Assessed coursework 1

ST221 Linear Statistical Modelling

Deadline: 28 February 2023, 1 pm

Please read these instructions carefully!

This assignment counts for **10%** of your final module mark. The maximum score for this coursework is 25 marks.

Your solutions must be produced using a word processor, R Markdown, or LaTeX. You may cut-and-paste R-output. Handwritten answers will not be awarded marks. Use a font size of 11pt or larger. Question sub-sections must be clearly labelled for ease of marking.

If you do not submit your solutions in a typed format, then this will not be accepted as a submission. You should convert your solutions into **one PDF file** to be submitted on the ST221 moodle page.

Please read Chapter 5 in the course guide which gives details around the procedures regarding coursework including applying for extensions and lateness penalties. Please ensure that you submit in good time before the deadline. Penalties will apply if work is submitted more than 1 minute after the deadline unless an extension or waiver is granted. Coursework is not eligible for mitigating circumstances due to the loss of work in progress. The penalty for late submission is 5% per 24 hour period encompassing a working day. However no submission will be accepted more than 5 working days after the original deadline unless there is a pre-approved extension extending past the cut off period.

If you have any queries about the coursework, please post them on the ST221 forum, but do not post any part of your solutions. You can also submit questions to the anonymous question form on moodle.

Please be aware that your work will be submitted to TurnItIn, a piece of plagiarism-detection software. Cases of suspected collusion or plagiarism will be followed up as outlined in Section 5.3 of the course guide. Note that detailed discussions of the assignment or comparisons of numerical/graphical results or computer code are **not permitted.**

Make sure to read questions carefully. If asked to produce a plot, then please include the plot in your report. Make sure it is of appropriate scale and the axes are clearly labelled. Include R code only if requested to do so.

Good luck with the assignment!

Download the file `dia.csv` from moodle and load it into `R`. The data is an adapted subset from the diamonds dataset in the `ggplot2` package. The dataset consists of information on 300 diamonds.

The variables are:

- `price`: the price of the diamond in US dollars ($);
- `weight`: the weight of the diamond in carat;
- `width`: the width of the diamond in mm.

Question 1

In this question you will be exploring the relationship between the price of a diamond and its width.

(a) **[2 marks]** Fit a simple linear regression of price on width. Produce a scatterplot of price against width and add the fitted regression line.

(b) **[4 marks]** Produce a residual plot for the fitted model in (a) and explain why this is not an acceptable plot. Identify which of the model assumptions are not appropriate.

(c) **[4 marks]** Suggest an improved model and fit it to the data. **Produce a residual plot** and comment on whether the residual plot for the improved model is now acceptable, justifying your answer.

(d) **[2 marks]** Give a quantitative interpretation of the estimated slope for the model in (c) in relation to the predicted price of a diamond.

Question 2

In Computer Practical 1 we considered a diamond dataset discussed in a paper[1] by Dr Singfat Chu. The paper develops a linear model that predicts the log-price of a diamond from its weight (`carat`) and other factors (`cut`, `clarity` and `color`). However, the initial model underestimates prices at both ends of the price range while it overestimates the midrange prices. As a remedy the paper suggests introducing an additional categorical predictor variable that divides the diamonds into groups according to weight. It then fits a linear model that includes an interaction between the new categorical predictor and the quantitative variable `carat`. In the following you will explore this approach for the dataset provided.

Use the code below to implement a factor variable `size` with levels

- `small` for diamonds of less than 0.5 carat,
- `medium` for diamonds with a weight of at least 0.5 carat and less than 1 carat, and
- `large` for diamonds of at least 1 carat.

[1] Chu, Singfat (2001) "Pricing the C's of Diamond Stones", Journal of Statistics Education, 9(2).

```
dia$size <- ifelse(dia$weight<0.5, "small",
                   ifelse(dia$weight<1, "medium", "large"))
dia$size <- factor(dia$size, levels=c("small", "medium", "large"))
```

(a) [**2 marks**] Fit a regression model of log-price on weight and produce the corresponding residual plot. Identify and describe any features that make this an unacceptable residual plot.

(b) [**2 marks**] Consider the linear model defined by the `R` model formula

$$\log(\text{price}) \sim \text{weight} + \text{size} + \text{weight:size}.$$

Write out the algebraic model equations for this model.

(c) [**1 mark**] Fit the model in `R` and produce its residual plot.

(d) [**2 marks**] When fitting the model in `R`, the summary output gives an estimate for the coefficient `weight:sizelarge`. Give an interpretation of the parameter that corresponds to `weight:sizelarge`.

(e) [**1 mark**] As an alternative model consider using the cubic root of `weight` as the only predictor variable, that is the linear model with model equations

$$\log(\text{price}_i) = \beta_0 + \beta_1 \sqrt[3]{\text{weight}_i} + \epsilon_i$$

where $i = 1, \ldots, 300$. Fit the model in `R` and produce its residual plot.

(f) [**5 marks**] In 5-6 sentences compare and contrast the model in (b) with the model in (e), discussing the relative advantages and disadvantages of the two models.