**Suicide Rate in the States**

Katherine Osadczuk

Capstone: Rsrch Data Analysis (Math-435-01X, Math-470-01X)

**Abstract**:

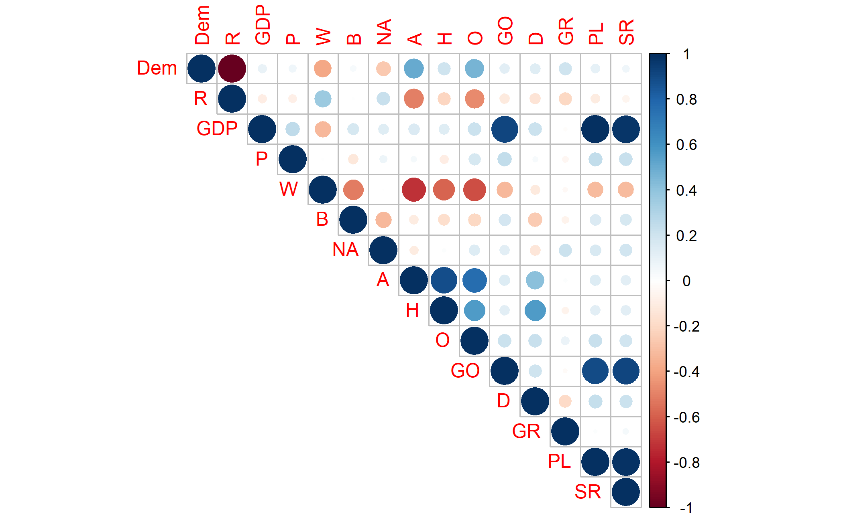
For this project, I wanted to explore the connections between different social issues and their possible connections to each other and/or other factors. I decided to do this because I am very interested to find out what issues cause higher suicide rates. Suicide is a real problem today, and I personally have been affected by it. I want to know if there are certain factors that increase the risk. One thing I am especially interested in is if there is any link towards gun ownership. I would also like to explore some other possible factors and see what effects they may have on a state's suicide rate. I will be doing this by performing a linear model analysis on the data, and then following that up with an evaluation of my final linear model using the Baye’s method to see how the results produced compared.

**Introduction**:

Suicide in the US has been at an all-time high, with a total of about 50,000 suicides and 1.7 million suicide attempts in 2021 (Suicide Statistics, 2023). By splitting the US into its respective 50 states, I hope to be able to analyze this data to be able to determine risk factors and causes of suicide. One factor that I believe will be important is gun ownership due to the fact that over 50% of suicides are committed using firearms (Suicide Statistics, 2023). I will also be looking into how political affiliation between the two main parties can increase or decrease suicide rate, or if political affiliation has no connection at all. Along with that I will be exploring the effects of various statistics, such as population and density, and the effects of the ethnic makeup of the population.

My hope is that this information can not only help us visualize what exactly causes suicide, but also lead to as helping change the relevant factors. While things such as population and density may not be factors one can easily change, gun ownership could easily be influenced via stricter gun control. While the model may not serve very well as a predictive model due to the fact we are not able to create or ‘test’ new states that are not already in the system, it’ll provide valuable information on the importance of the variables themselves.

I was able to obtain all the data I used for my analysis from the World Population Review. The data on this site was obtained from CBS, CNN, the Bureau of Economic Analysis, and the US Census Bureau and then put into a more digestible and visually appealing form that came with download csv files. The election results were from the 2020 presidential election (2020 Election Results, 2023), the gun statistics from 2022 (Gun Ownership by State 2023, 2023), GDP data from 2019-2022 (GDP by State 2023, 2023), the population, growth rate, and density data from 2023 (US States - Ranked by Population 2023, 2023), poverty rate data from 2022 (Poverty Rate by State 2023, 2023), and race/ethnic data from 2023 (US States by Race 2023, 2023).

**Exploratory:**

To start off, I cleaned up my data and ended up with fourteen different variables with data from each of the fifty states: percent voting democrat (Dem), percent voting republican (R), GDP per capita (in thousands) (GDP), population (in millions) (P), percent total white population (W), percent total black population (B), percent total Native American population (NA, written as NA1 in the code to prevent confusion with the value “NA”), percent total Asian population (A), percent total Hawaiian population (H), percent total other (ethnicities) population (O), percent gun ownership (GO), density (D), growth rate (GR), percent of population below poverty level (PL), and suicide rate (in suicides per 100k) (SR). After this I created a correlation graph (Graph 1) of y ~ x and found that GDP, GO, and PL all had significant correlation with the states suicide rate (Full table of correlation values is present in the appendix).

