BI5302 case study group work

28 October, 2020

Introduction

For this exercise you will work in small (3 - 5 people) groups to analyse data from two case studies (see below for details). The overall aim of this exercise is to challenge you with data typically obtained from biological and ecological studies, use R to fit suitable linear models to test specific research questions and then validate these models. As you might have guessed, data from both of these case studies contain a number of issues and as a result the standard linear modelling approach is unsuitable. Your job is to first identify what the problems are and then identify the causes of these problems.

So, for each case study you will need to:

1. Identify the specific hypothesis / research question.
2. Import the appropriate data file and perform graphical data exploration.
3. Fit a suitable linear model with appropriate explanatory variables and interaction terms.
4. Validate the linear model using appropriate residuals plots to identify violations of linear modelling assumptions (hint: there will be problems!).
5. Again using appropriate plots of the residuals, try to determine why the assumptions have been violated.

For all of the steps above, it would be great to document your workflow using an R markdown document in order to create a report detailing your approach and highlighting the issues (either pdf or html format is ok). We can then share our reports on MyAberdeen and learn from each other! Also, as the course progresses and you learn how to extend the linear model to deal with these issues we will return to these two datasets to analyse them correctly!

If you would prefer to work from the pdf version of this document you can find this here.

Benthic Biodiversity experiment

These data were obtained from a mesocosm experiment which aimed to examine the effect of benthic polychaete (*Hediste diversicolor*) biomass on sediment nutrient release (NH_4, NO_3 and

PO_3). At the start of the experiment replicate (n = 3) mesocosms were filled with homogenised marine sediment and assigned to one of five polychaete biomass treatments (0, 0.5, 1, 1.5, 2 g per mesocosm). The mesocosms were allowed to acclimatise for 24 h after which the concentration of either NH_4, NO_3 or PO_3 was measured in the water. The concentration of only one nutrient was measured in each mesocosm. The researchers were particularly interested in whether the nutrient concentration differed between polychaete biomass treatments and whether these effects were dependent on the type of nutrient.

Data for this case study is saved in a tab delimited file called `Hediste.txt` which you can download from the **Data** link. Save this data file in your `data` directory in your RStudio project.

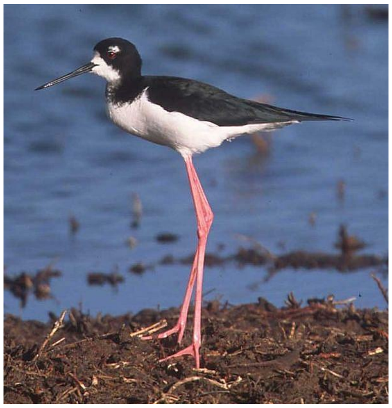
The data file contains the following variables:

- `concentration` : the nutrient concentration
- `biomass` : the polychaete biomass level
- `fnutrient` : the nutrient measured

As a hint, you should be looking out for violations of the homogeneity of variance assumption for this dataset.

Environmental impacts on Hawaiian black-necked stilt (*Himantopus mexicanus knudseni*) abundance

These data were collected from bird surveys conducted on two Hawaiian islands (Maui and Oahu) from 1956 - 2003. The annual abundance of black-necked stilts was measured each winter using transect surveys on each island. Along with bird counts, annual rainfall data for the region was also obtained from the National Climate Data Center. The researchers were interested in understanding whether levels of rainfall impacted on bird abundance and whether any impact was different between the two islands.



Data for this case study is saved in a tab delimited file called `hawaii2.txt` which you can download from the **Data** link. Save this data file in your `data` directory in your RStudio project.

The data file contains the following variables:

- `abund` : the abundance of stilts
- `rainfall` : the amount of annual rainfall
- `location` : the name of the island
- `year` : the year the data was collected

As a hint, you should be looking out for violations of both the homogeneity of variance assumption and independence of residuals assumption for this dataset.