Regression Analysis Final Report: Suicide Rates and Associated Social Demographics

Melissa L. Russell, Jonathan I. Voth, Nicholas J. Wawee

Bowling Green State University

Abstract

Preventing suicide is usually thought of as helping those that are contemplating suicide in the present moment; however, using decades of past data to help determine the characteristics of victims can lead to insightful information for preventing future suicides. This current study aimed to identify the most significant factors associated with suicide rates through a regression analysis. The dataset examined, *Suicide Rates Overview 1985 to 2016,* was found on the website Kaggle.com (Rusty, 2018). For the purpose of our analysis, only suicide rates in the United States were analyzed. Three significant regressors were included in our final model: age, sex, and GDP per capita. Our main findings are that men have higher suicide rates than women, men's suicide rates generally increase with age and are highest among the oldest age range, women's suicide rates peak around middle-age, and GDP per capita has an inverse relationship with suicide rates for both sexes. By uncovering this information, we have a better idea of the characteristics of those in the United States that are more at risk of committing suicide, so we can more effectively help those in need.

Introduction

Suicide is an issue that has plagued humanity since the beginning of time, yet it is impossible to eradicate. It often leaves loved ones behind with the unanswerable question: Why? Suicide has no vaccine, no remedy, no cure. Therefore, if we wish to fix the problem of suicide, we need to focus on preventative measures – we need to look to predictions. What do all those that commit suicide have in common? How can we predict the characteristics of those that have increased chances of committing suicide in the future, so that we can take preventative measures to help them? A good start is by looking at the data.

A study printed in the *Brazilian Journal of Psychiatry* in 2012 looked at suicide rates in São Paulo, Brazil (one of the highest-population cities in the world) from 1996 to 2009 (Bando et al., 2012). The authors of this study found that the overall suicide rate for this city during this timeframe was 4.6 per 100,000 inhabitants. Some of their most significant findings include that men committed suicide at a rate of three to four times higher than women, men typically committed suicide by firearm or hanging, while women typically committed suicide by poison, and finally, single/divorced/widowed people had higher rates of suicide than married people (Bando et al., 2012).

Similarly, a study conducted in the Republic of Korea also attempted to discover the factors associated with Korea's very high suicide rates (Kim, 2020). The suicide rate found was 26.5 per 100,000 persons, and like the Brazilian study, it found that males had much higher rates than females (Kim, 2020). Using a regression analysis, the author also discovered that income level and suicide rate had a negative correlation, while those that are heavy drinkers and those that are elderly had a positive association with suicide rates.

For our analysis, we will be focusing on suicides in the United States from 1985-2015 using a published dataset (Rusty, 2018). However, we do expect our findings to mirror the studies from Brazil and Korea. Specifically, we expect that men will likely have higher suicide rates than women and that low-income persons will have higher suicide rates that those in the middle or upper class. However, unlike the Korean study, we do not expect suicide rates and the elderly to have a positive association, due to the cultural differences between the United States and Korea. Examining the relationships between our regressors and our response variable, suicide rates, will allow us to be able to identify certain characteristics of those that make up the "high-risk" category, which in turn will aid preventative measures.

Methods

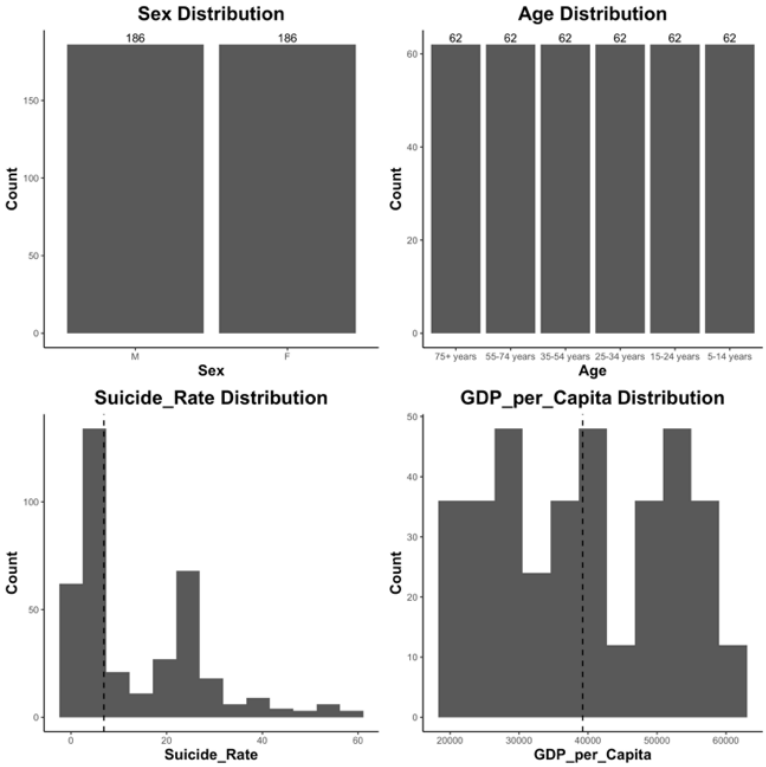


Figure 1: Distribution of All Variables

A multiple linear regression model was fit to the data using a backwards stepwise regression approach. The distribution of each variable is displayed in **Figure 1.** As shown in **Figure 1**, some variables are evenly distributed while others are not. The sex and age of each victim are evenly distributed. The numerical variables appear to have uneven distributions. A right skew is apparent in the response variable, and the GDP has a slight left skew. The dotted line indicates where the median lies for each variable.

