MATH20802 Computer Lab 2

Term Week 10

The topic of this computer lab is linear regression. In this session you will learn how to *exactly* reproduce the numerical output of the built-in R function lm() for fitting linear regression models.

Amitriptyline dataset

In this lab we use a dataset by Rudorfer (1982) used in Johnson and Wichern (2007, Ex. 7.25). You find the data in the file "amitriptyline.txt" on Blackboard.

Amitriptyline is prescribed by some physicians as an antidepressant. However, there are also conjectured side effects that seem to be related to the use of the drug: irregular heartbeat, abnormal blood pressures, and irregular waves on the electrocardiogram among other things.

The data were gathered on 17 patients who were admitted to the hospital after an amitriptyline overdose.

The dataset contains measurements of the following 7 variables:

- TOT: total TCAD plasma level.
- AMI: the amount of amitriptyline present in the TCAD plasma level.
- GEN: is gender (male = 0, female = 1)
- AMT: amount of drug taken at time of overdose
- PR: PR wave measurement
- DIAP: diastolic blood pressure
- QRS: QRS wave measurement

Tasks:

a. Load the file "amitriptyline.txt" in R using read.table()
b. Compute the summary statistics for all variables and plot pairwise scatter plots.
c. Fit a linear regression model using lm() with AMI as response and the variables GEN, AMT, PR, DIAP and QRS as predictors.
d. Use summary() to display the regression coefficients, t-scores etc.
e. Standardise the dataset so that each variable has mean 0 and variance 1, and fit another linear regression model. How does the output of the various coefficients change?
f. Estimate the means of the response (my) and predictor variables (mx), as well as the covariance among the predictors (Sxx) and the covariance between response and each predictor (Sxy).
g. Use these estimates to reproduce the output of summary(). In particular,
    - compute the regression coefficients,
    - the intercept,
    - the coefficient of determination R-squared, and
    - the residual standard error (consider a scaling factor if your results do not match the output by R!)
    - the error of the regression coefficients
    - the regression t-scores
    - the p-values
h. Which variable is least important and why? Fit another regression model without this variable and comment on the resulting output.