

Fjord Properties In Two Regions

Adam Pearce

KMA253 Data Handling and Statistics 2
School of Mathematics and Physics

Mentor: Dr Des Fitzgerald

Abstract

The data set titled *Fjords data1* contains two subsets, one dealing with fjords in New Zealand (NZ), and the other with fjords in the western Canadian province of British Columbia (BC). The recorded variables are: *catchment* = catchment area of the fjord, in sq. km; *length*, *width* = characteristics of the fjord valley, in km. The researcher is interested in the way in which catchment area influences the valley characteristics. The researcher is also aware from the research literature that power laws have been used as models in similar situations i.e. $\hat{y} = ax^b$ or $\log \hat{y} = \log a + b \log x$

Objectives

For the data on fjords in British Columbia and New Zealand, we are essentially interested in the relationships between catchment area (explanatory) and length and width of the fjord (response variables). This study will concentrate on the response *length*. We suspect a possible relationship of the form $y = ax^b$ according to parameters a and b , but we cannot be certain this applies. The underlying objective is then to find whether fjords in the two regions follow the same laws—in particular if their model parameters are indistinguishable.

Methods

Ignoring the suspected 'power' law, we first test the idea of a simple linear relationship for length in each region.

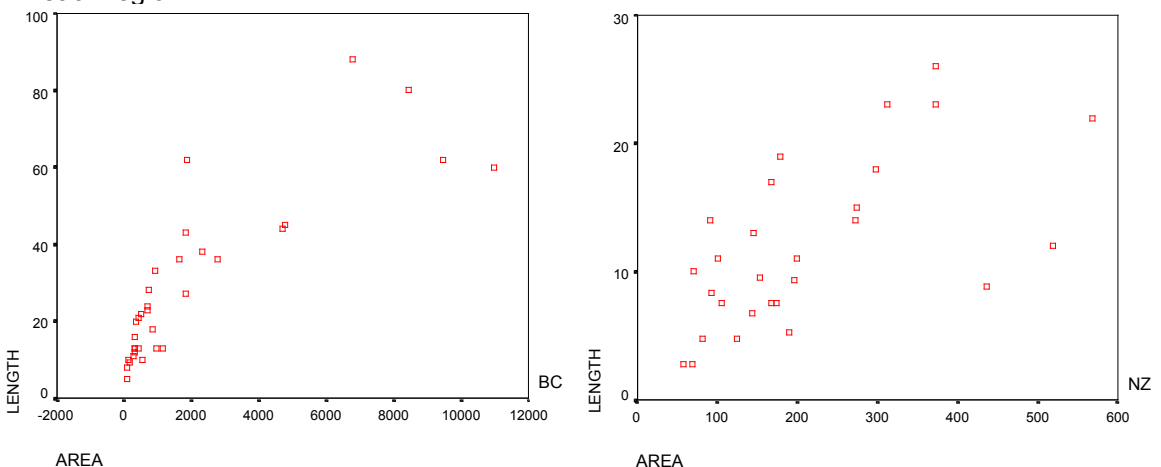


Figure 1: Length and area data for two sites.

1 Data by courtesy of Dr P Augustinus.

The scatterplots of Figure 1 show length against catchment area in New Zealand (NZ) and British Columbia (BC). If we are to model each region separately then we note by inspection, the plots do not show a very linear trend. The NZ plot has too few pairs to recognise, and the BC plot is steep and dense at low values, with some large large pairs that may not fit the same line very accurately. Note also the overall size difference in BC fjords compared with NZ ones.

