Ailun Huang, Judiah Lin, and Raymond Pang

Professor Elizaveta Levina

STATS 415 - Intro to Data Mining

23 April 2019

# Predicting Suicide Percentages in the United States

## Introduction

All over the world, suicide poses to be one of the most prominent issues. In a world filled with a lot of injustice and suffering, some people lose sight of hope in the future and believe it is better to take their own life. Depression and other psychological illnesses are huge contributors to the climbing suicide rates. Globally, someone dies by suicide roughly every 40 seconds. The big question is, how do we prevent suicide rates from climbing? In our research project, we hope to explore a recent suicide dataset to see if we can find any results that could lead us to make meaningful conclusions. The dataset we used for our project is *Suicide Rates Overview from 1985 to 2016*, which can be found on Kaggle. It contains 27,820 observations with 12 variables: country, year, age group, sex, number of suicides, population, suicides per 100k in population, country-year (ex. United States - 1985), HDI (Human Development Index), GDP, GDP per-capita, and generation (ex. Generation X, Silent Generation, Boomers, etc).

We are particularly interested in identifying key risk factors that contribute to the suicide percentages in the United States. Determining key factors in suicides is important for suicide prevention agencies because "suicide prevention activities, programs, and other efforts are most effective when they are guided by a strategic planning process" (Suicide Prevention Resource Center, 2019). For instance, if the age group for people 75 years old and older have the most suicides compared to all other age groups, then the Suicide Prevention Resource Center would consider allocating more resources in reaching people in the more at risk group. The question we are hoping to answer in our research is to see if there are significant predictors in our dataset that could lead us to new conclusions. After observing our results obtained from various linear and nonlinear regression methods, we found that age group and sex are significant contributors to suicide percentages. We hope to use these results and conclusions to benefit associations that are dedicated to raising suicide awareness.

**Methodology**

To begin, we created our response variable, suicide percentage, by dividing the suicide number in each age group for every year by the population in that age group in that year then multiplied by 100. We find that the suicide percentages range widely in different powers, which usually indicates a log transformation is needed. We confirmed our reasoning by looking at the residual plots (shown below); it does appear that a log transformation on suicide percentage would give us a more normal distribution.

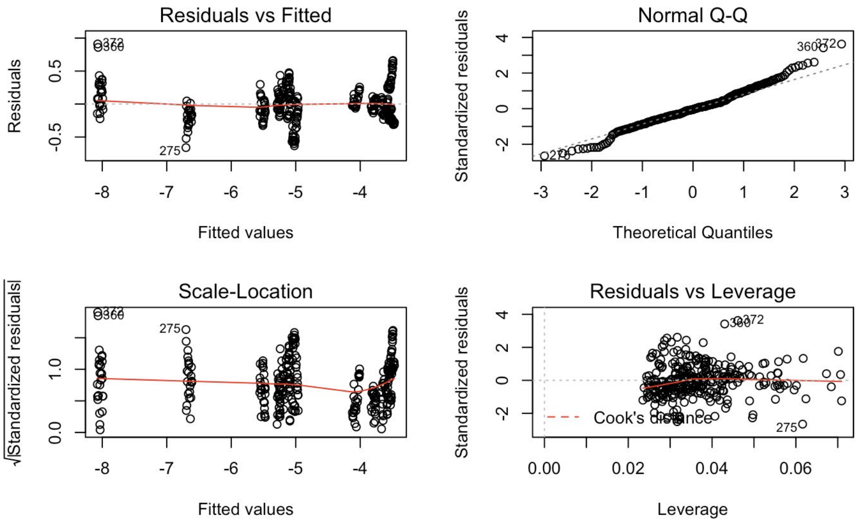


**Figure 1** Residual plots of suicide percentages (especially from Normal Q-Q plot) Indicates we need to use a log transformation for our data to be approximately normal.

