SDS 348 –Project 2

Modeling

Overview

The goal is to apply the skills you developed in this course to a topic of your interest. The second project focuses on modeling.

Instructions

You will submit a knitted R Markdown report (as a pdf) on Gradescope and a short video presentation on Canvas, both by **8am on 4/19/2021**.

Report:
The text of the document should provide a narrative structure around your code/output. All results presented must have corresponding code. Any answers/outputs/plots given without the corresponding R code that generated the results will not be considered for credit. Any outputs/plots that are not commented and referred to will not be considered for credit either. All code contained in your final report must work correctly (knit early, knit often, ask questions!). Please do not include any extraneous code or code which produces error messages. (Code that produces warnings is acceptable, as long as you understand what the warnings mean).

Presentation:
During the video presentation, you will share the title, a description of the dataset, and two main findings of your study. The videos will be peer reviewed meaning that you will watch 4 videos from other classmates and 4 other students will watch your video presentation. Video camera is not required for privacy reason but please consider having it turned on during your presentation. Your video does not have to be high quality, as long as we can hear you and see your presentation materials clearly throughout the entire video. The presentation should not last more than 2 minutes. Use Zoom (or any other software you want) to record yourself while sharing your screen so we can see your presentation materials while you present. To record on Zoom, start a new meeting with just yourself, share your screen and record your presentation. Do a trial run so you know that your audio and video are set up correctly. The intended audience for your presentation should be your classmates, peers who have some little background knowledge about statistics but have no background about your project topic. You can either use your report to present or use a slide presentation (with 5 slides maximum). You do not need to submit your slides.

Data

You can use the dataset you created for Project 1 (or use only one of them, that's fine too) or you can use a new dataset that should meet the following requirements: the dataset has at least **50 observations** (i.e., rows), at least **5 variables** with at least **2 numeric** and at least **1 categorical** (with no more than 4 categories, you can regroup some of the categories if necessary). Refer to the guidelines on Project 1 for where to find data (only datasets discussed in class or assignments are not allowed).

Guidelines

Here are some guidelines about the different steps to take for modeling your data:

1. **Introduction**. Introduce your dataset and each of your variables (or just the main variables if you have many). How was the data collected? How many observations? Did you have to tidy the data? Why are you interested in exploring this dataset? What do you expect to find?

2. **EDA**. Explore your main variables by producing univariate/bivariate statistics and graphs. In particular, investigate relationships that you are going to test about with MANOVA and a randomization test, and variables included in your regression models. For example, you could create a correlation matrix with univariate/bivariate graphs and correlation coefficients.

3. **MANOVA**. Perform a MANOVA to test whether any of your numeric variables (or a subset of them, if including them all doesn't make sense) show a mean difference across levels of one of your categorical variables.
   - If significant, perform univariate ANOVAs to find response(s) showing a mean difference across groups, and perform post-hoc t tests to find which groups differ.
   - Discuss the number of tests you have performed, calculate the probability of at least one type I error, and adjust the significance level accordingly (Bonferroni correction) before discussing significant differences.
   - Briefly discuss assumptions and whether or not they are likely to have been met.

4. **Randomization test**. Perform a randomization test on your data (that makes sense and is interesting to look at). This can be anything you want!
   - State the null and alternative hypotheses, perform the test, and interpret the results.
   - Create a plot visualizing the null distribution and the test statistic.

5. **Linear regression model**. Build a linear regression model predicting one of your **numeric** response variables from at least 2 explanatory variables, including their interaction.
   - Mean-center any numeric variables involved in the interaction.
   - Create a graph to visualize the interaction between 2 variables on the response.
   - Interpret the coefficient estimates in context (regardless of significance).
   - What proportion of the variation in the response does your model explain?
   - Check assumptions of linearity, normality, and homoscedasticity either graphically or using a hypothesis test (or both).
   - Regardless of meeting the assumptions, recompute regression results with robust standard errors. Discuss significance of results, including any changes from before/after calculating robust SEs if applicable.
   - Finally, compute bootstrapped standard errors. Discuss any changes you observe in SEs and p-values using these SEs compared to the original SEs and the robust SEs.

6. **Logistic Regression**. Build a logistic regression model predicting a binary categorical variable (if you don't have one, create one based on another variable or combination of other variables) from at least 2 explanatory variables (interaction is not necessary).
   - Interpret coefficient estimates in context (regardless of significance).
   - Report a confusion matrix for your logistic regression.
   - Compute and discuss the Accuracy, Sensitivity (TPR), Specificity (TNR), and Recall (PPV).

- Create a graph to plot density of log-odds (logit) by your binary outcome variable.
- Generate a ROC curve (plot) and calculate AUC (either manually or with a package) and interpret the results.

Rubric

For every step, document what your code does (with *#comments*) and make sure to interpret outputs with sentences.

Rubric Item		Points
Report	**Introduction** Give a title to your report and write a narrative introduction to describe the dataset, the variables, and the number of observations. Describe any step you had to take to tidy the data. Expand on potential associations you may expect, if any.	15
	EDA Create summary statistics and visualizations of important variables and relationships between your variables in the dataset.	10
	MANOVA Perform a MANOVA test and if significant perform ANOVAs and post-hoc t tests. Interpret p-values after correction and briefly discuss assumptions.	15
	Randomization Test Perform a randomization test on your data. Include hypotheses, the sampling distribution and test statistic, interpret the results.	10
	Linear Regression Build a regression model to predict a numeric response. Interpret the coefficients, check assumptions and calculate robust SEs and bootstrapped SEs.	35
	Logistic Regression Build a regression model to predict a binary response. Interpret the coefficients and generate a ROC curve, calculate AUC.	30
	Formatting Create the report using R Markdown, with headers for each section; include comments to the R chunk to describe what the code does; include references (datasets, context). The final report is no more than 30 pages.	10
Presentation	**Presentation materials** Texts and figures are easy to read, free of spelling errors, information is presented in a logical structure.	10
	Presenter Presenter is audible, shows enthusiasm for the topic and seems prepared; speech is jargon free and appropriate for the audience.	10
	Content Include title, description of the dataset, and two main findings in context.	5
Peer Review	**Grade and Comment** Peer review videos of 4 other students.	25
	Total	**175**

Undergraduate Statistics Project Competition

You can submit any of your project reports to a project competition organized by The Consortium for the Advancement of Undergraduate Statistics Education (CAUSE) and the American Statistical Association. The purpose of the competition is to encourage the development of statistics and data science skills, to enhance presentation skills, and to recognize outstanding work by undergraduate statistics and data science students.

More information here: https://www.causeweb.org/usproc/

Some considerations for submitting your research project to the USCLAP Competition:

- ✓ The deadline for the summer competition is **June 25, 2021**.
- ✓ You can visit the website, read or watch past winning projects:

https://www.causeweb.org/usproc/eusrc/2020

- ✓ Some examples of winning projects can be found here: https://www.causeweb.org/usproc/projects/winners. Be aware that these are winning projects and represent the best submissions. My expectations for the projects for this course are not for you to win (although it would be pretty awesome if you do!) but to explore a topic of your interest using some of the tools and concepts we have learned.
- ✓ If you collected data yourself over human subjects, you will need to apply for an IRB and be approved.