Exercises

Introduction to Linear Models: ANOVAs, Multiple Regression and ANCOVA

Saad Arif

Nov 2019

Exercise 1: Clinical trials

A clinical trial was conducted to study the effect 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, into an object called `drugTrials`, do the following in R:

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

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).

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?

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

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 put on one of three diets to encourage weight gain. Additionally the diets were trialled 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/JewAJ")
```

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 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), make sure to provide standard errors for means of all subgroups.

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?

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 surprsing 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?

Exercise 4: Stickleback Association Study

The dataset for this exercise consists of 1679 single nucleotide polymorphism (SNP) markers, typed across 647 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 differenct between two categories of forms is the number of armour plates along the lateral side of the fish.

A

Dorsal spines
Basal plates
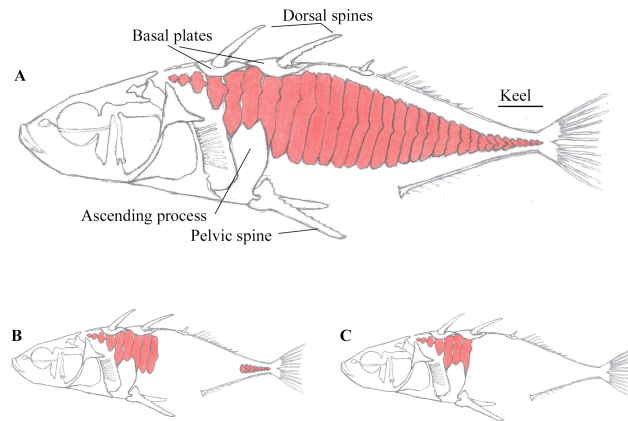Keel
Ascending process
Pelvic spine

B

C

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

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

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