STA2007H Project 1: Regression
Density of pollinating fynbos birds

Due Date: Friday 19th May before 4:00 pm

From wikipedia: “Pollination is the process by which pollen is transferred from the anther (male part) to the stigma (female part) of the plant, thereby enabling fertilization and reproduction.”

Creative Commons: Michael Hanselmann

The project is meant to be done in groups of 3 people and involves analysing the data described below and writing a short report (see the project brief below and the guidelines for Statistical report writing on Vula). Each group must hand in a type-written hard copy at Stats reception, as well as an electronic version including a pdf file of the report and the R script file via Vula.

The project uses data from a fictional study on the density of pollinating fynbos birds in the Western Cape. The data are on Vula in a file called ”birds.csv”. The response variable (*density*) is the density (per hectare) of pollinating birds. Other variables include:

Variable	Description	Codes/Values
`size`	size of fynbos patch (hectares)	1230 - 8310
`distance`	distance to the nearest patch (kms)	9.2 - 40.5
`altitude`	altitude (meters)	7.8 - 2211.0
`time`	time since last fire (years)	1.8 - 17.2
`alien`	presence of alien vegetation	absent ; present
`protea`	presence of key nectar-providing protea	absent ; present
`vegtype`	vegetation type	Mountain fynbos / Renosterveld / Strandveld

Project Brief and Suggested Layout

1. Introduction: brief introduction to the problem and the data.

2. Statistical methods: do not repeat your notes or text book. Just a brief explanation of what technique is used and why.

3. Data exploration (no formal hypothesis tests needed ie no p-values)

   (a) Begin with univariate exploration. Construct a table of appropriate descriptive statistics and interpret.

   (b) Explore and interpret the relationships between the outcome variable `density` and the other variables.

   (c) Remember that you don't have to include all output in the report and can put peripheral results in an Appendix.

4. Model building: build models that map on to the hypotheses below:

   - H1: food availability
     Density is a linear function of the presence of key protea species and the vegetation type.

   - H2a and H2b: fire's role on the fynbos
     Density is driven by the length of time that has elapsed since the last fire, the type of vegetation, and whether or not there is alien vegetation present. This hypothesis should be split into two and fitted with and without an interaction between `time` and `alien`. Can you explain in words what the model with interaction does?

   - H3: fragmentation
     Density is a linear function of the distance to the nearest patch as well as of the size of the patch.

   - H4a and H4b: environmental gradient
     Density is driven by the presence of alien species as well as by the altitude of the patch. This hypothesis should also be split into two and fitted with `altitude` as firstly a linear effect and secondly as a quadratic effect. Do you think it is sensible to include a quadratic effect of `altitude`? Explain your answer.

5. Use model selection tools to gauge the extent of relative support for the set of candidate models you have fitted.

6. Use appropriate plots to assess if the assumptions of the model hold. Discuss the characteristics of the three points that are marked in one of the plots (4, 44 and 63).

7. Interpretation:

   - Choose your optimal model. Use appropriate plots to illustrate the fitted regression lines from this model.
   - Fully interpret your chosen model. What inferences can you make about the effects of the predictors? Show that you can read the output from a statistical model and explain to somebody what it says about the data.

8. Discussion and Conclusions

Please note the following:

- An electronic version of the R script file must be provided in addition to the final report.

- In addition to content, the layout and presentation of your report are important. Do NOT cut-and-paste R output into the report. Rather present the output in neatly constructed tables. All graphs and tables are to be appropriately labelled and must be referred to in the text.

- The work that you will hand in must be your own. This means that you can discuss problems and difficulties or ideas with your tutor or an expert in the field. However, you must acknowledge any help received, and each group should do the analysis and write the report on its own. Your project will NOT be marked if you do not include a plagiarism declaration, or if you do not email me your script file.

- Although there is no formal page limit for this assignment, you should bear in mind that more does not mean better!

Hints

1. For Microsoft Word, save R graphs as metafiles, with .emf extension!! Everything else looks grainy and just not nice.

2. You can hand in a simple stapled report, with double-sided printing.

3. Harvard reference style: Give references in text, e.g. Erni et al. (2012) found that . . . (summarised in your own words), or: . . . (Erni et al. 2012).

4. All tables and figures should have a number and a caption and should be referred to in the main text.

5. The way to cite R:

   R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

   See citation() in R.

6. Avoid naked p-values! Always add the size of the effect as well, with standard error! We want to know how and by how much the explanatory variables influence the response. The significance level just ensures that you don't make claims that could have been just due to random fluctuations that occurred purely by change.