Suicide in the US: Identifying Variables to Improve Suicide Prevention Campaigns

**INTRODUCTION**

As a serious global health epidemic, suicide claims nearly 45,000 lives a year, and that number has been steadily increasing.1 This research study dives into a collection of potential variables from a compilation of reputable data sources to determine significant contributions to this issue. The National Suicide Prevention Lifeline has provided guidance to social media managers for effective suicide prevention messaging.2 This analysis is designed to create a predictive model and identify influential predictors of suicide rates for future monitoring. These results can be used as a guide to calibration suicide prevention campaigns to boost velocity of messaging in social media channels and other outreach venues based on increases in the contributing features of the model.

Our approach to this problem had three components. First, we reviewed the data gathered for data filtering, transformation, and feature analysis. Second, we ran the data through different linear regression models and determined the best performing linear model based on adjusted r2 and Akaike Information Criterion (AIC) values, which resulted in being our custom backwards selection based model. Lastly, we performed a 2-way ANOVA to examine the influence and interaction of age groups and gender on suicide rates.

**DATA DESCRIPTION**

The base data set came from a suicide data set on Kaggle. We added additional data we felt could be influential factors which included unemployment data5, S&P return data4, divorce rates6, and mass shooting data3. After collecting and aggregating these datasets, we were left with 27,820 global observations across 17 parameters.

We chose to limit the data set to the United States as the parameters we could track consistently were predominantly U.S. based, such as unemployment, S&P 500 annual returns, and divorce rates. The sex and generation fields were recoded for easier referencing, and we used the suicides per 100k respective population and GPD per capita. The data dictionary can be found in Table 13 in the appendix. Below, Table 1 shows the summary statistics of the final data set, comprising of 372 observations.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **VARIABLE** | **N** | **TRANSFORMATION** | **MINIMUM** | **MAXIMUM** | **MEAN** | **STD DEV** |
| **year** | 372 |  | 1985 | 2015 | 2000 | 8.956 |
| **population** | 372 |  | 4064000 | 43805214 | 21650610.760 | 9448629.560 |
| **suicides\_100k\_pop** | 372 | Log | 0.260 | 58.950 | 13.820 | 13.230 |
| **GDP\_Capita** | 372 |  | 19693 | 60387 | 39269.610 | 12334.120 |
| **MassShooting\_Victims** | 372 | Log | 7 | 444 | 96.032 | 95.095 |
| **Divorce\_Rate** | 372 |  | 3.100 | 5.000 | 4.074 | 0.558 |
| **Unemployment\_Rate** | 372 |  | 3.967 | 9.608 | 6.115 | 1.435 |
| **SP\_Return** | 372 |  | -38.490 | 34.110 | 9.817 | 16.522 |
| **gender\_code** | 372 | Recoded | 0 | 1 |  |  |
| **gencode** | 372 | Recoded | 0 | 5 |  |  |

*Table 1 Summary statistics*

Based on a visual representation of the data distribution, the following variables seemed to be highly skewed and thus were log-transformed: Suicide\_100k\_Pop, MassShooting\_Victims, and Unemployment\_Rate.

Testing transformations of both GDP\_Capita and SP\_Return did not provide enough beneficial evidence, so they were kept untransformed as a result. The scatterplot matrix of the raw data can be found in the appendix in Figure 7. The final scatterplot matrix is shown below in Figure 1. Strong correlation is also visible between Year and GDP, GPD and Divorce Rate, and Year and Divorce Rate.



Figure 1 Scatterplot Matrix after transformations

#### Objective 1: Regression Models

**REGRESSION MODEL**

For our linear regression analysis, we begin with a feature selection process. We first checked for multicollinearity based on the variable inflation factors (VIFs). We observed evidence of strong multicollinearity between the parameters year, GDP\_Capita, and Divorce\_Rate as shown in Table 2. This correlation is supported by our scatterplot matrix in the previous section, which displays a strong linear trend between these parameters. Redundant parameters, or those yielding high VIF values and correlations, have a strong association with each other, and therefore should be reduced accordingly. With this in mind, we elected to maintain year as a parameter, removing GDP\_Capita and Divorce\_Rate in the process.

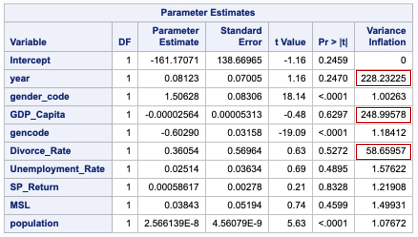


Table 2 Checking VIF for regression analysis

Meanwhile, the data was also put through an ordinary least squares (OLS) regression model. The analysis of variance (ANOVA) table shown in Table 3 below indicates that significant predictors exist within our reduced set of parameters (F Value=99.38, p-value < 0.0001).