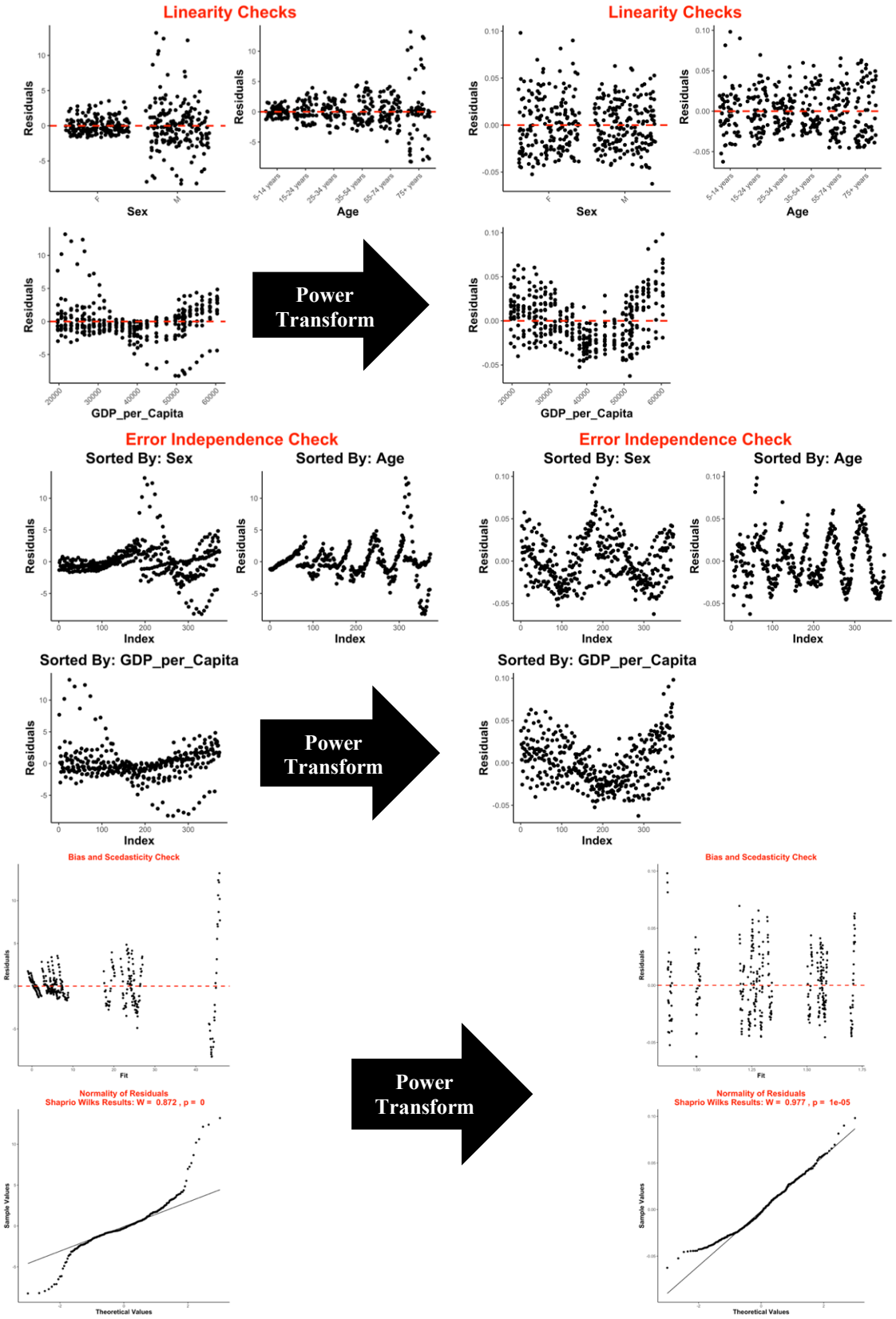


Figure 2: Transformation Effect on Model Assumptions

Figure 2 depicts the effect of the power transformation of the response variable on the model assumptions. A BoxCox (Box, 1964) transformation was employed to find the optimal power at which the log likelihood is maximized. An optimal power of 0.1414 was used to transform the response variable to alleviate model assumption violations.

The linearity plots reveal that the transformation removed the unevenness around zero for sex and age, but not for GDP. The transformed response variable allows for the errors to be more independent from each other, however, there still appears to be groups where residuals are weighted above and below zero for all regression variables. In both types of response variables, the model appears to be unbiased. The power transformation allowed for the model to transition from heteroscedastic to homoscedastic. Although the model fails to reject the Shapiro Wilks hypothesis (Royston, 1982), the qqplots reveal that the power transformation made the residuals more normal.

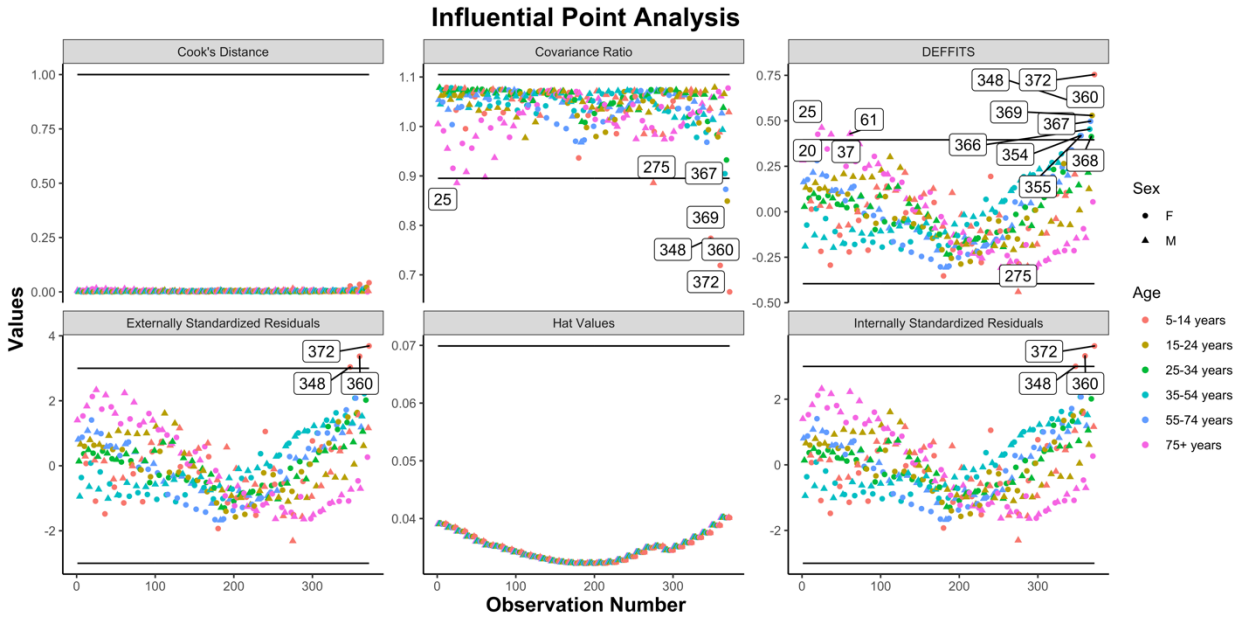


Figure 3: Influential Analysis of Transformed Residuals

Figure 3 depicts measures that elucidate outlier, leverage, and influential points. The hat values portion of the figure does not show any leverage points. Observation numbers 348, 360, and 372 show to be both considered as outliers and influential. Interestingly, they appear to be all in the 5-14 years age range and female. Another group of points that appear to be influential in at least one of the tests are males 75+ years old (Observations 25, 37, and 61). Further investigation of these suicides could be done to uncover similarities.

Hypothesis Testing

Multiple hypothesis tests were done to investigate if the final model is statistically significant. The goal of hypothesis testing is to investigate if the regressor variables significantly affect suicide rates and are significantly different from each other. The hypothesis tests that were conducted are listed below.

- H_o: $\beta 1 = 0$ vs. H_A: $\beta 1 \neq 0$
    - p-value = < 2e-16; We reject the null hypothesis. There is sufficient evidence to suggest $\beta 1$ is not equal to 0, meaning the effect of males is significant.
- H_o: $\beta 2 = \beta 3 = \beta 4 = \beta 5 = \beta 6 = 0$ vs. H_A: At least one $\neq 0$
    - p-value = < 2.2e-16; We reject the null hypothesis. There is sufficient evidence to suggest at least one of $\beta 2$, $\beta 3$, $\beta 4$, $\beta 5$, $\beta 6$ is not equal to 0, meaning the effect of age is significant.
- H_o: $\beta 8 = \beta 9 = \beta 10 = \beta 11 = \beta 12 = 0$ vs. H_A: At least one $\neq 0$
    - p-value = < 2.2e-16; We reject the null hypothesis. There is sufficient evidence to suggest at least one of $\beta 8$, $\beta 9$, $\beta 10$, $\beta 11$, $\beta 12$ is not equal to 0, suggesting evidence of interaction between sex and age in the model.
- H_o: $\beta 2 = \beta 3 = \beta 4 = \beta 5 = \beta 6$ vs. H_A: $\beta 2 \neq \beta 3 \neq \beta 4 \neq \beta 5 \neq \beta 6$

  - Each test resulted in a p-value < 0.05. We have enough evidence to reject the null hypothesis and conclude that suicide rates amongst the different age groups for females is significantly different.

- H_o: $\beta2+\beta8 = \beta3+\beta9 = \beta4+\beta10 = \beta5+\beta11 = \beta6+\beta12$ vs. H_A: $\beta2+\beta8 \neq \beta3+\beta9 \neq \beta4+\beta10 \neq \beta5+\beta11 \neq \beta6+\beta12$
  - Only two tests resulted in a p-value > 0.05. The difference between the suicide rate of males 25-34 years and 35-54 years is not significant (p-value = 0.05593). Additionally, the difference between the suicide rates in males 35-54 years and 55-74 years resulted in a p-value of 0.1359. This is not significant. This ultimately means that males in these age groups commit suicide at very similar rates.

Results

The final regression model has three significant regressor variables: sex, age, and GDP per capita. Sex and age are categorical variables with two and six levels, respectively. **Table 1** displays each variable estimate significant to the model, the corresponding test statistic and p-value, and a 95% confidence interval of the estimate.

