, it also shows that model doesn't meet the required assumptions for linear regression:

NZ3 Regression Residual Plot – Area=Width

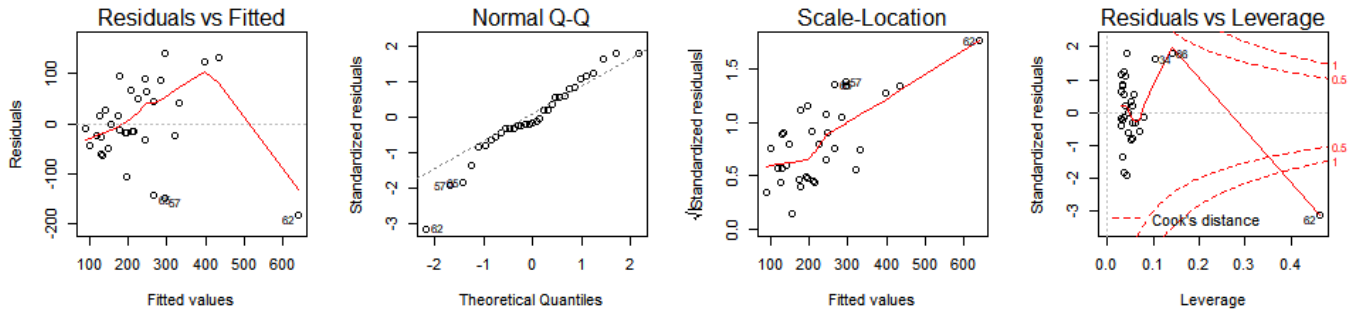

Figure 11: *The residuals vs fitted plot shows that the data isn't homoscedastic and it highlights point 62 as a potential outlier. The residuals vs leverage plot shows point 62 dramatic influence over the model, with its cook's distance well over 1.*

NZ3 Influence Measures Summary

	*dfb.1_*	*dfb.wdth*	*dffit*	*cov.r*	*cook.d*	*hat*
Point #62	**2.64**	**-3.35**	**-3.47**	0.92	**4.24**	**0.46**

Table 7: *This confirms thoughts from the previous figure, point 62 has massive influence over the model. Until the point's validity can be confirmed, the point will be left out of some the later models.*

If data point #62 is in fact legitimate, we require a more complicated model to accurately represent the data. The following analyses shows various transformation on the dependent variable as well as an added second order width term:

New Zealand – Linear Regression with Variable Transformations Summary

	NZ4 area=width+width^2			**NZ5** sqrt(area)= width+width^2			**NZ6** log(area)= width+width^2		
	B	*std. Err*	*p*	*B*	*std. Err*	*p*	*B*	*std. Err*	*p*
(Intercept)	-144.61	58.92	**0.02**	1.51	1.9	0.433	3.33	0.28	**<.001**
width	431.27	84.22	**<.001**	15.47	2.72	**<.001**	2.36	0.4	**<.001**
I(width^2)	-71.51	25.54	**0.009**	-2.84	0.82	**0.002**	-0.47	0.12	**<.001**
Observations	33			33			33		
R^2 / adj. R^2	.735 / .718			.738 / .721			.719 / .700		
F-statistics	41.680***			42.292***			38.315***		

Table 8: *The three models above achieve significantly higher adjusted R-squared values over the previous three models (NZ1, NZ2, NZ3).*

The residual plot for NZ4 (and Figure 12) show that the model is a tighter fit for the data:

