Final Project
Regression Fall 2024

Due: Saturday 10/12 @ 11:59pm

The final project is designed to test the students' ability to perform a complete linear/logistic regression analysis and communicate the results to a broader audience.

Option A:

Your task is to find or assemble a dataset suited to a regression problem which comes with at least 2 numeric variables, at least 2 categorical variables and at least 100 observations. Choose one variable as the outcome response (numeric or binary), perform a regression analysis using the data, then discuss the results in the final report.

Submission: A final report in .pdf, and a notebook with all your Python code. Each group will have one submission. You will also submit a breakdown of work and peer review score for each of your project partners individually and confidentially.

Report Description:
The final report will be a written report (submitted as a .pdf and a .ipynb supplement with relevant code) that clearly communicates the information below:

- A description of your dataset, including the data source, variable descriptions, context, motivation, etc.
- A clear statement of the research questions, and a summary of methods being used in the analysis.
- An exploratory data analysis which may include but not limited to plots, summaries, individual tests, etc. which motivate the hypotheses you explore in the ensuing analysis.
- Either a multiple regression analysis or a multiple logistic regression analysis (or both, depending on your scope), including a discussion of your model selection process and an assessment of diagnostics and what they imply about model assumptions.
- A discussion of your final model and a summary for your findings/results in the real-life context of the problem. Include a discussion of potential problems with this data and analysis, i.e problems like multicollinearity that you can not remedy further but may affect the sensitivity of your tests.

Option B:

Your task is to **1. write a blog post** and **2. create an accompanying interactive graphic or app** whose purpose is to educate a reader (e.g. a fellow MSDS student) and help them develop intuition about one of the topics we have covered in this course. For this option, you should focus on doing a "deep dive" into the topic – the scope your blog post/app must extend past the basic material we have covered in class. ***If you choose this option, I highly recommend discussing your topic with me before you get started to ensure your choice is well-suited for this project.***

With this option, you have freedom with where you want to go with your lesson—try to be creative and illustrate **your point of view**, not just what a textbook would say. In other words, ***show me something cool*** about your topic! Tell me **your thoughts** on the topic, how it can be used, what makes it interesting, what are its drawbacks, and so on. ***I am deliberately giving you room to explore the topic in an open and potentially unconventional way—if you just give me a boring, standard textbook lesson of the foundations without incorporating any personal perspective, then you are missing the point!***

Submission: Links to your blog post and app, and either a notebook or Github with all your Python code & scripts. Each group will have one submission. You will also submit a breakdown of work and peer review score for each of your project partners individually and confidentially.

Blog post & app description:
Your final project submission should have the following components, at minimum:

- A clear and motivating introduction to the topic which is accessible to a wide audience (in other words, don't jump straight into technical details).
- A discussion and illustration of the basics (i.e. what we've talked about in class) using either an example dataset or a simulation.
- Multiple sections *(at least 2)* extending the basics to more advanced or detailed aspects of the topic and illustrating them either through other data examples or simulations. You will need to do some of your own research for this part!
- An interactive graphic/app which a reader can use to build their intuition and explore the topic, either with their own data or through simulation.

Grading rubrics:

Option A:

Component	Points
Data Introduction, EDA, & Hypothesis Generation	/15
Model Selection	/15
Model Diagnostics & Assessment	/15
Model Interpretation & Conclusions	/20
Quality of Figures	/15
Collaboration Score (Peer Review)	/20

Option B:

Component	Points
Topic Intro & Motivation	/10
Lesson and Illustration of the Basics	/10
Topic Extensions (at least 2)	/30
Quality of Interactive Visuals/App	/30
Collaboration Score (Peer Review)	/20