6 Analysis of variance (ANOVA)

6.1 Oneway analysis of variance: Psoriasis

Prerequisites: Exercises 3.1 and 3.3 (reading and working with datasets)

Psoriasis is an immune-mediated disease that affects the skin. Researchers carried out an microarray experiment with skin from 37 people in order to examine a potential association between the disease and a certain gene (IGFL4). For each of the 37 samples the gene expression was measured as an intensity.

There were three different types of skin samples: 15 skin samples were from psoriasis patients and from a part of the body affected by the disease (`psor`); 15 samples were from psoriasis patients but from a part of the body not affected by the disease (`psne`); and 7 skin samples were from healthy people (`control`). The data is saved in the files `psoriasis.xlsx` and `psoriasis.csv` with variables `intensity` and `type`. There is also a variable `typeNum`, which is not used until the very last question.

The scientific question is whether the gene expression level differs between the three types/groups, and the natural type of analysis is thus a oneway analysis of variance (ANOVA).

1. *(Stripchart, boxplots)* Try the following commands to get an overview of the data, and explain what you learn from the plots:

   ```
   stripchart(intensity ~ type)
   boxplot(intensity ~ type)
   ```

2. *(Group means)* It is well-known that the group means are essential ingredients in the analysis. Find the groups means with the commands

   ```
   mean(intensity[type=="healthy"])
   mean(intensity[type=="psne"])
   mean(intensity[type=="psor"])
   ```

   Notice that you would rarely do this as part of the analysis, but the values are useful to understand what happens in the next questions.

3. *(Fit of oneway ANOVA, reference group)* The model for oneway analysis of variance is fitted with

   ```
   oneway1 <- lm(intensity ~ type, data=psoriasis)
   ```

   Notice that this is analogous to fitting a simple linear regression: `intensity` is the response variable; `type` is an explanatory variable.

   Then make a summary of the model:

   ```
   summary(oneway1)
   ```

Some explanation is most likely useful at this point. The “Coefficients” part of the output has three lines — one per group or parameter in the model.

The lines are denoted `(Intercept)`, denoted `typepsne` and `typepsor`. No line is denoted `typehealthy`. This is because R has selected `healthy` as the reference group, and compares the other groups to this reference group. More specifically the `(Intercept)` line concerns the expected value of the reference group, so the estimate is simple the mean of the 7 observations from healthy people, whereas the estimate in the `typepsne` line is the difference between the mean of the `psne` observations and the `healthy` observations. Similarly for the `typepsor` line.

All four values in the same line contain information about the estimation of the same parameter:

- `Estimate`: The estimated value of the parameter (mean for reference group or difference between group mean and reference group mean for the other groups)
- `Std. Error`: The standard error associated with the estimate in the first column
- `t value`: The *t*-test statistic for the hypothesis that the corresponding parameter is zero. Computed as the estimate divided by the standard error.
- `Pr(>|t|)`: The *p*-value associated with the hypothesis just mentioned.

In the `typepsne` line, for example, you find the estimate of the difference between the `psor` and `healthy` group, the correponding standard error, as well as information about the test for the hypothesis that this difference is zero, *i.e.* that the expected value is the same for `psne` observations and `healthy` observations.

At first glance, the parameterization may seem may seem annoying, but there is a point: Differences between groups means — not the groups means themselves — are the primary matters of interest in a oneway ANOVA, and the above version of the model gives directly the interesting quantities regarding comparison to the reference group.

Below the “Coefficients” part you find, among others, the “Residual standard error”, *i.e.* the estimated standard deviation for the observations (not the parametere estimates).

4. *(Parameterization in terms of group means)* Try

   ```
   oneway2 <- lm(intensity ~ type -1, data=psoriasis)
   summary(oneway2)
   ```

   and find the three group means as the estimates.

   Notice that the residual standard error in `oneway1` and `oneway2` are the same. This is because they fit the same model! We say that we have different versions, or parameterizations, of the model. The one fit is not more or less correct than the other, but they are useful for different purposes: `oneway1` gives you all the useful information about comparison to a reference group, whereas `oneway2` gives you all the useful information about the group means themselves.