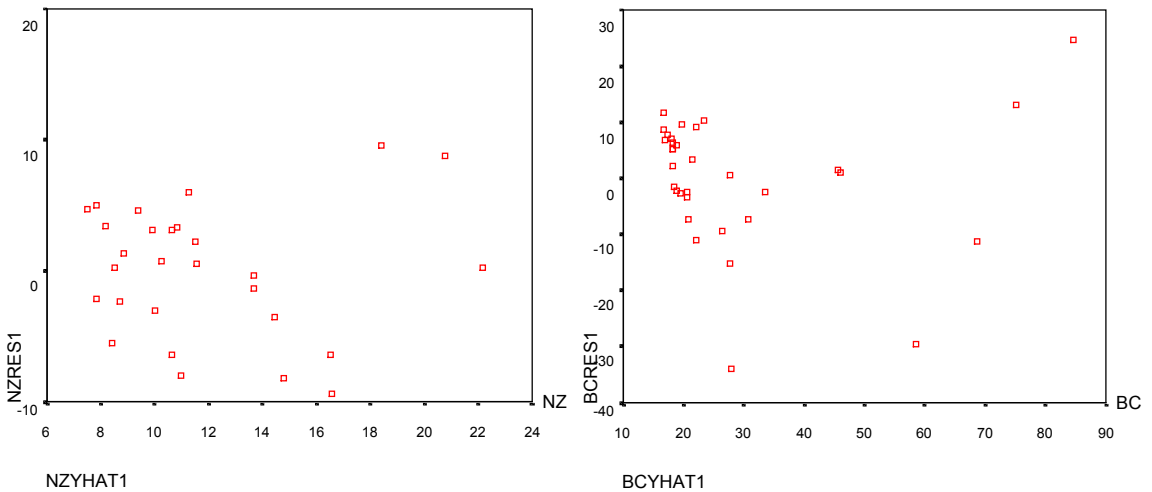
A 'least squares' method for simple linear regression for each region yielded the following models:

$$\text{NZ: LENGTH} = 2.8 \times 10^{-2} \times \text{AREA} + 5.82 ; \quad \text{SS}_{\text{RES}} = 690.88$$

$$\text{BC: LENGTH} = 6.2 \times 10^{-3} \times \text{AREA} + 16.28 ; \quad \text{SS}_{\text{RES}} = 4331.64$$

Thus, the SS_{res} for the whole model (both regions) is 5022.52. The R^2 value for NZ is 0.380, a poor proportion of explained variance for that model. With BC we have $R^2 = 0.704$, a lot better, but the poor shape of the models cannot be ignored.

The scatterplots show, for each respective model, the residual values against the model's expected values. In each case we should be seeing an even spread of residuals about a mean of 0, but the model might not be appropriate because of a well defined *shape* in each graph, specifically, curving downwards and, in the BC plot, fanning out towards larger 'yhat' values. The curve seems



worse in NZ (corresponding to a lower R^2).

Figure 2: Diagnostic plots for each region.

The original plots suggest a more appropriate model would be the suspected 'power law', where $y = ax^b$. With the apparent decreasing gradient we expect a model would yield a 'b' less than unity, and a low 'a'. We can test such a model by recognising the equivalent linear form

$$(\log y) = (\log a) + b (\log x) .$$

Hence, transforming the variables for a response 'log (length)' and explanatory 'log (area)', we see that Figure 3 indicates a more promising linear trend for each region.