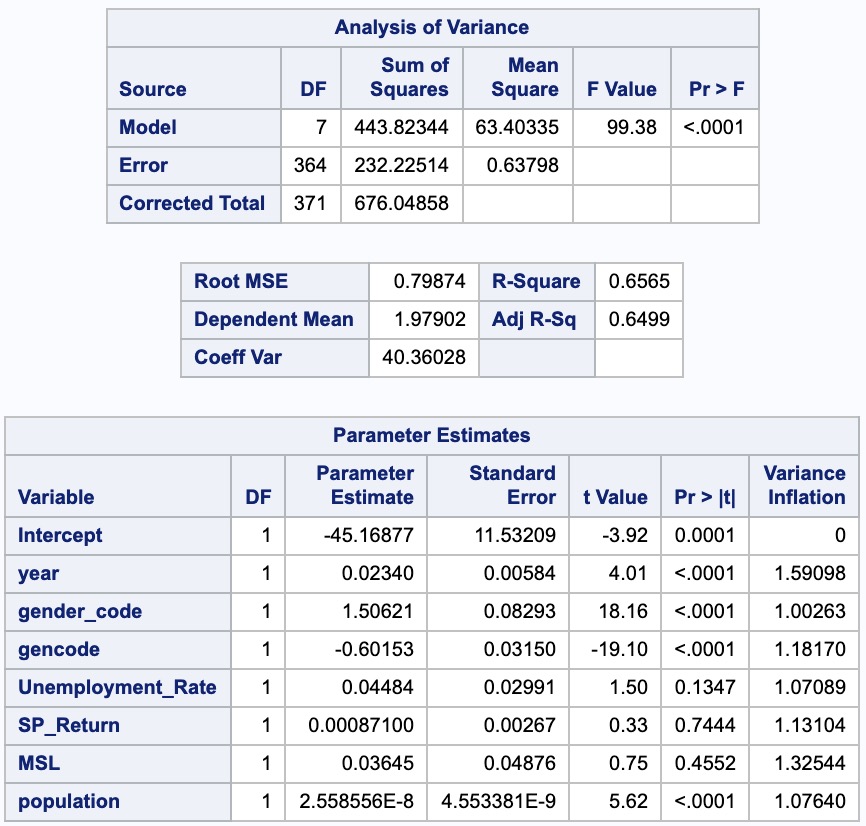


Table 3 Analysis of Variance and Parameter Estimates after multicollinearity removed

**MODEL SELECTION**

A total of 5 models were executed and the best model by our metrics of r2 and (AIC) values was our custom model (Table 4). These metrics and parameter estimates for the OLS, Forward, Backward, and Stepwise models can be found in the appendix.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **MODEL** | **AIC** | **BIC** | **ADJ R2** | **INTERVAL CV** |
| **OLS** |  |  | 0.650 |  |
| **Forward** | 125.092 | -209.719 | 0.726 | 190.838 |
| **Backward** | 125.092 | -209.719 | 0.726 | 190.838 |
| **Stepwise** | 125.374 | -213.356 | 0.725 | 192.224 |
| **Custom** | 106.099 | -209.118 | 0.743 | 182.507 |

Table 4 Model Comparison Data

While the unemployment rate is a figure that encompasses all unemployed job seekers in a given year, the unemployment rate will likely vary based on the age of the individual(s). Additionally, this allows us to include the impact of generational shifts on workplaces, with an increasing percent of the Baby Boomer generation leaving the workplace. To account for these, and to keep parameters consistent with the previously attempted models (OLS, forward, backward, and stepwise selections), we proposed an interaction term between the generation classification of an age group in a given year and the unemployment rate of that year.

For our custom model we elected to use a backwards selection model that includes the interaction between generation and unemployment. The adjusted r2 for this model is 0.7433 and the AIC value is 106.099. Our custom model found that all the interaction variables were significant and that the interaction between unemployment and generation was most significant for Generation X and Millennials.

**MODEL AssUMPTION CHECKS**

The fit diagnostics for LogSuicide matrix (figure 2) show the residuals are fairly randomly distributed, with some evidence of clustering around zero. The QQ Plot shows that the residuals follow a relatively normal distribution with no concerning deviations. Cook’s D values look extremely good with all values under 0.02. Our leverage plot does not contain any high leverage data points of concern.

