Multivariate Statistics: Exercise 1

Eduard Szöcs (szoecs@uni-landau.de)
If you have problems/questions contact me via email.

January 20, 2016

1 Part I - Data exploration and linear model.

1) Find, download and read the publication

Zuur, A.F, E.N Ieno, and C.S Elphick. 2010. "A Protocol for Data Exploration to Avoid Common Statistical Problems." Methods in Ecology and Evolution 1 (1): 3–14.

2) Answer the following questions:

q1_1 Outliers should always be removed from the dataset.

```
TRUE or FALSE
```

q1_2 Homogeneity of variance can be checked using residuals.

```
TRUE or FALSE
```

q1_3 Collinearity is the existence of correlation between covariates.

```
TRUE or FALSE
```

q1_4 Figure 1 below shows fitted vs. residuals of a model. Are there any problematic patterns?

```
TRUE or FALSE
```

q1_5 Figure 2 summarises the residuals of a model. Do you see any problematic patterns?

```
TRUE or FALSE
```

2 Part II - Applied linear model.

The red squirrel (*Sciurus vulgaris*) is an endangered species in Scotland. To setup a conservation plan, it is important to know the effect of forest composition on red squirrel abundance.
In this exercise we will use data of

Flaherty, S., Patenaude, G., Close, A., Lurz, P. W. W. (2012). The impact of forest stand structure on red squirrel habitat use. Forestry, 85(3), 437–444. doi:10.1093/forestry/cps042

Some informations about the study: 52 forest plots were setup and the number of trees, tree height, diameter of trees at breast height (DBH) and canopy cover in these plots was recorded. Feeding remains of cones were counted and serve as an index for red squirrel abundance. One research question was:

What habitat variables (no. trees, diameter and height of trees, canopy cover) influence squirrel abundance (an measured by the number of stripped cones observed)?

You can find the data in the *data* folder of this exercise (`RedSquirrels.txt`).

