STAT 6021: Linear Models for Data Science

Project Instructions

Important Dates:

Choose team members: 5 members (Recommended: Find team members by July 17th)

Data Source Selection (Recommended: Find dataset by July 19th)

Overview: this project should represent an impressive synthesis of your ability to find an interesting dataset, pose interesting research questions, explore the data using summary measure and visualizations, build a linear prediction model and present results professionally.

Learning Objectives:

1) Demonstrate the ability to find a reputable data source: the dataset should have at least 4 numeric variables and at least three categorical variables. Choose something your group is interested in – don't be afraid to pick unpopular topics.

2) Define specific questions that you would like to answer with the data: come up with important and interesting research questions that could be answered within the dataset using visualization/summary. Also, come up with two to three interesting research questions that could be answered with a linear predictive model. Questions should be nontrivial in nature.

3) Data preparation and cleaning: some data cleaning. Depending on how clean your dataset is, this might include: renaming columns, removing columns or rows, mutating columns, subsetting (or filtering) the data based on interesting questions to investigate, etc.

4) Visualize and summarize the pertinent variables in the data: your analysis should include a robust summary and visualization of pertinent variables from your dataset.

5) Regression: Use key variables in your dataset to build models and communicate the model

    - Build one or two multiple regression model to help answer an important research question. Use variable selection techniques to arrive at your final model. Check model assumptions. Assess the usefulness of the model.

    - Build one logistic regression model to help answer an important research question. Use variable selection techniques to arrive at your final model. Check model assumptions. Assess the usefulness of the model.

6) Present to the class: your group will do a **5-6-slide** presentation. The first slide will consist of a suitable topic for your project, group number and a list of project group members in alphabetical order of last names. The next slide will have key visualizations and variable summaries. The remaining slides will focus on research questions and how it was investigated using a linear predictive model. Presentations will be 8 mins long. Please practice to ensure you finish your presentations on time.

Project Deliverables:

Turn in the following via Canvas, under Assignments.

1. Presentation slides saved as a PDF or HTML, if possible.
2. Original Dataset in CSV or Excel format.
3. Compiled R codes: R Script or R Markdown file or HTML file.

Evaluation:

The project and presentation are worth 100 points. The scores for each group will based on the following rubric:

Final Project Rubric			
Points	**1**	**5**	**10**
Introduction and Dataset Summary	Complete lack of introduction and dataset summary.	Introduction and dataset summary are sparse and incomplete.	Effective and complete introduction and dataset summary that you would expect to see on a professional report.
Key visualizations/ Summaries	• Completely trivial exploratory data analysis. • Poorly designed chart.	• Decent exploratory data analysis. • Decently designed chart with some issues and poor reasoning for justification.	• Interesting and insightful exploratory data analysis with appropriate logic. • Well-designed chart. Chart design is well justified.
Research question for multiple regression model	• Research questions are not justified by model. • Model assumptions are not met. • Lack of assessment on how the model performs.	• Research questions are somewhat justified by model. • Model assumptions are somewhat met. • Decent assessment on how the model performs.	• Research questions are clearly justified by model. • Model assumptions are clearly met. • Model assessed on performance.
Research question for logistic regression model	" "	" "	" "
Prediction	Model is not used for prediction	Model is used for prediction but unreasonably.	Model is used to make a reasonable prediction including interval prediction.
Conclusion	Key findings are unjustified and not relevant to research questions.	Key findings are somewhat justified but not tied in with research questions.	Key findings are justified, coherent and ties with the research questions.
Timing/ Completeness	Presentation does not present significant portions of the project within the time limit.	Considerable portions of project not presented.	Summarizes all important information within the time limit.

Communication	Extremely difficult to follow the presentation.	Sections of the presentation are difficult to follow and understand.	Presenters demonstrate highly effective communication and presentation skills.
Ambition	Project scope is less than the bare minimum.	Project demonstrates reasonable work but there is little evidence of effort besides the absolute bare minimum.	Project scope, presentation, and difficulty is highly ambitious.
Organization	Poorly structured presentation and slides.	Decently structured presentation and slides.	Well-structured presentation and formatted slides that effectively presents and summarizes the data.