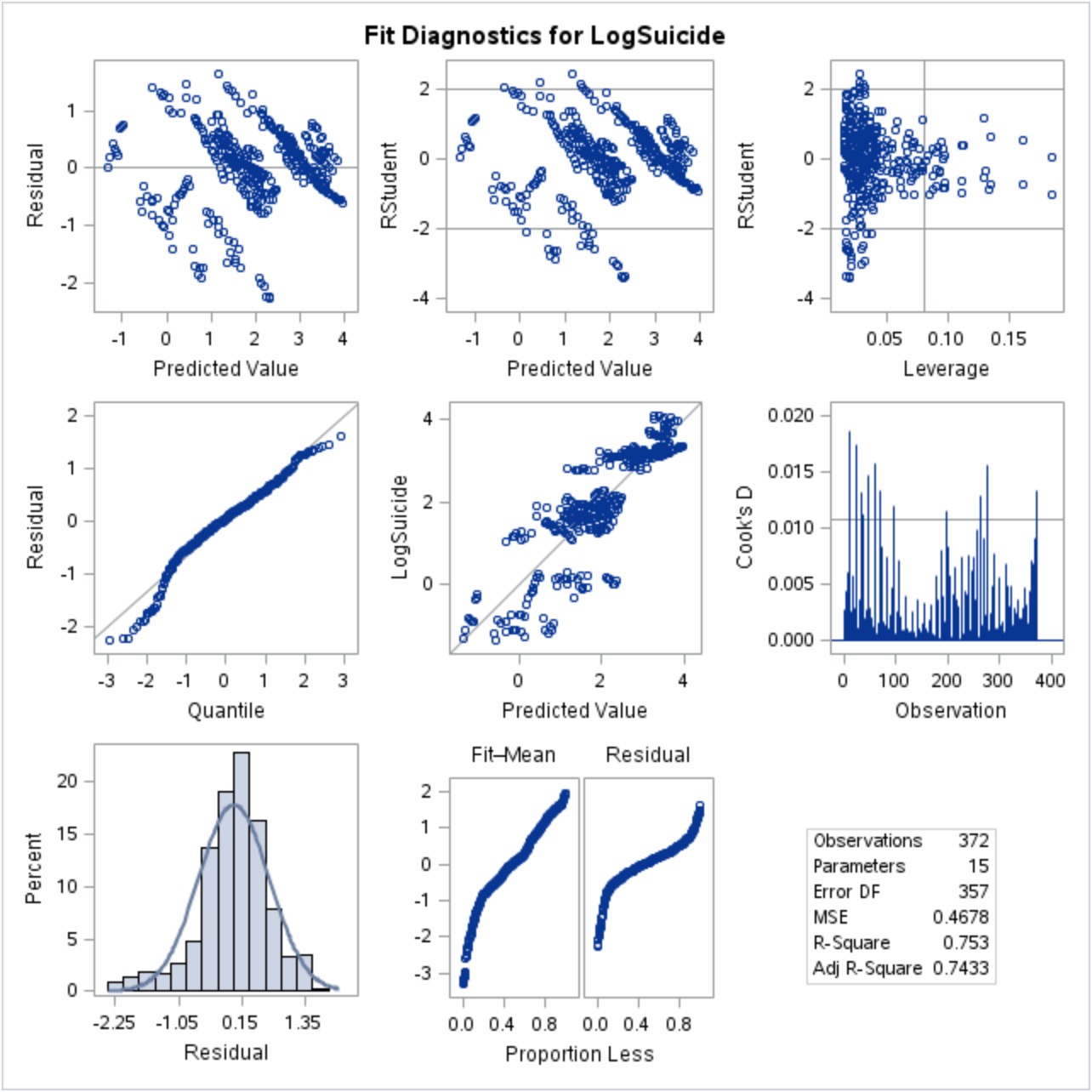


Figure 2 Custom Model fit diagnostics for LogSuicide

Building the Model

Predicted(LogSuicide) = β0 + β1\*(Year) + β2\*(Gender\_0) + β3\*(Population) + β4\*(Generation\_0) + β5\*(Generation\_1) + β6\*(Generation\_2) + β7\*(Generation\_3) + β8\*(Generation\_4) + β9\*(Unemployment) + β10\*(SP\_Return)+ β11 \*(Log(MassShooting\_Victims) + β12\*(Unemployment)(Generation\_0) + β13\*(Unemployment)(Generation\_1) + β14\*(Unemployment)(Generation\_2) + β15\*(Unemployment)(Generation\_3) + β16\*(Unemployment)(Generation\_4)

Fit the Model

|  |  |
| --- | --- |
| **Fitted Selection Models** | |
| Forward | Predicted(LogSuicide) = -83.612 + 0.041 (Year) – 1.490(Gender\_0) + 4.183 (Generation\_0) + 3.369(Generation\_1) + 3.432(Generation\_2) + 2.956(Generation\_3) + 1.747(Generation\_4) + 0.071(UnemploymentRate) + (8.156 x 10-9)(population) |
| Backward | Predicted(LogSuicide) = -83.612 + 0.041 (Year) – 1.490(Gender\_0) + 4.183 (Generation\_0) + 3.369(Generation\_1) + 3.432(Generation\_2) + 2.956(Generation\_3) + 1.747(Generation\_4) + 0.071(UnemploymentRate) + (8.156 x 10-9)(population) |
| Stepwise | Predicted(LogSuicide) = -86.741 + 0.043 (Year) – 1.482(Gender\_0) + 4.132 (Generation\_0) + 3.369(Generation\_1) + 3.554(Generation\_2) + 2.989(Generation\_3) + 1.757(Generation\_4) + 0.071(UnemploymentRate) |
| Custom | Predicted(LogSuicide) = -83.039 + 0.041 (Year) – 1.490(Gender\_0) + (8.554 x 10-9)(Population) + 2.978 (Generation\_0) + 3.479 (Generation\_1) + 3.290 (Generation\_2) + 3.024(Generation\_3) – 0.186 (Generation\_4) + 0.006(Unemployment) + 0.190(Unemployment)(Generation\_0) – 0.030(Unemployment)(Generation\_1) + 0.010(Unemployment)(Generation\_2) – 0.025(Unemployment(Generation\_3) + 0.304 (Unemployment)(Generation\_4) |

Table 5 Fitted Models

Custom Model: Parameters and Interpretation

Below is an interpretation of each parameter estimate used in our custom model in the following format: parameter estimate, 95% confidence interval, interpretation.

β0 = -83.039 (-102.276, -63.801) – The intercept in this model provides an estimated -83.039 Log(Suicide Rate) with all other variables being equal to zero. This is a pure extrapolation of our regression model and does not provide any clear nor practical meaning (as logs are bounded by (0, ∞).

β1 = 0.041 (0.032, 0.051) – This parameter indicates that a one unit increment in year influences the median Log(Suicide Rate) by a factor of e0.041, or 1.042 (4.2% increase). A 95% confidence interval for this multiplicative change is (e0.032, e0.051), yielding a multiplicative increase between 3.7% and 5.5% on a 95% confidence interval.

β2 = -1.490 (-1.630, -1.350) – This is an adjustment to the intercept for a female in reference to a male. A female is expected to influence the Log(Suicide Rate) by 10-1.490, or 0.032 (96.8% decrease). A 95% confidence interval for this decrease is (10-1.630, 10-1.350), which equals an estimated decrease between 95.5% and 97.7%.

β3 = 8.554 x 10-9 (0, 2x10-8) – This parameter indicates that a one unit increment in the population influences the median Log(Suicide Rate) by a factor of EXP(8.554 x 10-9). A 95% confidence interval for this multiplicative change is (0, EXP(2 x 10-8)).