Before we did any exploratory analysis, we decided to leave out some variables and many observations. Since we are primarily interested in suicide percentages in the United States, we filtered out other countries. Intuitively, planning prevention methods are based locally by country or even state, which made sense to leave out all other countries' observations. This reduced our dataset to 372 observations. Next, we left out suicides per 100k population because our response variable, suicide percentage, is calculated similarly. We also left out country-year because we are only looking at the United States and year is already a predictor we are interested in. Unfortunately, due to many missing values, we left out the HDI variable. We took out GDP per year because we were more interested in GDP per-capita. Including both variables would introduce collinearity in our model. Lastly, we are not including generation in our model because it changes very periodically. We don't have that many rows of data and we believe including the different generations could possibly make our analysis harder to interpret.

Before we explain the methods we used to model our data, we will explain how we split our data. First, we set a seed of 415 to ensure consistent results every time we ran our data. We want to reserve approximately 80% of the dataset to be our training data (298 observations) and 20% to be test data (74 observations). Our goal is to find the model that performs the best on our test dataset (has the lowest test error).

Next, we will focus on choosing the various ways of modeling our dataset. Initially, we were interested in performing linear regression methods on our dataset because we believed our data would be more linear after we log transformed suicide percentages. However, from our exploratory data analysis, we see that there may be signs of nonlinearity in our predictors. Therefore, we will be implementing both linear and nonlinear methods to our data so we can compare multiple methods. To note, we will be regressing on the logged transformation of suicide percentages. For linear regression methods, we will look at various best subset selection (BSS) regressions because we are interested in minimizing the number of predictors used in its respective models. We will use the following BSS techniques: forward, backward, AIC, BIC, and adjusted-$R^2$.

As for nonlinear regression methods, we decided to use polynomial regression, splines (natural and smoothing splines) and GAM to model the data. We believe these nonlinear techniques are the most appropriate for our data because we want to relax the linearity assumption, but still want to have as much interpretability as possible (Levina). In particular, we were interested in using GDP per-capita as our only predictor for natural and smoothing splines methods. It is the only continuous variable that suicide percentage is not derived from. Additionally, we were interested in the relationship between GDP per-capita and suicide percentages.

In addition to these nonlinear methods, we also fit a GAM because of its flexibility and support for categorical predictors like age and sex as well as the nonlinear behavior of GDP per-capita. We decided to fit two GAM models. Our first GAM model includes sex, age, and GDP per capita. For our second GAM model, we added an interaction term between age group and sex because we were interested in the relationship between age and sex. Intuitively, adding this interaction makes sense because females and males go through vastly different challenges at different ages, which could affect a person’s decision to commit suicide.

**Results**

After settling on the number of observations and the variables of interest, we looked at various exploratory plots that would allow us to better visualize data. First, we will look at the two predictors that seem to significantly predict suicide percentages at a first glance - age and sex. The following box plot below allows us to the visualize sex and suicide percentage:

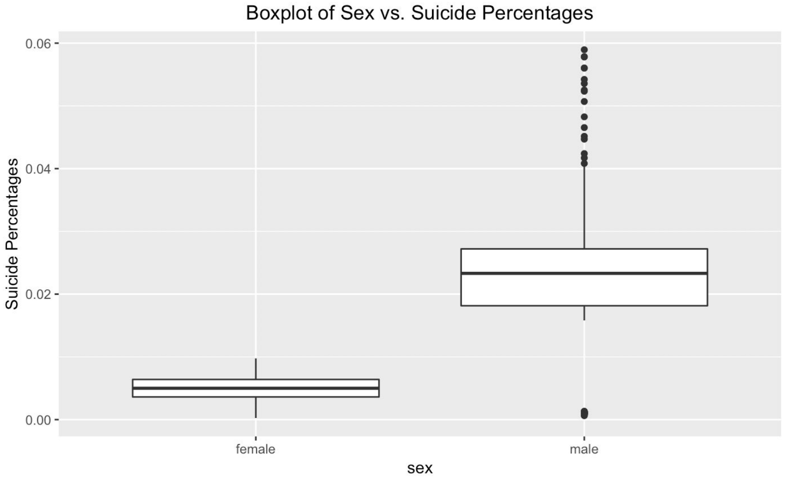


**Figure 2** Box plot of sex and suicide percentage. There is a clear distinct difference between suicide percentages in males and females.