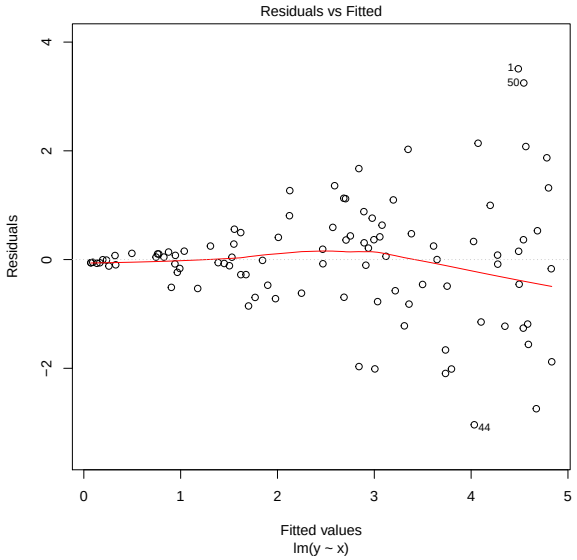

Figure 1: Residuals vs. fitted values of a model

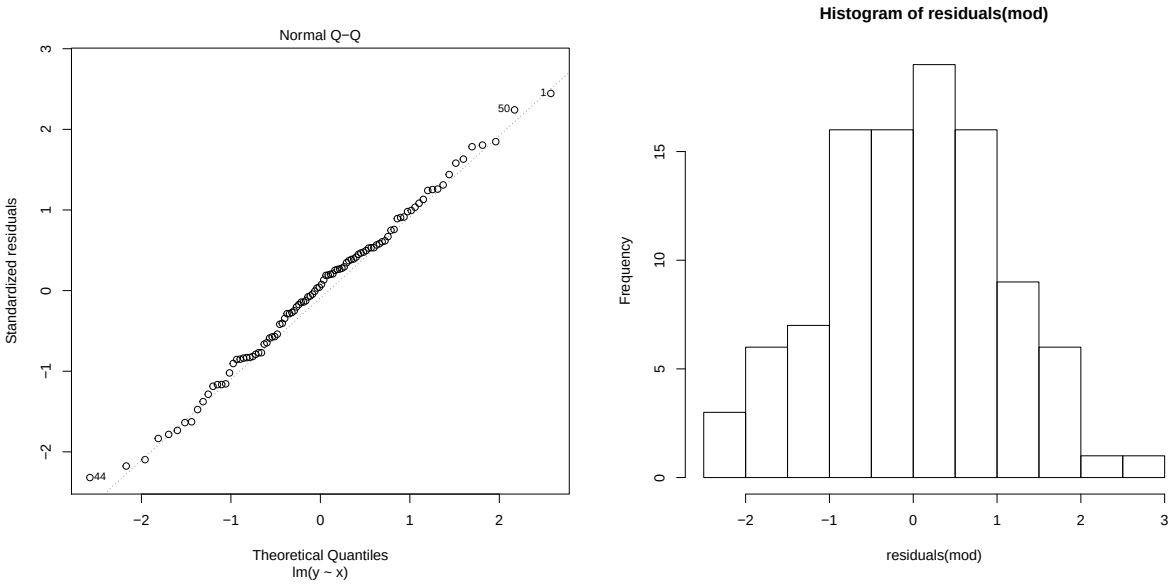

Figure 2: QQ-Plot and histogram of residuals of a model

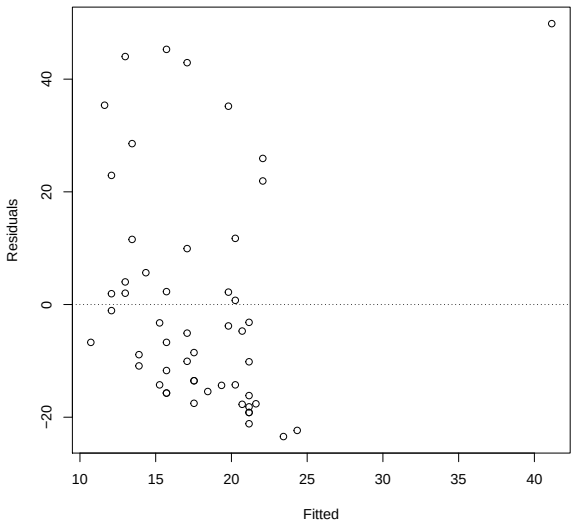

Figure 3: Residual vs. fitted values of the red squirrel model.

q2_1 Read the file `RedSquirrels.txt` into R.

You will need additional arguments. Have a look at the raw data file using a text editor and set them accordingly. Your table should look like:

```
## 'data.frame': 52 obs. of  6 variables:
##  $ Id         : Factor w/ 52 levels "Abern1","Abern10",..: 1 12 23 27 28 29 30 31 32 2 ...
##  $ SqCones    : int  61 4 15 9 42 4 12 27 0 4 ...
##  $ Ntrees     : int  32 4 34 22 22 21 19 15 12 9 ...
##  $ DBH        : num  0.23 0.27 0.17 0.23 0.18 0.23 0.22 0.26 0.23 0.12 ...
##  $ TreeHeight : num  20.4 15.2 16 22.4 19.4 ...
##  $ CanopyCover: num  91.3 61.5 91.4 92 93.2 93.5 88.5 88 89.8 73.3 ...
```

q2_2 Inspect the two variables `SqCones` (the number of stripped Cones) and `DBH` (diameter of trees at breast height). Use the methods described in Zuur (2010). Are there any outliers?

```
TRUE or FALSE
```

q2_3 Build a linear regression model, explaining `SqCones` with `DBH`.

q2_4 Shows the model normal distributed residuals?.

```
TRUE or FALSE
```

q2_5 Figure 3 shows a plot of residuals vs. fitted values of the fitted model. Is the assumption of variance homogeneity met?

```
TRUE or FALSE
```

q2_6 Are there any influential points?

```
TRUE or FALSE
```

q2_7 Is this a 'good' model (good = model fits the data)?

```
TRUE or FALSE
```

q2_8 Rerun the analysis, but omit the influential point! Does this results into a 'good' model?

```
TRUE or FALSE
```

3 Part III - Reading data into R!

You find four different data-files in the *Files* folder (`iris1.csv`, `iris2.csv`, ...). These contain all the same data, but stored in different formats and with some quirks.

1) Read these files into R using the `read.table()` function and store them in objects named according to the filenames.
All of these need some additional arguments for the function `read.table()`, to be read in correctly. The whole exercise should be answered using the `read.table()` function. Do not use wrappers like `read.csv()` or `read.delim()` - `read.table()` is a universal tool. Consult the help (`?read.table`) for possible arguments, their function and usage.

Look at the text-files in an **editor** to check what arguments are needed (column separator, decimal separator, NA-values etc).
Do not make any changes to the raw files! These would be not be reproducible by others. It's a good practice in science to keep raw data files unchanged.

Use `str()` to check whether the tables are read correctly. Especially check
(i) the number of columns,
(ii) the number of rows,
(iii) the data types of columns,
(iv) NA values.

The output of `str()` should look like:

```
## 'data.frame':  150 obs. of  5 variables:
##  $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
##  $ Sepal.Width : num  3.5 3 3.2 NA 3.6 3.9 3.4 3.4 2.9 3.1 ...
##  $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
##  $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
##  $ Species     : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1 1 1 1 ..
```