β4 = 2.978 (1.145, 4.807) – This is the adjustment of the intercept for the G.I. Generation with reference to Generation Z. The G.I. Generation is estimated to multiplicatively influence the median Log(Suicide Rate) by 102.978, or by 950.6. A 95% confidence interval for this multiplicative change is 13.96 and 64,121.

β5 = 3.479 (1.982, 4.976) – This is the adjustment of the intercept for the Silent Generation with reference to Generation Z. The Silent Generation is estimated to multiplicatively influence the median Log(Suicide Rate) by 103.479, or by 3,013. A 95% confidence interval for this multiplicative change is 95.9 and 94,624.

β6 = 3.290 (1.674, 4.906) – This is the adjustment of the intercept for the Boomers Generation with reference to Generation Z. The Boomers Generation is estimated to multiplicatively influence the median Log(Suicide Rate) by 103.290, or by 1,950. A 95% confidence interval for this multiplicative change is 47.2 and 60,538.

β7 = 3.024 (1.523, 4.525) – This is the adjustment of the intercept for Generation X with reference to Generation Z. Generation X is estimated to multiplicatively influence the median Log(Suicide Rate) by 103.024, or by 1,057. A 95% confidence interval for this multiplicative change is 33.3 and 33,497.

β8 = -0.186 (-1.704, 1.332) – This is the adjustment of the intercept for the Millennials Generation with reference to Generation Z. The Millennials Generation is estimated to multiplicatively influence the median Log(Suicide Rate) by 10-0.186, or by 0.652. A 95% confidence interval for this multiplicative change is (0.020, 21.5).

β9 = 0.006 (-0.176, 0.188) – A one percent increase in the unemployment rate is expected to change the median of Log(Suicide Rate) by a multiplicative factor of e0.006, or by 1.006 (0.6% increase). A 95% confidence interval for this multiplicative change is 0.839 and 1.21.

β12 = 0.190 (-0.86, 0.466) – For the G.I. Generation, every one percent increase in unemployment is expected to multiplicatively influence the median of Log(Suicide Rate) by a factor of e0.190, or 1.21, when compared to Generation Z. A 95% confidence interval for the multiplicative factor is (0.423, 1.594).

β13 = -0.030 (-0.237, 0.177) – For the Silent Generation, every one percent increase in unemployment is expected to multiplicatively influence the median of Log(Suicide Rate) by a factor of e-0.030, or 1.21, when compared to Generation Z. A 95% confidence interval for the multiplicative factor is (0.423, 1.594).

β14 = 0.010 (-0.215, 0.235) – For the Boomers Generation, every one percent increase in unemployment is expected to multiplicatively influence the median of Log(Suicide Rate) by a factor of e0.190, or 0.97, when compared to Generation Z. A 95% confidence interval for the multiplicative factor is (0.807, 1.265).

β15 = -0.025 (-0.234, 0.184) – For Generation X, every one percent increase in unemployment is expected to multiplicatively influence the median of Log(Suicide Rate) by a factor of e-0.025, or 0.975, when compared to Generation Z. A 95% confidence interval for the multiplicative factor is (0.791, 1.202).

β16 = 0.304 (0.093, 0.515) – For the Millennials, every one percent increase in unemployment is expected to multiplicatively influence the median of Log(Suicide Rate) by a factor of e0.304, or 1.36, when compared to Generation Z. A 95% confidence interval for the multiplicative factor is (1.10, 1.67).

**FINAL CONCLUSIONS FOR OBJECTIVE 1**

While the unemployment rate is a figure that encompasses all unemployed job seekers in a given year, the unemployment rate will likely vary based on the age of the individual(s). To account for this, and to keep parameters consistent with the previously attempted models (forward, backward, and stepwise selections), we proposed an interaction term between the generation classification of an age group in a given year and the unemployment rate of that year.

Given the fitted models, our analysis demonstrated that our custom fitted model, which introduced an interaction between the generation classification of a population and unemployment rate, yielded the best model to predict suicide rates. Other significant predictors that we found and included in this model are: gender, year, population, and generation. Using these parameters to establish a predictive model for the suicide rate in a given year, we established the following fitted model:

**Predicted(LogSuicide) = -83.039 + 0.041(Year) – 1.490(Gender\_0) + (8.554 x 10-9)(Population) + 2.978 (Generation\_0) + 3.479(Generation\_1) + 3.290(Generation\_2) + 3.024(Generation\_3) – 0.186 (Generation\_4) + 0.006(Unemployment) + 0.190(Unemployment)(Generation\_0) – 0.030(Unemployment)(Generation\_1) + 0.010(Unemployment)(Generation\_2) – 0.025(Unemployment(Generation\_3) + 0.304 (Unemployment)(Generation\_4)**

In this above model, it is estimated that the suicide rate in any given year, with an estimation for Generation Z and males, reduces to:

**Predicted(LogSuicide) = -83.039 + 0.041(Year) + (8.554 x 10-9)(Population) + 0.006(Unemployment)**

From the above regression, we see that the median Log(Suicide Rate) increases multiplicatively by a factor of 1.04 (e0.041) annually, a factor of 1.000000009 (EXP(**8.554 x 10-9))** with a one unit increase in population, and a factor of 1.006 (e0.006) with a one percent increase of unemployment rate.

For practical significance, we cannot control/influence the variable year which is included in our custom model. The other variables, gender, generation, population and unemployment rate are not bound in the same way, but there can be a limitation in how often data is updated, so the model will need to be updated when newer information is available, but the frequency would be in terms of months. This makes the model only as dynamic as the least frequently updated variable.

