Final Project

STAT 170 —Regression Analysis , Fall 2023

Due Date:

Background and Requirements

The Final Project is designed to give you practical experience in statistical modeling concepts and tools, using real-world data in R. Specifically, you will be picking a data set, analyzing some parts of the data, determining a few questions to ask regarding the data, running models, and analyzing those results.

In groups of 3-4, you will be tasked with turning in the following

- Introduction: due 11/16/2023
- Analysis 1: due 11/21/2023
- Analysis 2: due 12/6/2023 11:59pm
- Presentation: due 12/8 11:59pm
  - Watch presentations: 12/9 3:00-6:00pm
- Final Project Report: due 12/14/2023 at 11:59pm

Data

You will choose one of three datasets to analyze in your report. **Each dataset can by analyzed by a max of 9 groups on a first come first serve basis**. The datasets are the following:

- *insurance.csv*: 1,338 observations with 7 variables. Contains demographic information and costs regarding medical patients. You can find more information about the dataset here: Link
- *mlb2023.csv*: 766 observations with 31 variables. Contains hitting statistics of Major League Baseball players from the 2023 regular season. You can find more information about the dataset here: Link
- *usedcars.csv*: 301 observations with 9 variables. Contains general information about used vehicle sales listed on different websites. You can find more information about the dataset here: Link

Group Guidelines

- Groups will consist of 3-4 classmates.
- Each team member will be assigned a role: Script, Lead Data Analyst, Facilitator.

  - Script: Responsible for the compiling and submitting of weekly progress reports and final reports (Introduction, Analysis 1, Analysis 2, Presentation, Final Project Report)
  - Lead Data Analyst (LDA): Responsible for proposing and leading the data analysis. Note that this implies the LDA will have thought out what is missing from the project analysis and what is needed to make sure final project is submitted within the proposed deadline.
  - Facilitator: Responsible for facilitating the group meetings during discussion and hosting outside group meetings.

- Roles will be assigned by TAs and will change on a weekly basis.
- Contribution to project will be assessed via an anonymous evaluation required upon submission of final report. Make sure to play a significant role in the project development as the participation grade will constitute a non-trivial portion of your grade. Note that the participation grade will include attendance of discussions.

Discussions

All discussion sections remaining will be used as working time for the projects

- Discussions will be used to facilitate project progression.
- Submission of weekly progression reports are required and must include:
  - Assigned team member roles,
  - R-Code of analysis performed,
  - Summary of findings.
  - Goals for next week

Weekly Progress Report Guideline

- Introduction: due 11/16/2023 11:59pm
  - The research question you wish to explore, such as subject matter you're investigating, the motivation for your research question (citing any relevant literature), and your hypotheses regarding the research question of interest.
  - Describe the dataset inclduing the observations in the data, the variables, and how the data was originally collected (not how you found the data but how the original curator of the data collected it.)
- Analysis 1: due 11/21/2023 11:59pm
  - Description of the data and response variable
  - Exploratory data analysis- scatter plot of response variable with each independent variables under consideration, multicollinearity analysis, summary of data using descriptive statistics.

  - Brief explanation model selection procedure, Residual Diagnostics, Transformation if needed

- Analysis 2: due 12/6/2023 11:59pm
  - Brief explanation your exploration on seconder order and interaction terms to add to the final model, including multicollinearity analysis, the residual Residual Diagnostics, Transformation if needed for the final model.
  - output of the final model. Note that the final model will be the result of numerous iterations and trying different models. You can include the other models you consider in the "Additonal work" section.
  - Discussion of the assumptions for the final model
  - Interpretations and interesting findings from the model coefficients
  - Additional work of other models or analysis not included in the final model.

Presentation Guideline

The presentation will consists of two components: (1) Slide deck and (2) Video presentation

- Slide deck: The slide deck should have no more than 6 content slides +1 title slides
  - Titlle Slide
  - Slide 1: Introduce the topic and motivation
  - Slide 2: Introduce the data
  - Slide 3: Highlights from EDA
  - Slide 4: Final model
  - Slide 5: Interesting findings from the model
  - Slide 6: Conclusions + limitation/future work
- Video Presentation: The video should be no more than 3 minutes. The following are good examples, and not so good examples along with suggested rubric scores.
  Speech 1

  Organization – 4, Language – 4 , Delivery – 4 , Supporting Materials – 4 , Central Message – 4
  Speech 2

  Organization –4, Language – 3, Delivery – 4, Supporting Materials – 3, Central Message – 4

  Speech 3

  Organization –2, Language – 3, Delivery – 1, Supporting Materials – 2, Central Message – 2

Final Report Guidelines

The goal of the final write up is to demonstrate your ability to use regression analysis to answer meaningful questions,your proficiency in R, and your proficiency interpreting and presenting

results. Your write up should focus on the main conclusions and interesting findings that you derive from your analysis. It should not just be a list of every model you tried and interpretation of every model coefficient.

The final report should contain the following sections but **4-6 pages**:

- Introduction
  (1) clearly stated research question and why this question is of interest, (2) Describe the data and definitions of key variables. (3)exploratory data analysis and summary of data using descriptive statistics.
- Regression Analysis
  Final model proposed with rational as to why chosen including how you conducted model selection, interactions you considered, and any variable transformations. Also include a brief discussion of the model assumptions, diagnostics, and any model fit statistics (e.g. R2 etc.)
- Conclusion
  Any relevant prediction and/or conclusions from your model, i.e., answer to research question and recommendations. This should not just be a list of coefficient interpretations but rather use the interpretations from the model to support your conclusions.
- Limitations project Limitations, e.g., assumptions not met and how they affect the analysis, R2 low, variables statistically insignificant etc...
- Appendix

  Appendix with full R code showing analysis, model assumption checks, residual diagnostics etc.... There is no page limit on Appendix, but it should still be neatly organized and easy for the reader to navigate.

Report must be well organized and resemble reports for professional submission outside of academia.Put the final write up in a .Rmd file called "final-writeup-groupX.Rmd", and output it as a PDF. The document should be neatly organized, and all code and warning messages should be suppressed, i.e. not visible in the PDF.

Grading Guildeline

- Introduction:10 pt
- Analysis 1 : 10 pt
- Analysis 2 : 10 pt
- Final Report: 40 pt
- Presentation (slide deck): 10 pt
- Presentation (Video): 10pt
  Each group will be assigned 3 presentations prior to the viewing and will give 4 point score on Organization, Language, Delivery, Supporting Materials and Central Message. Evaluations will be due after watching all group's video (approximately 70 mins).

- Team peer evaluation 10 pt
  You will be asked to fill out a survey where you rate the contribution and teamwork of each team member out of 10 points. You will additionally report a contribution percentage for each team member. If any individual gets an average peer score indicating that they did less than 10% of the work, this person's project grade will be assessed accordingly.