NZ4 Regression Residual Plot – Area=Width+Width²

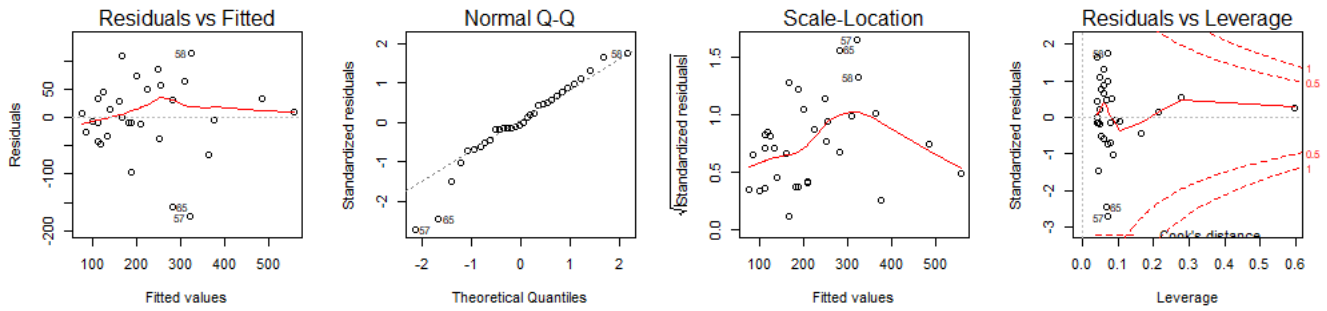

Figure 12: *This model improves over NZ3 in meeting assumptions required for linear regression: the homoscedasticity of variance and normality. The residuals vs leverage show that there are no points with significant influence over the model (all cook's distance values are below 0.5).*

Unfortunately, the models in Table 8 appear to overfit the data in the figure below. If point #62 is legitimate, more data is required to fully conclude whether these models accurately represent the relationship between fjord area and fjord valley width:

NZ3, NZ4, NZ5 & NZ6 - Area vs Width with Regression Lines

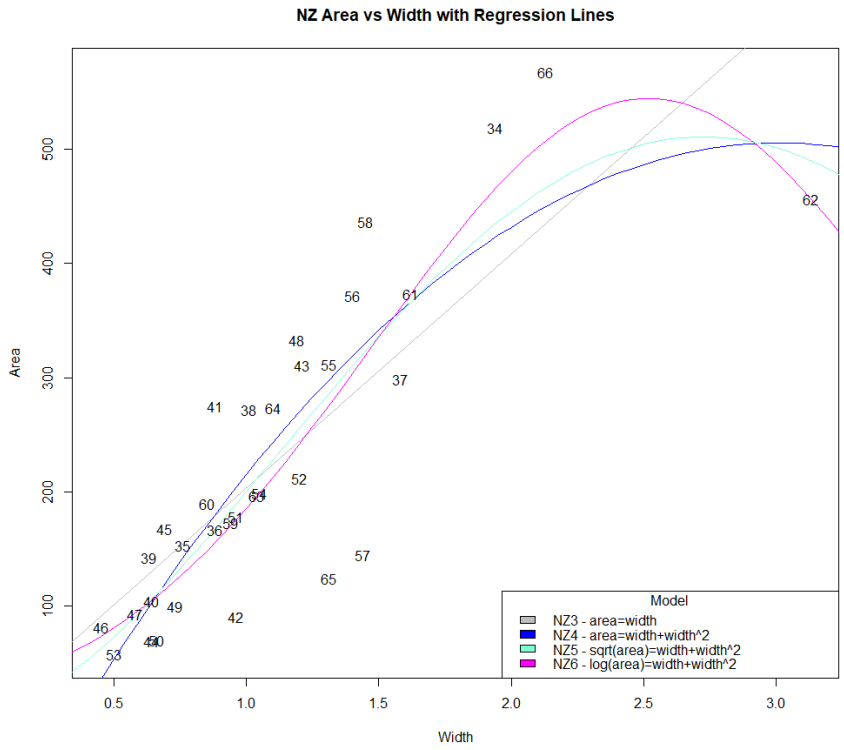

Figure 13: *The figure illustrates that all three models overfit the data, the final model doesn't make physical sense (intuitively). As the width reaches 2.5-3.0 the area starts to decrease, which is due to the significant influence point 62 has.*

More data is required to confirm or fully discard the three models. We need to verify that point #62 is legitimate before we can suggest the relationship between area and width follows models NZ4, NZ5 or NZ6.

