NATIONAL UNIVERSITY OF SINGAPORE

Department of Statistics and Data Science

ST1131 Introduction to Statistics and Statistical Computing

(Semester 2 : AY 2022/2023)

Individual Assignment

Due Date: 23:59 pm, Saturday 15 April 2023

INSTRUCTIONS TO STUDENTS

1. Students are supposed to submit the report 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. **Students are required to complete this assignment individually.**
4. All submission is done online.
5. Your submission has **two separate files**. One is a .pdf file of report, and the other one is an .R file of the R code. Make sure that there is no error when the graders open and run your R code file.
6. Be sure to lay out systematically the various parts and steps in your working.
7. Your submission files should be named as `A0123456B.pdf` and `A0123456B.R` where A0123456B is your student number.

The price of a HDB resale flat in Singapore depends on many factors. The data given in the file `hdb-2012-to-2014.csv` (Canvas/Files/Data) concern the selling price of HDB resale flats to some variables given in the data. This data set is extracted from published website [1], from 2012 to 2014.
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.

[1] `https://data.gov.sg/dataset/resale-flat-prices`

Variable	Description
month	the month when the flat was sold
town	the town where the flat belongs to
flat_type	the type of flat
street_name	name of the street on the address of the flat
storey_range	range of the storey where the flat is at
floor_area_sqft	the area of the flat's floor in square feet
floor_area_sqm	the area of the flat's floor in square meters
flat_model	model name of the flat
lease_commence_date	the year when the lease of the flat started$^{(*)}$
resale_price	resale price of flat, in SGD

Table 1: Variable description. (*) For Singapore HDB flats, the lease is limited to 99 years.

Suggestion for 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. For explanatory variable(s): you can remove the one(s) you think that it's not important; you also can re-categorize a variable if it's helpful for building model.

3. 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

4. Propose regressors for the starting model. Use R to fit and write down the fitted model (called M_1). Report the goodness-of-fit of this model and your comments.

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

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

7. 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(s), adding or removing regressor(s), etc.)?

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

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

10. 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.

11. 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.

12. 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. You must provide a **report**, not a list of the answers for the questions above.

2. Your report is a .pdf file, limited to **no more than SIX printing pages, font size 12**. Any parts from page 7 onward will not be graded.

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

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

END OF ASSESSMENT