Univariate Data and Modelling – Exercises

Session 7 – Two-way ANOVA

Exercise 1

After your first successful job on the FEV you become known as a good statistical consultant and you get more and more work for various clients. The last one seems like a big one, it is a big multinational pharmaceutical company. Doing a good job on this first assignment probably means that you are set for years. The problem they have is that they have worldwide twelve laboratories and that they want to select the most accurate one for the production of a new drug. As a test, all laboratories were given a set of standard solutions containing a predefined concentration of calcium. The labs had to determine the concentration of the calcium four times for each solution. The test was replicated two times, after which the labs send their results to head office. The dataset contains following variables:

ID	Identification Code
Lab	Laboratory (A - … - L)
Rep	Replication (1, 2)
Sol	Solution (1, 2, 3, 4, 5)
Target	(Unknown) target Ca concentration
M1	Measured concentration 1
M2	Measured concentration 2
M3	Measured concentration 3
M4	Measured concentration 4

a) How would you start to analyze this data?
   - How to translate the question of the client to a workable statistical question?
   - Do you need to perform actions on the data to achieve this?
   - What method would you use (regression, ANOVA, …)?
   - How does the full model looks like?
   - What hypothesis do you need to test? Anything else you need to test?

b) Due to randomization issues we cannot use all four measured concentrations per replication as such, with the techniques seen in this course. The only thing we can use is the mean measured value over these four observations.
   - Make a new column containing the mean observation per solution and replication

c) As response variable we will use the difference between the target Ca concentration and the mean observed value.
   - Make a new column containing the difference between the target and the mean observation per solution and replication

d) Now it is time to start comparing the different labs. The client wants to know which lab is giving the best overall performance

- Are we dealing with balanced or unbalanced data?
- Draw and interpret the interaction plot by using following function in R:

```
interaction.plot(lab.df$Lab, lab.df$Rep, lab.df$Diff,
type="b", pch=c(18,24), col=c(1,2))
```

- Run a two-way ANOVA with the difference between target and measured as response variable and the laboratory and replication as predictor variables. Include the interaction in the model. Interpret your results.
- Run some diagnostics on the proposed model. Adjust the model if needed.
- Do a multiple comparison on the means of the laboratory using the appropriate technique. Given the results from the diagnostics, can we use the Tukey method?
- Which lab would you recommend?