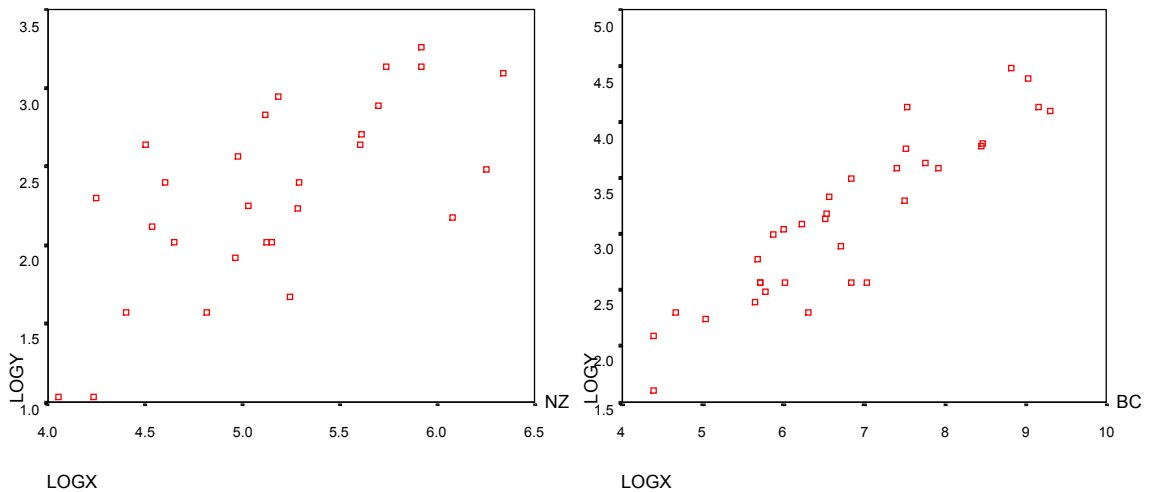


Figure 3: Log-transformed data for two regions.

The BC chart appears especially linear. Using these transformations we have ‘log (area)’ as a regressor, so each region yields the following models, where y = area, and x = length:

$$\begin{aligned} \text{NZ:} & \quad (\log y) = -0.980 + 0.640 (\log x) ; SS_{res} = 5.24 ; \\ \text{BC:} & \quad (\log y) = -0.171 + 0.486 (\log x) ; SS_{res} = 2.87 . \end{aligned}$$

Thus, the total SS_{res} for both regions is 8.12. (We cannot compare this with the first model because of the transformation.) Overall the models this time are seemingly more accurate. We have for BC an R^2 of 0.829, which is much better, and for NZ, R^2 is 0.456. The latter value is still not very high, but it is an improvement on the first model.

We now have a choice of rejecting the first model and working with the second, or improving on the second one. It seems appropriate that, given the difference in fit, the second model is better, but we might assume for now that there is a problem with the NZ data. In summary, we can say that it seems *likely*, given the precedents in research, that the power law can be applied to both regions, but there is the question about its accuracy for smaller fjords such as in NZ.

If we did accept these models, then here we have found the *best* case for the SS_{res} criterion, yielding different parameter values for each region, where

$$\begin{aligned} a_{\text{NZ}} &= e^{-0.980} = 0.375 ; & b_{\text{NZ}} &= 0.640 ; \\ a_{\text{BC}} &= e^{-0.171} = 0.843 ; & b_{\text{BC}} &= 0.486 . \end{aligned}$$

We have found the best model for each region, but the larger objective of the study is to examine the possibilities of whether fjords form in similar ways in different areas, so, if the power law is correct as we have seen, whether or not both regions have different parameters.

We can therefore form a complete model incorporating both regions and test it with ‘analysis of covariance’ techniques. We shall then inferentially test whether we can distinguish between models including different or equal parameters.

For the complete model of length against catchment area, we set up a binary variable I to indicate an NZ point (with value 1) or a BC point (with value 0). The problem is a multivariable one, where

in the data we expect the values to be given by the equation

$$(\log y) = (\log a_{BC}) + I (\log a)_{\delta} + b_{BC} (\log x) + I b_{\delta} (\log x) + \varepsilon .$$

That is, the regressor variables are I , $(\log x)$, and $I (\log x)$. ε is a normally distributed error term, and the coefficients we find will be expected to represent $(\log a_{BC})$, $(\log a)_{\delta}$ is the difference in $(\log a_{NZ}) - (\log a_{BC})$, b_{BC} , and b_{δ} is the difference in $b_{NZ} - b_{BC}$. We essentially have two lines being set up on the full plot of both regions, and submodels of this equation will shift the lines to force more parameters to be equal.

We denote the first model for uneven lines as the set $\{I, (\log x), I (\log x)\}$. Using the least squares method on this attains

$$(\log y) = -0.171 - 0.809 I + 0.486 (\log x) + 0.154 I (\log x) ; \quad SS_{res} = 8.12 .$$

(Note this exactly corresponds to the modelling of both regions separately in terms of coefficients and SSres.)

