Final Project Assignment

Binary Logistic Regression Analysis

TL; DR

1. **Project proposal:** Due Wednesday, March 2
    - One-paragraph summary of your project describing the topic and plan for collecting the data.
    - At least 5 cases from the dataset as proof of concept so that we can determine the viability of collecting the dataset you propose.

2. **Final Project write-up and Presentation: Due** Monday, March 14 @ 6:30pm. Also submit your data a a csv file, one submission per group.

Overview

The goal of this project is to work in pairs or triples (2 to 3 in a group) to carry out both simple (one predictor) and multiple (more than one predictor) logistic regression analyses using a data set that you collect either by observation (carry out an "experiment") or by finding cases from the Web. Guidelines for the data set are that it should have at least 100 observations, a single binary response variable, several (three to six) potential predictor variables. Of the predictor variables, one or two should be binary predictor variables. The model should deal with a topic and question that is interesting to you.

Since several of you have interest in sports statistics. Here is a link to an article written by one of our authors (Robin Lock) that discusses sports projects and has a section on binary logistic regression. One timely example might be to determine whether the outcome of a basketball game (won or loss) could be predicted by court (home or away) and several quantitative performance measures such as field goal percentage, assists, rebounds, etc. A reasonably sized data set on a team of interest could be efficiently collected. Another example of a dataset might be to look up election results by county/precinct from a recent election and identify whether or not the candidate of interest won (response) along with categorical and quantitative characteristics of the location (e.g., unemployment rate, average household income, etc.). You may also choose to analyze your trashball data but be sure that this is something your whole group would like to do. Those you in business and economics might be interested in the dataset described here in a paper about factors determining whether or not a loan should be approved.

Please see me to discuss any ideas you might have. I have a strong bias against surveys for projects like these because you will undoubtedly spend a disproportionately large amount of time on construction of the survey and sampling/interviewing. Email surveys also have notoriously low response rates.

Simple Logistic Regression

Choose a **single quantitative** predictor and carry out a simple logistic regression analysis.

(a) Include the summary output using R to estimate the coefficients of this model.

(b) Interpret the slope coefficient (in terms of an odds ratio) and interpret the test for the slope. Be sure to do these in the context of your data situation.

(c) Compute a 95% confidence interval for your slope and use it to find a confidence interval for the odds ratio. Does your interval include the value 1? Why does that matter?

(d) Show (by hand calculations) how to use the fitted model for predicting the proportion for a couple of typical cases, being sure to explain what you are finding in terms of your data situation.

(e) Include a plot of the logistic (curved) fit (with commentary) (see class notes for example of code).

(f) Show how to compute the Likelihood Ratio G-statistic and use it to test the effectiveness of your model. Be sure to indicate where the *p*-value for the test comes from.

Multiple Logistic Regression

Choose your best model using **at least two predictors**. Try to balance getting a good fit with keeping the model simple (but use multiple predictors, even if one or all are not very effective). This section will require some self-study of Chapter 10 on multiple logistic regression. Please see me if you need help clarifying topics or with R commands. Building on the what you've learned in this class and applying principles from other models (multiple regression) is a valuable skill to cultivate.

(a) Explain briefly the process that led to your choice for a final model. Include summary output from R for your final model.

(b) Show (by hand) how to use the fitted model to predict the proportion for one typical case in your dataset. Be sure to identify the relevant characteristics of that case.

(c) Comment on the effectiveness of each predictor in the model as well as the overall fit. Be sure to indicate what value(s) from the output lead to your conclusions.

Prediction Table

Use either your single or multiple model for this part, whichever you think is better.

(a) Store the predicted probabilities for all of your data cases in a variable. Hint: You can use `fitted(mymodel)` to obtain the values and `mutate()` to add them to the dataset.

(b) Classify each data point as being a predicted "success" (1) if the predicted $\hat{\pi}_i$ is greater than or equal to 0.5 and a predicted "failure" (0) if $\hat{\pi}_i$ is less than 0.5. (Note: The interpretation of success or failure depends on the definition of your binary response. You can add this success/failure variable to your dataset using `mutate()` and `ifelse()`. See me if you get stuck on this.

(c) Look at the classifications for each of your data points and create a 2 x 2 table showing counts of how the data are classified (predicted success or predicted failure) versus their

actual response values. The mosaic command `tally()` should be useful here. Look at the examples in the help for `tally()` if you need help with the syntax.

(d) Comment on the accuracy of the classifications for your data in the table. Did you have many cases that were “misclassified” (i.e., predicted to be “success" when actually a “failure” or vice versa)?

Report

Your report should be written as an RMarkdown document. I will ask you to submit 3 files: (i) PDF of the report, (ii) the Rmd file used to create the report ,and (iii) a CSV file containing your data. **One person from your group should submit these files on Moodle by 6:30pm Monday, March 14, the final exam time for our class.**

Presentation

Each group should also prepare a brief 10 minute presentation of your analysis to share with the class during the Final Exam Period.

Assessment

The project will be scored out of a total of **140 points** broken down as follows:

10 points: On-time submission of satisfactory proposal.

10 points: On-time submission of dataset.

100 points: RMarkdown report

20 points: Peer assessment of presentation.

If you feel a group member is not doing their share, first try to work it out within your group. If that doesn’t work, please speak to me about the situation and I will attempt to resolve it.