ST221: Assessed coursework 1

Linear Statistical Modelling

Deadline: 12 noon Thursday 12th March 2020

A printout of your solutions must be handed in to the support office by the deadline above. Your solutions should be produced using a word processor, Markdown, or LaTeX. Please remember to include only your **ID number** on your submission to allow anonymous marking. Please leave adequate time for printing and note that 'the printer broke' is not a valid reason for late submission. If you have any queries about the coursework please post them on the ST221 forum, but do not post any part of your solutions.

This assignment counts towards **15%** of your final module mark.

1. Download the file `courseworkData1.csv` from the module webpage and read it into R.

   The dataset consists of data on live births in a London hospital. The variables are:

   - `bweight` Birth weight of baby (in g).
   - `gestwks` Gestation period, *i.e.* the time between conception and birth (in weeks).
   - `hyp` Indicator for maternal hypertension (1:Present, 0:Absent).
   - `sex` Sex of baby (a factor with levels "male", "female").

   The aim of the study is to look at the effect of hypertension on birthweight. However, birth weight is also influenced by the other factors.

   (a) Calculate the mean birth weight among babies born to women with and without hypertension. [1]

   (b) Fit a linear model with birth weight as the outcome variable and maternal hypertension as the only explanatory variable. Write down the parameter estimates. Give an interpretation of the values of the two parameter estimates accessible to a non-statistician. [2]

   (c) Produce a scatter plot of birth weight against gestation period. Use different colours and/or plotting symbols to show babies born to mothers with and without hypertension. Your plot should be clearly labelled. [2]

   (d) Fit a second model with birth weight as outcome and hypertension, sex, and gestation period as explanatory variables. Report the effect of hypertension on birth weight in this model. Compare it with the result in part 1b. Why are the two estimates different? [2]

   (e) Give the first 5 rows of the design matrix for the model in part 1d. [1]

   (f) Give an unbiased estimate of the variance of the errors σ^2 in the model in part 1d and an estimate of the variance of estimator for the effect of hypertension. [2]

   (g) Give the expected birthweight of a female child born at 40 weeks to a mother with hypertension. [2]

   (h) A doctor suggests that hypertension might act on birthweight in two ways:

      - Directly, by reducing the birth weight of children born at the same number of gestational weeks.
      - Indirectly, through gestation period, by causing babies to be born earlier.

      Fit a model to investigate the hypothesis that maternal hypertension causes babies to be born earlier. Think carefully about which variables should and should not go in the model. Summarize your findings to the doctor. [3]

2. Download the file `courseworkdata2.csv` from the module webpage and read it into R.

   (a) Fit a simple linear regression model with `y1` as the response and `x` as the predictor. Make a plot of the data and add the fitted line. [1]

   (b) Produce a residual plot for the model in part 2a and comment on whether it is acceptable or not. If you think it is not acceptable describe which model assumptions are not appropriate. [3]

   (c) Suggest an improved model for the response `y1`, fit it to the data and produce a new residual plot. Comment on whether the residual plot for the improved model is acceptable. [3]

   (d) Fit a simple linear regression model with `y2` as the response and `x` as the predictor. Make a plot of the data, add the fitted line, produce a residual plot and comment on whether it is acceptable or not. If you think it is not appropriate describe which model assumptions are not appropriate. [3]

   (e) Suggest an improved model for the response `y2`, fit it to the data and produce a new residual plot. Comment on whether the residual plot for the improved model is acceptable. [3]

   (f) Assess whether the residuals from your model in part 2e appear to be normally distributed. [2]