Final Project of MA5771 – Applied Generalized Linear Models

Name: Brandon Coates

Project Title: Forever Alone: How Social and Biological Factors Influence Suicide Attempts

Section 1: Introduction

With the increasing prevalence of the mental health crisis that the world is facing, it is important to understand what factors are causing individuals to perform a suicide attempt. In order to derive some of the uniting factors, a survey was conducted on the subreddit /r/ForeverAlone, a website in which "people who have been alone most of their lives could come and talk about their issues."[1] A common topic within this community is the individual's struggles with mental health, some of which have led to suicide attempts.

The following study will aim to examine the results of the survey to identify key factors that may allow for the prediction of if an individual may attempt suicide. These factors may also be useful in identifying reasons to why an individual may be suffering from poor mental health conditions or may even undergo regular mental health crises. The biological factors that are examined are gender with levels female, male, transgender female, transgender male, sexuality with levels bisexual, gay/lesbian, and straight, bodyweight with factors normal weight, obese, overweight, and underweight, age measured in years, and if the individual suffers from depression with two levels of yes and no. The social factors that are examined are yearly income with factors low ($0 - $39,999), middle ($40,000 - $99,999), and upper ($100,000 - $200,000 or more), if the individual is a virgin with levels of yes or no, the amount of friends the individual has, and if the individual has social fears with levels of yes or no. The response variable that is examined is the response of the survey taking individual on whether or not they have attempted suicide before.

The data that is used in this study can be found on a public repository on Kaggle[2]. It should be noted that this survey was open to the public, so there may be erroneous results that were collected. The results of this study should be taken with caution since there was no control over the surveyed sample. Though there was no control over the sample, the results of this study may shed a light on factors that should be explored further in a controlled study with proper statistical design of experiment concepts put in place.

Section 2: Statistical Methods

The dataset that will be examined was downloaded as a .csv file, this file was then imported into RStudio. The following statistical analysis will be conducted using R programming language through the RStudio environment. For these analyses, the significance level that will be used will be 0.05.

Section 2.1: Exploratory Data Analysis

Before any analysis of interest is conducted on the data set, plots will be constructed that examine the relations of the explanatory variables to identify any possible trends that may need to be addressed when building a generalized linear model to fit the data. The relationships between the response variable and the explanatory variables will also be examined through the means of histograms, box plots, and scatterplots.

Section 2.2: Generalized Linear Model

In order to examine the probability that an individual may attempt suicide based on the examined factors, a binomial generalized linear model with the logit link function will be constructed from the data. The binomial GLM was chosen for this dataset due to the yes or no nature of the response variable. Once the model has been fit, which contains all factors considered in this study, model selection tests such as

[1] https://www.reddit.com/r/ForeverAlone/comments/byosqf/about_the_sub_common_misconceptions_and_an/
[2] https://www.kaggle.com/antonaks/suicide-attempt-prediction-foreveralone-dataset/data

likelihood ratio tests and analysis of deviance tests will be used to downselect the significant factors in the model. Tools such as the AIC, BIC and residual deviance will also be used to select a model with the most accuracy. Once a model has been selected, diagnostic analyses will be conducted to ensure that the model follows all of general linear models, such as plots of Cook's distance, quantile-quantile plots, and plots of the deviance residuals.

Section 3 Results and Conclusions

Before any models were fit to the data, histograms were constructed to examine possible relations between the explanatory variables and the individuals that had reported that they had attempted suicide before. The exploratory plots can be seen in **Figure 1**. From these plots several relations could be observed, when examining the relation between suicide attempts and gender, it appears that there were more males that attempted suicide, however the ratio of females that had reported a suicide attempt was greater than the males surveyed. There also appeared to be a larger ratio of bisexuals that had reported a suicide attempt than any other sexuality. When examining the reported incomes, it was found that suicide attempts were more common in individuals that were low income. There appeared to be a larger amount of suicide attempts in individuals that were in the age range of 20-30. It was also found that the majority of individuals surveyed had a low number of friends, which can be observed by the right skew in the data. Lastly, it was found that individuals that reported that they were depressed led to higher suicide attempts than those that did not report that they were depressed.

