Analytical Methods I (ANLY 502) Course
Project: Tell your data story!

We are surrounded by data. As future data analysts, you will be called upon to tell a story, answer questions, and make predictions from a data set. This is what you will do with this project.

Tasks

Data set: Find a data set with the following characteristics

DO NOT USE KAGGLE DATA SETS!

- 1 response variable
- At least 10 explanatory variables
- At least 100 observations

EDA: Do some exploratory data analysis to tell an "interesting" story about data. Instead of limiting yourself to relationships between just two variables, broaden the scope of your analysis and employ creative approaches that evaluate relationships between two variables while controlling for another one.

Inference: Come up with a research question that can be answered with a hypothesis test or a confidence interval. Your question could be used to shed some light on your choice of the "best" linear model. Carry out the appropriate inference task to answer your question.

Modeling: Develop the "best" multiple linear regression model to explain your response variable values.

Prediction: Based on your model, make a prediction using the predict function in R. Also quantify the uncertainty around this prediction.

Deliverables

A. <u>Data set description submitted to Moodle</u>
   - Due by Due Date in Moodle

   General description of data
   - Link to data set
   - List of explanatory variables
   - Size of data set

B. Report

- Due by midnight of Due Date In Moodle
- 3 pages max @ 12 pt font size, arial or times new roman font type
- Your report should be organized with the following parts included and clearly labeled:
    1. ***Introduction:*** a summary of the data set and your goal.
    2. ***EDA:*** any univariate or bivariate summaries worth reporting.
    3. ***Inference:*** Answer the research question you have posed using a hypothesis test or a confidence interval.
    4. ***The "Best" Model:***

        What is the "best" linear model for predicting the response variable? You do not need to explain every step you took to arrive at this model, but should give some indication of why you chose the model you did. If you tried a few different models, how did you settle on one?

        - How well does your model do? What is the percent variation explained?
        - What does your model tell you about relationships between your explanatory variables and your response variable?
        - What conditions do you need for your analysis to hold? What are the implications if some of those conditions are violated.

    5. ***Prediction:***

        Using your best model, make a prediction about a future event from your response variable.

        Include a description of the uncertainty of your prediction.

    6. ***Conclusion:***

        - What is the bottom line from your analysis?
        - How well can you predict your response variable?
        - What are the caveats to your analysis?
        - Does this data set lack information that you would have liked to use?

C. Code

- Due by midnight Due Date in Moodle
- Additional details will be provided later

D. Presentation

- Slides due by midnight Due Date in Moodle
- To be given in during Date Specified in Moodle

- 15 minutes max
- Live synchronous delivery on Adobe connect with ALL team members present
- Scheduling instructions will be provided later

Tips for your report and presentation

This project is an opportunity to apply what you have learned about descriptive statistics, graphical methods, correlation and regression, and hypothesis testing and confidence intervals.

The goal is not to do an exhaustive data analysis i.e., do not calculate every statistic and procedure you have learned for every variable, but rather to show that you are proficient at using R at a basic level and that you are proficient at interpreting and presenting the results.

You might consider critiquing your own method, such as issues pertaining to the reliability of the data and the appropriateness of the statistical analysis you used within the context of this specific data set.

Grading

Grading of the project will take into account:

Correctness: Are the procedures and explanations correct?
Presentation: What was the quality of the presentation and poster?
Content/Critical thought: Did your think carefully about the problem?

Your grade will be roughly based on the following components:

30% report
30% presentation
25% code
10% team peer evaluations
` 5% data set description submitted on Moodle

Team peer evaluation: You will be asked to fill out a survey where you rate the contribution of each team member. Filling out the survey is a prerequisite for getting credit on the team member evaluation. For grades less than 3.0, please provide some explanation. If any individual gets an average peer grade less than 2.0, this person will receive half the grade of the rest of the group.

Honor code

You may not discuss this project in any way with anyone outside your team, besides the professor. Failure to abide by this policy will result in a 0 for all teams involved.