The above box plot indicates a clear distinction in suicide percentage between males and females. We notice that the maximum of female suicide percentages is far below the minimum of male suicide percentages (excluding outliers). A possible, reasonable explanation for this may be because males experience a lot of pressure to be able to provide for the family, which leads to high levels of stress and potentially suicide (Hawton 484). Next, we will look at box plot of age and suicide percentage:

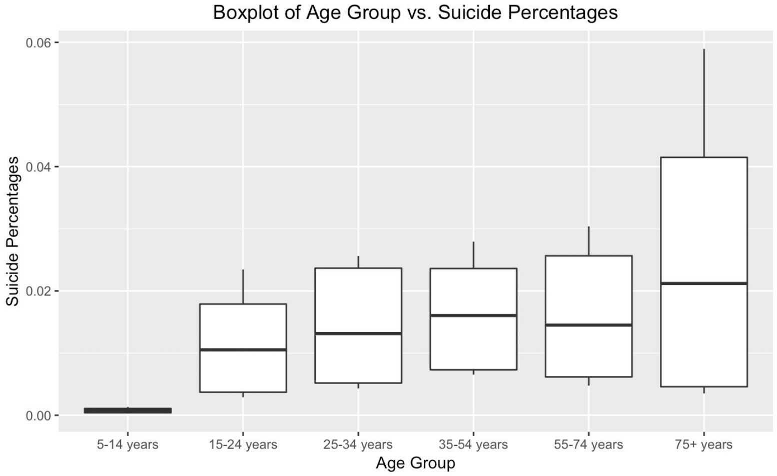


**Figure 3** Box plot of various age group and suicide percentages. There are some differences between suicide percentages in age groups.

There appears to be a distinction in suicide percentage between 5-14 years and older age groups. This may make sense because many children are living with their families and have generally less worries than older aged people.

Additionally, we looked at how year and GDP per-capita impacted suicide percentages, respectively. We notice that both of these variables have a polynomial trend. Furthermore, there appears to be a strong correlation between these two predictors:

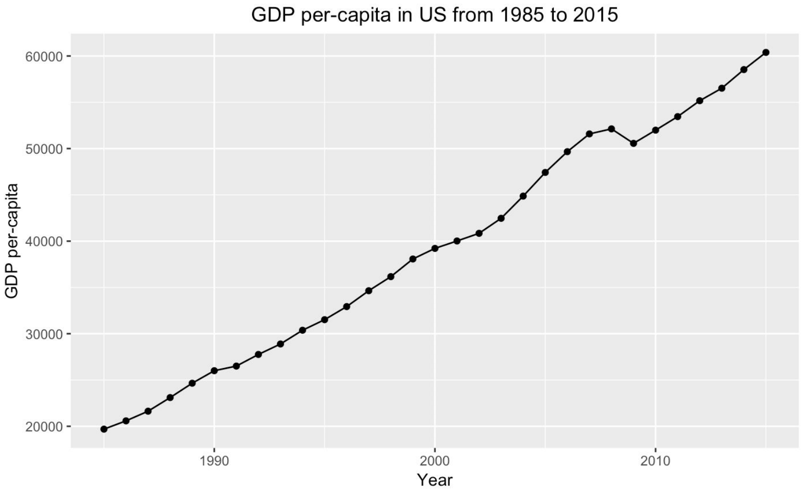


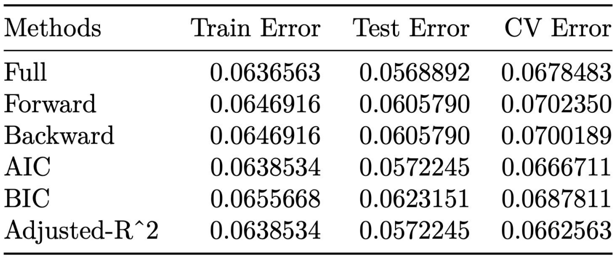
**Figure 4** Line plot of GDP per-capita and year - there appears to be a positive correlation.