Table 1: Regression Coefficients Details

Coefficient	Estimate	Test Statistic	P-Value	95% CI
Intercept	0.8987	133.140	< 2e-16	(0.885, 0.912)
sexmale	0.1274	18.134	< 2e-16	(0.114, 0.141)
age15-24 years	0.3267	46.515	< 2e-16	(0.313, 0.341)
age25-34 years	0.3868	55.068	< 2e-16	(0.373, 0.401)
age35-54 years	0.4530	64.494	< 2e-16	(0.439, 0.467)
age55-74 years	0.4133	58.842	< 2e-16	(0.400, 0.427)
age75+ years	0.3656	52.046	< 2e-16	(0.352, 0.379)
GDP_per_capita	-5.207e-7	-4.473	1.04e-5	(-7.5e-7, -2.9e-7)
sexmale: age15-24 years	0.1815	18.271	< 2e-16	(0.162, 2.010)
sexmale: age25-34 years	0.1652	16.633	< 2e-16	(0.146, 0.185)
sexmale: age35-54 years	0.1125	11.324	< 2e-16	(0.093, 0.132)
sexmale: age55-74 years	0.1627	16.377	< 2e-16	(0.143, 0.182)
sexmale: age75+ years	0.3364	33.868	< 2e-16	(0.317, 0.356)

Because of the multiple levels within the categorical variables, there are 12 total regression equations in the model, one for each sex and age combination. As previously mentioned, a transformation was applied to the response variable. As a result, the response variable (suicide rates) was raised to the power of 0.1414. The intercept of each equation changes for each sex and age combination. The slope is GDP per capita and remains the same throughout. The equations are outlined in **Table 2** below.

Table 2: Model Equation Differences in Sex and Age

Group	Intercept
Male, 5-14 years	1.0261
Male, 15-24 years	1.5343
Male, 25-34 years	1.5781
Male, 35-54 years	1.5916
Male, 55-74 years	1.6021
Male, 75+ years	1.7218
Female, 5-14 years	0.8987
Female, 15-24 years	1.2254
Female, 25-34 years	1.2855
Female, 35-54 years	1.3517
Female, 55-74 years	1.3120
Female, 75+ years	1.2643

Response = Suicide Rate$^{0.1414}$, Slope (GDP_per_capita) = -5.207e-7

The R^2 and adjusted R^2 values for the final model are 0.9872 and 0.9868, respectively. In every age group, males have a higher suicide rate (per 100K population) than their female counterparts. This is consistent with previous models (Kim, 2020). The suicide rate amongst males increases with age and is highest in the 75+ years age group. However, among a few age groups, the difference was not significant. For females, the suicide rate peaks in the 35-54 years age group. It is important to note that this does not mean the total number of suicides is the highest in these groups. GDP per capita has an inverse effect on suicide rates. In the United States, GDP per capita has steadily increased since 1985 and has decreased the suicide rates for both sexes and across all age groups.

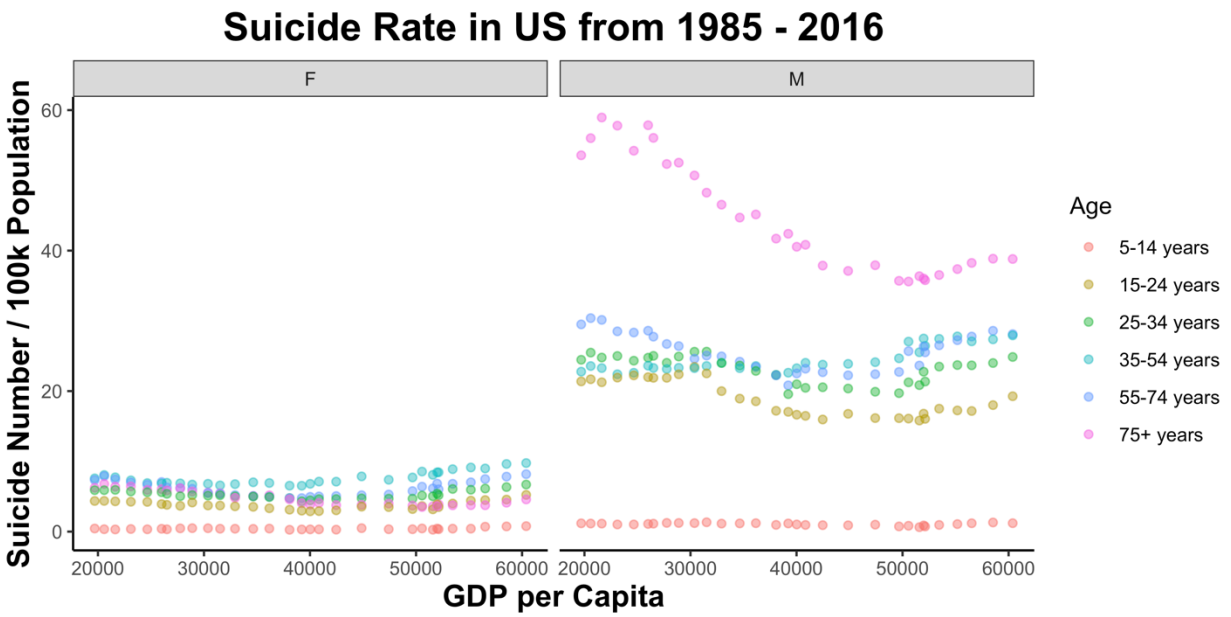


Figure 4: United States Suicide Rates in Women and Men

Discussion

Our regression analysis provides many interesting and important results. Some results were expected; some were surprising. Our overall model shows a very strong fit to the data – over 98% in the variance of suicide rates are explained by our regressor variables. This tells us that the combination of sex, age, and GDP per capita are responsible for predicting suicide rates in the United States. Before running any analyses, we predicted that men would likely have higher suicide rates than women based on the studies in Brazil (Bando et al., 2012) and Korea (Kim, 2020), as both of these studies found significantly higher suicide rates for men than women. As shown in our results, this prediction was accepted - men did have significantly higher suicide rates than women. **Figure 4** clearly shows the right side (men) having almost all age groups far higher than the highest age group on the left side (women). The only age group not higher than women's is the 5-14 years, which looks very similar to the women's side for the same age group. This is to be expected, as children rarely have the motivation or means to

commit suicide. However, it should be noted that three subpopulations of women in the 5-14-year age range were found to be outliers and influential.

The other major prediction we had prior to the analysis was that the elderly would not have higher suicide rates than younger adults. Although the Korea study did show this in their results (Kim, 2020), we believed that this was due to cultural differences in Korea compared to the U.S., so we did not think that this would be shown in our results. This turned out to be somewhat true. Suicide rates for males generally increase with age (although not all age groups are significantly different) and actually were highest among the 75+ years age group. For females, however, suicide rates peak around mid-life (35-54 years), meaning that elderly females do not have higher suicide rates than younger adult women. Thus, our prediction was accurate for females, but completely wrong for males. As shown in **Figure 4**, the 75+ years age group for men is very far above all the rest of the age groups, which was a surprise.

Conclusion

Our goal of this analysis was to identify significant factors in predicting suicide rates in the United States. With a model that explained over 98% of the variance in suicide rates, we were able to achieve this. Sex, age, and GDP per capita all were shown to be significant regressors. This is important, because it shows the characteristics of those that are in the higher-risk category for suicide. For example, the highest category we found in this study was men aged 75+ years old with lower GDP per capita. Because we know this, we can target preventative measures more efficiently. We can make psychological help and treatment more accessible to this group in hopes of preventing future suicides. Our regression analysis of suicide rates in the United States attempts to help thousands of victims each year in the most effective way possible – by using past data to help predict the future.

Limitations & Future Work

A limitation of this study include that the data was not taken at an individual level and was already binned. Our model was fit to data that was taken at a population level. Fitting a model to data that has characteristics of individuals that have and have not committed suicide within a population may lead to a more informed result. Additionally, income information was not provided as a population. It is assumed in our model that GDP per capita directly reflects the economic status of individuals within the population.

After power transformation of the response variable, the errors failed to show to be independent. Since the data was taken at yearly intervals, the suicide rate of one year may be influenced by the previous year. This phenomenon is called autocorrelation. Our data show that autocorrelation is present because of the continued violation of the independence of errors assumption.

Future work would include the removal of autocorrelation to create a better model. One avenue to remove autocorrelation is to make the data stationary. This is done by differencing or subtracting the mean of the observations of the not stationary variable. To validate the data is stationary, we would inspect the autocorrelation and partial autocorrelation function graphs for evidence of immediate truncation of autocorrelation. We would the employ the KPSS test (Kwiatkowski, Phillips, Schmidt, & Shin, 1992) to statistically validate that the data is stationary. The final step would be to verify that the errors are independent as shown in **Figure 2**.