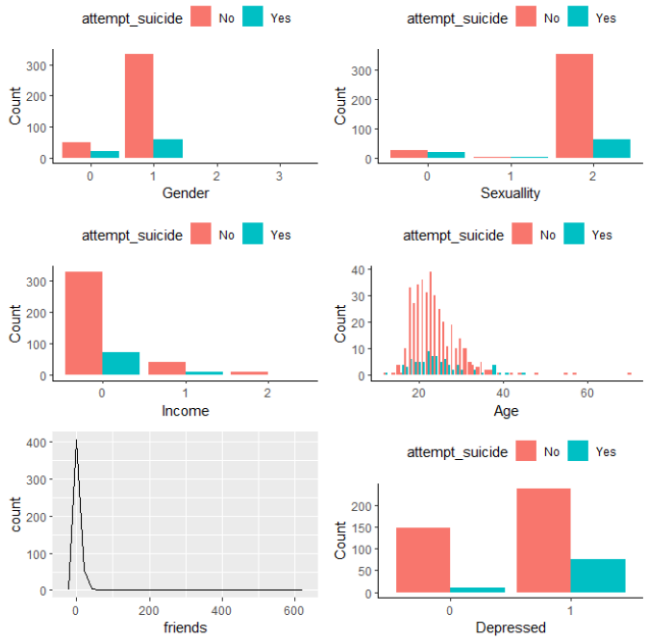

Figure 1: Exploratory Plots Examining Relations with Suicide Attempts

Once initial data exploration was completed, a Binomial GLM with a logit link function was fit to the data that included all of the exploratory variables that were included in the data set. Before any conclusions were drawn from the model, the assumptions for the GLM were examined to determine if this model approach was appropriate for the data set. The diagnostic plots for the full model can be seen in **Figure 2**. The plot of the quantile residuals versus the scaled fitted values was used to test the assumption that the proper systematic component of the GLM was being used. From the plot, it was found that this assumption holds true since there is no observable pattern to the data. The next assumption that was tested was the assumption that the appropriate link function was being used, which was done by examining the working residuals versus the linear predictors plot. In the plot generated from the model, the assumption does not appear to be satisfied, however, multiple models were built and tested to determine if the link

function was incorrect and similar results were found for each model, so it is assumed that this assumption is satisfied. The assumption that the random component distribution is appropriate was tested by examining the quantile-quantile plot of the quantile residuals. From the plot generated, it was found that this assumption was satisfied since most of the points fall closely on the linear trendline that is in the plot. The last assumption that was examined was the assumption that there were no outliers or highly influential points in the data, which was done by examining the Cook's Distance Index. For this model, this assumption also appears to be satisfied since there were no points on the plot that were greater than 1.

