ST221: Assessed coursework 2

Linear Statistical Modelling

Extended Deadline: 12 noon Friday 29 May 2020

Your solutions should be submitted electronically in the form of a PDF document using the submission portal on the ST221 Moodle page. Please remember to include only your **ID number** on your submission to allow anonymous marking.

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.

The maximum score for this coursework is 25/25 (15 marks for part 1; 10 marks for part 2).

1. Download the file `courseworkData3.csv` from the module web page and read it into R.

   An experiment was conducted to determine the yield produced by 4 varieties of wheat (labelled A, B, C, D) under different conditions. For each experimental unit, data were recorded on the average amount of rainfall per month (`rainfall`), measured in mm, and the soil type (`soil`), as well as the yield (`yield`) in tonnes per hectare.

   (a) Fit the model below and report the parameter estimates. For $i = 1, \ldots, n$,

   $$\texttt{yield}_i = \alpha_{\texttt{variety}_i} + \gamma\ \texttt{clay}_i + \delta\ (\texttt{rainfall}_i - \overline{\texttt{rainfall}}) + \epsilon_i,$$

   where α_k represents the intercept for wheat of variety $k \in \{A, B, C, D\}$; $\overline{\texttt{rainfall}}$ is the mean rainfall; and $\texttt{clay}_i$ is an indicator variable that takes the value one if the wheat is grown in clay soil and zero otherwise.

   For each parameter, report a 95% confidence interval along with the parameter estimate. [2]

   (b) Write down the interpretation of the parameter α_A in the model from part 1a. [1]

   (c) Refit the model with a global intercept:

   $$\texttt{yield}_i = \mu + \beta_{\texttt{variety}_i} + \gamma\ \texttt{clay}_i + \delta\ (\texttt{rainfall}_i - \overline{\texttt{rainfall}}) + \epsilon_i,$$

   What is the mathematical relationship between the parameters $\alpha_A, \alpha_B, \alpha_C, \alpha_D$ in the model in part 1a and the parameters $\mu, \beta_A, \beta_B, \beta_C, \beta_D$ in this model? [1]

   (d) Using the model in part 1c complete the following ANOVA table for the test for the existence of regression, then calculate the p-value and report your conclusion. [3]

Source	d.f.	SS	MS	F
Regression				
Residual				
Total				

   (e) Using the model from part 1c, test the hypothesis that the yield does not depend on the whether the wheat is grown in clay soil (State the hypotheses under consideration; calculate an appropriate test statistic; give the corresponding p-value and state your conclusion) . [2]

   (f) Using the model from part 1c, test the hypothesis that all of the varieties of wheat have the same effect on yield. [3]

(g) A colleague observes from the output of the model in part 1a that variety A has higher yield than variety C. After adjusting for rainfall and clay soil, assess the evidence that variety A produces a better yield than variety C. [3]

2. Download the file `courseworkData4.csv` from the module web page and read it into R.

   The spreadsheet describes the amount of profit made by a small shop on each day, alongside some weather data collected in the same village. The potential predictor variables include grass minimum temperature[1] (`GRSMIN`), minimum and maximum temperature (`MIN`, `MAX`), the relative humidity (`HUM`), rainfall in mm (`RAIN`) and barometric pressure (`BAR`).

   (a) Fit the full model, with `profit` as the response and all weather variables as predictors (plus an intercept). Report the parameter estimates and an unbiased estimate for the variance of the errors. [2]

   (b) Using the model fitted in part 2a, create a table ranking the weather variables in order of statistical significance, from most important (i.e. should not be removed from the full model) to least important (i.e. may be removed from the full model). Report the p-value for each variable in your table to 3 decimal places. (For p-values less than 0.001 we usually write "$p < 0.001$", but you may leave this as "p=0.000" if you wish). [1]

   (c) For each weather variable, fit a univariate model that uses only a single predictor (plus intercept) to model profit.

   Using the results of these univariate models, create another table ranking the variables from most important to least important, along with their p-values, as you did in part 2b. [1]

   (d) Consider the variables `GRSMIN` and `MIN`. Calculate the correlation between these variables.

   Explain the different p-values for these two variables in the univariate models in 2c and the multivariate model in 2b. [2]

   (e) Create a table ranking the univariate models from best-fitting to worst fitting, along with an appropriate statistic justifying your answer. [1]

   (f) Suppose that tomorrow we observe the weather data below.

GRSMIN	MIN	MAX	HUM	RAIN	BAR
10.5	11	16	84	30.1	987

   Calculate the expected profit from:

   i. the full model, and

   ii. the best univariate model in question 2e.

   Include also a 95% prediction interval for each model. [3]

[1] Grass minimum temperature can be defined as "the temperature recorded in open air ground on short turf, with the bulb of the thermometer just in contact with the tips of the blades of grass" (`https://www.weatheronline.co.uk/reports/wxfacts/Grass-Minimum-Temperature.htm`)