As these data are observational, and rightfully so, there are no causal inferences drawn from these conclusions or relationships between any significantly tested parameter and suicide rates. Due to this constraint, our scope of inference is limited to the analyzed data set. Although we cannot draw casual inferences from this study, we can conclude that the included parameters may have some significant influence on the rates of suicide. Any trends found above should be extrapolated with caution, as this analysis was conducted only for the United States within the years of 1985 and 2015.

Suicide is a very complex issue to discuss and model, as there are countless significant explanatory parameters that are not included in our analysis. In addition, there are numerous variables which may not be capable of direct measure. It is also very possible that the parameters of significance here are not significant when accounting for other significant parameters excluded in this data.

If this data is to be used to determine when to post suicide prevention campaigns, additional research would be needed to test the effectiveness and build a more accurate predictive model as what is being measured is dynamic.

#### Objective 2: 2-Way ANOVA

**OBJECTIVE**

Our second objective of the analysis is a 2-Way ANOVA and started with studying the relationship between suicides and two categorical predictors; gender and generation. Like the previous analysis, we used the logged suicides per 100k population. Gender has two levels and they were recoded to female = 0 and male = 1. Generation had 6 levels and their recodes are: G.I. Generation = 0, Silent = 1, Boomers = 2, Generation X = 3, Millennials = 4 and Generation Z = 5.

Table 6 and Figure 3 show males have higher logged suicides rates than females and the oldest generation has the lowest logged suicide rate. The assumption of homoscedasticity is not met so we changed our analysis from generation to age groups.

|  |  |
| --- | --- |
| Table 6 Means for Gender and Generation Code | Figure 3 Boxplots of logged suicide rate by Gender and Gencode |

The age groups recodes are: 5-14 years = 0, 15-24 years = 1, 25-34 years = 2, 35-54 years = 3, 55-74 years = 4, 75+ years = 5.

**ANALYSIS**

Right away we can see the ANOVA is balanced, as evidenced by the box plots in Figure 4 which show a noticeable improvement with heteroscedasticity. There is no change with observing males having higher logged suicide rates than females when reviewing the means in the Table 7 or Figure 4. The age group 5-14 years has the lowest logged suicide rate and this is expected.

|  |  |
| --- | --- |
| Table 7 Means for Gender and Age Group | Figure 4 Boxplots of logged suicide rate by Gender and Age Group |

Using Figure 5 below, we see that the updated model meets the required assumptions for a 2-way ANOVA. The residual plot show that variances appear constant, the residuals are relatively normal as seen by the Q-Q plot and independence exists between gender and generation. Neither the Leverage nor the Cook’s D plots show any influential points that need further investigation.

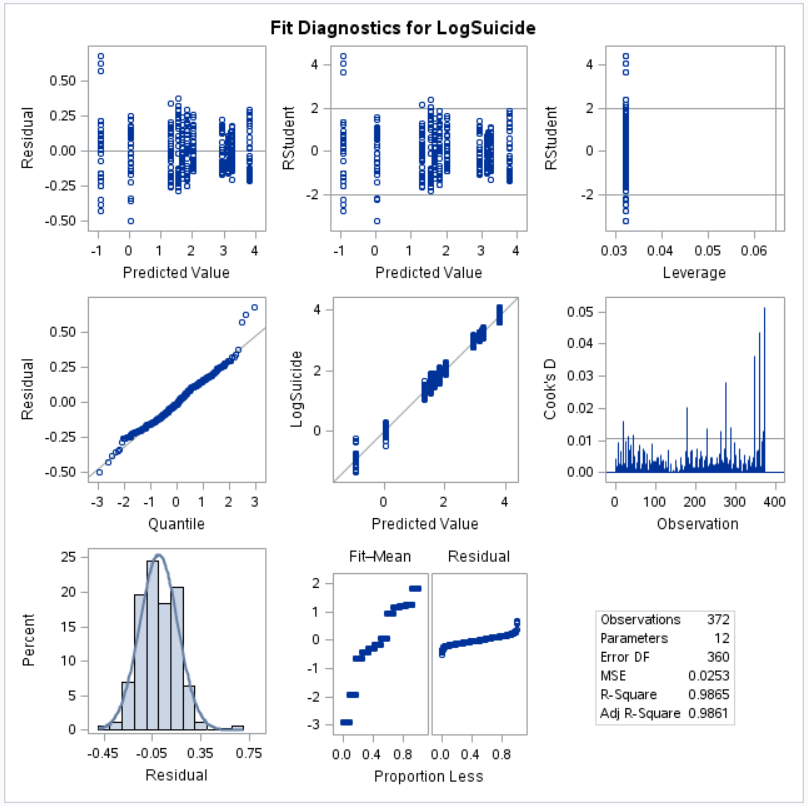


Figure 5 Fit diagnostics for LogSuicide

To evaluate the model results, we review the Type III Sum of Squares F-Table in Table 8. The data not only shows there is a significant influence with gender and age group but that the interaction between gender and age group is significant all having p-values <.0001. Interaction is also visible in the interaction plot in Figure 6. Since the interaction terms are significant, we reject the null hypothesis that an additive model is sufficient and conclude the effect of one factor depends on the other.

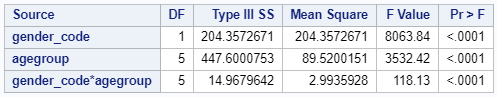


Table 8 Type III Sums of Squares

The interaction plot in Figure 6 shows several lines intersecting which indicates there is an interaction between gender and age group.

