Statistics for Data Analysis: R Exercises 10

1. **Birthweight data.**

   This is the full data set containing birthweight data for both boys and girls. You have analysed the girls data in the previous practical.

   ```
   > birth <- read.csv("birthweight.csv",header=TRUE)
   > birth
   > age <- birth$week
   > weight <- birth$weight
   > sex <-birth$sex
   >
   > # plot
   > plot(age[sex=="M"],weight[sex=="M"],xlab="age",ylab="weight",
   +      pch="M",col=3,ylim=c(2300,3500))
   > points(age[sex=="F"],weight[sex=="F"],pch="F",col=1)
   ```

   We have a factor `sex` in addition to the regression variable `age` that we had for the `girls` data.

   - We can fit the model we have seen in the lecture with the command.

     ```
     > birth1.lm <- lm(weight˜age+sex)
     ```

     Find the estimates of the parameters. How can you interpret them?

   - We may want to allow also a different rate of increase (slope) between boys and girls. This can be done by including an additional predictor (called *interaction*) whih is given by the product between the age and the indicator variable:

     ```
     > birth2.lm <- lm(weight˜age*sex)
     > # or, equivalently
     > birth2.lm <- lm(weight˜age+sex+age:sex)
     ```

     Does this suggest that the boys' and girls' rate of increase of weight with age are the same? What do you conclude? Check the assumption of the model with the appropriate diagnostic plots.

2. **Savings data**
   We consider now an economic dataset including 50 different countries. This dataset is available as example within R.

   ```
   > data("LifeCycleSavings")
   > head(LifeCycleSavings)
   > plot(LifeCycleSavings)
   ```

   The data are averages between 1960 and 1970 of three economic indices: `dpi` is the per capita disposable income in US dollars, `ddpi` is the percentage rate of change in the per capita disposable income and `sr` is the aggregate personal saving divided by disposable income. The data include also the proportions of the population under 15 and over 75, `pop15` and `pop75` respectively. We want to investigate the relationship between savings and the other variables in the dataset. The description of this dataset that you can find in the help states that

> ”Under the life-cycle savings hypothesis as developed by Franco Modigliani, the savings ratio (aggregate personal saving divided by disposable income) is explained by per-capita disposable income, the percentage rate of change in per-capita disposable income, and two demographic variables: the percentage of population less than 15 years old and the percentage of the population over 75 years old. The data are averaged over the decade 1960–1970 to remove the business cycle or other short-term fluctuations.”

We want to fit a model to see if this dataset supports this theory.

```
> attach(LifeCycleSavings)
> lm1<-lm(sr~pop15+pop75+ddpi+dpi)
> summary(lm1)
> plot(lm1$fitted.values,lm1$residuals)
```

(a) Write down the mathematical expression of the model. Based on the output of the `summary` command and on the residuals plot, is the model a good fit for the data?

(b) The following instruction carry out the F-test for the sequence of nested models (including one predictor at the time):

```
> anova(lm1)
```

Which model should be preferred? Note that the answer depends on the order in which the predictors are added into the model. Fit the chosen model and write down the estimates for the parameters.

```
> lm_R<-lm(sr~pop15+pop75+ddpi)
> summary(lm_R)
```

(c) We can also test if two (or more) variables can be excluded together from the model, using the F-test of their residual sums of squares. To implement this in R, we need first to fit the two models (with and without the variables in question) and then use the `anova` command. For example:

```
> lm2<-lm(sr~pop15+ddpi)
> anova(lm2,lm1)
```

Is the addition of the percentage of population over 75 and of the growth in the income needed in the model?

(d) What can we conclude about the economic theory?

3. **A two-way ANOVA example**

In this exercise we consider the so-called two-way ANOVA, i.e. the comparison between the mean of different groups when the groups are identified by the combination of two factors. We analyse the same dataset used for one of the question of the tutorial, where three treatments (control, A and B) are applied to the plants in two different fields 1 and 2. The file `"plants.csv"` contains the mass of each plant (`mass`), the treatment (`treatment`) and the field where the plant grew up (`field`).

```
> data<-read.csv("plants.csv",header=TRUE)
> head(data)
> plot(data)
> attach(data)
```

We can fit the model easily by specifying to `R` to treat the factors as such:

```
> treatment<-as.factor(treatment)
> field<-as.factor(field)
```

The linear model can now be fitted with the same command we used for quantitative variables:

```
> ANOVA_TWO_WAY<-lm(mass~treatment*field)
> summary(ANOVA_TWO_WAY)
> plot(ANOVA_TWO_WAY)
```

alternatively, we can define a model through indicator variables:

```
> x1<-as.double(treatment=="A")
> x2<-as.double(treatment=="B")
> x3<-as.double(field==2)
>
> # model
> model_dummy<-lm(mass~x1*x3+x2*x3)
> summary(model_dummy)
```

- Even though the two models are equivalent, you should obtain different numerical values with respect to the `ANOVA_TWO_WAY` model, can you explain why?

- Define a new set of indicator variables that allows to fit exactly the same model as in `ANOVA_TWO_WAY`.

```
> x1<-as.double(treatment=="control")
> x2<-as.double(treatment=="B")
> x3<-as.double(field==2)
>
```

```
> # model
> model_dummy2<-lm(mass~x1*x3+x2*x3)
> summary(model_dummy2)
```