Graph 1: Correlation Matrix of all variables using y ~ x

My next step was to graph each of the variables against suicide rate to see if I could notice any patterns. I split them into three categories: election results (Image 1), general statistics (Image 2), and ethnicities (Image 3). I graphed each variable individually and created a line of best fit if there appeared to be a trend. In the end, I didn’t end up finding anything interesting. There were no other obvious visible trends other than the previously stated linear ones.

A graph with blue dots

Description automatically generated with medium confidence

Image 1: Suicide Rate per 100k vs Election Results

A graph of different types of data

Description automatically generated with medium confidence*Image 2: Suicide Rate per 100k vs Various Statistics*

A group of blue dots

Description automatically generated

Image 3: Suicide Rate per 100k vs Ethnicities

**Linear Models:**

I decided to begin my linear model creating with a full model, so that I could then compare any subsequent models to compare to. Any models that had similar or greater R2 could then be considered for use as my final model. Due to extreme correlation between some variables, O was left out, and due to having a coefficient value of less than a thousandth, P was left out, leaving me with

– 6.719

with a final adjusted R2 of 0.973. The model itself proves to be fairly accurate, but with eleven different variables used it is very messy and difficult to understand. Only two of these variables were found to have a significant p-value, those being the GO and PL.

With the previous results in mind, I decided to create a minimum model using the significant variables. The model

had a final adjusted R2 of 0.960. As this was the simplest and easiest to understand model, I would want to ensure that whatever I selected as my final model would have an equal or larger R2 value because otherwise any new variables would be presenting no beneficial information.

My next step was to take my full model and run both a step forward and step backward analysis. The final models were similar, with the only difference being which of the two parties was used a variable. The step forward model had an adjusted R2 of 0.977 and was

with the backwards model having an identical adjusted R2 of 0.977 using

Step Backwards 1:

Both models proved to be more accurate than the minimum model and thus valid to be used as a final model. What I found interesting is that the coefficients for PL, GO, and GR were virtually identical, and the different variable also had a very similar coefficient but with an opposite sign. Seeing this I wanted to see what a model with both parties looked like. The model I got was

with an adjusted R2 of 0.976. This election percent model had many similarities with the last two, once again, PL, GO, and GR all had nearly identical values. While the R and Dem coefficients change by a decent amount, their unique dichotomy remains. This made me wonder if the R and Dem variables would serve better as a single yes/no variable instead of two integer values.

After combining the republican and democratic variable into a single variable I created a box plot (Graph 2); however, this didn’t really show me much and both sides seemed very similar, so I went on and created a step forward model using all the variables. The output of the step wise model was

A graph with red and blue lines

Description automatically generatedwith an adjusted R2 of 0.9749. The election party model that was output was not what I expected. Not only did my new variable not have any presence, but the GR value was lost and two of the ethnicities were added in. The two ethnicities presented a promising new lead in what could contribute to suicide rates. Both coefficients were in the same relative size as the PL and GO ones, thus they would have a significant effect on the over SR. With this in mind, I decided that the best model would have to be one using numeric variables for the parties. Since the accuracy was unchanging between the two parties and that Dem and R can easily be predicted from each other (In general, with some fluctuation, and due to the prominence of the US’s two-party system), I decided to settle on the previous step forward 1 model, but I also added in the variables O and NA

Graph 2: Boxplot of Democratic vs Republican and Suicide Rate per 100k

Final Model:

**Bayes Model:**

A group of blue lines with white text

Description automatically generatedWith a final model in mind, I decided to try refining this model by running a Baye’s analysis on it. Using a glm model with 4 chains and 5000 \* 2 iterations, I was able to get a model. The chains (Graph 3) all appeared to be consistent and there was no noticeable leaning one way or the other, the Neff ratios (Table 1) were all above 50%, while some values, like GO and PL, are dangerously low, it’s not so bad that they need to be thrown out. The rhat value (Table 1) across the board is close to one, with the highest presenting its first post decimal figure at the ten thousandths place, showing that the chains did indeed mix well since the standard is that rhat should be less than 1.01. By all accounts the model creation was successful and viable for further use.