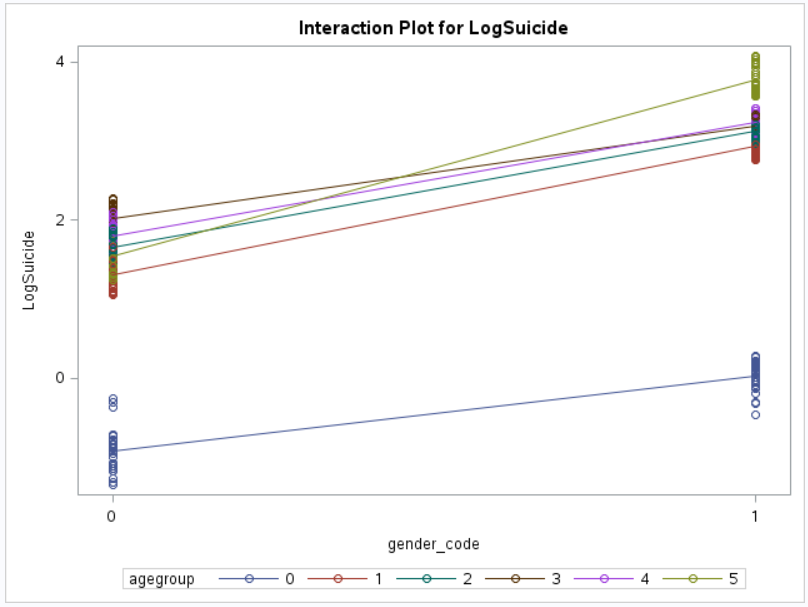


Figure 6 Interaction Plot for Log Suicide

The differences in means by factor and levels can be seen in Table 10. A cross-reference indicator (LSMEAN Number) is also provided in Table 10 to relate to Table 11 which identifies the significant interactions. The results include a Bonferroni correction to account for multiple tests.

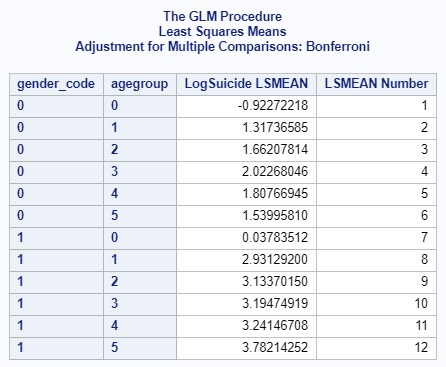


Table 5 LS Means Comparisons

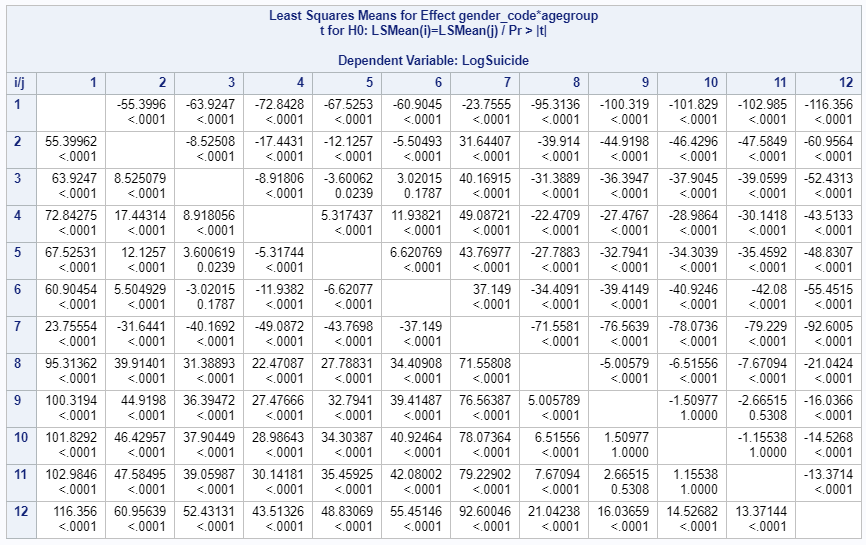


Table 6 LS Means interactions

Nearly all interactions between Gender and Age Groups were significant with p-values < .05 (at a 95% confidence level). Table 12 below shows the non-significant interactions in which there are three for males and one for females. Suicide rates for those in age group one are not statistically significantly different from that of those age group two for their respective genders at a 95% confidence level (Table 12).

|  |  |  |
| --- | --- | --- |
| **Gender** | **Age Group 1** | **Age Group 2** |
| **Female** | 15-24 yrs | 55-74 yrs |
| **Male** | 15-24 yrs | 25-34 yrs |
| 15-24 yrs | 35-54 yrs |
| 25-34 yrs | 35-54 yrs |

Table 7 Non-Significant Interaction Summary of Gender and Age Groups

**CONCLUSION**

The 2-Way ANOVA started with trying to understand the influence of Gender and Generation on logged suicide rates. However, using Generation did not meet the assumptions required to do a 2-Way ANOVA so the objective changed to understanding the influence of Gender and Age Group on logged suicide rates. Notably, males have higher logged suicide rates than females regardless of age group. There is sufficient evidence to suggest a difference in means within Genders, Age Groups, and the interaction between Gender and Age Groups. Since the interaction term Gender\*Age Group is significant with a p-value < .0001 we reject the null hypothesis that an additive model is sufficient and conclude the effect of one factor depends on the other (at a 95% confidence level).

The predicted means with 95% confidence intervals for each interaction is shown in the appendix and the influential interactions are portrayed in Table 11 above. Since suicides cannot be randomized, no causation can be implied. This was purely an observational study so the results cannot be generalized beyond the data in this analysis.

