MATHS 7107 Data Taming Assignment Three Questions

Due date: 5pm, Wednesday 15th March 2023.

Scenario

You are performing data science for a major media company, Masthead Media, in the United States, determining the direction of their flagship newspaper, the Boston Sun-Times. The Boston Sun-Times is known for investigative journalism, for which it has won a number of Pulitzer Prizes. The newspaper's editorial staff believe that this unrelenting focus on truth and justice is responsible for the newspaper's sterling reputation. The Boston Sun-Times has received, on average, a Pulitzer Prize every year for the past 25 years.

Masthead Media is deciding whether to continue to invest in the Sun-Times' investigative journalism, or to encourage the newspaper to take a more populist, tabloid slant, in a bid to arrest a recent decline in readership. The Boston Sun-Times currently has a circulation of 453,869.

Masthead Media wants to know:

- Whether publications that win more Pulitzer Prizes have a smaller, or a larger, average circulation;
- Whether publications that win more Pulitzer Prizes see a percentage increase, or decrease, in circulation, during the period that they win the prizes; and
- If these relationships exist, how the trajectory of the Boston Sun-Times's circulation might change depending on the newspapers strategic direction.

To this end, using provided data on newspaper circulation and number of Pulitzer Prizes in a 25 year period, you should answer Questions 1–5 below. You will also need to report the results to the company, using an executive summary (at the start of your assignmnet) and a conclusion (at the end of your assignmnet).

Some rules about your submissions:

- **You must complete this assignment using R Markdown**;
- Your assignment must be submitted as **pdf only** on MyUni with Rmd file (R markdown file);
- 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 pdf 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.

Executive summary

You need to write an executive summary to report your findings to Masthead Media. This summary should include:

- A recap of the project outline including what the purpose of your project is and what problem you are solving.
- A very brief outline of what you did.
- Your final recommendation to Masthead Media.

Your executive summary should be written as if you are presenting it to Masthead Media. They will read the executive summary to get a quick overview; to evaluate the quality of the report; or even to make a decision. It should be no more than half a page, and contain no figures, plots or code. It should be written in plain English with no or minimal terminology. If terminology is included, it must be explained.

This section will be easier to write once you have completed questions 1-5.

[Total: 4 marks]

Question One: Reading and Cleaning

Load the data contained in `pulitzer.csv` into `R`. A summary of the variables as represented in the csv file are below.

Variable	Description
`newspaper`	The name of one of the United States' 50 largest newspapers, as at 2004
`circ_2004`	The newspaper's circulation in 2004
`circ_2013`	The newspaper's circulation in 2013
`change_0413`	The percentage change in the newspaper's circulation, between 2004 and 2013
`prizes_9014`	The number of Pulitzer Prizes won by the newspaper's journalists between 1990 and 2014

For our analysis, we would like to predict either average circulation between 2004 and 2013, or change in circulation between 2004 and 2013, using the number of Pulitzer Prizes between 1990 and 2014.

(a) Recode the `change_0413` variable so it represents the percentage change in circulation between 2004 and 2013 as an integer. This will require manipulating the strings in change_0413.

[2 marks]

(b) Append a new variable to the tibble which contains the average of `circ_2004` and `circ_2013`.

[2 marks]

[Total: 4 marks]

Question Two: Univariate Summary and Transformation

(a) Describe the distribution of the variable representing average circulation, including shape, location, spread and outliers (Reminder: plots and summary statistics are useful here). [5 marks]

(b) Describe the distribution of `change_0413`, including shape, location, spread and outliers. [5 marks]

(c) Do either of `change_0413` and the variable representing average circulation have a skew that could be resolved by a log transform? For each variable, select whether it should be transformed. [2 marks]

[Total: 12 marks]

Question Three: Model building and interpretation

(a) Build a model predicting the variable representing a newspaper's circulation using `prizes_9014`, incorporating a log transform for the average circulation if you decided this was necessary. State and interpret the slope and intercept of this model in context. Is there a statistically significant relationship between the number of Pulitzer Prizes, and average circulation?

[7 marks]

(b) Build a model predicting `change_0413` using `prizes_9014`, incorporating a log transform for `change_0413` if you decided this was necessary. Is there a statistically significant relationship between the number of Pulitzer Prizes, and change in circulation?

[7 marks]

(c) Check the assumptions of the linear models. Recall that there are four assumptions for each model.

[12 marks]

[Total: 26 marks]

Question Four: Prediction

Masthead Media is considering three *strategic directions* for the Boston Sun-Times. These are:

- Investing substantially less in investigative journalism than present. In this case, Masthead Media projects that the newspaper will be awarded 3 Pulitzer Prizes in the next 25 years.
- Investing the same amount in investigative journalism than present, leading to the award of 25 Pulitzer Prizes in the next 25 years.
- Investing substantially more in investigative journalism, leading to the award of 50 Pulitzer Prizes.

For the following questions, assume that the projected number of prizes under each possible strategic direction is known; that is, do not incorporate any uncertainty in the number of Pulitzer Prizes.

(a) Using the model from Question 3(a), calculate the expected circulation of the newspaper under each of the three proposed strategic directions and represent these in a table. How does this compare with the current circulation?

[5 marks]

(b) Using the model from Question 3(b), calculate the change in circulation of the newspaper, across the next decade, under each of the three proposed strategic directions and represent these in a table. Comment on whether the projections of each of the two models are consistent.

[4 marks]

(c) Using the model from Question 3(a), calculate 90% confidence intervals for the expected circulation of the newspaper under each of the three proposed strategic directions. Place these confidence intervals in a table, and contrast them in context.

[5 marks]

(d) Using the model from Question 3(b), calculate 90% prediction intervals for the expected change in circulation of the newspaper under each of the three proposed strategic directions. Place these prediction intervals in a table, and contrast them in context.

[5 marks]

[Total: 19 marks]

Question Five: Limitations

(a) Discuss what limitations there might be to each of the models. Why might this model be insufficient for its application? You should discuss **at least** two limitations of these models in application.

[4 marks]

[Total: 4 marks]

Conclusion

You need to write conclusion to recap your findings to Masthead Media. This will go at the very end of a report and should include:

- A recap of the project outline including what the purpose of your project is and what problem you are solving.
- A very brief outline of what you did.
- Your final recommendation to Masthead Media.
- Any limitations that need to be addressed or future work to be done.

Your conclusion should be written as if you are presenting it to Masthead Media. They will read the conclusion to get a recap of everything outlined in the full report. It should be no more than 3/4 of a page, and contain no figures, plots or code. It should be written in plain English with no terminology.

This section will be easier to write once you have completed questions 1-5.

[Total: 5 marks]

[Assignment Total: 74 marks]