Statistic for AI and Data Science

Coursework 2 (Version 1)

1 Introduction

This document outlines the requirements for coursework 2, which is based on notebook 3 and uses the Texas Bridge data. The coursework tests whether you can:

- Carry out an exploratory analysis of data with both continuous and categorical variables
- Use regression modelling to look at the effect of predictor variables on a target variable.

1.1 Scenario

The management of the Texas Department of Transportation wishes to investigate the use of the following variables to predict the current condition of bridges:

1. Age (derived from variable `Year`)
2. average use (variable `AverageDaily`)
3. percent trucks (variable `Trucks_percent`)
4. material (variable `Material`)
5. design (variable `Design`)

The current condition is derived from variables `Deck_rating`, `Superstr_rating` and `Substr_rating` of the bridges. They wish to know:

- How well the proposed variables can predict the bridge condition.
- Which of the proposed variables has more influence on the current condition.

The use of regression has been agreed in advance.

Your report is to be submitted to a representative of the Texas Department of Transportation. She is a specialist in metal corrosion and the deterioration of concrete and wants to understand "what the data says" but also "how you have processed it" (for example so that she can check that your assumptions are sensible). She is not interested in understanding how your code works. She understands that you do not have a detailed knowledge of bridges (or corrosion) so will be content provided that any assumptions you have made are clearly explained: if any of them turn out to be inappropriate she can ask you to modify the analysis.

2 Coursework Requirements

The requirements are in three parts. These requirements should be read alongside the submission requirements and the mark scheme.

2.1 Part 1: Data Preparation

You must prepare the data by deriving new variables. You should look at distributions, simplify categories and review any outliers (*are they data errors or just extreme but plausible values?).*

- There isn't an age variable, so derive one from the Year variable with age in years.
- You are recommended to exclude very old bridges (possibility the historic ones).
- You are also recommended to reduce the number of categories of materials and design by merging some of the very small categories.
- The current condition should be derived from the combination of the three main condition variables (ignore 'scour'), by treating each as an integer score (0 for failed)

and adding the three scores. This means that the regression (part 2) is on a continuous variable.

Be sure that any changes you have made to the data are carefully justified.

2.2 Part 2: Exploratory Analysis

You should look at the relationship between the five predictor variables and the target variable. You should also look at and comment on the relationship between the predictors. You should draw preliminary conclusions at the end of this part of the analysis on the answers to the questions asked by the Texas Department of Transportation.

Hints

- The variables are a mix of continuous and categorical. You need to use appropriate techniques to explore the relationship between continuous variables, between continuous and categorical variables and between categorical variables.

2.3 Part 3: Regression Modelling

You are required to construct a linear regression to look at the effect of the five predictor variables on the target variable.

- You should record the R2 (coefficient of determination) and comment on the value.
- You should show and comment on the distribution of residuals (errors).
- You should use the regression coefficients to compare the influence of the different predictors.
- You should draw final conclusions at the end of this part of the analysis on the answers to the questions asked by the Texas Department of Transportation. You should include brief suggestions for further analysis.

Hints

- Remember that the beta coefficients have units and the range of the predictor needs to be considered when making comparisons.
- Confidence intervals are not required.
- It is vital not to include any metrics (such as R2, RMSE or others) without a thorough explanation of their meaning. You will lose marks if you write in a way that might confuses your client.

3 Submission Requirements

The following additional requirements are about how your work should be submitted.

1. You must submit a single .ipynb' file only.
2. The notebook must be executable without errors. It must read the original data file of exits (from the same directory as the notebook); the data file must not be changed. Rerun the complete notebook before submission, so that the cells are executed in order.
3. The notebook must be readable. Markdown cells should be used to organised the notebook with a title and section headings. The code cells should be short, alternating with text cells (using markdown). The markdown text should be written to a ‘domain expertise’ interested in how the data is being manipulated (rather than in how the code works).
4. You can use material from notebook 3 (e.g. code to load the CSV) and from ‘other topic notebooks’, but do include code that is not relevant to the requirement above. You are

strongly advised against using a complex regression library: if you do so, be sure to explain all the outputs generated as you will lose marks if your notebook contain unexplained information.

4 Mark Scheme

The following table shows the mark scheme. However, the detailed criteria on presentation and code correctness are essential and a poor grade on these will have a disproportionate effect on the overall mark.

Section	Weighting	Criteria	Detailed Criteria
All	20%	Presentation of the document	The notebook has a clear structure, with a title and sections, all in the style of the notebooks provided on the module
			Document includes markdown text cells interleaved with code, suitably formatted. Writing addresses a 'domain expert' – a reader interested in transport patterns
All	20%	Correctness and clarity of the code	The notebook shows the code executed in order without errors. Code runs when all cells are executed in order
			Code 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	20%	Data preparation	Creating derived variables
			Distributions of variable. Simplification of categorical variables and review of outliers. Clearly described (and reasonable) assumptions
Part 2	20%	Quality of the exploratory analysis	Appropriate visualisation and analysis of the relationship between continuous and categorical predictors and of their relationship with the target variable.
			Appropriate use and interpretation of measures of correlation and clear conclusions on finding of exploratory analysis
Part 3	20%	Quality for the regression analysis	Correct construction of the regression, distribution of residuals distribution and R2 measure.
			Comparison of the influence of the predictors and overall conclusions of the analysis.

5 Feedback

No feedback is provided until after the last hand-in date (i.e. one week after the deadline). The following forms of feedback will be provided:

1. A sample answer.
2. General review comments on the main strengths and weaknesses for the class as a whole (written or in a review lecture)
3. The grades obtained by your answer on each of the detailed criteria. Clarifying comments may be added.

6 Available Resources

You can use material from notebook 3 (e.g. code to load the CSV), but do include code that is not relevant to the requirement above. A separate notebook is available on regression: other notebooks may also be useful.