**APPENDIX**

Data Dictionary

This data dictionary provides details on the data used in our study from all data sources.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Variable Name** | **Description** | **Levels** | **Recode** | **Class** | **Unique values** | **Missing** |
| country | Country Name |  |  | factor | 101 | 0% |
| year | Year data is recorded |  |  | integer | 32 | 0% |
| sex | Sex | 2 | male = 1,  female = 0 | factor | 2 | 0% |
| age | Age |  |  | factor | 6 | 0% |
| suicides\_no | Number of suicides |  |  | integer | 2084 | 0% |
| population | Population of age groups (15-24, 25-34, 35-54, 55-75, 75+) |  |  | integer | 25564 | 0% |
| suicides\_100k\_pop | Suicides per 100k people in the population of the age group |  |  | numeric | 5298 | 0% |
| country.year | Country & Year Concatenated |  |  | factor | 2321 | 0% |
| HDI.for.year | Household disposable income |  |  | numeric | 306 | 69.94 |
| gdp\_for\_year. | GDP per Year |  |  | factor | 2321 | 0% |
| GDP\_Capita | GDP per Capita |  |  | integer | 2233 | 0% |
| gencode | Generation code | 5 | G.I. Generation = 0, Silent = 1, Boomers = 2, Generation X = 3, Millennials = 4, Generation Z = 5 | factor | 6 | 0% |
| MassShooting\_Victims | Total killed and injured victims of mass shootings in the US |  |  | integer | 31 | 0% |
| Divorce\_Rate | The divorce rate in the US per year |  |  | numeric | 19 | 0% |
| Unemployment\_Rate | The US unemployment rate per year |  |  | numeric | 32 | 0% |
| SP\_Return | The average annual S&P 500 returns |  |  | numeric | 32 | 0% |

Table 8 Data Dictionary

Scatterplot Matrix

This is the scatterplot matrix performed on the US data before log transformations were applied. Skews can be observed in Suicide\_100k\_Pop, MassShooting\_Victims, Unemployment\_Rate, and Population.



Figure 7 Scatterplot Matrix before transformations

Forward

|  |
| --- |
|  |

Backward

|  |
| --- |
|  |

Stepwise

|  |
| --- |
|  |

Custom

|  |
| --- |
|  |

2-Way ANOVA

|  |
| --- |
|  |
|  |

### **REFERENCES**

1. Suicide Facts

<https://www.nimh.nih.gov/health/statistics/suicide.shtml>

1. Support for Suicidal Individuals on Social and Digital Media Tool Kit

<https://suicidepreventionlifeline.org/wp-content/uploads/2018/09/lifeline_socialmedia_toolkit.pdf>

1. MassShooting\_Victims

Zeeshan-ul-hassan Usmani. (2017; November). Mass Shootings Dataset, Version 5. Retrieved February 10, 2019 from <https://www.kaggle.com/zusmani/us-mass-shootings-last-50-years/downloads/us-mass-shootings-last-50-years.zip/4>.

4 S&P 500 Historical Annual Returns

<https://www.macrotrends.net/2526/sp-500-historical-annual-returns>

1. Civilian Unemployment Rate: -

U.S. Bureau of Labor Statistics, Civilian Unemployment Rate [UNRATE], retrieved from FRED, Federal Reserve Bank of St. Louis; <https://fred.stlouisfed.org/series/UNRATE>, February 10, 2019.

1. Divorce rate in the United States from 1990 to 2017 (per 1,000 of population) <https://www.statista.com/statistics/195955/divorce-rate-in-the-united-states-since-1990/>

SAS CODE:

PROC IMPORT OUT = Master

dataFILE='/home/mtieu0/Project\_1/Master2.csv'

dbms=csv replace;

getnames=YES;

dataRow=2;

/\* Add codes to categorical variables, log skewed data \*/

data stats;

set Master (keep = country year sex population suicides\_100k\_pop GDP\_Capita

generation MassShooting\_Victims Divorce\_Rate Unemployment\_Rate SP\_Return);

if sex = "female" then gender\_code = 0;

else if sex = "male" then gender\_code = 1;

if suicides\_100k\_pop > 0 then LogSuicide = log(suicides\_100k\_pop);

else LogSuicide = 0;

if generation = "G.I. Generation" then gencode = 0;

else if generation = "Silent" then gencode = 1;

else if generation = "Boomers" then gencode = 2;

else if generation = "Generation X" then gencode = 3;

else if generation = "Millenials" then gencode = 4;

else if generation = "Generation Z" then gencode = 5;

else gencode = 6;

MSL = log(MassShooting\_Victims);

run;

/\*Scatterplot matrix\*/

proc sgscatter data = stats;

matrix suicides\_100k\_pop year gender\_code GDP\_Capita

gencode Divorce\_Rate

Unemployment\_Rate SP\_Return population

MassShooting\_Victims / diagonal=(histogram);

run;

/\*Scatterplot matrix transformed data\*/

proc sgscatter data = stats;

matrix LogSuicide year gender\_code GDP\_Capita

gencode Divorce\_Rate Unemployment\_Rate

SP\_Return MSL population / diagonal=(histogram);

run;

/\*VIF\*/

proc reg data = stats plots = all;

model LogSuicide = year gender\_code GDP\_Capita

gencode Divorce\_Rate Unemployment\_Rate

SP\_Return MSL population / vif;

run;

/\* will remove GDP and divorce rate based on high VIFs\*/

/\*VIF, OLS with GDP & Divorce\_Rate removed \*/

proc reg data = stats plots = all;

model LogSuicide = year gender\_code

gencode Unemployment\_Rate

SP\_Return MSL population / vif;

run;

/\*Mallows CP\*/

proc reg data = stats plots = all outest=MallowCP;

model LogSuicide = year gender\_code

gencode Unemployment\_Rate

SP\_Return MSL population / cp;

run;

proc print data = MallowCP;

run;

