Statistic for AI and Data Science

Coursework 4

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).

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' [50 Marks]

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

2.1 Review Questions

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.

3 Part 2: Additional Analysis of the Storks Data [50 Marks]

The data shown in the table in the paper is available as a CSV file, with an additional variable – the percentage of land area suitable for storks. Complete two analyses of the data as described below. Your work should be submitted as Jupyter notebook.

3.1 Requirements for Analysis 1

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

Part 1.1: Implement two regression models for the number of births

- Model 1: predictor: the number of storks (as given in the paper)
- Model 2: predictor is the population size

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

Part 1.2: Use the bootstrap technique to estimate the distribution of the difference in the r2 parameter for the two models.

- Resample the data; fit both models; calculate the difference in the r2 parameter for the two models
- Repeat these steps many times and plot a distribution of the differences

Estimate appropriate confidence intervals for the difference in the r2 values. Explain whether we can be confident that one of the models explains more of the variability in the number of births than the other model.

3.2 Requirements for Analysis 2

The aim of this analysis is to compare the performance of two regression models to predict the number of storks

Part 2.1: Implement two regression models for the number of storks

- Model 1: predictor is the land area
- Model 2: predictors are i) the population density and ii) the area of suitable land

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

Part 2.2: Use the bootstrap technique to estimate the difference in the root mean squared error (RMSE) between the predicted and actual values for the two models

- Resample the data and fit both models; calculate the difference in the RMSE parameter for the two models
- Repeat these steps many times and plot a distribution of the differences

Estimate appropriate confidence intervals for the difference in the RMSE values. Explain whether we can be confident that one model predicts the number of storks better than the other model.

3.3 Mark Scheme

Section	Weight	Criteria	Detailed Criteria
All	20%	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 transport patterns
			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
Part 1.1	20%	Implementation of the regression models for the number of births	The regressions are implemented correctly.
			The fit of the regression models is shown by appropriate plots and metrics, support by brief explanation
Part 1.2	20%	Use of bootstrap to compare r2	The bootstrap distribution is constructed correctly and use to estimate a confidence interval.
			There is a clear explanation of the findings of the analysis.
Part 2.1	20%	Implementation of the regression models for the number of storks	The regressions are implemented correctly.
			The fit of the regression models is shown by appropriate plots and metrics, support by brief explanation
Part 2.2	20%	Use of bootstrap to compare RMSE	The bootstrap distribution is constructed correctly and use to estimate a confidence interval.
			There is a clear explanation of the findings of the analysis.

4 Available Resources

Notebooks are available on the QMPlus site covering

- Regression modelling
- The bootstrap technique