MM916 Regression Modelling Project

06/11/2023

Project Overview

In this project you will analyse and explore a data set that examines various factors that may influence house prices in Chicago. You will be expected to use your own judgement, as well as the course content in order to come up with a predictive model for house price. The main purpose of this project is to generate an optimal linear model that can predict the house prices using all the predictors provided in the dataset. In addition, you will also need to discuss and interpret the model outputs as fully and clearly as possible.

Data

The CSV file **real_est.csv** contains data on prices and related information for 157 houses in Chicago:

- Price : price of house (unit: 10K US Dollars)
- Bedroom: number of bedrooms
- Room: number of rooms
- Space: size of house (in square feet)
- Lot: width of a lot
- Tax : amount of annual tax (unit: US Dollars)
- Bathroom : number of bathrooms
- Garage : number of garage
- Condition: condition of house (1 if good, 0 otherwise)

Report guideline

The report should be concise and include only relevant information and output. It should be produced in the format of a standard scientific report (maximum 25 pages including R output and code; font size 12; single spacing; 2 cm margin). i.e.

- Introduction (5 marks)

This should include brief details on the study, the main objectives, and hypotheses of the study.

- Materials and Methods (10 marks)

Include information on how you explore and visualize the key relationships in the data (e.g., via scatterplots and correlation coefficients), how you perform variable selection and the underlying rationale and criteria, the details of the statistical models employed (including model assumptions), as well as the statistical software being used and the significance level set for any statistical tests.

Include information on how you check the model assumptions and consider whether any data transformations are needed to satisfy the model assumptions.

- Results and Discussion (25 marks)

Describe clearly about the steps you perform variable selection to build appropriate models for prediction, how you assess the validity of your model assumptions and how you select the final model.

Provide a clear description of the output including the key relationships in the data and model coefficients.

Discuss whether the assumptions of your statistical model have been met and how you modify the model if necessary.

Provide the mathematical equation of the final model and interpret the model.

Compare the performances of different models using appropriate metrics.

- Conclusion (5 marks)

This should concisely summarise the findings with some comments and/or suggestions.
Specifically, your work should address the following questions:

1. Which predictors have significant effects on the house prices?
2. Does any predictor have nonlinear effects on house price?
3. Is there a significant interaction between **Condition** and any other predictor on the house prices?

Marks are also awarded for clarity and presentation of the report (5 marks). Figures should be suitably labelled, sensibly scaled and cropped with appropriate figure captions. Numerical R outputs used to answer questions should be neatly presented in tables or in the text. Remember to explain and interpret each table and figure shown in the report.

You should submit your R script along with your report in the assignment submission on MyPlace. You may use an appendix for supplementary plots and tables, however these *must not* contain information that is relevant to the key points being made in the report.