/\*Forward Selection\*/

proc glmselect data = stats seed=1 outdesign(addinputvars)=Forward\_GLM;

class gender\_code gencode;

model LogSuicide = year gender\_code

gencode Unemployment\_Rate

SP\_Return MSL population

/ selection = forward(stop=cv choose = AIC)

showpvalues cvmethod=random(4) stats=adjrsq;

output out = Suicide\_Forward;

run;

/\*Generates 95% confidence intervals for parameter estimates\*/

%put &=\_GLSIND;

proc glm data = stats;

class gender\_code gencode;

model LogSuicide = &\_GLSIND / solution CLPARM;

run;

/\*Generates residual, Cook’s D, and leverage plots\*/

proc reg data=Forward\_GLM

plots=(CooksD RStudentByLeverage DFFITS DFBETAS);

model LogSuicide=&\_GLSMOD / ;

run;

/\*Backward\*/

proc glmselect data = stats seed=1 outdesign(addinputvars)=Backward\_GLM;

class gender\_code gencode;

model LogSuicide = year gender\_code

gencode Unemployment\_Rate

SP\_Return MSL population

/ selection = backward(stop=cv choose = AIC)

showpvalues cvmethod=random(4) stats=adjrsq;

output out = Suicide\_Backward;

run;

/\*Generates 95% confidence intervals for parameter estimates\*/

%put &=\_GLSIND;

proc glm data = stats;

class gender\_code gencode;

model LogSuicide = &\_GLSIND / solution CLPARM;

run;

/\*Generates residual, Cook’s D, and leverage plots\*/

proc reg data = Backward\_GLM

plots=(CooksD RStudentByLeverage DFFITS DFBETAS);

model LogSuicide=&\_GLSMOD /;

run;

/\*Stepwise\*/

proc glmselect data = stats seed=1 outdesign(addinputvars)=Step\_GLM;

class gender\_code gencode;

model LogSuicide = year gender\_code

gencode Unemployment\_Rate

SP\_Return MSL population

/ selection = stepwise(stop=cv SLE=0.1 SLS=0.05 choose = AIC)

showpvalues cvmethod=random(4) stats=adjrsq;

output out = Suicide\_Stepwise;

run;

/\*Generates 95% confidence intervals for parameter estimates\*/

%put &=\_GLSIND;

proc glm data = stats;

class gender\_code gencode;

model LogSuicide = &\_GLSIND / solution CLPARM;

run;

/\*Generates residual, Cook’s D, and leverage plots\*/

proc reg data = Step\_GLM

plots=(CooksD RStudentByLeverage DFFITS DFBETAS);

model LogSuicide=&\_GLSMOD /;

run;

/\*Custom, added Unemployment vs Generation Interaction \*/

proc glmselect data = stats seed=1 outdesign(addinputvars)=Custom\_GLM;

class gender\_code gencode;

model LogSuicide = year gender\_code

SP\_Return MSL population gencode | Unemployment\_Rate

/ selection = backward(stop=cv)

showpvalues cvmethod=random(4) stats=adjrsq;

output out = Suicide\_Custom;

run;

/\*Generates 95% confidence intervals for parameter estimates\*/

%put &=\_GLSIND;

proc glm data = stats;

class gender\_code gencode;

model LogSuicide = &\_GLSIND / solution CLPARM;

run;

/\*Generates residual, Cook’s D, and leverage plots\*/

proc reg data = Custom\_GLM

plots=(CooksD RStudentByLeverage DFFITS DFBETAS);

model LogSuicide=&\_GLSMOD /;

run;

/\* --2-Way ANOVA-- \*/

/\* Get Basic Stats \*/

PROC MEANS DATA=stats;

CLASS gender\_code gencode;

VAR LogSuicide;

RUN;

/\* Box Plot of generations by gender\*/

PROC SGPLOT DATA=stats;

VBOX LogSuicide / CATEGORY=gencode GROUP=gender\_code;

RUN;

/\* 2-Way ANOVA Model \*/

proc glm data=stats PLOTS=(DIAGNOSTICS RESIDUALS);

class gender\_code gencode;

model LogSuicide = gender\_code gencode gender\_code\*gencode;

/\* Data showed a bit non-normality with residuals so using Age Group instead of Generation \*/

/\* Add recode for Age Groups \*/

data stats;

set stats;

if age = "5-14 years" then agegroup = 0;

else if age = "15-24 years" then agegroup = 1;

else if age = "25-34 years" then agegroup = 2;

else if age = "35-54 years" then agegroup = 3;

else if age = "55-74 years" then agegroup = 4;

else agegroup = 5;

run;

/\* Get Basic Stats \*/

PROC MEANS DATA=stats;

CLASS gender\_code agegroup;

VAR LogSuicide;

RUN;

/\* Box Plot of age groups by gender\*/

PROC SGPLOT DATA=stats;

VBOX LogSuicide / CATEGORY=agegroup GROUP=gender\_code;

RUN;

/\* 2-Way Model with Interaction \*/

proc glm data=stats PLOTS=(DIAGNOSTICS RESIDUALS);

class gender\_code agegroup;

model LogSuicide = gender\_code agegroup gender\_code\*agegroup;

lsmeans gender\_code\*agegroup / pdiff tdiff adjust=bon; /\* lsmeas with Bonferroni correction \*/

/\* Using proc mixed to get table format and fit stats\*/

proc mixed data=stats PLOTS=(RESIDUALPANEL);

class gender\_code agegroup;

model LogSuicide = gender\_code agegroup gender\_code\*agegroup;

lsmeans gender\_code\*agegroup / pdiff tdiff CL adjust=bon; /\* lsmeas with Bonferroni correction \*/

run;