This correlation is understandable because we expect GDP per-capita to increase yearly (largely due to inflation). Because we do not want to introduce collinearity to our model, we will only fit some of our models using GDP per-capita. Additionally, because of this nonlinear predictor, this further indicates that modeling our data using nonlinear methods is appropriate.

Now that we have explored our data, we will move on to introducing the results of our linear methods. We notice that every subsetted models includes the two predictors, sex and age. From this, we are able to conclude that these are two very meaningful attributes to suicide percentages. As for nonlinear methods, we chose the predictors for our models. As mentioned above, we are using GDP per-capita. From the polynomial regression, we can see that the degree that gives us the smallest error (based off of cross-validation) is 1. This indicates that a model other that a polynomial regression is not the best. For natural and smoothing splines, the results indicate that a nonlinear model is appropriate. We find that the C-V error for natural splines is slightly better than smoothing splines, so we will be using natural splines (degree of freedom of 2) in both of our GAM models. Furthermore, the results of both of our GAM models indicate a nonlinear model is valid. The next section of our paper will analyze the various errors (training, test and C-V) for all of models and see which model performs the best on our test dataset.

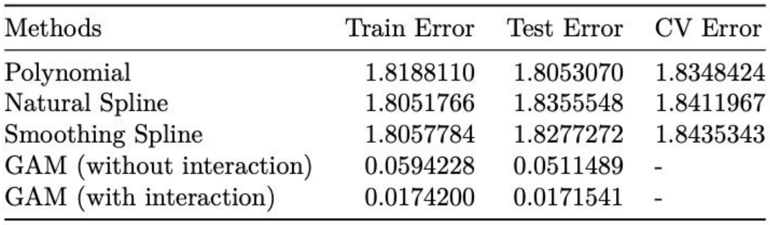
## Analysis

The following two tables display the three errors of all linear and nonlinear models:



| Methods | Train Error | Test Error | CV Error |
|---|---|---|---|
| Full | 0.0636563 | 0.0568892 | 0.0678483 |
| Forward | 0.0646916 | 0.0605790 | 0.0702350 |
| Backward | 0.0646916 | 0.0605790 | 0.0700189 |
| AIC | 0.0638534 | 0.0572245 | 0.0666711 |
| BIC | 0.0655668 | 0.0623151 | 0.0687811 |
| Adjusted-Rˆ2 | 0.0638534 | 0.0572245 | 0.0662563 |

**Table 1** Table of errors of the linear models. For the most part, the errors are relatively similar throughout the various models. From the train and test errors, we can see that forward and backward selections picked the same predictors while BIC and adjusted-$R^2$ techniques picked the same predictors.



| Methods | Train Error | Test Error | CV Error |
|---|---|---|---|
| Polynomial | 1.8188110 | 1.8053070 | 1.8348424 |
| Natural Spline | 1.8051766 | 1.8355548 | 1.8411967 |
| Smoothing Spline | 1.8057784 | 1.8277272 | 1.8435343 |
| GAM (without interaction) | 0.0594228 | 0.0511489 | - |
| GAM (with interaction) | 0.0174200 | 0.0171541 | - |

**Table 2** Table of errors of the nonlinear models. Besides the two GAM models, we can see that the other three methods perform poorly, as all three errors for each method is extremely large.

Looking at the linear regression methods, it appears that the full model has the lowest test error of 0.0568. The full model is the best performing linear model on our test dataset, with sex and age being good predictors of suicide percentage because of its significant p-values. In our polynomial regression, natural splines, and smoothing splines for GDP per-capita, we observed test errors of 1.8053, 1.8356, and 1.8278, respectively. A possible reason why the errors are so large is GDP per-capita is not a good sole predictor for suicide percentages. We do not expect there to be a strong relationship between a country's GDP per-capita (wealth) and one's decision to commit suicide. Despite having the lowest test error, we ruled out polynomial regression because the lowest C-V error of this particular model was at a degree of 1. As mentioned before, this indicates polynomial regression is not a good model.

