Practical Exercises for **Day 6**

Sonja Hartnack, Terence Odoch & Muriel Buri

October 2017

Exercise 17

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

```
data(ToothGrowth)
ToothGrowth$dose.fac <- factor(ToothGrowth$dose, levels = c(0.5, 1.0, 2.0),
                               labels = c("low", "med", "high"))
```

```
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.

```
mod1 <- lm(len ~ dose.fac, data = ToothGrowth)
mod2 <- lm(len ~ supp, data = ToothGrowth)
mod3 <- lm(len ~ dose.fac + supp, data = ToothGrowth)
# mod4 <- lm(len ~ dose.fac + supp + dose.fac:supp, data = ToothGrowth)
mod4 <- lm(len ~ dose.fac*supp, data = ToothGrowth)
summary(mod1)
summary(mod2)
summary(mod3)
summary(mod4)
# Check model assumptions
par(mfrow=c(2, 2))
plot(mod1)
plot(mod2)
plot(mod3)
plot(mod4)
```

- How do you interpret the model coefficients?
- Which model is best?

```
AIC(mod1, mod2, mod3, mod4)
# t.test(ToothGrowth$len ~ ToothGrowth$supp) # not significant
# mod4 is the best model, because it has the smallest AIC.
# THE SMALLER THE AIC, THE BETTER THE MODEL!
```

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.

```
library("HSAUR3")
data("water")
str(water)
head(water)
lm.mod <- lm(mortality ~ hardness + location,
             data = water)
summary(lm.mod)
```

- Check the underlying model assumptions.

```
par(mfrow=c(2,2))
plot(lm.mod)
```

- Add an interaction term between `hardness` and `location` to the above estimated multiple linear regression model.

```
lm.mod.int <- lm(mortality ~ hardness + location + hardness:location,
             data = water)
summary(lm.mod.int)

mod1 <- lm(mortality ~ hardness + location,
             data = water)
mod2 <- lm(mortality ~ hardness + location + hardness:location,
             data = water)
AIC(mod1, mod2)
```

- Interpret the interaction coefficient `hardness:locationSouth`.
- Check the underlying model assumptions.

```
par(mfrow=c(2,2))
plot(lm.mod.int)
```

- Which one is the better model? With or without the interaction term?

```
AIC(lm.mod, lm.mod.int)
summary(lm.mod)
summary(lm.mod.int)
```

- How to derive confidence intervals for the regression coefficient of `hardness` and `location`?

```
confint(lm.mod)
```

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.

```
lepto <- read.csv("~/Dropbox/201710_Makerere/03_Exercises/data/lepto.csv", sep = ";")
# SONJA
# lepto <- read.csv("C:\\Users\\admin\\Dropbox\\201710_Makerere\\03_Exercises\\data\\lepto.cs
# sep = ";")
head(lepto)
str(lepto)
```

- Encode the numeric variable `antibodies` as a factor with levels 0 and 1.

```
table(lepto$antibodies)
class(lepto$antibodies)
lepto$antibodies <- factor(lepto$antibodies, level = c(0, 1),
                           labels = c("no", "yes"))
# Many different ways how to encode a numeric variable into a factor:
# lepto$antibodies <- factor(lepto$antibodies,
#                             levels = c(0, 1),
#                             labels = c("no", "yes"))
# lepto$antibodies <- factor(lepto$antibodies,
#                             levels = c(0, 1),
#                             labels = c("NO", "YES"))
table(lepto$antibodies)
class(lepto$antibodies)
```

- Make a crosstable with the risk factor `exposure` and `antibodies`.

```
table(lepto$exposure, lepto$antibodies)
```

- Run a Chi-squared test, a Fisher's exact test and a logistic regression (`glm`) to assess if the `exposure` (living in rural vs. urban areas) is a risk factor.

```
chisq.test(lepto$exposure, lepto$antibodies)
fisher.test(lepto$exposure, lepto$antibodies)
# fisher.test(table(lepto$exposure, lepto$antibodies))
glm.mod <- glm(antibodies ~ exposure, data = lepto,
               family = "binomial")
summary(glm.mod)
confint(glm.mod)
```

- Create a subset for `male` and `female` based on the variable `gender`.

```
lepto.fem <- subset(lepto, gender == "female")
lepto.male <- subset(lepto, gender == "male")
```

- Repeat the crosstable, Chi-squared test, Fisher's exact test and a logistic regression (`glm`) for the subsets **separately**.

```r
# FEMALES
table(lepto.fem$exposure, lepto.fem$antibodies)
chisq.test(lepto.fem$exposure, lepto.fem$antibodies)
fisher.test(lepto.fem$exposure, lepto.fem$antibodies)
glm.mod.fem <- glm(antibodies ~ exposure, data = lepto.fem,
               family = "binomial")
summary(glm.mod.fem)
confint(glm.mod.fem)
# MALES
table(lepto.male$exposure, lepto.male$antibodies)
chisq.test(lepto.male$exposure, lepto.male$antibodies)
fisher.test(lepto.male$exposure, lepto.male$antibodies)
glm.mod.male <- glm(antibodies ~ exposure, data = lepto.male,
               family = "binomial")
summary(glm.mod.male)
confint(glm.mod.male)
```

- 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.

```r
glm.mod.final <- glm(antibodies ~ exposure + gender, data = lepto,
               family = "binomial")
summary(glm.mod.final)
# Check that exposure and gender is also associated.
glm.exp.gen <- glm(exposure ~  gender, data = lepto,
               family = "binomial")
summary(glm.exp.gen)
```

- **SPECIAL FOR GUMA**: Is `exposure` being from an urban area a risk factor?