3 Regression and ANOVA (AJ Chapter 10; Faraway)

The (*) questions from sheets 1 to 4 form homework 1.

1. This is an introduction to some features of R that you will need in the other questions. Download lab3.R from the website and load it in RStudio.

   (a) Working with tables.

   (b) QQ-plots

   (c) Regression towards the mean. The term regression comes from the phenomenon of "regression towards the mean". If you carry out a pair of test where the observation are related, the correlation coefficient is normally less than one.
   For example, if students take two tests, some students with do particularly well/badly on the first test, so there second exam result will be closer to the average grade than you might expect.
   Another example: tall fathers tend to have (in an average sense) tall sons; height has a genetic component. However, the sons are not as tall as you might expect them to be just based on their fathers' heights. Other factors having a role to play, including the height of the mother, environmental factors and biological randomness.

      i. Suppose students take two exams, and we want to model the second set of exam results in terms of the first set.

      ii. Suppose students take four exams, and we want to model the fourth set of exam results in terms of the first three.

   (d) The `anscombe` dataset (run `?anscombe` in R). The four graphs are all fit by the same linear model. Is regression an appropriate tool for analyzing the four datasets?

2. (*) Simpson's paradox:

   (a) Load the admission statistics for UCB: `data(UCBAdmissions)`. This a "famous" dataset. It count the number of applications to 6 different UC Berkeley departments in 1973.

   (b) First calculate some summary statistics:

      i. the total number of people accepted/rejected,

      ii. the number of men and women in the dataset,

      iii. and the number of applications per department.

   (c) Produce a 2x2 table showing the numbers of men/women whose applications where accepted/rejected. Do the proportions seem to differ? (Do a χ^2-test.)

(d) Repeat part (c) 6 times for each of the 6 departments. Which of the individual departments display the same pattern of behavior seen in part (c). Can you explain this?

(e) Produce a 2x6 table showing, separated by gender, the fraction of people applying to the six departments.

(f) Produce a 2x6 table showing the fraction of acceptances for each of the six departments.

3. (*) Load the faithful dataset (?`faithful`). Try to model the waiting time against the length of the eruptions.

   (a) Consider the models:

   ```
   waiting~0
   waiting~eruptions
   waiting~eruptions+I(eruptions^2)
   ```

   How many degrees of freedom are there when fitting the different models? Do the resulting residuals seem to be normally distributed? With variance independent of the predicted value? Independent of the predicted values? Use the output of `lm`, `summary` and `anova` to compare the models and interpret the results.

   (b) Plotting the data reveals two quite distinct clusters. Add an extra variable to the dataframe (i.e. `faithful$extraVariable` = ... that is zero for the left cluster and one for the right cluster. Does the improvement in fit justify the added complexity of having an extra variable?

4. Load the `savings` dataset. Start off looking at the model

   ```
   lm(sr~pop15+pop75+dpi+ddpi,data=savings)
   ```

   Does it seem like a reasonable model. Do any of the variables play a similar role (i.e. are they meaningfully correlated)? Can any of the variables be removed from the model without significantly degrading the fit?

5. Load the `aatemp` dataset. Can you fit it to a linear model?