References

Bando, Daniel H., Brunoni, André R., Fernandes, Tiótrefis G., Benseñor, Isabela M., & Lotufo, Paulo A.. (2012). Suicide rates and trends in São Paulo, Brazil, according to gender, age and demographic aspects: a joinpoint regression analysis. *Brazilian Journal of Psychiatry*, *34*(3), 286-293. https://doi.org/10.1016/j.rbp.2012.02.001

Box, G. E. P. and Cox, D. R. (1964) An analysis of transformations (with discussion). Journal of the Royal Statistical Society B, 26, 211–252.

Kim, Agnus M. "Factors Associated with the Suicide Rates in Korea." *Psychiatry Research*, vol. 284, no. 112745, Feb. 2020. *Elsevier*, doi:https://doi.org/10.1016/j.psychres.2020.112745.

Kwiatkowski, D., Phillips, P. C., Schmidt, P., & Shin, Y. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root. *Journal of Econometrics, 54*(1-3), 159-178. doi:10.1016/0304-4076(92)90104-y

Royston, Patrick (1982). An extension of Shapiro and Wilk's W test for normality to large samples. Applied Statistics, 31, 115–124. doi: 10.2307/2347973.

Rusty (2018). *Suicide Rates Overview 1985 to 2016*, Version 1. Retrieved from https://www.kaggle.com/russellyates88/suicide-rates-overview-1985-to-2016.

Acknowledgements

We would like to thank Dr. Shuchismita Sarkar for her help with the regression analysis in general by teaching the MSA 5020 course at Bowling Green State University.

Appendix A: Code for Model Fitting, Distribution Check, Assumption Checking, Power Transformation, Influential Point Analysis, and Scatter Plot Generation

```
df = read.csv("/Users/nickwawee/Desktop/BGSU/MSA_5020/Project/master.csv", st
ringsAsFactors = T)
str(df)
```

```
## 'data.frame':    27820 obs. of  12 variables:
##  $ country           : Factor w/ 101 levels "Albania","Antigua and Barbuda
",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ year              : int  1987 1987 1987 1987 1987 1987 1987 1987 1987 1
987 ...
##  $ sex               : Factor w/ 2 levels "female","male": 2 2 1 2 2 1 1 1
2 1 ...
##  $ age               : Factor w/ 6 levels "15-24 years",..: 1 3 1 6 2 6 3
2 5 4 ...
##  $ suicides_no       : int  21 16 14 1 9 1 6 4 1 0 ...
##  $ population        : int  312900 308000 289700 21800 274300 35600 278800
257200 137500 311000 ...
##  $ suicides.100k.pop : num  6.71 5.19 4.83 4.59 3.28 2.81 2.15 1.56 0.73 0
...
##  $ country.year      : Factor w/ 2321 levels "Albania1987",..: 1 1 1 1 1 1
1 1 1 1 ...
##  $ HDI.for.year      : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ gdp_for_year....  : Factor w/ 2321 levels "1,002,219,052,968",..: 727 7
27 727 727 727 727 727 727 727 727 ...
##  $ gdp_per_capita....: int  796 796 796 796 796 796 796 796 796 796 ...
##  $ generation        : Factor w/ 6 levels "Boomers","G.I. Generation",..:
3 6 3 2 1 2 6 1 2 3 ...
```

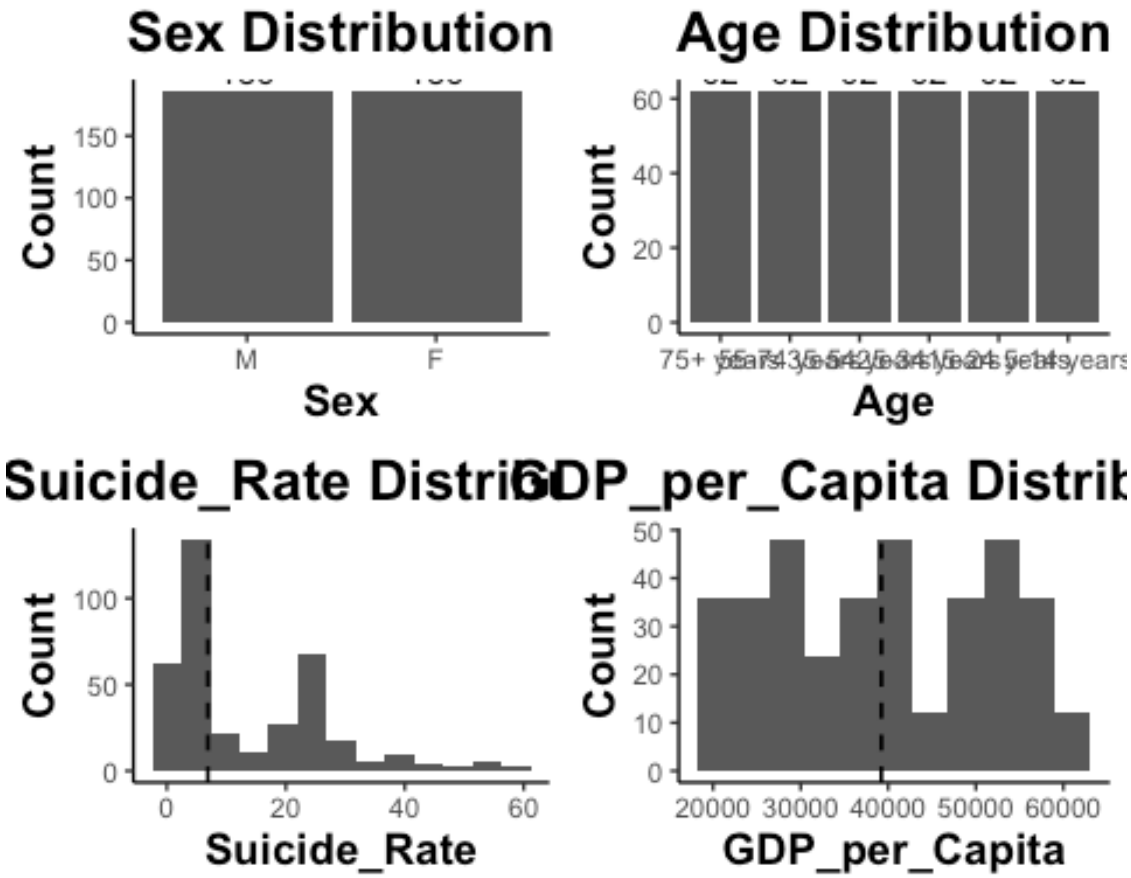
Below will examine the na values in each column

```
for (col in colnames(df)){
  print(length(which(is.na(df[,col]))))
}
```