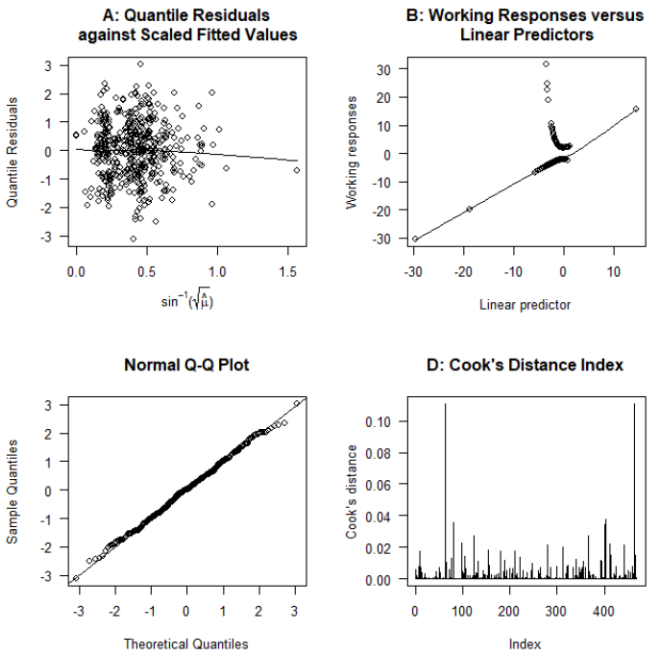

Figure 2: Diagnostic Plots to Test Assumptions of GLM

Once it was found the appropriate GLM was being used, model down-selection was conducted by removing the least significant variables from the full model one at a time. Likelihood ratio tests were conducted to determine which variables should be included in the final model by testing a model with a specific variable in question to a model without the variable. Once a preliminary final model was found, the AIC and BIC values of all of the generated models were calculated and examined to find the most accurate binomial GLM for this data set. By employing these techniques, a final binomial GLM was created using a logit link function. The linear model is summarized in **Table 1**. The GLM assumptions were once again checked and it was found that there were no violations to this assumptions.

Table 1: Summary of Final Binomial GLM Model

Variable	Estimate(β_j)	Exp(β_j)	Standard Error	Wald Test Statistic	p-value
(Intercept)	-1.0522	0.3492	0.5016	-2.098	0.0359
Gender Male	-0.7412	0.4765	0.3308	-2.241	0.0250
Gender Transgender Female	15.3121	4466430	882.7434	0.017	0.9862
Gender Transgender Male	1.6653	5.2873	1.5375	1.083	0.2788
Sexuality Gay/Lesbian	-0.1737	0.8405	0.8168	-0.213	0.8316
Sexuality Straight	-1.0646	0.3449	0.3647	-2.919	0.0035
Friends	-0.0359	0.9647	0.0232	-1.549	0.1214
Depressed Yes	1.5503	4.7129	0.3795	4.085	4.4e-5
Null deviance: 443.92 on 468 degrees of freedom **Residual deviance: 389.51 on 461 degrees of freedom** **AIC: 405.51**					

From this model it was found that the significant factors in the data set were the gender of an individual, the sexuality of the individual, the number of friends an individual has, and also if the individual reported having depression or not. It was found that when comparing to female individuals, the chance that an individual attempts suicide decreases by a factor of 0.4765 when the individual is male, increases by 4466430 when they are a transgender female, and increases by a factor of 5.2873 when they are a transgender male when all other variables are held constant. When examining how sexuality affects the chance someone may attempt suicide, when comparing to bisexual individuals, there is a decrease of a factor of 0.8405 for gay/lesbian individuals and a decrease of a factor of 0.3449 for straight individuals. The chance of an individual attempting suicide decreases by a factor of 0.9647 for each single increase in the number of friends they report having. Lastly when comparing to individuals that reported to not being depressed, the chance an individual attempts suicide increases by a factor of 4.7129 when they report that they are depressed.

Section 4 Discussion

From the work conducted, it was found that the significant factors in if an individual would attempt suicide or not are their gender, sexuality, number of friends, and if the individual is depressed. The results showed that the gender that has the highest risk of attempted suicide was transgender females, followed by transgender males, females, and lastly males. Though the results for the transgender individuals may be skewed since there was only one individual surveyed that reported being a transgender female and only two responders that reported being transgender males. It was also found that bisexual individuals were found to have the highest chance of attempting suicide, followed by gay/lesbian individuals. The number of friends an individual reported was also found to have an impact on the chance an individual would attempt suicide, it was found that the chance decreases the more friends an individual reported having. Lastly, it was found that if an individual reported being depressed, the chance they would attempt suicide greatly increases. Though most of the social factors that were considered were not found to be significant, this study showcases how the acceptability of certain genders and sexualities may help lessen the chance an individual may attempt to commit suicide. With more acceptance of the LGBT community, the number of suicide attempts may be decreased overall since more acceptance may lead to these individuals also making more friends. Though this survey was conducted on a possibly skewed population, there are still important takeaways that were found in the data.