RESEARCH SCHOOL OF FINANCE, ACTUARIAL STUDIES AND STATISTICS

REGRESSION MODELLING

(STAT4038/STAT6038)

Assignment 1 for Semester 1, 2023

Due date: 3:00 pm (Canberra time) on Friday, 31 March 2023

INSTRUCTIONS:

- This assignment is worth 15% of your overall marks for this course.
- You must complete this assignment by yourself. If you copy someone else's work or allow your work to be copied, you will receive a mark of zero for the assignment and risk very severe academic consequences.
- Your report should be submitted to Turnitin on Wattle as a single pdf document (less than 25MB), including the following:
    1. The assignment cover sheet (available to download from Wattle).
    2. Your assignment (a maximum of 10 pages).
    3. An appendix including the R codes you used. Failure to upload the R code will result in a penalty.
- When submitting your assignments, please ensure that they are typed. You may include some edited R output, such as graphs and tables, showing the results of your data analysis and a discussion of these results. Additionally, you may include carefully selected code to support your analysis. It is important to be selective about what you present and only include as many pages and as much R output as necessary to justify your solution. Please label each part of your report with the corresponding question number to make it easier for the reader to follow your work.
- Unless otherwise advised, use a significance level of 5%. Round numeric answers to 4 decimal places (e.g., 0.0012).
- To ensure that your report is concise and meets the expectations of the assignment, please follow the instructions carefully. Note that marks may be deducted if the instructions are not strictly followed. Specifically, the report should be no longer than 10 pages, including graphs and tables. However, you may include an appendix in addition to the 10-page limit. Please note that the appendix will not be assessed but will be checked only if there is a question about what you have actually done.
- Name your report "Course code-Uid", e.g., "STAT2008-u1234567".
- To avoid any unexpected issues, such as internet connectivity problems, please aim to submit your assignment at least 15 minutes before the deadline.
- Please note that late submissions will NOT be accepted. If you require an extension, it will be granted on medical or compassionate grounds, provided that appropriate evidence is produced. However, obtaining the lecturer's permission for an extension at least 24 hours before the deadline is essential.

Analysing Used Car Prices with Simple Linear Regression

You have been provided with a dataset containing information on 818 used Toyota cars. The data includes the year of manufacture (Year), model name (Model), odometer reading in kilometres (Odometer), transmission type (Transmission), power type (Power), and the asking price in $AUD for each car (AskPrice). Your goal is to use simple linear regression to analyze the data and answer the following questions:

1. Is a seller asking a reasonable price for the car?
2. Is the seller asking a lower price than the market price for the car?
3. Is the seller asking a higher price than the market price for the car?

Therefore, we will be using the asking price of a vehicle as the response variable and the odometer reading and the year it was manufactured as the predictor variables in this assignment.

To complete the assignment, you need to perform the following tasks:

(a) [5 marks] Exploring the predictor variable, Odometer, through appropriate graphic diagnostics and descriptive statistics can assist in identifying any outliers in the variable and provide insight into the range of validity for the regression analysis based on the concentration and range of the variable levels. Explore the predictor variables, Odometer and Year, one at a time, using appropriate graphical diagnostics and descriptive statistics. Based on your analysis, what conclusions can you draw about these variables?

(b) [5 marks] Diagnostic plots for the response variable are usually not very informative in regression analysis because the values of the response variable are influenced by the predictor variable, making it challenging to evaluate the goodness of fit of the regression model by examining the response variable directly on itself. Nonetheless, exploring the response variable can still yield useful insights, particularly in identifying outliers or anomalies. Explore the response variable, AskPrice, using appropriate graphic diagnostics and descriptive statistics. Based on your analysis what conclusions can you draw?

(c) [5 marks] Perform an exploratory data analysis to assess the correlation between the two variables, AskPrice and Odometer. Based on your prior knowledge or assumptions, what was your expectation regarding the relationship between the two variables? Does the observed relationship match your expectation?

(d) [10 marks] Fit a simple linear regression (SLR) model to the data, with AskPrice as the response variable and Odometer as the predictor variable. Then, create several plots to assess the model's assumptions and identify any unusual data points. These plots should include a plot of the residuals against the fitted values, a normal Q-Q plot of the residuals, a bar plot of the leverages for each observation, and a bar plot of Cook's distances for each observation. Use these plots, along with any other relevant means, to comment on the model assumptions and identify any potential outliers or influential observations.

(e) [5 marks] Produce the ANOVA table for the fitted model in part (d) and conduct the F-test based on the output. Please provide your interpretation of the ANOVA results and the F-test. What is the coefficient of determination (R-squared) for this model, and how should it be interpreted as a summary measure?

(f) [10 marks] What are the estimated coefficients and their standard errors for the fitted model in part (d)? Provide an interpretation of these coefficient values and perform t-tests to determine if they differ significantly from zero. Based on the results of these tests, what conclusions can you draw?

(g) [5 marks] Perform an exploratory data analysis to assess the correlation between the two variables, AskPrice and Year. Based on your prior knowledge or assumptions, what was your expectation regarding the relationship between the two variables? Does the observed relationship match your expectation?

(h) [10 marks] Fit a simple linear regression (SLR) model to the data, with AskPrice as the response variable and Year as the predictor variable. Then, create several plots to assess the model's assumptions and identify any unusual data points. These plots should include a plot of the residuals against the fitted values, a normal Q-Q plot of the residuals, a bar plot of the leverages for each observation, and a bar plot of Cook's distances for each observation. Use these plots, along with any other relevant means, to comment on the model assumptions and identify any potential outliers or influential observations.

(i) [5 marks] Produce the ANOVA table for the fitted model in part (h) and conduct the F-test based on the output. Please provide your interpretation of the ANOVA results and the F-test. What is the coefficient of determination (R-squared) for this model, and how should it be interpreted as a summary measure?

(j) [10 marks] What are the estimated coefficients and their standard errors for the fitted model in part (h)? Provide an interpretation of these coefficient values and perform t-tests to determine if they differ significantly from zero. Based on the results of these tests, what conclusions can you draw?

(k) [10 marks] Do you think that the simple linear regression models fitted in parts (d) and (h) are suitable? If not, what variable transformations could you apply to enhance the models? Please provide an explanation for your choice of transformation. Using the transformation of your choice, refit the model(s) and select the best one for interpretation and prediction. Please provide reasoning for your selection and interpret its coefficients.

(l) [10 marks] After analysing the data, select the best model from parts (d), (h), and (k) and explain your reasoning for choosing that model. Based on your analysis of the data, recommend a car for your friend to purchase and explain why you chose that particular car. Additionally, suggest a car that you would advise your friend not to buy and provide reasons for your decision.

(m) [10 marks] Your friend is selling a Toyota Corolla with a 2015 year of manufacture and an odometer reading of 100,000 km. Using the model of your choice, construct a 95% confidence interval for the expected selling price and a 95% prediction interval for the potential selling price of the car. Explain how these intervals can assist in determining an appropriate asking price for the car. Based on your analysis, what would be your recommended asking price for the car?