5. *(Interpretation of output)* Let us make sure that we understand the output correctly:

   - Is there a significant difference between the `healthy` group and the `psne` group?

- Is there a significant difference between the healthy group and the psor group?
- Is the *p*-value for the comparison between the psne and psor groups available in the current output?
- Which hypothesis is tested in the typepsor line in oneway2? Is this a relevant hypothesis?
- Why are the standard errors for psne (and psor) not the same in oneway1 and oneway2?
- Why are the standard errors for healthy and psne not the same in oneway2 even though they both concern group means?

6. *(Test for homogeneity between groups)* It is standard to carry out an F-test for the overall effect of the explanatory variable. To be precise, the hypothesis is that the expected values are the same in all groups. One way to carry out the test is to fit two models — the model with as well as the model without the variable in question — and compare them with the anova function:

   ```
   noEffect <- lm(intensity ~ 1, data=psoriasis)
   anova(noEffect, oneway1)
   ```

   Notice the way the model with no type effect is fitted: with 1 on the right-hand side of the ~, meaning that there are no explanatory variables in the model, such that all observations are assumed to have the same distribution (this is true under the hypothesis).

   Is there a significant difference between the three type of skin samples?

7. *(Model validation)* Try the commands

   ```
   par(mfrow=c(2,2))
   plot(oneway1)
   ```

   Does the model seem to be appropriate for the data? Check also the stripchart and the boxplot from question 1 again, and discuss how they can be used to assess the validity of (some of) the assumptions behind the analysis.

8. *(Change of reference group)* As default R sorts the groups in alphabetical order and chooses the first one as the reference group. In our case this happened to be the healthy group, but this was a coincidence. Luckily, it is easy to change reference group. Try the following commands and explain what you see:

   ```
   type
   newType <- relevel(type, ref="psor")
   newType
   newType
   oneway3 <- lm(intensity ~ newType, data=psoriasis)
   summary(oneway3)
   ```

9. *(Categorical variables coded with numeric values)* Finally, we are going to consider the variable typeNum. Try the following commands:

```
typeNum
table(type, typeNum)
reg <- lm(intensity ~ typeNum, data=psoriasis)
summary(reg)
```

Notice how typeNum gives the same group structure as type, but with numbers instead of letter names.

Which model is fitted here in reg? Why does the model not make sense in the current set-up? *Hint:* Look at the name of the model.

So what should you do if a categorical variable is coded with numeric values? Try the following commands:

```
newVariable <- factor(typeNum)
newVariable
newModel <- lm(intensity ~ newVariable -1, data=psoriasis)
summary(newModel)
```

Explain the difference between the variables typeNum and newVariable. Compare the estimates from newModel and oneway2.

6.2 Oneway ANOVA: Pillbugs*

Prerequisites: Exercises 3.1, 3.3 (reading and working with datasets), and 6.1 (ANOVA)

An experiment on the effect of different stimuli was carried out with 60 pillbugs. The bugs were split into three groups: 20 bugs were exposed to strong light, 20 bugs were exposed to moisture, and 20 bugs were used as controls. For each bug it was registered how many seconds it used to move six inches. The data are saved in the files pillbugs.xlsx and pillbugs.csv with variables time and group.

1. Make stripcharts and/or parallell boxplots where you use time as response. Do the same where you use log(time) as response. Explain why it is more reasonable to use log(time) than time as the response in a oneway ANOVA.
2. Fit a oneway ANOVA model with log(time) as response, and carry out model validation. Is the model appropriate?
3. Is the expected value for log-time the same for all three treatment groups?
4. What is the estimated expected log-time for the control group? For the group with light exposure? For the group with moisture exposure?
5. What is the estimated difference in expected log-time between the group with light exposure and the control group? What is the interpretation of the exponential of this value?
6. Does the light exposure have a significant effect on log-time? How about the moisture exposure?

7. Finally, fit also the oneway ANOVA with `time` (not `log(time)`) as response, and carry out model validation. What do you see?

