Univariate Data and Modelling – Exercises

Session 3 – Hypothesis Testing, Correlation, and Simple Linear Regression

Exercise 1

The annual rainfall (in centimeters) was measured in Iran for the last 10 years and for Belgium in the last 18 years:

Iran: 128,125,133,104,146,132,125,118,129 and 124
Belgium: 160,128,169,105,151,164,162,177,185,150,182,158,156,123,141,176,162 and 172

a) Which assumptions have to be checked before testing?
b) Give the hypotheses to test whether both countries have the same annual rainfall;
c) Give a 90% confidence interval for the difference in the means;
d) Test this hypothesis with 90% confidence and compare the result with c);
e) Give the hypothesis and test if there is significant less rainfall in Iran than in Belgium;

Exercise 2

Import the BLOOD.DAT dataset.

a) Make two subsets: the persons younger than 50 and older than 68 years;
b) Give the hypotheses to test if both age groups have a significant different testosterone level;
c) With which test would you test these hypotheses?
d) Test this hypotheses with 95% confidence.

Exercise 3

Install the “faraway” package and load the “tvdoctor” dataset. It gives the life expectancy, no. of doctors and no. of televisions in 38 countries collected in 1993.

a) Give the correlation coefficient between the life expectancy and the no. of televisions.
b) Give the correlation coefficient between the life expectancy and the no. of doctors.
c) Give the correlation coefficient between the no. of doctors and the no. of televisions.
d) What can you conclude from this? (Hint: look at the scatterplot)

Exercise 4

Load the SENIC.DAT dataset from Toledo. This is a historic study (1975 – 1976) on the reduction of hospital-acquired infections by means of infection surveillance and control programs. Each observation is the data of one US hospital. It has the following variables:

ID	Identification Number
length	Average length of stay of all patients
age	Average age of all patients
risk	Risk of acquiring infection during stay (percentage)
cult	Number of routine cultures performed on patients without symptoms (times 100)
xray	Number of routine X-rays performed on patients without symptoms (times 100)
beds	Number of beds in hospital
meds	Medical school affiliation - 1 = Yes; 2 = No
reg	Region - 1 = NE; 2 = NC; 3 = S; 4 = W
cen	Average number of patients in hospital per day
nur	Average number of nurses in hospital per day
fac	Available facilities and services at hospital (percent)

a) What can you tell about the following variables by means of descriptive statistics ? Do you expect any outlying values? Any expected correlations by looking at the scatterplots?
   - Average length of stay;
   - Infection risk;
   - Available facilities;
   - Routine X-rays;

b) Calculate the appropriate correlation coefficient between the infection risk and the other three variables. What do you expect regarding linear regression?

c) Regress the infection risk on the average length of stay;
   - Give the appropriate hypotheses to test by means of an F-test;
   - Look at the F-test;
   - Interpret the estimated value for β_1;
   - Interpret the R^2 value;
   - Plot the data points together with the regression line. Do you think that linear regression was appropriate?

d) Regress the infection risk on available facilities and do the same exercises as in c);

e) Regress the infection risk on routine X-rays and do the same exercises as in c).

f) How would you try to reduce the infection risk given the information you collected above?