STAT3022/3922/4022: Project

Deadline: 23h59 Sunday 29 May 2022

The project aims to help you apply knowledge and techniques covered in the class to analyze a real dataset. You will work on a group of three to four. The group members and the datasets have been randomly assigned and they are available on Canvas.

- Please inform the lecturer if you are not able to contact your group members.
- The name of the csv data files assigned to your group should have your group number at last. Please make sure you work on the right dataset.
- Please take a look at the data dictionary to know more about the variables that are included in the dataset.

1 Project guidelines

Your group will analyze the assigned dataset by building multiple linear regression models with the indicated outcome variable in the dictionary. You are not allowed to change this outcome variable.

Together, your group should submit a report written on Rmarkdown and present your report in a 10 minutes video. More details on it can be found on the next session. Below is the minimum requirement for a complete report:

1. Data description and visualization: Please provide and describe summary statistics, correlation matrix/plot, and any other characteristics of the variables in the dataset that can be relevant to model building (for example, missing data, if any). You are encouraged to search the literature for domain knowledge. Please be creative in using visualization techniques in this section.
2. Model building: Please describe in detail every step that you build your models. Justify all your steps carefully. Below are minimum expected content that should be included in the model building process:

- Model inference, including confidence interval, hypothesis testings, etc.
- Variable selection: use either forward, backward, or bidirectional search.
- Model evaluation: you should fit at least two multiple linear regression models and compare them based on different metrics (prediction, information criteria, etc.)
- Identification of unusual observations, including high-leverage, outliers, and influential points and discuss your findings. Please note that this analysis may be repeated for each model.
- Polynomial or interaction terms: should be explored and discussed whenever necessary.

3. Explain and implement at least one statistical method/model to improve the multiple linear regression model that is not covered in the lectures. In presenting this method, you need to explain:

    - How does your chosen method conceptually improve the estimator/model? Please focus more on the concept, rather than the technical details.
    - How does a certain characteristics of the dataset/current model motivate the use of the chosen method? For example, you may say "the dataset has a high level of multicollinearity, so this new method can lessen the multicollinearity issue."
    - Note that it is not sufficient to just write "we implemented the method as suggested from a reference paper" - you need to explain according to the two above points.

    Some suggestions for the methods that are used to improve linear regression models are provided include weighted least square, robust regression, Box-Cox transformation, ridge regression and lasso, additive model, projection pursuit regression, and multiple imputation (for missing data). These methods are extensions of the ordinary least square estimators or the MLR itself. They are widely discussed on the internet or in reference books, and are readily implemented in some R packages.

2 Report and video

Your group should submit a report in Rmarkdown, with all the codes being embedded in the document (i.e don't create an Appendix to put all the codes there). Please include in

your report the declarations of individual contribution for each group member at the end of the report.

Furthermore, your group will need to create a video presentation of your report with all your group members.

- The video should be no more than 10 minutes long, and every group member has to present in the video.
- While there is no page limit for the report, your group must go through all parts of the report in the video, highlight notable findings and explain every step *conceptually* rather than in mathematical details.
- Separate presentation slides are not allowed.

Both the report and the video should be uploaded to Canvas before the deadline.

3 Evaluations

Unless special circumstances arise, everyone in the group will get the same mark for the project. Your report and video will be evaluated based on:

- Technical details (70%), including how you build, compare regression models and discussion of model assumptions and limitations.
- Visualization and presentation of the report and the video (30%).