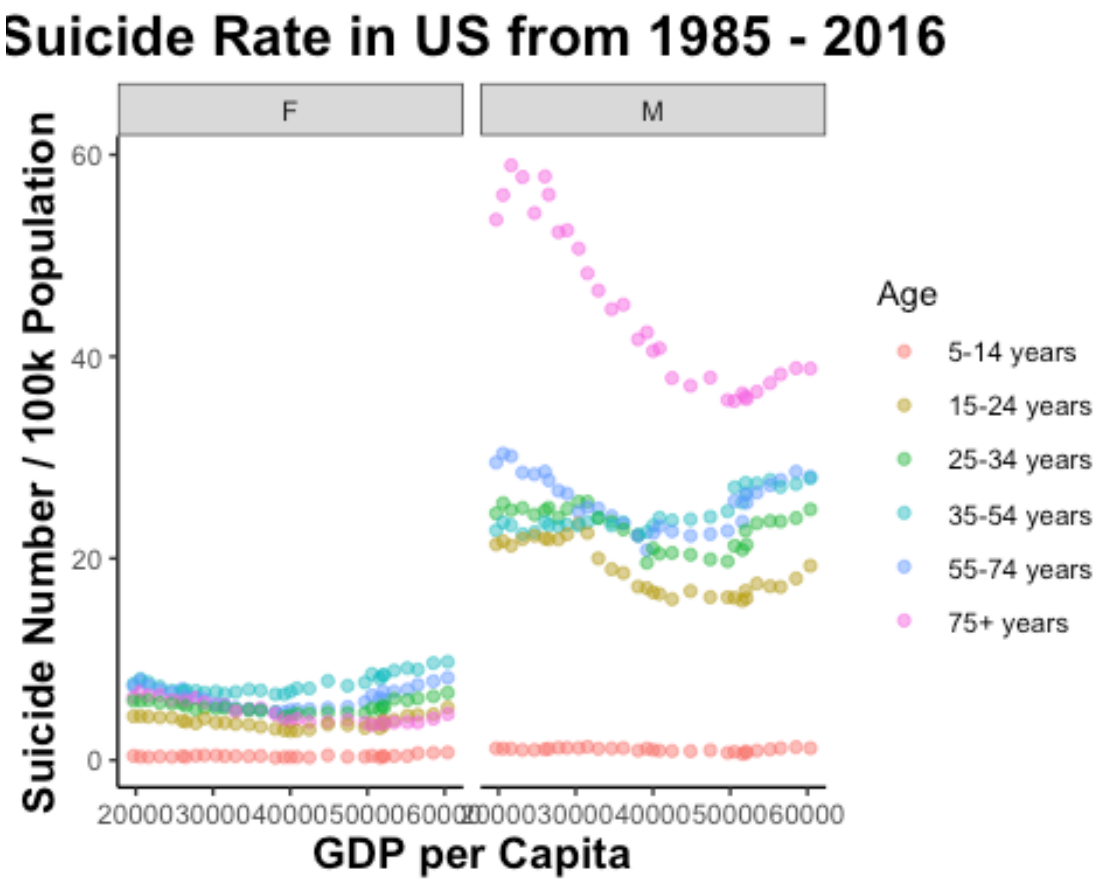
```
## [1] 0
## [1] 0
## [1] 0
## [1] 0
## [1] 0
## [1] 0
## [1] 0
## [1] 0
## [1] 19456
## [1] 0
## [1] 0
## [1] 0
```

It appears that the HDI for year column has the majority of NAs and will be discluded from the remaining of the analysis. The data will be filtered for the United States only.

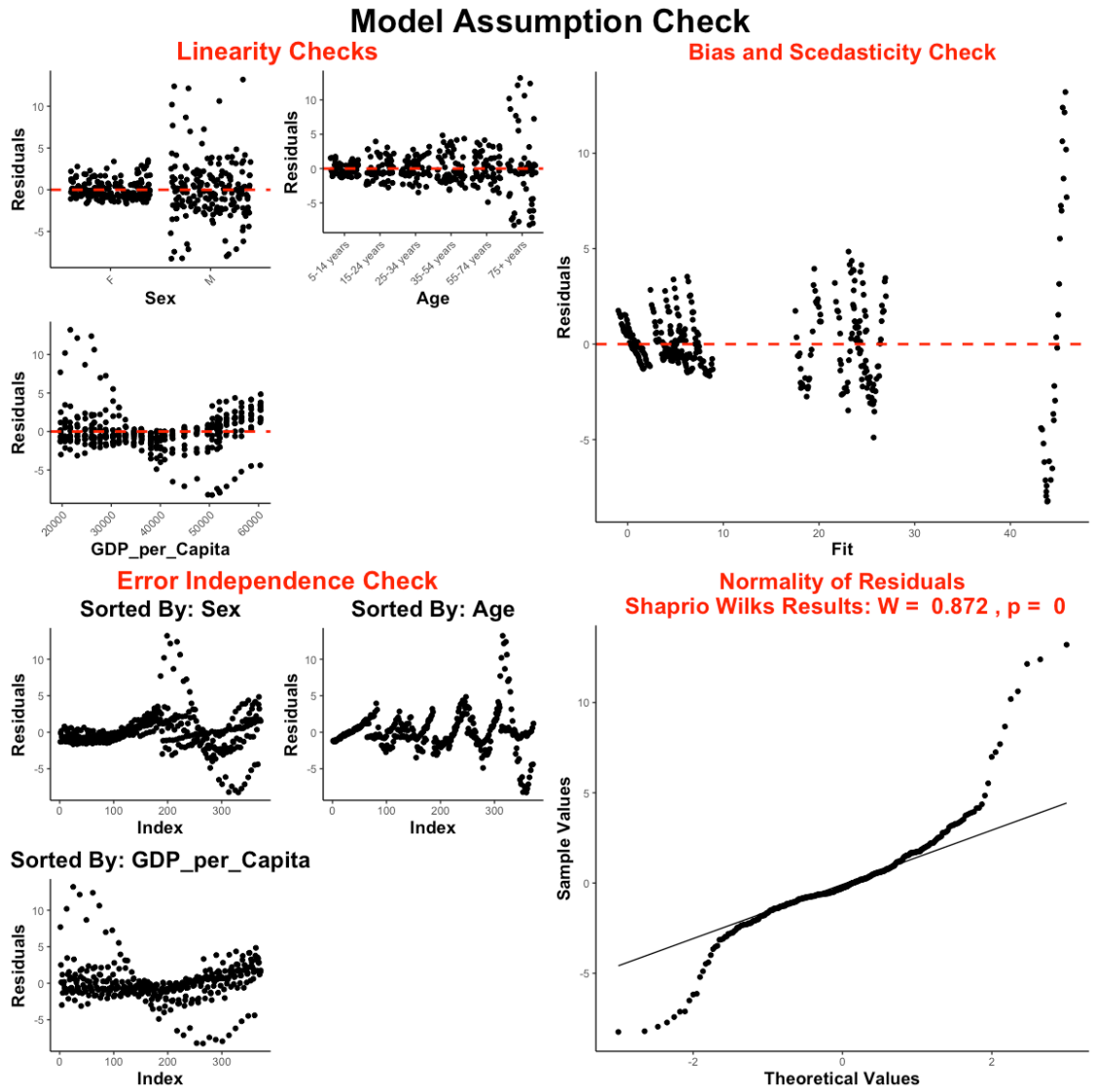
```
## 'data.frame':    372 obs. of  6 variables:
##  $ year              : int  1985 1985 1985 1985 1985 1985 1985 1985 1985 1
985 ...
##  $ sex               : Factor w/ 2 levels "female","male": 2 2 2 2 2 1 1 1
1 1 ...
##  $ age               : Factor w/ 6 levels "15-24 years",..: 6 5 2 3 1 3 5
6 2 1 ...
##  $ suicides.100k.pop : num  53.6 29.5 24.5 22.8 21.4 ...
##  $ gdp_for_year....  : num  4.35e+12 4.35e+12 4.35e+12 4.35e+12 4.35e+12 .
..
##  $ gdp_per_capita....: int  19693 19693 19693 19693 19693 19693 19693 1969
3 19693 19693 ...

colnames(df) = c('Sex', 'Age', 'Suicide_Rate', 'GDP_per_Capita')
df$Age <- relevel(df$Age, ref = '5-14 years')
df$Sex = factor(paste(toupper(strsplit2(df$Sex, split ="")[,1])))
path = "/Users/nickwawee/Desktop/BGSU/MSA_5020/Project/distplots.png"
plotdists(df, path)
```



