STAT 303-2 Assignment 3 Template

January 30, 2022

1 STAT303-2: Assignment 3

1.1 Instructions:

a. You may talk to a friend, discuss the questions and potential directions for solving them. However, you need to write your own solutions and code separately, and not as a group activity.

b. Do not write your name on the assignment. (1 point)

c. Export your Jupyter notebook as a PDF file. If you get an error, make sure you have downloaded the MikTex software (for windows) or MacTex (for mac). Note that after installing MikTex/MacTex, you will need to check for updates, install the updates if needed, and re-start your system. Submit the PDF file. (1 point)

d. Please include each question (or question number) followed by code and your answer (if applicable). Write your code in the 'Code' cells and your answer in the 'Markdown' cells of the Jupyter notebook. Ensure that the solution is well-organized, clear, and concise (3 points)

1. It's easy enough to identify different sections of the homework assignment (e.g., if there are different sections of an assignment, they're clearly distinguishable by section headers or the like)

2. It's clear which code/markdown blocks correspond to which questions.
3. There aren't excessively long outputs of extraneous information (e.g., no printouts of entire data frames without good reason)

This assignment is **due at 11:59pm on Wednesday, January 9th**. Good luck!

Submissions will be graded with a maximum of **52 points** – 47 points for code & answers, 5 points for anonymity and proper formatting.

1.2 Part 1

The datasets *house_feature_train.csv*, *house_price_train.csv*, *house_feature_test.csv*, and *house_price_test.csv* provide data on housing features and prices.

This part is worth 23 points overall.

(1a) Using house_feature_train.csv and house_price_train.csv, fit a multiple linear regression model without transformation to predict *house_price* based on *distance_MRT*, *latitude*, and *longitude*, *house_age*, and *number_convenience_stores*.

Print the model summary. What is the R^2 value?

(2 points for code, 1 point for answer)

(1b) Obtain the residuals and plot them separately against fitted values and each of the five feature variables. Make one plot including the 6 subplots.

(3 points for visualization)

(1c) Comment on the plot of residuals against fitted values. Does the model violate the assumption of linearity? Does the model violate constant variance assumption?

(3 points for answer)

(1d) Comment on the plot of residuals against the predictor variables. On the basis of these plots, should any further modifications of the regression model be attempted?

(2 points for answer)

(1e) Calculate the RMSE using the test datasets for the model constructed in (a).

(2 points for code)

(1f) Using appropriate transformation(s) and/or variable interaction(s), update the model in (a) to obtain a model that has an R-squared of at least 80%, and a RMSE (Root mean squared error) of at max $350k on test data.

Print the model summary and report the R-squared, and RMSE on test data.

(5 points for code)

Note:

(1) House prices are provided in thousands of dollars. A value of 556 in the *house_price* column indicates a house price of $556k.

(2) The test datasets are *house_feature_test.csv* and *house_price_test.csv*.

(3) R-squared is computed on training data, and RMSE is computed on test data.

(1g) Are the assumptions of linearity and constant variance of errors satisfied in the model developed in the previous question? Make the appropriate plot and use it to answer the question.

(3 points for visualization, 2 points for answer)

1.3 Part 2

The datasets `Austin_Affordable_Housing_Train.csv` and `Austin_Affordable_Housing_Test.csv` provide data on housing development projects that have received funding from the Affordable Housing Development Fund in Austin, Texas. The city provides property developers with tax credits and other forms of funding in exchange for agreements to set housing prices (e.g. rent) below market rate.

Each row represents a housing development in Austin. Variables include the amount (USD) provided by the city, the status of the housing project, the number of housing units, the period of affordability, and more. The data provided is a modification of an Affordable Housing Inventory found at https://data.austintexas.gov/Housing-and-Real-Estate/City-of-Austin-Affordable-Housing-Inventory/x5p7-qyuv.

Let's say that you're hired by the city as a consultant to work with subject matter experts in their Housing and Planning Department.

General Hint: For written sections, writing "it depends" (along with an explanation) often characterizes a good answer.

This part is worth 24 points overall.

(2a) Suppose you run the line `status_vars = pd.get_dummies(housing_dataframe["Status"])`, append the columns of `status_vars` to your original data frame, and use the columns as predictors in a linear regression model. What potential problem would you likely be introducing into the model? How could it affect your results?

(2 points for answers)

(2b) Suppose that a subject matter expert recommends using the variables *Total_Units*, *Total_Affordable_Units*, *Total_Accessible_Units*, and *Market_Rate_Units* as predictors in your model. From a regression modeling standpoint, does this sound advisable? Produce metrics to quantify the potential impact of including the four predictors in a model. Interpret at least one of the metrics you provide, both statistically and in the context of the problem.

(2 points for code, 2 points for answer)

(2c) Say that the subject matter expert agrees to use *Total_Affordable_Units*, *Affordability_Expiration_Year*, and *Units_Under_50_Percent_MFI* as predictors for *City_Amount*. Fit the appropriate model (without transformations). Then interpret the results associated with *Total_Affordable_Units*, as well as the overall model fit.

(1 point for code, 2 points for answer)

(2d) Using visualizations, investigate whether the model you fit in (2c) yields outlying observations. What count and proportion of observations would you classify as outliers?

Note: Show separate plots for both - residuals and studentized residuals. However, consider studentized residuals when identifying outliers.

(2 points for code, 1 point for answer)

(2e) Based on your results in (2d), would you choose to remove outlying observations? Briefly, why or why not?

(1 point for answer)

(2f) Consider a scenario in which the model will be used by property owners seeking to predict the amount of money they may receive from the city of Austin. How would this change, support, or complicate your answer in (2e), if at all?

(1 point for answer)

(2g) Say that the model will be used by a team of sociologists seeking statistical evidence at the $\alpha = 0.01$ significance level that a property's affordability expiration year has an effect on the amount of money issued by the city of Austin? How would this change, support, or complicate your answer in (2e), if at all?

(1 point for answer)

(2h) Determine whether the model you fit in (2c) contains any high-leverage points. Produce a visualization, then report the count and proportions of observations that are high-leverage (defining an observation as "high-leverage" if its leverage statistic is greater than four times the average leverage statistic).

(2 points for code, 1 point for answer)

(2i) Based on your results in (2h), would you choose to remove high-leverage observations? Briefly explain, why or why not?

(1 point for answer)

(2j) Identify and remove any influential points from the training data and refit the model. How does removing influential affect the model, if at all?

Think about using the model summary, and the test data provided.

(3 points for code, 2 points for answer)