Practical Exercises for Introduction to Linear Models: ANOVAs, Multiple Regression and ANCOVA

Saad Arif

21st Nov 2019 - Afternoon

For most of the following exercises, you should be able to copy the code from the lecture and modify it for the specific datasets in the exercies. Please make sure to ask questions if you get stuck. You may want to visit exercise 4 tomorrow after we have covered the lecture on ANCOVA's in the morning.

Exercise 1: Clinical trials

A clinical trial was conducted to study the effect of the levels of a drug on some measure of well-being (called `SCORE` in this case). Both male and female subjects were randomly choosen for the study and then randomly assigned to to either a low or a high `DOSE` treatment.

To read this data, directly from the web (the csv file is stored at the web location below, you can verify this by copying and pasting the url below into a new browser window/tab), into an object called `drugTrials`, do the following in R:

```
drugTrials <- read.csv("https://git.io/JeoOu")
```

Make sure the data is read in correctly. If it is read correctly, there should be 48 rows and 4 columns (we don't need the first column which is unique ID for each individual). Please make sure to have a look at the data first using `head()` to see the first 5 rows or the `View()` function to see the dataset in all its glory.

1.) Perform some exploratory analysis on the data set and answer the following questions: (a) Do you think the SCOREs are different between genders? (b) are SCOREs different between differend doses of the drug? (c) is there an interaction between GENDER and DOSE? (d) is this data set balanced (i.e. equal observations across all subgroups)?

2.) Carry out a two-way ANOVA with interaction on this data set and answer the following questions (a) Is the interaction effect signficant? (b) are any of the main effects signficant? (c) if the interaction is signficant can you still easily interpret the results for any significant main effects? Why or why not?

3.) Carry out the appropriate Tukey Post-hoc tests and examine the results. Do they contradict your interpretation of 2(c) (if you attempted one) ?

4.) Evaluate the assumptions of the anova model for the drug trial data? Are there any assumptions violated? Which one might be the most problematic?

Exercise 2: The Diet Experiment

This data for this study comes from an experiment in which people were randomly selected and then randomly assigned to one of three diets to encourage weight gain. Additionally the diets were trialled independently in three different countries.

To read this data, directly from the web, into an object called `dietData`, do the following in R:

```
dietData <- read.csv("https://git.io/JeoO4")
```

Make sure the data is read in correctly. If it is read correctly, there should be 27 rows and 3 columns.

1.) Perform some exploratory analysis on the data set and answer the following questions: (a) Do you think the different diets are equally effective for weight gain? (b) How do you think country of origin might influence weight gain? (d) is this data set balanced (i.e. equal observations across all subgroups)?

2.) Carry out a two way ANOVA with interaction and Tukey post-hoc tests to determine which diet is most effective for weight gain? Is there a single best answer to the previous question? Why or why not?

3.) Evaluate the assumptions of the anova model for the drug trial data. Are there any assumptions violated? Which one might be the most problematic?

4.) Visualize your results as interaction plot (as in the lecture slides), but unlike the example on the lecture slides, provide 95% CI for means of all subgroups (instead of the standard errors, which are rougly 68% CIs).

Exercise 3: Swiss Fertility data

For the follow exercise we will use a built-in data set in R. To load the dataset, type the following in R

```
data(swiss)
```

Briefly, this data includes a standardized fertility measure from 47 french-speaking provinces in Switzerland from 1888. It also includes 5 socio-economic indicators of the provinces as well. For further details type `help(swiss)`. All variables have been scaled to a numerical continuous scale.

We would like to understand what influence, if any, the 5 socio-economic factors have on fertility.

1. Use the `pairs()` function to draw scatterplots of all 6 variables with one another: (a) Which variables seem most correlated with `Fertility` (b) Which variables seem highly correlated with one another? (c) Which variable would you choose to model changes in fertility and why?
2. Use the following code to fit all 5 variables as the explanatory variables for `Fertility` and the save the fit in '`fit1`:

```
fit1 <- lm(Fertility ~ ., data = swiss)
```

Call `summary()` on the saved model. Are all the slopes/coefficients signficant? Are there any results *surprising* based on your plots from (1) above?

