|  |  |
| --- | --- |
|  | **Faculty of Environment and Technology**  **Academic Year: 2021/2022**  **End of Teaching Block: 1** |

**Module Leader: Deirdre Toher**

**Module Code UFMFHR-M-15**

**Module Title: Statistical Inference**

**Examination Duration: 24 hours**

**ONLINE EXAM**

**Instructions to Students:**

**Answer all questions.**

**The dataset – called Jan\_2022\_Exam\_Data.csv is available on blackboard.**

**Download the file currently named Exam\_Template.Rmd. Change this filename to your StudentID.Rmd. Complete all your work in this file. Compile your markdown document to Word and include that file in your submission.**

**There are additional instructions to students in the file Exam\_Template.Rmd.**

**Your submission should include all of the following in a zipped folder which must be named StudentID\_StatInf\_Jan22Exam.zip replacing “StudentID” with your UWE student ID.**

* **Your R markdown file, with filename StudentID.Rmd, containing your answers.**
* **Any other files used in your analysis (including any additional R files used).**
* **A knitted version of your R markdown document – in Word format.**

This is an individual assessment: do not work with any other person during this exam. Text-matching software will be used on all submissions.

# Instructions for submission

You must submit your assignment *before* the stated deadline by electronic submission through Blackboard.

* Multiple submissions can be made to the portal, but only the final one will be accepted. Please save your work frequently.
* **It is your responsibility to submit the exam in a format stipulated above**

Your marks may be affected if your tutor cannot open or properly view your submission.

* **Do not leave submission to the very last minute**. Always allow time in case of technical issues.
* The date and time of your submission is taken from the Blackboard server and is recorded when your submission is complete, not when you click Submit.
* **It is essential that you check that you have submitted the correct file(s),** and that each complete file was received. Submission receipts are accessed from the Coursework tab.

**There is no late submission permitted on this timed assessment.**

**Data description**

This dataset is part of a larger dataset that has been collected to help to estimate the price of used cars.

It contains the following variables:

* brand (manufacturer)
* model (of car)
* year (of registration of the car)
* price (in GB pounds)
* transmission (type of gearbox)
* mileage (total distance covered by the car)
* fuelType (type of fuel used by the car)
* tax (annual cost of vehicle tax)
* mpg (miles per gallon - a measure of fuel efficiency)
* engineSize (size of the engine in litres)

You are interested in modelling the price of vehicles that have all of the following properties:

* mileage less than 60000
* Manual transmission
* Petrol engine (fuelType)
* Costing less than £200 in annual Vehicle Tax.

Once you have selected the rows of data with these properties, then you must *use your studentID* to select a random sample of 2000 rows of the data to perform the rest of your analysis with.

You should remove any redundant variables (where only one value remains in that variable).

This subset of the data is what you should use for the rest of this assessment.

**Question 1: Data Preparation (11 marks)**

1. Explain what data preparation is required in order for the data in **Jan\_2022\_Exam\_Data.csv** to be suitable for analysis.

**(4 marks)**

1. Implement the required data preparation in the code chunk below:

**(7 marks)**

**Question 2: Exploratory Data Analysis (22 marks)**

**Descriptive Statistics**

1. What descriptive statistics would be appropriate for this dataset? Explain why these are useful in this context.

**(2 marks)**

1. Produce those descriptive statistics in the code chunk below:

**(4 marks)**

1. What have those descriptive statistics told you – and how does this inform the analysis that you would undertake on this data or any additional data cleaning requirements?

**(4 marks)**

**Exploratory Graphs**

1. What exploratory graphs would be appropriate for this dataset? Explain why these are useful in this context.

**(2 marks)**

1. Now produce those exploratory graphs in the code chunk below:

**(4 marks)**

1. Interpret these exploratory graphs. How do these graphs inform your subsequent analysis?

**(4 marks)**

**Correlations**

1. What linear correlations are present within this data?

**(2 marks)**

**Question 3: Bivariate relationship (14 marks)**

1. Which of the potential explanatory variables has the strongest linear relationship with the dependent variable?

**(1 mark)**

1. Create a linear model to model this relationship.

**(2 marks)**

1. Explain and interpret the model:

**(3 marks)**

1. Comment on the performance of this model, including comments on overall model fit and the validity of model assumptions. Include any additional code required for you to make these comments in the code chunk below.

**(4 marks)**

**Bootstrap**

1. Use bootstrapping on this model to obtain a 95% confidence interval of the estimate of the slope parameter.

**(4 marks)**

**Question 4: Multivariable relationship (10 marks)**

Create a model with all of the appropriate (non-constant) remaining explanatory variables included:

1. Explain and interpret the model:

**(4 marks)**

1. Comment on the performance of this model, including comments on overall model fit and the validity of model assumptions. Include any additional code required for you to make these comments in the code chunk below.

**(4 marks)**

1. What general concerns do you have regarding this model?

**(2 marks)**

**Question 5: Model simplification (8 marks)**

1. What approaches for model simplification would you consider implementing and why?

**(4 marks)**

1. What are the potential advantages of simplifying a model?

**(2 marks)**

1. What are the potential disadvantages of simplifying a model?

**(2 marks)**

**Question 6: Reporting (35 marks)**

A client is looking to purchase a used Skoda Superb (registration year either 2018 or 2019, manual transmission, petrol engine) and wants to understand what factors influence the expected price of a used car, and how they influence the price.

Write a short report of 300-500 words for the client.

Furthermore, include an explanation as to which model you would recommend, and why you have selected that model.

Comment on any suggestions for alterations to the model that would be appropriate to consider.

Highlight what may or may not be directly transferable from the scenario analysed in Questions 1 to 5.

**END OF QUESTION PAPER**