MAS programmes - Statistical Analysis (Autumn Term)

Exercises 5

The data for question 1 is available as a `.csv` file `water1.csv` on Moodle, and can be entered into R as the data frame `water`.

1. Consider again the water mortality example of Lecture 5. Split the data frame `water` into two, according to the levels of the factor `north`, to obtain separate data sets, say `water.n` for towns in the North and `water.s` for towns in the South. Fit separate regression models for `mortality` on `calcium` to each of these reduced data sets, in order to obtain two separate regression models for towns in the North and South. How do the parameter estimates compare with those found during the lecture. What advantages/disadvantages are there between the two approaches?

You may use, for example, the following code:

```
> water.s<-water[north==0,]
> water.n<-water[north==1,]
```

2. Recall the two-sample t-test of Exercises 1 (question 2(c)), concerning the weights of female sparrows recorded by Bumpus. By forming a single vector for `weight` and a dummy variable to indicate the category `survived` or `died`, carry out a simple linear regression of `weight` on the (single) dummy variable. How does your output from this regression compare with your previous findings?

3. Recall the paired t-test of Exercises 1 (question 1), concerning the percentage of solids in grapefruit halves which have been exposed to sunlight or kept in the shade. By forming a single vector of the response `solids` and suitable *factors* for `exposure` (with categories `shaded` or `exposed`) and for `fruit` (`1, ..., 25`), fit a suitable regression model which recovers the t-test previously found.

If you saved your script files for these exercises, you will be able to make use of the commands you used previously to enter the data into R.