Graph 3: Visual depiction of all chains ran

|  |  |  |
| --- | --- | --- |
| Variable | Neff Ratio | Rhat |
| Intercept | 0.647 | 1.000 |
| GO | 0.548 | 1.000 |
| PL | 0.537 | 1.000 |
| GR | 0.876 | 1.000 |
| R | 0.662 | 1.000 |
| O | 0.726 | 1.000 |
| NA | 0.750 | 1.000 |
| Sigma | 0.805 | 1.000 |

*Table 1: Table of Neff and rhat values*

**Comparison:**

My last step was to test my Bayes model and linear model to see how they performed. Since the dependent variable wasn’t categorical, a straight up confusion matrix was not an option, so I had to get a little creative. I decided to find the mean and standard deviation for the difference between the actual value and predicted value. For both tests (Table 2), the mean difference rounded to 0.691, with the Bayes having a higher accuracy by 0.051%, which is negligible. The standard deviations were also similar, with Bayes having a value of 0.441 and the linear model having a value of 0.440 with a higher accuracy of about 0.143%, yet again being a negligible amount. The similarity of the two models and how each one has its own “better” value shows that both models are, for analytical purposes, identical. This being the case, the best model would then be the linear one because it is easier to understand.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Mean Difference | Standard Deviation | % Mean Change  (Linear - Bayes) / Bayes \* 100 | % Standard Deviation Change  (Linear - Bayes) / Bayes \* 100 |
| Linear | 0.691 | 0.440 | 0.051 | -0.143 |
| Bayes | 0.691 | 0.441 |

*Table 1: Table of comparisons between the Bayes and Final Model*

**Conclusion**:

All of the models I was able to create had relatively high accuracies, which shows overtraining, but it is not possible to create “new states” to have a test set. Instead of being used as a stern prediction model, I would argue that the models are more useful in understanding the effects of the individual variables. If we understand what can contribute to SR, we can work on either lowering or raising these factors in accordance with their effect. My analysis has shown that GO, PL, GR, Dem, R, O, and NA have measurable effects on a state’s suicide rate.

Starting with gun ownership, the effect was much smaller than I thought. While the correlation may be high, the actual effect is lower than I expected for every percent of the population that owns a gun, only 0.1 more suicide occur, which would mean that 10% ownership is about equivalent to 1 additional suicide per 100,000. PL is shown to have the highest effect. With a coefficient of 1.656, for every percent of the population below the poverty level, there is an increase of over 1 on the suicide rate. Next, we have GR, while the value is larger than that of GO, the overall effect is actually much smaller, with this actual value ranging from a decrease of 0.281 to an increase 0.575, leading to a total range of change less than 1. While the fact that an increasing growth rate can cause higher suicide rates conflicts with the idea that an increase in PL causes more suicide, taking the actual numbers into effect, the difference is so small that this inconsistency makes sense. While the party correlation was high, in the end the effect was minimal. The two ethnicities, while showing important information, both have minor effects.

With the knowledge of what variables provide the largest impact are GO and PL, we can look into ways to decrease the suicide rate via gun control and laws concerning minimum wage and poverty. Growth rate shouldn’t ever have a significantly large change, so looking into changing it is unimportant. Along those lines we have Dem and R, both values that are extremely difficult to change. Political lobbying is expensive, beliefs are difficult to change, and there is just an overwhelming number of factors that can change how someone votes. Finally, we have the O and NA ethnicities, the O ethnicity is actually shown to be in less danger of committing suicide, thus no need in input there, However, the NA population is shown to have an increased suicide rate. This is like due to the difficulty that Native Americans face in everyday life, such as racism, poverty, isolation, a lack of education, and so on. Efforts could contribute to the overall living situation of this group to reduce the rate of suicides. Overall, while some effects may be interesting, they are not all necessarily important, but it is important to take not of the GO and PL values due to the fact that the variables themselves can be directly affected using laws and the value of the NA coefficient, showing a general worse outcome for natives, meaning that would could work on attempting to directly change that coefficient instead by improving their overall living circumstances.

# **References:**

*2020 Election Results*. (2023, December 12). Retrieved from World Population Review: https://worldpopulationreview.com/state-rankings/2020-election-results

*GDP by State 2023*. (2023, December 12). Retrieved from World Population Review: https://worldpopulationreview.com/state-rankings/gdp-growth-by-state

*Gun Ownership by State 2023*. (2023, December 12). Retrieved from World Population Review: https://worldpopulationreview.com/state-rankings/gun-ownership-by-state

*Poverty Rate by State 2023*. (2023, December 12). Retrieved from World Population Review: https://worldpopulationreview.com/state-rankings/poverty-rate-by-state

*Suicide Statistics*. (2023, December 12). Retrieved from American Foundation for Suicide Statistics: https://afsp.org/suicide-statistics/

*US States - Ranked by Population 2023*. (2023, Decemer 12). Retrieved from World Population Review: https://worldpopulationreview.com/states

*US States by Race 2023*. (2023, December 12). Retrieved from World Population Review: https://worldpopulationreview.com/states/states-by-race

**Appendix**:

https://github.com/osadcook/Suicide-in-the-States