**Assignment #2:**

**Advanced Quantitative Methods**

**Logistic Regression**

This data set is a representative of the titanic shipwreck from Kaggle. Kaggle is an online platform for predictive completions, including machine learning and statistics. The below detail is from the [competition page](https://www.kaggle.com/c/titanic):

*The sinking of the RMS Titanic is one of the most infamous shipwrecks in history.  On April 15, 1912, during her maiden voyage, the Titanic sank after colliding with an iceberg, killing 1502 out of 2224 passengers and crew. This sensational tragedy shocked the international community and led to better safety regulations for ships.*

*One of the reasons that the shipwreck led to such loss of life was that there were not enough lifeboats for the passengers and crew. Although there was some element of luck involved in surviving the sinking, some groups of people were more likely to survive than others, such as women, children, and the upper-class.*

*In this challenge, we ask you to complete the analysis of what sorts of people were likely to survive.*

You will be using logistic regression to create a model predicting whether a passenger on titanic would survive or die, based on factors such as gender, class, # of siblings, and age. A description of the dataset is included below. I’ll provide more overview in class.

**Data Set Description**

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | Definition | Key | Notes |
| survival | Survival | 0 = No, 1 = Yes |  |
| pclass | Ticket class | 1 = 1st, 2 = 2nd, 3 = 3rd | A proxy for socio-economic status (SES)  1st = Upper  2nd = Middle  3rd = Lower |
| sex | Sex |  |  |
| Age | Age in years |  | Age is fractional if less than 1. If the age is estimated, is it in the form of xx.5 |
| sibsp | # of siblings / spouses aboard the Titanic |  | The dataset defines family relations in this way... Sibling = brother, sister, stepbrother, stepsister Spouse = husband, wife (mistresses and fiancés were ignored) |
| parch | # of parents / children aboard the Titanic |  | The dataset defines family relations in this way... Parent = mother, father Child = daughter, son, stepdaughter, stepson Some children travelled only with a nanny, therefore parch=0 for them. |
| ticket | Ticket number |  |  |
| fare | Passenger fare |  |  |
| cabin | Cabin number |  |  |
| embarked | Port of Embarkation | C = Cherbourg,  Q = Queenstown, S = Southampton |  |

**Analysis Process**

*Preliminary steps*:

1. Before we do any statistical analysis we should get to know our dataset and our sample. This typically involves running some frequencies and means. The teaching assistant will show you how to do this.
2. Summarize the variables in the dataset. Discuss measures of central tendency and variation.
3. Run frequencies on the following: Sex, sibsp, Pclass, Survived. Report the basic frequencies. Use the description of the dataset to help you interpret the output. This will explain what “0” and “1” codes mean.
4. Analyze how the various variables differ by levels of the Survived variable. This will give you an overview for the modelling.

*Core Assignment*:

Develop a logistic regression model to determine whether class matters and what else predicts likelihood of surviving. Use the model building process described in the lecture notes (the same basic process used to develop a “regular” regression model—that is, brainstorm predictors, put them in order, add one variable at a time, etc). As with the regression assignment include a 2-3 page front memo with a light literature review motivating the problem and analysis and highlighting your bottom line, actionable findings. In the appendix portion you should be sure to explain the substantive findings and their meaning, as well as all of the key statistical attributes (for example the odds ratios, the various measures of goodness of fit, what was significant and what was not in predicting the outcome). Again: be sure to explain the statistics in some detail (though 5 pages for the appendix should be more than enough).

What do the results imply to you? (include all output from final model!!). What are the main predictors of surviving? Who is more likely to make it onto a lifeboat? Does siblings or spouses on board matter? Does this vary by class? What interactions do you see and what do they imply?

Post your code on canvas at the end of the assignment. Be sure to interpret all of the logistic statistics that we talk about in class (I just want to see your final model, but I do want to hear about variables which you tried which were not statistically significant). Remember, there are no standardized coefficients in logistic regression. Use the absolute value of the Z statistic to describe the strongest relationships. Use graphics to reinforce your arguments if you wish (not required)

*Advanced extensions:*

If you want to go beyond the base assignment to expand your skills, here are some ideas. You are required to try at least (2) of these, or just #5 – machine learning extension.

1. Try to fit an ROC curve and try some interactions (many are included on the dataset). The ROC curve is described in the Logistic Regression with R materials.
2. Find the individual or group with the highest and lowest probability of survival. What are their characteristics? Who are they? Does this make sense to you?
3. Analyze the false positive, false negative rate, the sensitivity and the specificity. What happens to the classification accuracy and these statistics if you move down the classification cutoff? What happens to them if you move it up?
4. Analyze the interactions and consider creating more of them. Remember from regular regression that interactions involves multiplying two dummy variables to create a new variable and then adding this interaction variable to see if the two variables depend on each other to create differential relationships—it is a way of testing if the interaction is more than the sum of the two variables. If you try the interactions then be sure to test for multicollinearity. Often interactions cause multicollinearity. If this happens, remove the individual variables that you interacted from the model and just put the interaction in.
5. Test your most successful model with new testing data provided. The testing data is located in canvas as “test\_set”. This is the base of supervised machine learning – we’ve created a model and now we apply it to a test set and check how successful it was. As we discussed in class, often perfect accuracy of our training set can result in overfitting. We see this when we apply it to new data. Did your accuracy improve, or decline? **This exercise alone is sufficient to complete the extension requirement**. You can also submit your model to kaggle on their [submission page](https://www.kaggle.com/c/titanic/submit), however this is not required.

Submit this assignment on time, even if you have some questions.