MATHS 7107 Data Taming Assignment Four Questions

Due date: 5pm, Wednesday 29th March 2023.

Scenario

You have been approached by the Melbourne Water Corporation ('MWC'), which manages the supply of water in Melbourne, Australia, to produce a report relating to evaporation. In light of recent changes in Melbourne's climate, MWC's previous estimates of rates of evaporation at their reservoirs are unreliable. You are thus asked to build a new model assessing the effects of Melbourne's day-to-day weather's on evaporation, to aid in their management of their Cardinia Reservoir, in the city's South East.

You are given Melbourne's weather observations, including evaporation, for the previous financial year, to build a model predicting evaporation, in order to ensure the stability of the city's water supply. This data can be found in the file `melbourne.csv`. You will need to produce a `report` outlining which temporal and meteorological factors have a significant impact on the amount of evaporation in a given day, the manner in which these factors affect the amount of evaporation, as well as demontrate your ability to make predictions for individual days at Cardinia reservoir. MWC expect your model to be justified. This means you should present confidence intervals alongside your forecasts of the expected evaporation for days of a given character, you should test all the assumptions of your model, and you should provide a foundation for your model by summarising the variables included in your analysis. A more complete breakdown of MWC's expectations can be found under the heading 'Your tasks'.

Your tasks

Components of the Case Study

Bivariate summaries

In your analysis, we are interested in the following potential influences on amount of evaporation in a day:

- Month,
- Day of the week,
- Maximum temperature in degrees Celsius,
- Minimum temperature in degrees Celsius, and
- Relative humidity, as measured at 9am.

You should *not* consider any other variables in your analysis.

Produce plots to explore the relationship, or lack thereof, between your response variable, and each of your potential predictors. Describe any relationships that arise.

In order to do this, you may need to edit strings within existing variables, recode existing variables, or create new variables. All R code used in this and other questions should be contained in your appendix.

You do not need to build any linear models for this part of the case study. This component should be in the methods section of your report, with code in your appendix. [15 marks]

Model selection

Build a model to predict evaporation, in mm, in a given day in Melbourne, using all of the predictors listed in the *Bivariable summaries* paragraph above. Also consider an interaction term between month and 9am relative humidity, to determine whether humidity has a different effect in different months. Your model should include all predictors with a significant effect on evaporation, and no predictors that are not significant. In order to do this, produce your model according to the following process:

1. Fit a model containing all the possible predictors.
2. Determine the p-value for inclusion of each predictor:

   P-values for quantitative variables can be determined using the linear model summary.

   P-values for categorical variables, or interactions containing categorical variables, can be determined using an ANOVA.
3. Remove the predictor with the highest p-value for inclusion, unless all remaining predictors are significant at the 5
4. Update your model to include only the remaining predictors.
5. Repeat Steps 2-4 until only significant predictors remain.

State the significant terms in your final model. Do these terms differ from what you concluded from your bivariate analyses? Why might this be the case?

This component should be in the *methods* section of your report, with code in your appendix. [7 marks]

Model interpretation

Interpret the coefficients of your model in context. This includes the intercept, and the coefficients relating to each predictor. Your interpretation should be done in a manner that can be interpreted by your client.

For categorical predictors, you do not need to explain all coefficients, but provide an overview of how the model operate in relation to these terms, with an example of one of the coefficients.

This component should be described in your *results* section. [10 marks]

Model diagnostics

Test all of the assumptions of your linear model.

This component should be referred to in your methods section, with code and assessment of your assumptions done in an appendix. [8 marks]

Prediction

MWC is interested both in the general application of your model, and in some particular extreme scenarios that it envisages. They thus seek your predictions for the amount of evaporation, in mm, for days of the following character:

- February 29, 2020, if this day has a minimum temperature of 13.8 degrees and reaches a maximum of 23.2 degrees, and has 74% humidity at 9am.
- December 25, 2020, if this day has a minimum temperature of 16.4 degrees and reaches a maximum of 31.9 degrees, and has 57% humidity at 9am.
- January 13, 2020, if this day has a minimum temperature of 26.5 degrees and reaches a maximum of 44.3 degrees, and has 35% humidity at 9am.
- July 6, 2020, if this day has a minimum temperature of 6.8 degrees and reaches a maximum of 10.6 degrees, and has 76% humidity at 9am.

Provide a table containing appropriate intervals for making forecasts on these particular days, and explain the intervals in context. Compare and contrast potential amounts of evaporation on different days.

If there is more than 10mm of evaporation at MWC's Cardinia Reservoir, the corporation takes temporary measures to ensure a continuous supply of water, including transferring water from its Silvan Reservoir upstream.

- For which of the predicted days can we say with 95% confidence that this will occur?
- For which of the predicted days can we say with 95% confidence that this will not occur?

This component should be described in your discussion section, with code in your appendix. [10 marks]

Format of Assignment

This assignment should be written in the format of a report. Your submission should comprise:

- An executive summary, with key results outlined in plain English;
- A methods section, outlining what data was analysed, including summary statistics, and stating the software used to analyse it;
- A results section, interpreting your models in language your client will understand;
- A discussion section, in which you discuss your models' outcomes and projections, with specific relevance to your client's objectives;
- A conclusion, in which you summarise your findings and recommendations to your client, written in plain English; and
- An appendix, including all R code you used to perform your analysis, as well as any technical material beyond what you might expect your client to understand.

Marks for the format will be allocated as follows:

(a) The assignment is in a report format, including summary, methods, results, conclusions and an appendix. [4 marks]

(b) The executive summary, discussion and conclusion are written in plain English, with all terminology used explained.

[2 marks]

(c) All figures in the report are captioned.

[2 marks]

(d) All statements are justified by reference to the relevant figures, tables or models.

[2 marks]

[Total: 10 marks]

Some rules about your submissions:

- **You must complete this assignment using R Markdown**;
- Your assignment must be submitted as **pdf only** on MyUni;
- You must include **units** when providing solutions;
- Include any working when providing solutions;
- Provide all numerical answers to **3 decimal places**;
- Make sure you include both your code and R output / plots in your answers;
- Make sure any tables or plots included have captions;
- Do not write directly on the question sheet;

- You can submit more than once if you find errors and your latest submission will be marked;
- Make sure you only upload one document for your final submission. If you submit multiple pages (i.e. one per question) you will be deducted 10% per page submitted;
- Penalties for late submission - within 24 hours 40% of final mark. After 24 hours, assignment is not marked and you get zero; and
- Finally, make sure you check your submitted assignment is the correct one, as we cannot accept other submissions after the due date.

[Assignment total: 60 marks]