BIO782P Introduction to General Linear Models Cribsheet

Rob Knell / Joe Parker

9 November 2017

Basic GLMs Part A: Multi-factor ANOVA

Part A has two options for you to choose from. Either:

Option 1: Bushmeat example

The dataset ***bushmeat.txt*** contains some of the data from a study published in 2013 by Effiom **et. al.** to investigate the effects of bushmeat hunting on vegetation regeneration in African Rainforests. The authors compared regeneration of experimentally cleared plots in sites where there is little hunting of primates and sites where primates are rare because of hunting. There are four variables in the data file. ***Dispersal*** is the method of seed dispersal (abiotic, other or primates), ***Hunting*** is whether the plot in question is hunted or not and ***Forest*** refers to the specific forest that the plot in question was located in. ***Number*** is the number of seedlings per category a year after the plot was cleared.

1. Load the data into R. Note that this is a csv (comma separated values) file rather than tab-separated text so you'll need to use the `read.csv()` function, like this:
2. Fit a two-factor ANOVA to the data with Number as a response variable and Dispersal and Hunting as explanatory variables, plus the interaction between the two.
3. Is the interaction between the two significant? Check the diagnostic plots to see if the residuals are well-behaved. If they are not, then you might be able to correct the problem by transforming the Number variable. If you need to do this, re-fit the model and check the diagnostic plots again.
4. Examine the nature of the interaction between Dispersal and Hunting by using the `interaction.plot()` function. Read the help file to find out how it works. Use Dispersal as the x.factor, Hunting as the trace.factor and Number as the response. What do you see? What does this mean?
5. Another way of visualising these data is to draw a boxplot with the data categorised by more than one grouping factor. See if you can manage to do this. You can align the text on the x-axis to vertical using the `las=2` argument in the `boxplot()` function, but you'll also have to adjust the margins of the figure to make the x-axis text visible:

```
## Adjust margin sizes to fit the axis labels

par(mar=c(8,4,1,1))

## Once you've finished, reset margins to defaults

par(mar=c(5,4,4,2)+0.1)
```

Effiom, E.O., Nuñez-Iturri, G., Smith, H.G., Ottosson, U., and Olsson, O. (2013). Bushmeat hunting changes regeneration of African rainforests. Proc. R. Soc. B *280*:20130246.

Or:

Option 2: Burying beetle gene expression example

The dataset caring.csv contains a set of gene expression data for 867 genes (identified as a "caring" set of genes) from burying beetles ***Nicrophorus vespilloides***. Male and female beetles which were engaged in parental care of their offspring either as part of a pair of beetles or as a single parent had their transcriptomes sequenced and compared with transcriptomes from control beetles that were not engaged in caring for offspring. The data for the 867 pairs of genes are presented as the log2-fold change in expression in the beetles that were actively caring for their offspring.

1. Load the data into R. Note that this is a csv (comma separated values) file rather than tab-separated text so you'll need to use the `read.csv()` function.
2. Check the data set using `str()`. `Sex` and `Biparental` should be factors with 2 levels each. `Log_2_fold_change` should be a numeric variable.
3. The data as they are provided are signed but we want to analyse the absolute (unsigned) values since decreasing gene expression can be as important as increasing gene expression. You can convert your values to absolute ones using the `abs()` function.
4. Draw a boxplot showing the absolute value of `Log_2_fold_change` according to each level of `Sex`, and then another showing it according to each level of `Biparental`. What do you see? Do you see anything that might make you cautious about analysing these data with a normal ANOVA?
5. Fit a two-factor ANOVA to the data with the absolute (unsigned) value of `Log_2_fold_change` as a response variable and Sex and Biparental as explanatory variables, plus the interaction between the two.
6. Is the interaction between the two significant? Check the diagnostic plots to see if the residuals are well-behaved. If they are not, then you might be able to correct the problem by transforming the Number variable. If you need to do this, re-fit the model and check the diagnostic plots again.
7. Examine the nature of the interaction between Sex and Biparental by using the `interaction.plot()` function. Read the help file to find out how it works. What do you see? What does this mean?
8. Another way of visualising these data is to draw a boxplot with the data categorised by more than one grouping factor. See if you can manage to do this.

Parker DJ, Cunningham CB, Walling CA, Stamper CE, Head ML, Roy-Zokan EM, McKinney EC, Ritchie MG, Moore AJ. 2015. Transcriptomes of parents identify parenting strategies and sexual conflict in a subsocial beetle. Nat Commun 6:8449.

Lab Part B continues on next page…

Basic GLMs Part B: A continuous explanatory variable and a factor

The deleterious effects of inbreeding are well known as problems in small populations. In a study published in 2014, van Bergen and colleagues described the results of an experiment to investigate the effects of inbreeding on flight performance and pheromone production in the butterfly ***Bicyclus anynana***. We will analyse one part of their data to look at how flight performance relates to thorax size and to inbreeding.

1. The file `van_Bergen_Bicyclus.txt` has data from this experiment. Save it as an object in R and check the data frame using `str()`. There should be three variables. `Inbreeding` is a factor with three levels indicating the number of generations of sib-matings in a butterfly's recent family history. There are three levels: none, one and two. `Drythor` is the dry weight of the butterfly's thorax, and `FII` is the flight inhibition index – the number of times the butterfly settled during a two-minute period when it was being stimulated to take off immediately once it settled. Butterflies that are less able to fly for long bouts have higher FII measurements.
2. For some exploratory data analysis, draw a boxplot of `FII` grouped by `Inbreeding` level, and a scatterplot of `FII` against thorax weight (`Drythor`). What do you see? Is there any reason to think these data might not be suitable for a standard parametric analysis?
3. Fit a model with `FII` as the response variable and `Inbreeding`, `Drythor`, and the interaction between the two as explanatory variables.
4. Check your diagnostic plots. How do they look? As with the previous exercise, if the residuals are not well behaved then you might be able to correct the problem by transforming the FII variable. If you need to do this, re-fit the model and check the diagnostic plots again.
5. This time we're going to use a deletion test to assess whether our interaction term is statistically significant. Use the `drop1()` function with `test="F"` to do this.
6. If the interaction term is not significant, re-fit the model without it and repeat the process until you have a minimal adequate model. Look at the summary table and try to work out what the coefficients mean.
7. Plot the data with the fitted model. You might need to think about what the best approach might be for this.
8. Explain the model output in words.

Bergen, E. van, Brakefield, P.M., Heuskin, S., Zwaan, B.J., and Nieberding, C.M. (2013). The scent of inbreeding: a male sex pheromone betrays inbred males. ***Proc. R. Soc. B*** **280**:20130102.

End of lab