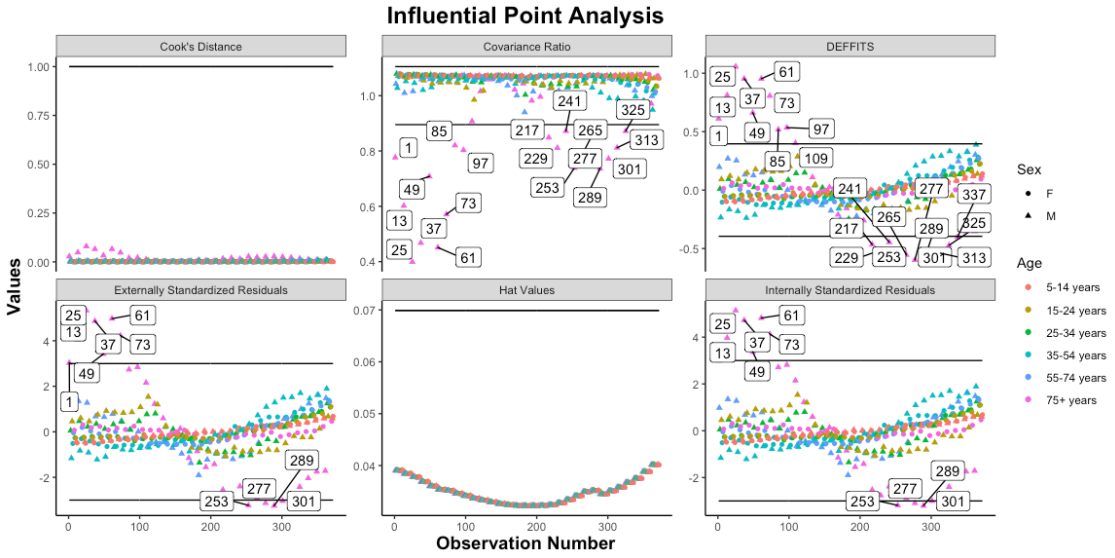
```
mlr2 = lm(Suicide_Rate ~ Sex*Age + GDP_per_Capita, data = df)
summary(mlr2)

##
## Call:
## lm(formula = Suicide_Rate ~ Sex * Age + GDP_per_Capita, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.2498 -1.0883 -0.2779  0.9381 13.2023
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)          2.989e+00  6.393e-01   4.675 4.16e-06 ***
## SexM                 6.410e-01  6.653e-01   0.963    0.336
## Age15-24 years       3.361e+00  6.653e-01   5.052 6.96e-07 ***
## Age25-34 years       4.889e+00  6.653e-01   7.349 1.36e-12 ***
## Age35-54 years       7.200e+00  6.653e-01  10.822  < 2e-16 ***
## Age55-74 years       5.767e+00  6.653e-01   8.669  < 2e-16 ***
## Age75+ years         4.363e+00  6.653e-01   6.558 1.90e-10 ***
## GDP_per_Capita      -6.560e-05  1.102e-05  -5.950 6.37e-09 ***
## SexM:Age15-24 years  1.450e+01  9.409e-01  15.414  < 2e-16 ***
## SexM:Age25-34 years  1.710e+01  9.409e-01  18.174  < 2e-16 ***
## SexM:Age35-54 years  1.621e+01  9.409e-01  17.234  < 2e-16 ***
## SexM:Age55-74 years  1.889e+01  9.409e-01  20.072  < 2e-16 ***
## SexM:Age75+ years    3.917e+01  9.409e-01  41.636  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.619 on 359 degrees of freedom
## Multiple R-squared:  0.9621, Adjusted R-squared:  0.9608
## F-statistic: 758.9 on 12 and 359 DF,  p-value: < 2.2e-16

outp = "/Users/nickwawee/Desktop/BGSU/MSA_5020/Project/"
p = ggplot(data = df, aes(x = GDP_per_Capita, y = Suicide_Rate))+
  geom_point(aes(color = Age), alpha = 0.5)+plot_opts+facet_wrap(~Sex)+
  labs(x = 'GDP per Capita', y = 'Suicide Number / 100k Population', title =
'Suicide Rate in US from 1985 - 2016')
p
```




```
ggsave(filename = paste(outp,'scatter.png',sep=""), plot = p, dpi = 600, widt
h = 8, height = 4, units = 'in')

pf = Assumption_Check(mlr2, outp)

## Saving 12 x 12 in image
## Saving 12 x 12 in image
## Saving 12 x 12 in image
## Saving 12 x 12 in image

pf
```



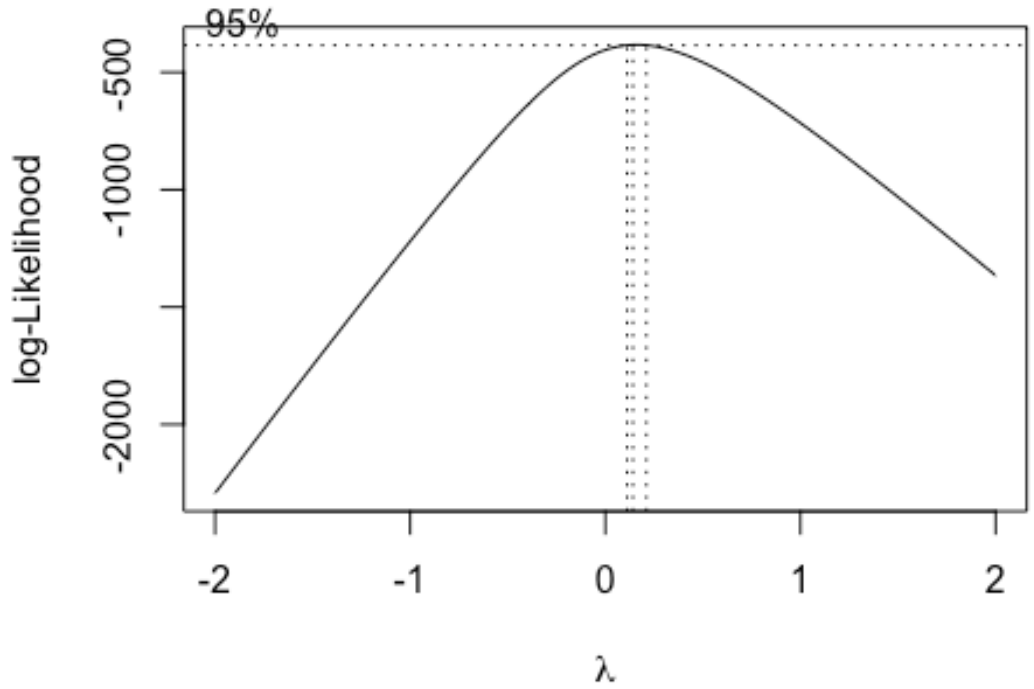

```
ret_df = infl_analysis(mlr2, df =df)
ret_df = cbind(ret_df, df)
p = ggplot(data= ret_df, aes(x= Row_Num, y = Values))+
  geom_point(aes(color = Age, shape = Sex))+
  facet_wrap(~Type, scales = "free_y")+plot_opts+geom_line(aes(y=Bound1))+geo
m_line(aes(y=Bound2))+
  geom_label_repel(aes(label=Label))+
  labs(title = 'Influential Point Analysis', x = 'Observation Number')
p
```




```
ggsave(filename = paste(outp,'influential.png',sep=""), plot = p, dpi = 600,
width = 12, height = 6, units = 'in')
```

Boxcox

```
bc = boxcox(mlr2, data = df)
```



```
p = bc$x[which.max(bc$y)]
df$Suicide_Rate = df$Suicide_Rate**p
```

