MATH20811 Practical Statistics: Coursework 1 (November 2020)

The marks awarded for this coursework constitute 30% of the total assessment for the module.

Your solution to the coursework should be fairly concise (maximum of about 10 pages) and it should take, on average, about 15 hours to complete.

Please read all the instructions and advice given below carefully.

The submission deadline is 10:00 am on Monday 16 November 2020.

Late Submission of Work: Any student's work that is submitted after the given deadline will be classed as late, unless an extension has already been agreed via mitigating circumstances or a DASS extension.

The following rules for the application of penalties for late submission are quoted from the University guidance on late submission document, version 1.3 (dated July 2019):
"Any work submitted at any time within the first 24 hours following the published submission deadline will receive a penalty of 10% of the maximum amount of marks available. Any work submitted at any time between 24 hours and up to 48 hours late will receive a deduction of 20% of the marks available, and so on, at the rate of an additional 10% of available marks deducted per 24 hours, until the assignment is submitted or no marks remain."

Your submitted solutions should all be in one document. This must be prepared using LaTeX. For each part of the question you should provide explanations as to how you completed what is required, show your workings and also comment on computational results, where applicable.

When you include a plot, be sure to give it a title and label the axes correctly.

When you have written or used `R` code to answer any of the parts, then you should list this `R` code after the particular written answer to which it applies. This may be the `R` code for a function you have written and/or code you have used to produce numerical results, plots and tables. `R` code should also be clearly annotated.

Do not use screenshots of `R` code/output. Instead, to include `R` code use the *verbatim* environment and summarise `R` output in tables using the *table* environment, as demonstrated in the solution of Example Sheet 2.

Your file should be submitted through the module site on Blackboard to the Turnitin assessment in the Coursework folder entitled "MATH20811 CW1" by the above time and date. The work will be marked anonymously on Blackboard so please ensure that your filename is clear but that it does not contain your name and student id number. Similarly, do not include your name and id number in the document itself.

There is a basic LaTeX template file on Blackboard which you may choose to use for typing-up your solutions. The file is called `CW1_submitted_work.tex`.

Turnitin will generate a similarity report for your submitted document and indicate matches to other sources, including billions of internet documents (both live and archived), a subscription repository of periodicals, journals and publications, as well as submissions from other students. Please ensure that the document you upload represents your own work and is written in your own words. The Turnitin report will be available for you to see shortly after the due date.

This coursework should hopefully help to reinforce some of the methodology you have been studying, as well as the skills in `R` you have been developing in the module. Correct interpretation and meaningful discussion of the results (i.e. attempt to put the results into context) are as important as correct calculation of the results, in order to achieve a high mark for the coursework.

The data are in the file `white_wine.csv` (Cortez *et al*, 2009) contain various measurements on white wine variants of the Portuguese Vinho Verde wine. Import the data into `R` from your default folder using the command:

```
white_wine=read.table("white_wine.csv", sep = ";", header = TRUE)
```

The object `white_wine` contains measurements on 11 continuous variables: `fixed.acidity`, `volatile.acidity`, `citric.acid`, `residual.sugar`, `chlorides`, `free.sulfur.dioxide`, `total.sulfur.dioxide`, `density`, `pH`, `sulphates`, `alcohol` plus one discrete, ordinal variable: `quality`.

For the purposes of this coursework we will just use the variables in columns $7, 8$ and 12 which are:

```
total.sulfur.dioxide
density
quality
```

Note that `total.sulfur.dioxide` and `density` are both numeric variables, `quality` is a discrete, ordinal variable

1. (i) Using selected summary statistics and graphical displays from those discussed in weeks 1 and 2 of this module, explore the univariate empirical distribution of `total.sulfur.dioxide`. Comment on your results.

   [4]

   (ii) Using box-plots, look at the distributions of the `total.sulfur.dioxide` data at the different values of `quality`. Comment on the results.

   [4]

   (iii) Produce a scatterplot of the `total.sulfur.dioxide` and `density` data with the contours from a bivariate Normal density having appropriately estimated parameters. Comment on your impression of the bivariate Normal distribution as a suitable probability model for these data.

   [4]

2. Using the function `cor`, calculate both Pearson's and Spearman's correlation between:

   - `total.sulfur.dioxide` and `density`
   - `log(total.sulfur.dioxide)` and `log(density)`

   Comment on the results and give an explanation for any discrepancies or similarities between the various correlation estimates.

   [3]

3. Let ρ_1 denote the correlation in the joint distribution of `total.sulfur.dioxide` and `density`. Based on using Pearson's correlation coefficient, perform a DIY (ie. write your own code to do the calculations) hypothesis test of

   $H_0 : \rho_1 = 0.5$ vs $H_A : \rho_1 \neq 0.5$

   at the 5% significance level using Fisher's Z-transform. Compute the p-value and use it to decide whether to reject the null hypothesis in favour of the alternative.

   Calculate DIY an approximate 95% confidence interval (CI) for ρ_1 based on Fisher's Z-transform and verify that your calculations agree with the CI produced by `cor.test`.

   [6]

4. Write a **function** in `R` to verify via simulation that the distribution of Fisher's Z-transform statistic, for a given sample size n, is approximately Normal. Your function should produce a plot comparing the sampling distribution of Fisher's Z-transform statistic and the appropriate approximate Normal distribution the statistic has under the assumption that the true correlation parameter equals zero. In your simulation, you may assume sample data pairs (x, y) come from independent Normal distributions having user-input parameter values.

   As your solution to this part, please submit the code for your function and also run it in R to produce the plot described in the paragraph above.

   [6]

5. Using the function `cor.test`, test null the hypothesis:

   H_0: there is no monotonic association between `total.sulfur.dioxide` and `density` in their joint distribution vs H_A: there is a monotonic association between these two variables.

   Describe how the test has been carried out for these data and report your conclusions.

   [3]

References

[1] P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. *Modeling wine preferences by data mining from physicochemical properties.* In Decision Support Systems, Elsevier, 47(4):547-553. ISSN: 0167-9236.