The results of both of our GAM models prove to be more interesting. In our GAM model without the interaction term, we observed a test error of 0.0511. Out of all the methods, linear and nonlinear, this model appears to have the lowest test error, indicating it best performs on our test data. Because of this, we will only be adding an interaction term between age and sex to the GAM model to see if this will reduce the test error. From this, we observed a test error of 0.0172.

The GAM model with the interaction term performed the best based on test error. Even without the interaction term, the GAM model had a lower, but approximately equal test error compared to the full model. We noticed age and sex were significant predictors based on their p-values from linear regression, and we were interested to see if there is a significant relationship between them. Consequently, adding the interaction term played a significant role in lowering the test error. Clearly, this interaction term is extremely important. So, to answer our main question of our research (*what are the key predictive factors in suicide percentage?*) we conclude from the GAM model that strong predictors are age, sex, their interaction, and GDP per-capita.

While we conclude that the GAM model with the interaction term is the best performing model on our test dataset, we want to acknowledge that the test errors seem low but may indicate otherwise. Our response variable, suicide percentage (without the log transformation), is usually a very small number for most observations. Because of this, a small error may be because the suicide percentages are close to zero. So, we want to bring attention to this detail and explain why we are drawing conclusions cautiously.

**Discussion**

Although we were able to identify some key predictors, we have several suggestions for improvements. First, we would suggest looking more into the importance of the interaction between age and sex. To do this, we can collect more data about various groups that victims fall under. These can include ethnicity, sexuality, and so forth. Additionally, we suggest collecting more data on individuals in our sample so that we may better predict suicide percentages, as there are many other factors not included in our dataset that explains why an individual would commit suicide. Future studies should also look into previous medical histories in order to predict and identify those who may be at the greatest risk for committing suicide.

Furthermore, knowing that GDP per-capita is a key predictor in suicide percentages suggests that wealth may play a role in suicide percentage; as a result, we would suggest that future studies pay particular attention to socioeconomic status when making predictions for suicide percentages in other countries. A final suggestion to further studies would be to implement other methods not included in our study. In doing so, we can potentially provide insight on what a "low" test error means in relation to suicide percentage. Since data analysis like these are important in planning effective suicide prevention programs, we would recommend exploring different predictors like the ones we suggested.

Overall, despite our limitations, we believe our main research questions have been answered by our methods and results. We found significant predictors - age, GDP per-capita, and sex - for the log transformation of suicide percentages. Because we know there is significance between age and sex, we want to cater more awareness and mental health programs for specific age/sex groups that have higher suicide rates.

**Contributions**

Ailun Huang and Raymond Pang were responsible for writing the majority of the code for this project and working on the PowerPoint presentation together. Judiah Lin was responsible for assisting with the PowerPoint and writing the majority of the paper. The paper was edited by Ailun and Raymond before it was finalized. The three of them brainstormed ideas together to come up with the content for the report and presentation, meeting weekly and frequently checking in with one another in order to update one another on progress.

# Works Cited

Hawton, Keith. “Sex and Suicide: Gender Differences in Suicidal Behaviour.” British Journal of Psychiatry 177.6 (2000): 484-85. Print.

Lahiri, Rupa. *Suicide Rates Overview 1985 to 2016 (Version 1)*. Kaggle, 2018. Web. 21 Apr 2019.

Levina, Elizaveta. “Splines.” STATS 415 (Winter 2019). Splines Lecture, 21 Apr. 2019, Ann Arbor, University of Michigan.

"Strategic Planning." *Strategic Planning | Suicide Prevention Resource Center*. Web. 21 Apr. 2019.

"Suicide Statistics and Facts – SAVE." *SAVE*. Web. 21 Apr. 2019.