The following models show that a simple linear model represents the data well if point #62 is ignored:

New Zealand – Linear Regression with Fjord #62 Removed Summary

	NZ7 – area=width			**NZ8** – sqrt(area)=width			**NZ9 (Robust)** area=width		
	B	*std. Error*	*p*	*B*	*std. Error*	*p*	*B*	*std. Error*	*p*
(Intercept)	-72.86	33.15	**0.036**	4.8	1.13	**<.001**	-45.68	26.18	**0.045**
width	278.25	29.38	**<.001**	8.95	1	**<.001**	254.76	21.15	**<.001**
Observations	32 (#62 removed)			32 (#62 removed)			33 (all data included)		
R^2 / adj. R^2	.749 / .741			.726 / .717			-		
F-statistics	89.709***			79.559***			-		
Res. std. err.	66.96			-			60.04		

Table 9: *Model NZ7 improves on model NZ3 (the simple linear model with point #62), the adjusted r-squared value increases from .65 to .74. The robust regression model NZ9 slightly improves on model NZ7 in terms of residual standard error.*

NZ7 Regression Residual Plot – Area=Width (with Fjord #62 Removed)

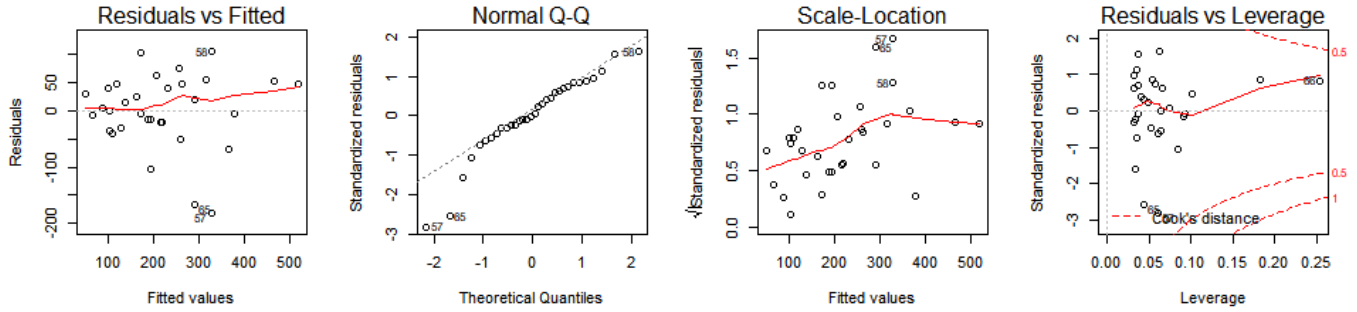

Figure 14: *Homoscedasticity has improved from plot NZ3 (area=width with all data), but the Normal Q-Q plot shows that there are a few which break the normality assumption. The residuals vs leverage plot shows that there aren't any points with significant cook's distance.*

NZ8 Regression Residual Plot – sqrt(Area)=Width (with Fjord #62 Removed)

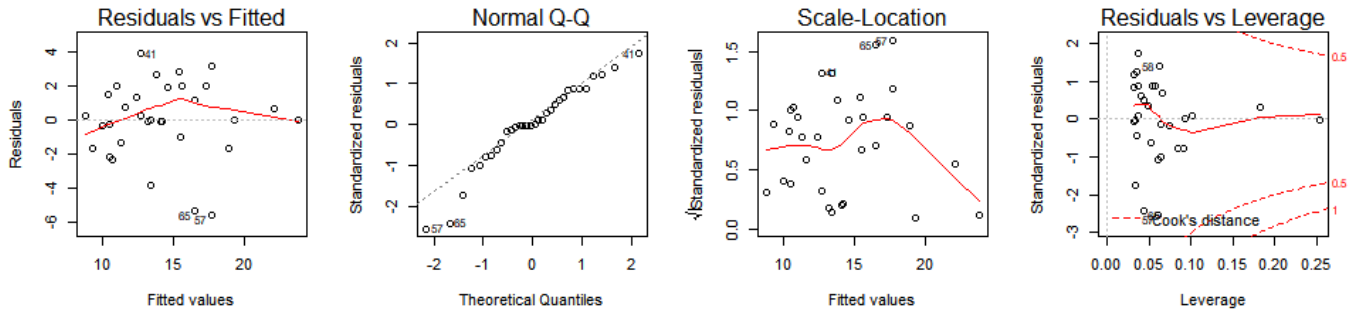

Figure 15: *The plot shows that this model has similar characteristics as the previous model.*

