NATIONAL UNIVERSITY OF SINGAPORE

Department of Statistics and Data Science

ST1131 Introduction to Statistics and Statistical Computing

(Semester 1 : AY 2023/2024)

Individual Assignment

Due Date: 23:59 pm, Sunday 12 November 2023

INSTRUCTIONS TO STUDENTS

1. Students are supposed to submit the answers on time. Any submission after the due time of the due date are marked as late.
2. 10% of the given mark will be deducted for each 2 hours late in submission.
3. **No extension on the deadline for any circumstances.**
4. Students are required to complete this assignment individually.
5. Submission is done online (under Assignments on Canvas).
6. Your submission has **two separate files**. One is a .pdf file of report, and the second file of the R code. Make sure that there is no error when the graders open and run your R code file.
7. Be sure to lay out systematically the various parts and steps in your working.
8. Your submission files should be named as `A0123456B.pdf` and `A0123456B.R` where A0123456B is your student number.

The price of a house depends on many factors. The data given in the file `house_selling_prices_OR.csv` (Canvas/Files/Data) concern the selling price of a house relating to other variables given in the data. Description of the columns in the file is given in Table 1.

Purpose of this assignment: Write a report to propose a linear regression model for the response variable. Investigate if the proposed model is adequate. Propose and fit a new model with a transformation on the response or regressor(s) if it is needed.

Variable	Description
House.Price	original house price in dollar
HP.in.thousands	house price in thousand dollars
House.Size	house size, in square feet
Acres	size of the whole lot of land including the house, in acres
Lot.Size	size of the whole lot, square feet
Bedrooms	number of bedrooms
T.Bath	number of bathrooms, (0.5 means bathroom without toilet)
Age	house's age, in years
Garage	1 = yes, 0 = no
Condition	0 = good; 1 = not good
Age.Category	O = old; M = medium; N= new; VO = very old

Table 1: Variable description

Suggestion for the main part of the report

Part I Exploring the variables

1. Summarize the response variable using summary statistics, figures and/or plots. Comment if it is suitable to fit a linear regression model for this response.

2. Check the association between the response and other variable (using tests and/or plots where it is needed). Comment on the strength of the association if possible. This step is to identify the potential regressor(s) for the model.

Part II Building Model

3. Propose regressors for the starting model. Use R to fit and write down the fitted model (called M_1).

4. Check if model M_1 is adequate using residual plot. Does it have any outlier or influential point?

5. Check if each regressor in model M_1 is significant. Any regressor that is highly non-significant? If yes, what is your proposal?

6. What is/are the next step after assessing the adequacy and the goodness-of-fit of your starting model (such as: transforming response, transforming regressor, adding or removing regressor, etc.)?

7. State clearly what is the choice of your final model (called M_n). Interpret the effect of each variable on the response in the model M_n.

8. Note 1: Each student must report at least two different models: initial model (M_1) and final model (M_n).

9. Note 2: Step 4-5 above should be repeated for each model that you consider, however you just need to report and show your analysis of step 4-5 for the initial model and the final model.

10. Note 3: Each student might have few models in between the starting model M_1 and the final model M_n, however you don't have to report all the between-models. Choose to report one to two between-models only.

11. Note 4: Each student might have a different starting model M_1 and might choose different model to be the final one. However, you need to justify your choice clearly.

Format of the report

1. Your report is a .pdf file, limited to **no more than SIX printing pages, font size 12**.

2. Table and/or figure in the report should be numbered clearly.

3. If you submit the report without submitting R code file, your mark will be deducted by half of the mark given to your report.

4. If you add any R code into your report, it will still be counted within the six pages allowed. Hence, it's advised not to add R code into your report.

END OF ASSESSMENT