MATH224 - Data Analysis Project

Part III: Regression Modeling and The Whole Story
Total 40 Points Possible
Due Date: 05/08/2022

Instructions: Use the R Markdown report outline template your created in RStudio cloud during the lab to write a final report that addresses the following tasks. Knit your R Markdown code as a **PDF** report and submit your report (which should contain all code lines and answers to the questions below) under "Data Analysis Project - Part III" in Blackboard by the due date.

In Part I of the project, understood the dataset, identified research questions and performed exploratory data analysis. In Part II, you continued to explore your data and used methods of statistical inference to answer questions about the population parameters based on your dataset. In this part of the project, you will focus on the following tasks:

1. **(25 points)** Use correlation and linear regression to describe and model the associations between the outcome (response) variable that you identified in Part I of the project and the explanatory (predictor) variables you identified in the same part of project. Make sure to address the following:

    a. **(6 points)** Identify all numerical (quantitative) variables in the dataset. Then construct a Scatter Plot to show the relationship between the outcome (response) variable (e.g., house price, body weight, income, etc.) and each explanatory variable. First use `select()` to select only the outcome variable and the numerical explanatory variables from the dataset. Then use the function `ggpairs()` from package `GGally` (you need to install this package first) to create the scatter plots and compute corresponding correlation coefficients. Describe the associations between the pairs of variables based on the scatter plots. Describe the associations between the pairs of variables based on the correlation values. Which pair has the strongest linear association compared to the others?

    b. **(5 points)** Develop a multiple linear regression model to predict the outcome (response) variable using all the relevant explanatory variables. Write the model equation. For example,

    $$\widehat{CarPrice} = \beta_0 + \beta_1 \times \text{City MPG} + \beta_2 \times \text{Horsepower} + \cdots + \beta_p \times \text{Engine Size}.$$

    Use the `lm()` function to estimate the model using your dataset and report the summary of the model using the function `summary()`.

    c. **(4 points)** Write the estimated regression equation. For example,

    $$\widehat{CarPrice} = -8.81 - 0.34 \times \text{City MPG} + 0.13 \times \text{Horsepower} + \cdots + 1.74 \times \text{Engine Size}.$$

    Interpret the results of the regression model obtained in the context of your dataset. Make sure to describe how each explanatory variable affects/correlates with the outcome variable (the estimates of coefficients would show this effect).

    d. **(4 points)** Describe the quality of the model by reporting and commenting on the percentage of total variation in the outcome (response) variable that is explained by the explanatory variables collectively. Is the whole regression model statistically significant?

    e. **(3 points)** Are there any explanatory variables that should be removed from the regression model you built above? If so, explain why and re-estimate the regression model after removing such variable(s). Report the summary of model results using the function `summary()`. How does the reduced model compare with the full model from above: compare values of Adjusted R^2 for the two models?

f. **(3 points)** Use the function `plot()` to report residuals plots of the final model and comment on these plots. Do the plots show that all assumptions of the regression model are satisfied?

2. **(15 points)** Put together all your project work from Parts I, II and task one above into a single **story-telling report** and submit it as your final project report. Make sure to use the R Markdown report outline template your created in RStudio cloud during the lab at the beginning of semester. The template is also available in Blackboard along with a sample final report.