6.3 Twoway ANOVA: Growth of soybean plants

Prerequisites: Exercises 3.1, 3.3 (reading and working with datasets), and 6.1 (ANOVA)

An experiment with 52 soybean plants was carried out in order to examine the effect of light and stress on plant growth. There were two different levels of light exposure (`low` and `moderate`), and two different levels of stres (`no` or `yes`, where `yes` means that the plant has been shaken daily for 20 minutes). The 52 plants were divided into four groups corresponsing to the combinations of the light and stress treatments. After a periode the leaf areas was measured for each plant. The data is saved in the files `soybean.xlsx` and `soybean.csv`.

1. It is natural to start with a twoway ANOVA with interaction between stress and light. This model can be fitted in several different ways, for example:

   ```
   twowayWithInt <- lm(leafarea ~ stress * light, data=soybean)
   ```

   Fit the above model, and carry out model validation.

2. Make a summary of the model, and make sure you understand the estimates. In particular: Find, for each of the four stimuli combinations, the expected value of leafarea.

   Hint: Notice how reference levels are selected for each of the two factors (`low` and `no`, respectively), such that the intercept is to be interpreted as the expected value for this combination of stimuli. This estimate should be "corrected" for the other stimuli groups.

3. Make an interaction plot as follows:

   ```
   interaction.plot(stress, light, leafarea)
   ```

   Make sure you understand what has been plotted. Does the graph indicate an interaction effect between light and stress stimuli or not?

   A test of the hypothesis that there is no interaction is carried out by fitting the model without interaction (with main effects only), and comparing the two models with `anova`:

   ```
   twowayWithoutInt <- lm(leafarea ~ stress + light, data=soybean)
   anova(twowayWithoutInt, twowayWithInt)
   ```

   What is your conclusion regarding interaction? Make a summary of the model without interaction, and make sure you understand the estimates.

4. Is there a significant difference between the two light exposure levels? Is there a significant effect of stress?

   Hint: For the light exposure, say, make a model with `stress` as the only explanatory variable, and compare to the model without interaction. Alternatively, look at the summary from the model without interaction.

6.4 An analysis with categorical as well as quantitative variables: FEV

Prerequisites: Exercises 3.1, 3.3 (reading and working with datasets), 5.1 (linear regression), and 6.1 (ANOVA)

In Chapter 5 we considered models with only quantitative explanatory variables, whereas in this chapter we have so far considered categorical explanatory variables only. In this exercise we would like to use both types.

The primary objective of the analysis is to examine if children exposed to smoking have lower respiratory function than children who are not exposed, but we will also account for other variables that may influence respiratory function. The dataset contains information on more than 600 children. The measured outcome of interest is forced expiratory volume (FEV), which is, essentially, the amount of air an individual can exhale in the first second of a forceful breath. The data is saved in the files `fev.xlsx` and `fev.csv` and include the following variables: `FEV` (liters), `Age` (years), `Ht` (height, measured in inches), `Gender`, and `Smoke` (exposure to smoking, 0 = no, 1 = yes).

1. Read the data into R, and make a scatterplot matrix of the data: If you have called the dataset `fevdata`, then use the command `plot(fevdata)`.

2. Which of the variables in the dataset are quantitative and which variables are categorical? Make factor-type versions of the categorical variables, *i.e.* define

   ```
   GenFac <- factor(Gender)
   SmokeFac <- factor(Smoke)
   ```

3. Fit a model as follows:

   ```
   fevdata <- transform(fevdata, HtSqr=Ht^2)
   fevModel0 <- lm(FEV ~ Age + Ht + HtSqr + GenFac + SmokeFac +
                   SmokeFac*GenFac + GenFac*Age, data=fevdata)
   ```

   Make sure that you understand all the terms in the model (including the interactions and the term `HtSqr`).

4. Is the model appropriate for the data, or is some transformation of the response needed?

5. Simplify the model (possibly after transformation) as much as possible, *i.e.*, test the significance of the terms in the model and remove non-significant terms. Remember that you use should only remove one term (variable or interaction) from the model at a time.

6. What is your conclusion regarding smoking: Does smoking status influence respiratory function? If yes, how much? Discuss also he effect of the other variables.