3. Use the `vif()` function from the `car` package, on `fit1` to find the variable that leads to the most variance inflation. Fit a new linear model that omits this variable but retains the other 4 variables, save this model as `fit2`. Perform a nested likelihood ratio to test which model `fit1` or `fit2` is more appropriate. **Note** if the `car` package is not installed, you will have to install it yourself.
4. Repeat the process above to find the the variable that leads to the most variance inflation in `fit2`. Fit a new linear model that omits this variable but retains the other 3 variables, save this model as `fit3`. Perform a nested likelihood ratio to test which model `fit1`, `fit2` or `fit3` is more appropriate.
5. For the best fitting model from (4) check all the assumptions of the multiple linear regression model. Do any of them seem particularly problematic?
6. Plot the model coeffecients (from the best fit model in part 5) with 95% confidence intervals. Which variables appear to have the largest effect on fertility?

Exercise 4: Stickleback Association Study

The dataset for this exercise consists of 1682 single nucleotide polymorphism (SNP) markers, genotyped across 328 adult threespine stickleback fish (*Gasterosteus aculeatus*). There is an enormous range of morphological variation present within three-spined sticklebacks. These fall into two main categories, the **anadromous** and the **freshwater** forms. One major difference between the two categories of forms is the number of armour plates along the lateral side of the fish. Anadromous fish are heavily plated and have upwards of 25 plates on each side, whereas freshwater fish only have about 5 on each side (see the image below)

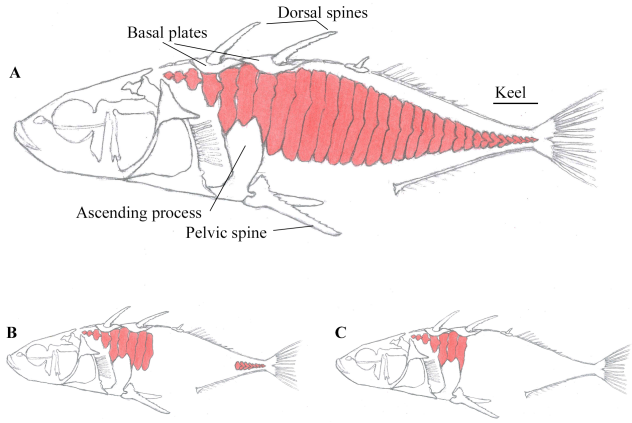

Fig Source: Wiig E, Reseland JE, Østbye K, Haugen HJ, Vøllestad LA (2016) Variation in Lateral Plate Quality in Threespine Stickleback from Fresh, Brackish and Marine Water: A Micro-Computed Tomography Study. PLoS ONE 11(10): e0164578. https://doi.org/10.1371/journal.pone.0164578

The data for the 328 fish comes from adults caught in an **admixture zone**. In the admixture zone, anadramous and freshwater types meet and breed freely, allowing for recombination and linkage disequilibrium between genetic markers and phenotypes. Hence, these *hybrid* individuals are ideal for identifying the genetic basis of traits that differ in the parental popualtions (anadromous *vs.* freshwater). **We will use this dataset to find the genetic loci associated with the number of armour plates**

Read in the dataset as follows (this make take some time):

```
stickle <- read.csv("https://git.io/Jer8h")
```

Make sure the data set has 328 rows (adult fish) and 1686 columns. Here is a brief description of the variables:

The first column `ind` is a unique identifier for each fish. You likely will not need this for anything.

Columns 2:1683 are all the genetic marker data. The name of each genetic marker starts with the chromosome number followed by a . then the marker position in bases followed by some additional information that is not relevant here.

`Ancestry` is the predicted ancestry for each individual fish. This number ranges from 0-1. A number closer to zero means the fish has mostly a freshwater genetic makeup, while a number close to 1 means the fish has mostly an anadromous genetic makeup.

`Std.Length`is a numerical variable the represents the size of the fish (in cm)

`No.PLATES` is the number of armour plates along the lateral side of the fish

1.) Explore the relationship between Number of plates and the other numerical variables (Ancestry and Std. length). Would you expect a relationship between any of these variables? What does the data suggest?

2.) Based on your exploratory analysis above, which, if any of Std.length or Ancestry seem correlated with No.Plates? Fit a linear model with one or both of these as explanatory variables and No.Plates as the response variable and call it `baseline`. How much variance in No.Plates does your model explain (Adjusted R Squared)?

3.) Iteratively add each SNP marker to the `baseline` above to create another model. Compare the two models using the likeleihood ratio test from the `anova` function. This comparison will till you whether adding that particular SNP is a better model fit than jus the baseline model, or in other words if that SNP is signficantly associated with plate number or not *after* accounting for variable(s) in the baseline model. Make sure you do this for each of the 1682 markers and store the p-value of the likelihood ratio test in another dataframe along with the name of the marker. After doing this for all markers, use the bonferroni correction ($\frac{\alpha}{\text{no. markers}}$) to find all markers with signficant association with the number of plates.