Practical Exercises for **Day 6**

Sonja Hartnack, Terence Odoch & Muriel Buri

October 2017

Exercise 14

- Load the below data set and for further information check the command `?water`.

```
# install.packages("HSAUR3")
library("HSAUR3")
data("water")
str(water)
head(water)
summary(water)
```

- Try to plot the variables `mortality` against `hardness` from the `water` data set.
- Add a main title to the above plot (`mortality` against `hardness`).
- Change the …

  1. font size of the axis annotation
  2. font size of the x- and y-axis labels
  3. the point sizes within the plot

  … of the above plot (`mortality` against `hardness`).

- Looking at the above plot: Do you think the two variables `hardness` and `mortality` correlate? What function do you use to find out the correlation coefficient? Do they have a positive or a negative correlation coefficient? How do you interpret the correlation coefficient in your own words?
- In the `water` data set, can you graphically find out if there is a difference between the the two variables `hardness` and `mortality` conditional on the `location` (`North`, `South`).
- Add a legend to the above plot so that you can easily differentiate the locations (`North` or `South`) of the observations.

- Do a barplot of the variable location from the water data set.

- ADDITIONAL: Try if any of these following plotting functions can be applied to the data sets perulung or ToothGrowth.

```
install.packages("graphics")
library("graphics")
?coplot
#
# install.packages("lattice")
library("lattice")
?xyplot
#
?interaction.plot
```

Exercise 15

- Download the .R file ANOVA_with_chickwts.R from the switch drive and have another look on how we applied the anova to the chickwts data set.

- Load the ToothGrowth data set into R and encode the numeric variable dose as a factor variable. Define the new factor variable as dose.fac with the three levels low, med and high and add it to the data frame of ToothGrowth.

- Visualize the variable len per dose level in a boxplot.

- With the help of the R-commands written in the ANOVA_with_chickwts.R file, apply a analysis of variance (ANOVA) to the data set ToothGrowth

Exercise 16

- Download the .R file LM_with_water.R from the switch drive and have another look on how we applied the linear model to the water data set.

- Reuse these commands to fit a simple as well as multiple linear regression model to the data set of perulung_ems. Use fev1 as your response variable y.

Exercise 17

- Load the `ToothGrowth` data set and run the following four linear regression models.

```
mod1 <- lm(len ~ dose.fac, data = ToothGrowth)
mod2 <- lm(len ~ supp, data = ToothGrowth)
mod3 <- lm(len ~ dose.fac + supp, data = ToothGrowth)
```

- Have a look at the summary of these models.
- How do you interpret the model coefficients?
- Which model is best?

Exercise 18

- Load the `water` data set and fit a multiple linear regression model. Use `mortality` as your response variable and add `hardness` and `location` as an explanatory variable.
- Check the underlying model assumptions.
- Add an interaction term between `hardness` and `location` to the above estimated multiple linear regression model.
- Interpret the interaction coefficient `hardness:locationSouth`.
- Check the underlying model assumptions.
- Which one is the better model? With or without the interaction term?
- How to derive confidence intervals for the regression coefficient of `hardness` and `location`?

Exercise 19

Hypothetical example - from Kirkwood and Sterne, Medical Statistics, 2nd ed., p. 177

- Read in the data set `lepto`. This study presents a serology survey of leptospira sero-prevalence in rural and urban areas of the west indies.
- Encode the numeric variable `antibodies` as a factor with levels 0 and 1.
- Make a crosstable with the risk factor `exposure` and `antibodies`.

- Run a Chi-squared test, a Fisher's exact t-test and a logistic regression (`glm`) to assess if the `exposure` (living in rural vs. urban areas) is a risk factor.
- Create a subset for `male` and `female` based on the variable `gender`.
- Repeat the crosstable, Chi-squared test, Fisher's exact t-test and a logistic regression (`glm`) for the subsets **separately**.
- Does the conclusion of your research question change with the analysis of the subsets? (Research question: Is the `exposure` (rural and urban areas) a risk factor?)
- Fit a logistic regression model (`glm`) with `exposure` and `gender` as explanatory variables.
- **SPECIAL FOR GUMA**: Is `exposure` being from an urban area a risk factor?