Team Project

Regression Data Analysis Project

1 Background

In this project you will analyze a data set of your choice using multiple regression. Regression is applicable to a wide range of data sets, so you can select data that answers questions that you are interested in!

2 Selecting the data set

You can select any publicly available data set (many are available online) that meets the requirements below. You can also combine data from multiple sources to create your data set for analysis.

Your data set must:

- Have at least 100 cases. Recall that for a given effect size in the population, the p-value will tend to decrease as as the sample size increases. As a result, your project will be stronger if you have a larger sample. However, beware: when the sample size is very large, you will see statistically significant results that may not be practically significant.
- Have at least 8 predictor variables.
- Have at least one quantitative and one categorical variable.

You may not use any of the following types of data sets for your project:

- A data set for which an existing analysis is published online.
- A data set that is built in to R or from the R `dataset` package.
- A data set that is more than 10 years old, or a data set for which more current data is readily available.

I can make exceptions on a case-by-case basis if you have a data set that you are very interested in using but that does not meet all of these requirements.

3 Proposal

A brief proposal is due on **Monday, April 10, at 11:59 PM** and must describe the following. There is no specific format required for the proposal. You will receive rapid feedback on your proposal so we can help you steer away from questions or data sets that will be too complex or time-consuming to work with. Your proposal is graded all-or-nothing; if your proposal is not accepted then you will receive no credit on the proposal but will have the opportunity to resubmit within one week of receiving the initial feedback so your group can receive full credit.

The proposal should include the following:

1. The research questions you want to answer (i.e., what you are trying to predict, and what variables you think might be helpful in making that prediction).
2. A description of the data set and its source (provide a URL to the data set if found online).
3. The data set you have chosen to address these questions, in CSV format.

4 Report

A full report of your findings is due at **11:59 PM on May 5**. **The paper can be no more than 1500 words.** This is a hard limit: the TA may stop grading after 1500 words! It will be submitted in Canvas and can be in Microsoft Word, PDF, or LaTeX. The rubric below outlines the required sections of your report and what each section should contain. Your report should be accompanied by an R script file and a CSV file of your data set that together reproduce all of the results in your report and presentation. **One** member of your team should submit the report and supplementary files in Canvas.

5 Grading Rubric

This rubric will be used to arrive at your team's project grade. Each member of the team will receive the same grade, unless a member of the team is not participating fully in the work.

Note that no credit is awarded based on the statistical or practical significance of your model, but rather the correct application of the analytical process.

1. Proposal (5 pts)

   Full credit is awarded for an accepted proposal; no credit is given if the proposal is not accepted. If a proposal is not accepted, your group has one week from getting the proposal back to revise and resubmit for full credit.

2. Report (30 pts)

   (a) Background (7 pts)

   (2 pts) Report introduces topic to the reader

   (2 pts) Report describes the data set, including the source of the data and how it was collected

(1 pt) Report indicates what variable is to be predicted and what variables will be considered as predictors

(2 pts) Report poses one or more questions that can be answered by the data set selected

(b) Methods (10 pts)

(2 pts) Report describes the regression model used and the process for arriving at the final model

(2 pts) Data cleaning process is described, and any outliers reported on

(2 pts) Report verifies that all regression assumptions have been met; transformations are described and applied if necessary

(2 pts) The regression model addresses the questions described in the Background section

(2 pts) Explanations are clear and make it easy to reproduce the regression results

(c) Results (8 pts)

(1 pt) Report includes descriptive statistics (e.g., means, SDs) and basic graphs (e.g., boxplots or histograms) are of each variable used in the regression model

(1 pt) Report includes graphs (e.g., boxplots or scatterplots) showing the relationship between the response variable and each predictor variable

(2 pts) Report interprets the individual coefficients of the final regression model, along with the confidence intervals and statistical significance of each

(2 pt) Report interprets R^2 and the residual standard error

(1 pt) Report considers and interprets the overall p-value of the model

(1 pt) The tables and graphs are clear, selected appropriately, and professionally formatted

(d) Conclusions (5 pts)

(3 pts) Report ties back specific results about the coefficients and R^2 to the questions described in the Background section

(1 pt) Report considers the limitations of this data set and of its conclusions

(1 pt) Report considers future work that might be an appropriate extension of this study