Coursework 4: Statistic for AI and Data Science

1 Introduction

This document outlines the requirements for coursework 4. This coursework has two parts:

1. Answer questions relating to a short paper (submit a PDF).
2. Carry out further analysis of the data from the paper (submit a Jupyter notebook).

The two parts have equal marks.

Your work should be submitted as two files, one for each part. Section 2 outlines part 1; part 2 is outlined in Section 3. The final section notes the relevant resources that are available on the QMPlus site for the module.

2 Part 1: Review of the paper 'Storks Deliver Babies'

This part of the coursework requires written answers to questions relating to the paper.

2.1 Review Questions [Equal Marks]

Write brief answers to the following questions about points made in the following paper. The questions carry equal marks.

> *Robert Matthews. "Storks Deliver Babies (p = 0.008)". Teaching Statistics. Volume 22, Number 2, Summer 2000, p36-8.*

Question 1: The paper explains that the p-value can be misunderstood. Explain the misunderstanding described in the paper and give the correct interpretation of the p-value in the context of the analysis in the paper.

Question 2: Explain how the correlation coefficient and the p-value relate to the question 'how good is my regression model?', making clear the difference between them.

Question 3: Using the example from the paper, explain the difference between causation and correlation, covering possible relationships between them.

Question 4: Explain what is meant by a confounding variable and suggest possible confounders for any relationship between storks and births. Draft and describe a diagram of causes (see lecture topic 13) that you believe is most likely to explain the relationship between the 4 variables Area, Storks, Humans and BirthRate used in the paper.

2.2 Submission requirements

- Your answers should be presented in a document submitted in PDF.
- The document should be no more than 2 pages long. Only the first two pages will be marked; any extra pages will be ignored.
- You are expected to answer the questions in your own words. If you choose to quote text from other sources, you must clearly indicate this and reference the source document. Overuse of quotations will be marked down.
- You should include a reference to the paper by Matthews.

3 Part 2: Additional Analysis of the Storks Data

The requirements in this section are not detailed. Use your judgement where there are no explicit steps itemised. Also review the mark scheme.

Complete an analysis of the data as described below. Submit your work as a Jupyter notebook.

3.1 Load and Review the Data

The data shown in the table in the paper is available as a CSV file, with an additional variable – the GDP per capita. Load and review the data.

- Add a new variable showing the population density (millions per Km^2)
- Document the variables (including the units)
- Briefly review the distributions and correlations.

3.2 Implement Two Regression Models for the Number of Births

The aim of this analysis is to complete and compare two regression models to explain the variability of the number of births.

- Model 1: predictor: the number of storks (as given in the paper)
- Model 2: predictors are:
    - The population size (millions of people) – the variable is 'Human'.
    - The GDP per Capita (in dollars) – the variable is 'GDP_per_capita'
    - The population density – a derived variable from the 'Area' and the 'Humans'.

The idea of the model 2 predictors are:

- Population size: there are more births in a larger population.
- GDP per capita: an increase in prosperity has been associated with a reduction in family size.
- Population density: a crowded country may discourage large families.

Show the fit of the two models with suitable plots and metrics; explain these briefly.

3.3 Use the Bootstrap Technique to Construct Confidence Intervals

We can use the bootstrap technique to estimate confidence intervals. The first step is to implement functions to create a bootstrap distribution:

1. A function for re-sampling (see the Bootstrap notebook)
2. A function that creates the model(s) (from Section 3.2 above), given a bootstrap sample. This function is passed to the re-sampling function (see the Bootstrap notebook for this approach).

Confidence Intervals for the Predictors Weights ('beta') in Model 2

Use the bootstrap technique to evaluate the multipliers (often called 'beta' values) for the predictors 'GDP per capita' and 'population density'.

- Fit the second regression model (model 2) for each bootstrap sample; find the multipliers (beta values).
- Repeat these steps many times and plot a distribution of the two beta values.
- Estimate appropriate confidence intervals for the two values. Select suitable 'alpha' values (e.g. 90%, 95%).

Explain what should be inferred about the predictors.

Confidence Interval for the Difference in Performance of the Two Models
Use the bootstrap technique to estimate the CI for the difference in the root mean squared error (RMSE) between the predicted and actual values for the two models:

- Calculate the difference in the RMSE parameter for the two models for each bootstrap sample.
- Repeat these steps many times and plot a distribution of the differences.
- Estimate appropriate confidence intervals for the difference in the RMSE values. Select suitable 'alpha' values (e.g. 95%, 99%).

Explain whether we can be confident that one model predicts the number of storks better than the other model.

3.4 Overall Conclusions

It is claimed that the data analysis shows that storks do not deliver babies. Comment on this claim.

3.5 Mark Scheme for Part 2

Section	Weight	Criteria	Detailed Criteria
All	10%	Presentation of the document and code	The notebook has a clear structure, with a title and sections; suitably formatted markdown cells are interleaved with code. Writing addresses a 'domain expert' – a reader interested in birds and bees.
All	10%	Code quality	Code, executed in order without errors, is organised in short segments, alternating with text explaining the operations on data. All the code presented in the notebook is needed. Appropriate use of library code (e.g. pandas), avoiding unnecessarily complex code
Section 3.1	10%	Data loading and review	The review of the data is clear and concise.
Section 3.2	10%	Implementation of the regression models for the number of births	The regression models are implemented correctly, and their fit is shown by appropriate plots and metrics, support by brief explanation
Section 3.3	30%	Use of bootstrap to estimate CI for predictor weights	The bootstrap distribution is constructed correctly.
			The bootstrap distribution is used to estimate CIs for the two weights.
			Clear explanation of the implications of the analysis.
Section 3.3	20%	Use of bootstrap to estimate CI for comparing the two models	The bootstrap distribution is correctly constructed and used to estimate a CI for the difference in RMSE of the two models.
			Clear explanation of the implications of the analysis.
Section 3.4	10%	Overall conclusions	Clear, concise and accurate conclusion.

4 Available Resources

Notebooks are available on the QMPlus site covering

- Regression modelling
- The bootstrap technique