The model we are interested in is, firstly, whether both lines have the same gradient. Thus, here $b\delta$ is zero, so the 'interaction' term of that coefficient is gone, leaving the model denoted by the set $\{I, (\log x)\}$. Regression analysis on this model achieves

$$(\log y) = -0.332 + 2.31 \times 10^{-2} I + 0.510 (\log x) ; \quad SS_{res} = 8.32 .$$

As expected, SS_{res} is higher, but we have one less variable. Note the new (common) gradient is roughly between the gradients of each line before. We can check the feasibility of this new model, but first we notice the coefficient 2.31×10^{-2} is so low, the two lines must be almost equal in this model. It may even be feasible to use a submodel with $\{(\log x)\}$. That is, this equates both regions in relationship. This regression finds

$$(\log y) = -0.290 + 0.504 (\log x) ; \quad SS_{res} = 8.33 .$$

Note the SS_{res} here is very close to the previous submodel.

Now, we can check these nested models to find which one is most likely.

	SS_{RES}	df_{RES}		
FULL (3 var)	8.106	57		
SIMPLE (2 var)	8.321	58	$\Rightarrow f$	= 0.085/(8.321/58)
Diff from full	0.085	1		= 0.59
SIMPLE (1 var)	8.326	59	$\Rightarrow f$	= 0.09/(8.326/59)
Diff from full	0.090	2		= 0.64

Table 1: Comparison of models with f-test.

The simple feature of the third (last) model has important scientific implications, and may be preferred in the long term with respect the study of these fjord regions. We can also check *adjusted* R^2 values for each of the submodels.

In the full model, R^2 adjusted is 0.763; in the second, 0.761; in the third, 0.765. For the differences here, and for its simplicity, it may be that the final model is most favourable.



Conclusions

Therefore, with a generally high accuracy we can say that for these regions, NZ and BC fjords indicate that they follow a power law $y = ax^b$ for the response 'length' against 'catchment area'. There is no evidence to contradict the suggestion that they follow similar gradients for $(\log x)$, indicating an equal parameter b .

We conclude that both regions follow exactly the same relationship, indicating the parameters

$$a = e^{-0.290} = 0.748 ; \quad b = 0.504 .$$

Hence, we can say that the parameters for each region are generally indistinguishable.

A final diagnosis of this model shows that for the data range we have, these parameters seem to fit the trend fairly well.

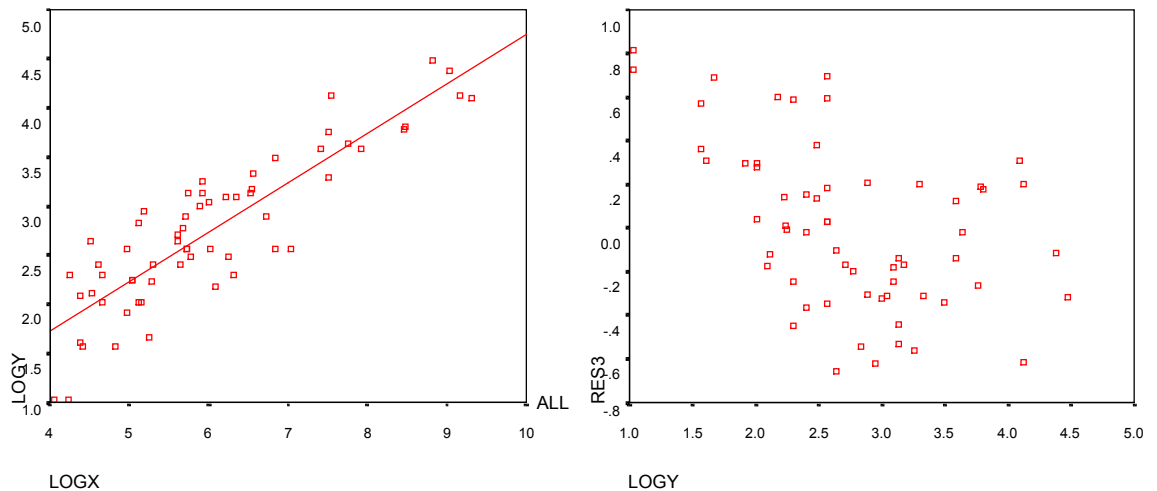


Figure 4: Diagnosis plots of the log-transformed model.

Figure 4 shows the final trendline plotted on the data set, and the residuals against predicted values on this model. Thus, to a limited extent the model seems to fit.