Fitting linear models

QERM 514 - Homework 2

10 April 2020

Background

The goal of this assignment is to familiarize yourself with fitting linear models in **R**. We will be working some data from nearby Lake Washington that is part of a long-term monitoring program begun in the 1960s by the late, and rather famous, Dr. W.T. Edmondson and since continued by Dr. Daniel Schindler. The accompanying data file `L_Washington_plankton.csv` contains information on the following four variables:

- `Daphnia`: index of the density of the cladoceran *Daphnia* (unitless)
- `Greens`: index of the density of green algae (unitless)
- `Cyclops`: index of the density of the copepod *Cyclops* (unitless)
- `Temp`: water temperature (C)

Daphnia are an effective grazer on phytoplankton and green algae make up a large proportion of their diet. *Cyclops* are an inferior grazer compared to *Daphnia*, but a competitor nonetheless. *Daphnia* growth rates are also affected by water temperature.

As you work through the following problems, make sure to explain your thought process and show your code, so Mark can give you partial credit, if necessary.

Question 1

a) Write out the equation for a linear regression model that expresses *Daphnia* abundance as a function of its preferred prey, green algae, and describe the terms in your model.

b) Produce a scatterplot that shows the relationship between `Daphnia` and `Greens`. Make sure to label your plot accordingly and give it an informative caption. Describe the relationship between `Daphnia` and `Greens`. Does a linear model seem reasonable here?

c) Produce the step-by-step **R** code required to fit your model **via linear algebra** to generate estimates the model parameters and the data. Be sure to show the construction of the design matrix ($\mathbf{X}$), the calculation of the parameter estimates ($\hat{\beta}_i$), the calculation of the hat matrix ($\mathbf{H}$), and the calculation of the model predictions ($\hat{y}_i$).

d) Calculate and report your estimate of the residual variance (σ^2).

e) Give a prediction of what you might expect the specific abundance of *Daphnia* to be on the next sampling occasion if the abundance of green algae is 1.5 units. Also provide an estimate of the interval around your estimate that conveys 95% confidence in your prediction. Again, do so via direct calculations rather than relying on **R**'s built-in functions.

Question 2

a) Expand upon your model from Question 1 to include the additional effects of *Cyclops* and water temperature on *Daphnia*. Write out your equation and describe the terms in the model.

b) Using **R**'s built-in functions, fit the model from (a) and show the resulting table of results. For each of the p-values shown in the table, describe the null hypothesis being tested.

c) Test the hypothesis that $\beta_{Greens} = \beta_{Cyclops} = \beta_{Temp} = 0$. What is the F-statistic, the associated df, and the p-value? What can you conclude from this test?

d) It has come to your attention that someone has done lab experiments suggesting the effect of temperature on *Daphnia* abundance is 0.4 per degree Celsius after controlling for the effects of prey (green algae) and competitors (*Cyclops*). Create a null hypothesis test to evaluate the evidence for this finding from the data collected in the field. Specify H_0 and report the results of your test. What do you conclude?