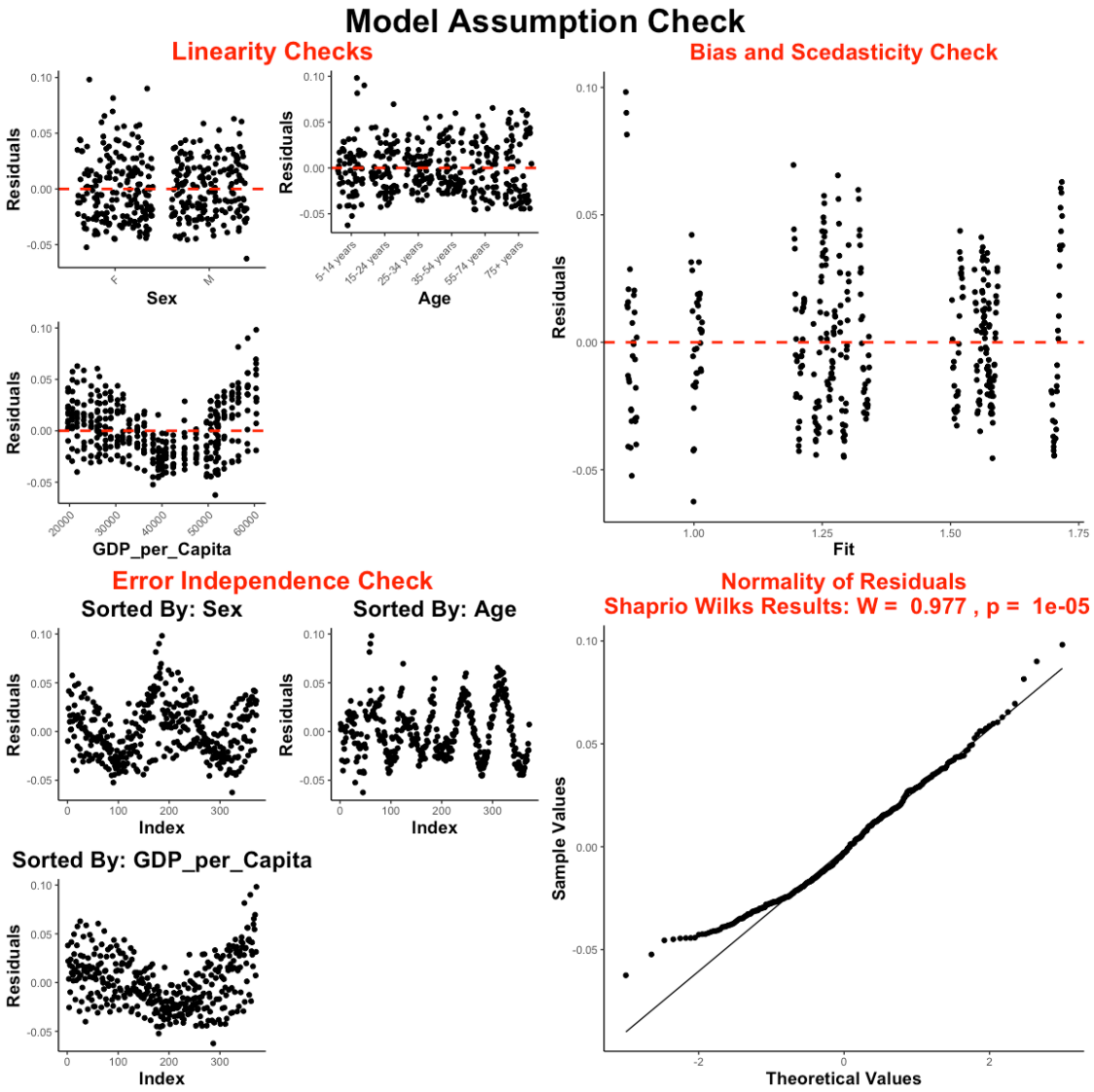
Transformed Response Variable

```
##
## Call:
## lm(formula = Suicide_Rate ~ Sex * Age + GDP_per_Capita, data = df)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.062511 -0.021512 -0.002907  0.018208  0.098173
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)           8.987e-01  6.750e-03 133.140  < 2e-16 ***
## SexM                  1.274e-01  7.024e-03  18.134  < 2e-16 ***
## Age15-24 years        3.267e-01  7.024e-03  46.515  < 2e-16 ***
## Age25-34 years        3.868e-01  7.024e-03  55.068  < 2e-16 ***
## Age35-54 years        4.530e-01  7.024e-03  64.494  < 2e-16 ***
## Age55-74 years        4.133e-01  7.024e-03  58.842  < 2e-16 ***
## Age75+ years          3.656e-01  7.024e-03  52.046  < 2e-16 ***
## GDP_per_Capita       -5.207e-07  1.164e-07  -4.473 1.04e-05 ***
## SexM:Age15-24 years   1.815e-01  9.934e-03  18.271  < 2e-16 ***
## SexM:Age25-34 years   1.652e-01  9.934e-03  16.633  < 2e-16 ***
## SexM:Age35-54 years   1.125e-01  9.934e-03  11.324  < 2e-16 ***
## SexM:Age55-74 years   1.627e-01  9.934e-03  16.377  < 2e-16 ***
## SexM:Age75+ years     3.364e-01  9.934e-03  33.868  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.02766 on 359 degrees of freedom
## Multiple R-squared:  0.9872, Adjusted R-squared:  0.9868
## F-statistic:  2314 on 12 and 359 DF,  p-value: < 2.2e-16
```

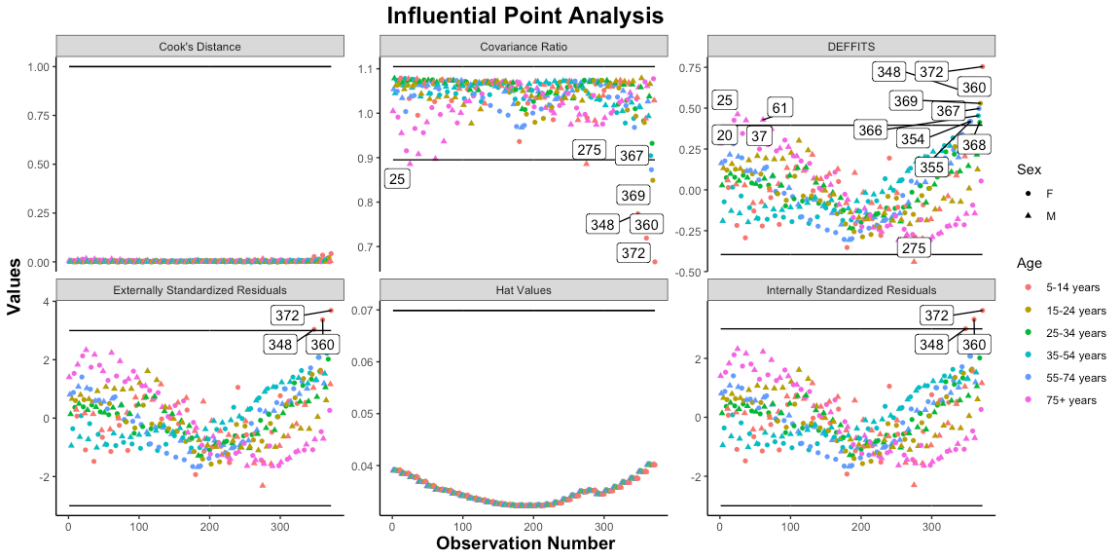
```
outp = "/Users/nickwawee/Desktop/BGSU/MSA_5020/Project/trans_"
pf = Assumption_Check(mlr3, outp)
```

```
## Saving 12 x 12 in image
## Saving 12 x 12 in image
## Saving 12 x 12 in image
## Saving 12 x 12 in image
```

```
pf
```



