The Likelihood of Death Using Logistic Regressions

Julie Osborne

DSP 539 Big Data Analysis

Dr. Rachel Schwartz

5/10/19

Abstract

This paper used the National Health and Nutritional Examination Survey Epidemiologic Follow-Up Study dataset, to create a model that was able to predict the likelihood of death in a subject based on certain covariates of interest that are included in the dataset. The data was divided into a train set to create the model and a test set to observe the predictive capabilities of the model.

Introduction

There are many causes that can influence the death of a person such as education, alcohol usage, level of exercise, etc. In this dataset, the area of interest is the relationship these variables have with death. This relationship is particularly intriguing because the dataset does not provide information regarding the cause of a person's death. By using logistic regression, it will be able to capture the categorial nature of the variables and see if there is a relationship between the chosen variables and death.

Data Description

This dataset is based off the National Health and Nutritional Examination Survey (NHANES I) Epidemiologic Follow-up Study (NHEFS). This study evaluated the relationships between clinical, nutritional, and behavioral factors and various outcomes such as morbidity, mortality, and hospital utilization. The individuals surveyed were aged 25-74 years when examined during NHANES I between 1971-1975, and were subsequently followed until 1992.

Methods

Overall, the aim was to evaluate which variables have a statistically significant relationship with death as well as test the accuracy of the model in predicting death.

Since the response variable, death is a binary categorical variable, a logistic regression model was used. The predictor variables that were used were sex, education, alcohol frequency (how often a person drinks alcohol in a year), exercise (how often a person exercises), and race. The glm() function was used to create the model. The model was first run on a train data set, which will come into play when we run the model for prediction.

Once the model was assessed to determine which variables were statistically significant in predicting the likelihood of death, the test data set was used to predict if someone died or not. The predict() function was used. The prediction from the test dataset was compared to the actual results in the test dataset in order to determine the accuracy in prediction of the model.

Results

Before any analysis can be done, each of the response and predictor variables needed to be factored into levels since they are all categorical variables. **Death:** 1: yes, 0: no; **Sex**: 0: male, 1: female; **Education:** 1: 8^{th} grade or less, 2: HS dropout, 3: HS, 4: College dropout, 5: College or more; **Alcoholfreq:** 0: Almost every day, 1: 2-3 times/week, 2: 1-4 times/month, 3: < 12 times/year, 4: No alcohol last year/unknown; **Exercise:** 0: much exercise, 1: moderate exercise, 2: little or no exercise; and **Race:** 0: white, 1: black or other.

Before running any statistical modeling, it is good practice to do data visualization in order to get a feel for the data and examine trends that may not be apparent from just looking at the data. All variables were plotted against the number of deaths as a single variable. Then, every two variable combination was also plotted against the number of deaths. In this paper, only the most interesting graphs will be highlighted. All graphs can be found in the Final_Graphs.pdf. Figure 5 represents the breakdown of number of deaths based upon race. There is large difference in between the number of white deaths vs. the number of black or other deaths. This

unevenness could influence the significance of race as a predictor. Another interesting graph is Figure 7, which shows the number of deaths by sex and alcohol frequency. It is interesting to see that more males die whom drink at higher alcohol frequency, more females die whom have a lower alcohol frequency, and its approximately equal between the sexes at a moderate alcohol frequency. Figure 10 displays the number of deaths by alcohol frequency and level of education. Of the five alcohol frequency levels, subjects with a high school education level have the greatest number of deaths by a large margin. By doing these visualization, it provides even further insight into the dataset, and how that may influence the model.

Once all the variables were factored and visualized, a logistic regression was run using the glm() function. The glm model was first run on a train dataset, which represents 70% of the data, selected using the createDataPartition() function. In this case, the response variable is death and the predictors are education, sex, alcoholfreq, exercise, and race. From running the model summary, the variables that were significant at a 5% level are sex1, education2, education3, education4, education5, and alcoholfreq2. It is interesting to note that none of the exercise or race variables were deemed significant in this model. Then, an ANOVA test was run to determine which variables were significant to the model when not in their individual levels. The significant predictors were education, sex, and alcoholfreq, which match up with the results obtained from the model summary. To assess the model's predictive ability, a reduced model was used to compare the results. The first reduced model tested, used the significant predictors from the ANOVA test: education, sex, and alcoholfreq. However, when this first reduced model was compared to the original model, using ANOVA, it was deemed not to be a better model than the full model. The next reduced model test, used just the variables in which all the levels were significant in the full model, which were sex and education. The results of the ANOVA test

showed that this reduced model was a better fit. Both the full model and the reduced model were used to run prediction on the test set to see how well each model predicted the likelihood of death. The predictions were then compared to the actual results in the test dataset. This is how to the determine the accuracy of the models. Both the full model and reduced model have very high accuracy of about 80%.

Conclusion and Future Work

The full model considered all five variables, education, sex, alcoholfreq, exercise, and race, in predicting the likelihood death. The reduced model that was deemed to be a better fit from the chi-squared test was a model with just education and sex. When going from the full to the reduced model it was interesting to see that exercise and race were not deemed to be statistically significant. One possibility of exercise not being chosen is because the three levels: much exercise, moderate exercise, and little or no exercise, are very subjective. The codebook for this dataset does not provide any indication about what would be considered ‘much’ or ‘moderate’, so it is left to the subject to determine what would they would qualify as those values. This could lead to much variation because each subject could have a different interpretation about what falls into each of those categories. As for race, the distribution is quite uneven between white and black or other levels. This might be a contributor to why it was not deemed significant. I was surprised to see that alcoholfreq was not considered to be a significant predictor. I would think that the more a person drinks in a year, it would increase the likelihood of death. In the future, one could use a different set of predictors and see if alcholfreq is significant in that situtation. Since both the full and reduced model, had a prediction accuracy of about 80%, this confirms that race, exercise, and alcoholfreq did not contribute anything in

predicting the likelihood of death. In the future, interactions could be added to the model to see if it improves the model's accuracy in predicting death.

Author's Note

The author has used this dataset for previous courses, using SAS as well as Bayesian approaches, but has never done this set of visualizations and analyses.