Dataset

Refer to the Fjords dataset ('fjords.csv'). The file contains data on fjords 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 (NZ or BC), categorical variable;
- area - catchment area of the fjord, in sq. km.;
- length, width - characteristics of the fjord valley, in km.

Research question

The researchers are interested to know if valley characteristics (width or length) can be used to get accurate predictions of catchment area. The researchers are aware from the literature that power laws have been used as models in similar situations, i.e.,

$$y = ax^b \text{ (equivalently } \log(y) = \log(a) + b\log(x)\text{) or } \log(y) = a + bx.$$

Analysis tasks

1. Visualise how area changes with length and width for each region.
2. Conduct a simple linear regression for each region and valley characteristic separately. Are the same variables (width and/or length) significant in the models for BC and NZ? *Note that you are only expected to do a simple linear regression so do not fit models that use both length and width as predictor variables.*
3. Investigate the diagnostic plots for each regression model. Do the simple linear models fit well?
4. Based on the original plots, and also on the residual diagnostics from the simple models, investigate models that use transformations of the variables and/or polynomial models. Which transformations do you suggest (if any) and is the transformation the same for both regions?
5. Propose a "best" model for each region. How well does this model fit? How good is its predictive accuracy?