

13 April 2009

Hampton

Doing logistic regression in the R software package

R is a free software that is handy for statistics and for graphing. It is rapidly becoming the standard for doing statistics in ecology and environmental sciences, as well as in other disciplines.

There are many reasons to prefer “**scripted analysis**”, like what you are doing when you use **R**, over the point-and-click software packages that are a little easier to use. First, scripting forces you to write down every step in your analysis – this is great for keeping a record of what you have done in analysis, for sharing with colleagues, and for doing analyses over and over when there are many steps involved.

Languages like R also expose the guts of your analysis, so that it is not a black box – programs like Excel hide statistical errors and assumptions in their formulations that you don’t see.

If you are doing research, scripted analysis in R is the industry standard. You will ultimately save time in your research by investing time in learning R now.

When you are not doing research yourself, but you are interacting with others doing scientific analysis, you will benefit from the openness and meticulous record-keeping in which scripting forces your colleagues to engage as they perform research.

Other packages that are common in ecology and environmental science that involve scripting analyses are Matlab and SAS – both are excellent statistical platforms, but both are very expensive. JMP can provide a “script” of the actions performed at the end of your analysis – if you must do analysis in JMP, rather than R, it is a good idea to ask it to produce a script for your final analyses for you to keep in your files, so that you maintain a good record of your work.

Today’s lab is conceptually much more simple than previous labs, because a major objective is to familiarize you with a scripted analysis in R.

Please read this short paper about using best practices in data management and analyses:

<http://www.nceas.ucsb.edu/files/news/ESAdatamng09.pdf>

Downloading the free R program onto your computer

<http://cran.cnr.berkeley.edu/>

Choose your computer’s platform (Windows or Mac?)

Choose “base”

Choose “Download R 2.8.1 for Windows”

The .exe that you download will have a set-up wizard to walk you through the steps.

Today's lab

Objectives: 1) compare results of linear and logistic regression models for the cobra lily data discussed in lecture, and 2) become familiar with a scripted analysis in R

Cobra lily data were retrieved from:

<http://harvardforest.fas.harvard.edu/personnel/web/aellison/publications/primer/primer.html>

- 1) On your desktop, **create a folder called "Logistic"** – place the data file (Ellison_DarlingtoniaData3.csv) and the R script (LogisticRegression.R) in this folder
- 2) **Open R** – you should see a window called "R Console" – this is the window where the action will occur, but first you need to know how to tell it to do stuff. We wrote a script that you will use, we'll open that in a few steps.
- 3) Change directory, so that R will know where to look for your script file and your data file – **File - change dir** – choose the folder "Logistic" on your desktop
- 4) Open the script that's been prepared for lab. **File – Open script – "LogisticRegression.R"** in your Logistic folder on your desktop. This opens a new window – this window is called a text editor. It holds the script that will ultimately be executed in the R Console window.
- 5) Look at the script in your text editor.
 - a. The symbols **##** tell R that the stuff in that line is just your notes – R ignores everything you write behind a **#**
 - b. Everything that is not behind a **#** is code for R
 - c. Take a few minutes to read through this script
 - d. There are a couple different ways to use the script to run the regressions in the R Console – you can copy and paste or highlight and right-click...
- 6) Load the data and make sure it loaded correctly. To do this, **Highlight text** from "**##R code for comparing...**" through "**head(lilydata,5)**" – **copy and paste** into the R Console, hit **Return** – you should see the first 5 rows of your Darlingtonia data – if you don't, then hail Nick
- 7) From here on out, you should look at the code in your text editor. **Read the comments (##)** in the script for each block of text, then **highlight** the block of text and **copy and paste** it from your text editor window into the R console,
- 8) What is the AIC from the linear model?
- 9) What is the AIC from the logistic model?
- 10) You can run any part of this script again if you missed something.
- 11) Does the AIC tell you that one model is better than the other?
- 12) What other reasons are there to choose a logistic or a linear model for these data?