**37252 Regression and Linear Models**

**Assessment Task 2: Assignment**

This assessment task is marked out of 60.

It is worth 30% of the marks for this subject.

**Due: 12 noon Thursday 9 May 2024**

**This assignment is a compulsory group assignment.** Groups must consist of **three to five** students. The assignment is expected to be the joint work of all of the members of the group.

Each group is to write a report answering the questions in this document. You should approach the writing of this report as would an employee in industry or a researcher in academia; i.e. your report should be structured and formatted in a professional manner that makes it easy to read, follow and understand. **The document should be self-contained; all R output that is referred to should be included as items in the body of the report. Please retain the question numbering in your report.**

After submission, the contribution of each group member will be assessed by the other group members using an online rating system. (Details on how to do this will be posted on Canvas nearer to the submission date.) **This MAY result in the adjustment of the mark awarded to the individual group members.**

You must attach the assignment cover sheet to your assignment. **Students who have not signed the assignment cover sheet will not be given a mark for the assignment.**

Your group report should be submitted on Canvas in PDF format before the date given above. You must also submit the R data file used in Q1 together with any R codes you have relied on for your analysis.

In the **week prior to the deadline**, the group must send an email to joanna.wang@uts.edu.au. The email should include a list of the names and student IDs of the group members. Each group will then be created on the online SPARK rating system.

**QUESTION 1. Simple linear regression [20 marks]**

Find a dataset suitable for demonstrating simple linear regression. It should contain two numerical variables, one that can be a response variable ($y$) and one an explanatory variable ($x$). There should be at least 50 observations.

1. **[3 marks]** Use a scatter plot to explore the direction, type and strength of the relationship between the two variables you have identified to be $y$ and $x$ (include the scatterplot with your answer).
2. **[5 marks]** Obtain and write-down the fitted regression line and comment on whether $x$ is a useful predictor using a T-test. Make sure you clearly state the hypotheses, test statistic, test result and your conclusion in plain English.
3. **[2 marks]** Interpret the estimated values of $β\_{0}$ (the intercept) and $β\_{1}$ (the slope) of the regression line in the context of the dataset.
4. **[2 marks]** Find the value of the coefficient of determination $R^{2}$and interpret its value.
5. **[2 marks]** Choose a new value for $x$ that is not in your current data. Find the 95% confidence interval for the predicted mean value of $y$ at your chosen value of $x$.
6. **[3 marks]** Using appropriate plots, perform a visual analysis of the standardised residuals. Assess the assumptions made about the error terms in the model.
7. **[3 marks]** Use Cook’s D to identify the most influential observation in your data. State the observation number and remove it from the regression. Discuss any impact this has in terms of regression coefficients and $R^{2}$.

**QUESTION 2. Multiple linear regression [20 marks]**

In this question we model fuel consumption. The data are observations from the forty-eight contiguous US states taken in 1980. The variables we consider are summarised in the table below.

|  |  |  |
| --- | --- | --- |
| **Name** | **Type** | **Description** |
| $$consumption$$ | response | state fuel consumption |
| $$miles$$ | predictor (continuous) | miles of paved highway |
| $$proportion$$ | predictor (continuous) | proportion of population with driver's license |

The data are available in “37252\_AssessmentTask2\_Autumn2024.csv\*.

1. **[2 marks]** Construct a linear regression model with $consumption$ as response and $miles$ and $proportion$ as predictors. Write down the estimated regression equation and provide interpretations of the estimated coefficients.
2. **[4 marks]** Test if the regression model is significant at the 0.05 significance level. Write down the hypotheses, the test statistic and p-value, the result of the test and conclusion in plain English.
3. **[2 marks]** What is the percentage of the total variation in $consumption$ that can be explained by using this multiple linear regression model?
4. **[2 marks]** Which predictor variable is more important for explaining $consumption$? Explain your answer (Hint: check p-values).

Below is a table of some quantiles from the relevant Student’s T distribution.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| $$t\_{0.005}$$ | $$t\_{0.01}$$ | $$t\_{0.025}$$ | $$t\_{0.05}$$ | $$t\_{0.1}$$ | $$t\_{0.9}$$ | $$t\_{0.95}$$ | $$t\_{0.975}$$ | $$t\_{0.99}$$ | $$t\_{0.995}$$ |
| -2.69 | -2.41 | -2.01 | -1.68 | -1.30 | 1.30 | 1.68 | 2.01 | 2.41 | 2.69 |

1. **[4 marks]** Is there enough evidence to conclude that the coefficient for $proportion$ is less than 1750 at the 0.05 significance level? Write down the hypotheses, calculate the test statistic, report the test result and write a conclusion in plain English.
2. **[3 marks]** State the assumptions made about the error terms in the model. Using appropriate plots, perform a visual analysis of the standardised residuals.
3. **[3 marks]** Determine if the residuals are normally-distributed at the 0.05 significance level. Write down the hypotheses, the test statistic and p-value, the result of the testand a conclusion in plain English.

**QUESTION 3. Regression with categorical predictor [20 marks]**

In this question we extend the model built in Question 2. The variables we now consider are summarised in the table below.

|  |  |  |
| --- | --- | --- |
| **Name** | **Type** | **Description** |
| $$consumption$$ | response | state fuel consumption |
| $$miles$$ | predictor (continuous) | miles of paved highway |
| $$proportion$$ | predictor (continuous) | proportion of population with driver's license |
| $$income$$ | predictor (continuous) | per capita income |
| $$taxBracket$$ | predictor (categorical) | petrol tax bracket: low (1), medium (2), high (3) |

1. **[5 marks]** Construct a linear regression model with $consumption$ as response and $miles$, $proportion$, $income$ and $taxBracket$ as predictors, also include interaction between $taxBracket$ and $income$. Hint: create dummy variables for $taxBracket$ with $taxBracket=3$ as reference category. Write down the estimated regression equation and interpret coefficients of the two binary dummy variables and two interaction terms.
2. **[2 marks]** Using R to make the calculations (i.e. without using the regression equation directly), find predicted fuel consumption and 95% individual confidence interval associated with this prediction when $miles=697$, $proportion=0.56$ and $income=4568$ for low petrol tax bracket states.
3. **[2 marks]** Comment on the statistical significance of the interaction terms. What does the result imply?
4. **[2 marks]** Write down the adjusted R2, compare with the one in Question 2 and comment.
5. **[2 marks]** By performing an appropriate regression, calculate the VIF for the predictor $miles$ in the model in part (a). Make sure you show all working.
6. **[3 marks]** Using appropriate plots, perform a visual analysis of the standardised residuals. Assess the assumptions made about the error terms in the model.
7. **[2 marks]** Determine if there is any statistical evidence against the assumption of independence of the error terms.
8. **[2 marks]** Identify any potentially influential points by calculating the appropriate statistic.