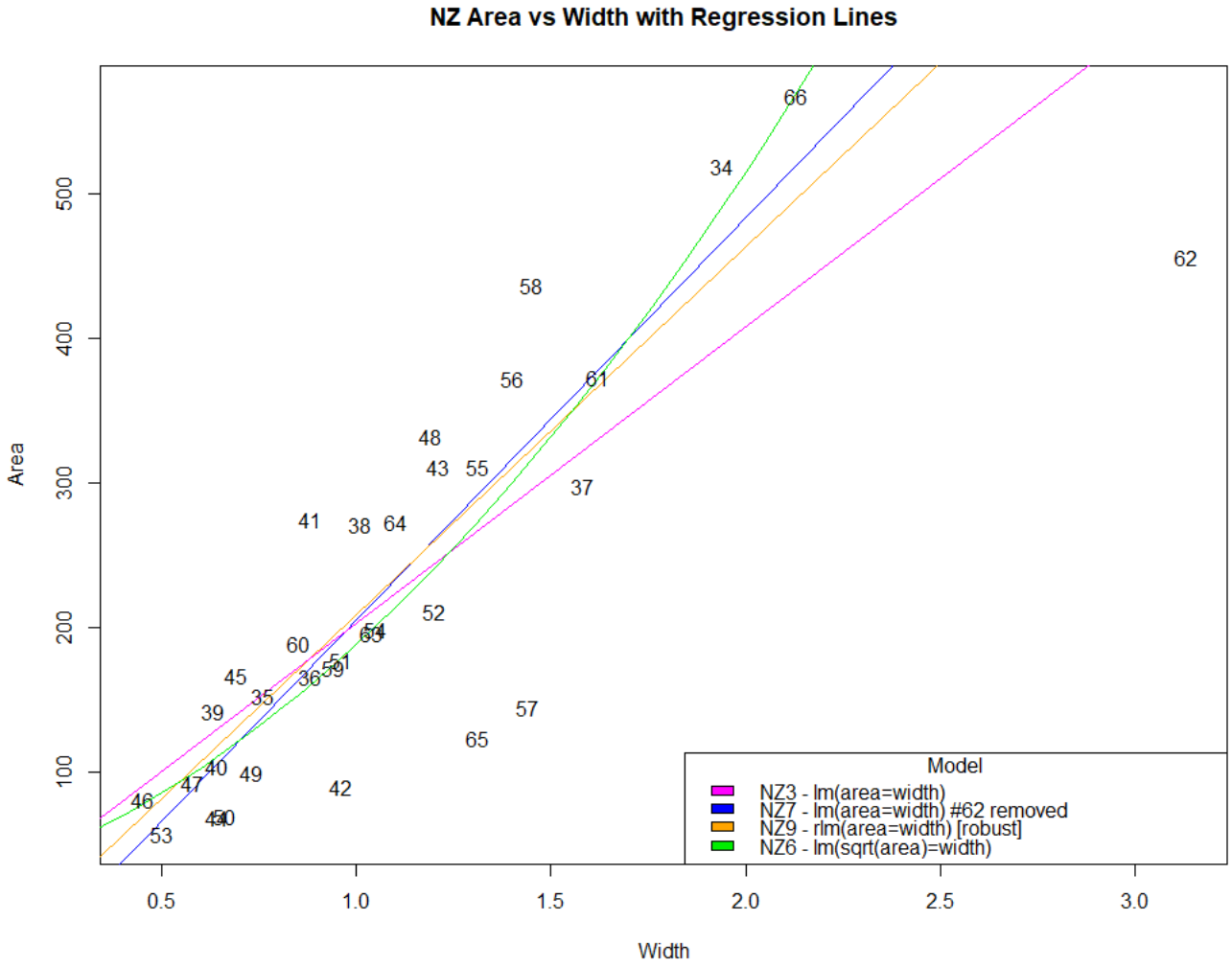

Figure 16: *The plot shows that the NZ3 regression line doesn't adequately represent the whole data set (point #62 included). The NZ9 robust regression line is a much better representation and is somewhat a middle ground between the original linear model (NZ3) and the linear model without point #62 (NZ7). The NZ6 model with a square root transformation has a slight upwards inflection, but more data is needed to confirm or discard this model.*