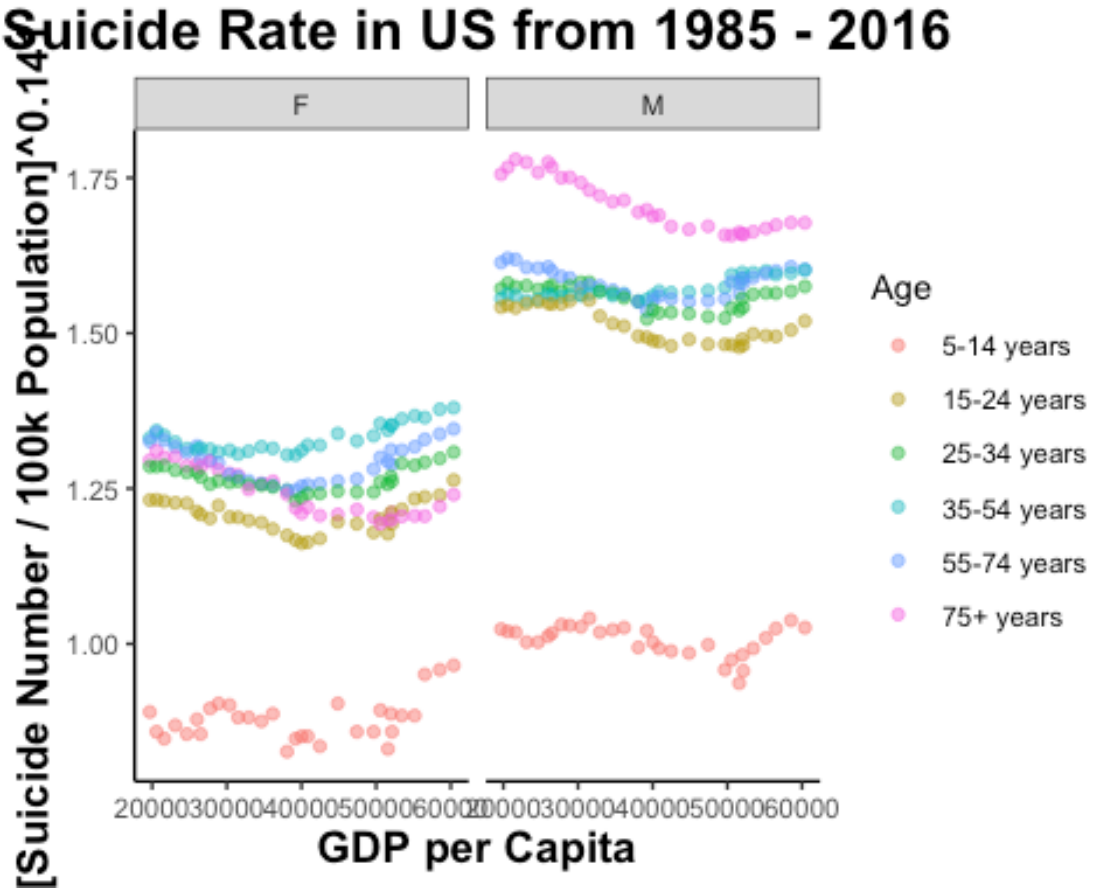

```
ret_df = infl_analysis(mlr3, df =df)
ret_df = cbind(ret_df, df)
p = ggplot(data= ret_df, aes(x= Row_Num, y = Values))+
  geom_point(aes(color = Age, shape = Sex))+
  facet_wrap(~Type, scales = "free_y")+plot_opts+geom_line(aes(y=Bound1))+geo
m_line(aes(y=Bound2))+
  geom_label_repel(aes(label=Label))+
  labs(title = 'Influential Point Analysis', x = 'Observation Number')
p
```




```
ggsave(filename = paste(outp,'influential.png',sep=""), plot = p, dpi = 600,
width = 12, height = 6, units = 'in')
```

Final Plot

```
p = ggplot(data = df, aes(x = GDP_per_Capita, y = Suicide_Rate))+
  geom_point(aes(color = Age), alpha = 0.5)+plot_opts+facet_wrap(~Sex)+
  labs(x = 'GDP per Capita', y = '[Suicide Number / 100k Population]^0.141',
title = 'Suicide Rate in US from 1985 - 2016')
p
```




```
ggsave(filename = paste(outp,'scatter.png',sep=""), plot = p, dpi = 600, widt
h = 10, height = 5, units = 'in')
```

Appendix B: Code for Hypothesis Testing and Confidence Intervals

```
# Hypothesis testing
library(car)
linearHypothesis(mlr4, c('age15-24 years = age25-34 years', 'age25-34 years = age35-54 years', 'age35-54 years = age55-74 years', 'age55-74 years = age75+ years', 'age75+ years = 0'))

## Linear hypothesis test
##
## Hypothesis:
## age15 - 24 years - age25 - 34 years = 0
## age25 - 34 years - age35 - 54 years = 0
## age35 - 54 years - age55 - 74 years = 0
## age55 - 74 years - age75  + years = 0
## age75  + years = 0
##
## Model 1: restricted model
## Model 2: suicides.100k.pop^p ~ sex * age + gdp_per_capita
##
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    364 4.4685
## 2    359 0.2746  5    4.1939 1096.7 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

linearHypothesis(mlr4, c('sexmale:age15-24 years = sexmale:age25-34 years', 'sexmale:age25-34 years = sexmale:age35-54 years', 'sexmale:age35-54 years = sexmale:age55-74 years', 'sexmale:age55-74 years = sexmale:age75+ years', 'sexmale:age75+ years = 0'))

## Linear hypothesis test
##
## Hypothesis:
## sexmale:age15 - 24 years - sexmale:age25 - 34 years = 0
## sexmale:age25 - 34 years - sexmale:age35 - 54 years = 0
## sexmale:age35 - 54 years - sexmale:age55 - 74 years = 0
## sexmale:age55 - 74 years - sexmale:age75  + years = 0
## sexmale:age75  + years = 0
##
## Model 1: restricted model
## Model 2: suicides.100k.pop^p ~ sex * age + gdp_per_capita
##
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1    364 1.19660
## 2    359 0.27456  5   0.92204 241.12 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

linearHypothesis(mlr4, c('age25-34 years = age35-54 years'))
```

```
## Linear hypothesis test
##
## Hypothesis:
## age25 - 34 years - age35 - 54 years = 0
##
## Model 1: restricted model
## Model 2: suicides.100k.pop^p ~ sex * age + gdp_per_capita
##
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1    360 0.34252
## 2    359 0.27456  1  0.067956 88.854 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
linearHypothesis(mlr4, c('age35-54 years = age55-74 years'))
```

```
## Linear hypothesis test
##
## Hypothesis:
## age35 - 54 years - age55 - 74 years = 0
##
## Model 1: restricted model
## Model 2: suicides.100k.pop^p ~ sex * age + gdp_per_capita
##
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1    360 0.29899
## 2    359 0.27456  1  0.024429 31.942 3.241e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
linearHypothesis(mlr4, c('age55-74 years = age75+ years'))
```

```
## Linear hypothesis test
##
## Hypothesis:
## age55 - 74 years - age75  + years = 0
##
## Model 1: restricted model
## Model 2: suicides.100k.pop^p ~ sex * age + gdp_per_capita
##
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1    360 0.30989
## 2    359 0.27456  1  0.035328 46.193 4.477e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
linearHypothesis(mlr4, c('age25-34 years = age55-74 years'))
```

```
## Linear hypothesis test
##
## Hypothesis:
```

```
## age25 - 34 years - age55 - 74 years = 0
##
## Model 1: restricted model
## Model 2: suicides.100k.pop^p ~ sex * age + gdp_per_capita
##
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1    360 0.28546
## 2    359 0.27456  1  0.010896 14.247 0.0001875 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
linearHypothesis(mlr4, c('age25-34 years = age75+ years'))
```

```
## Linear hypothesis test
##
## Hypothesis:
## age25 - 34 years - age75  + years = 0
##
## Model 1: restricted model
## Model 2: suicides.100k.pop^p ~ sex * age + gdp_per_capita
##
##   Res.Df     RSS Df Sum of Sq      F   Pr(>F)
## 1    360 0.28155
## 2    359 0.27456  1 0.0069843 9.1321 0.002692 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
linearHypothesis(mlr4, c('age15-24 years + sexmale:age15-24 years = age25-34
years + sexmale:age25-34 years'))
```

```
## Linear hypothesis test
##
## Hypothesis:
## age15 - 24 years - age25 - 34 years  + sexmale:age15 - 24 years - sexmale:
age25 - 34 years = 0
##
## Model 1: restricted model
## Model 2: suicides.100k.pop^p ~ sex * age + gdp_per_capita
##
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1    360 0.30430
## 2    359 0.27456  1  0.029739 38.885 1.264e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
linearHypothesis(mlr4, c('age25-34 years + sexmale:age25-34 years = age35-54
years + sexmale:age35-54 years'))
```

```
## Linear hypothesis test
##
## Hypothesis:
```

```
## age25 - 34 years - age35 - 54 years  + sexmale:age25 - 34 years - sexmale:
age35 - 54 years = 0
##
## Model 1: restricted model
## Model 2: suicides.100k.pop^p ~ sex * age + gdp_per_capita
##
##   Res.Df     RSS Df Sum of Sq      F  Pr(>F)
## 1    360 0.27738
## 2    359 0.27456  1  0.002813 3.6781 0.05593 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
linearHypothesis(mlr4, c('age35-54 years + sexmale:age35-54 years = age55-74
years + sexmale:age55-74 years'))
```

```
## Linear hypothesis test
##
## Hypothesis:
## age35 - 54 years - age55 - 74 years  + sexmale:age35 - 54 years - sexmale:
age55 - 74 years = 0
##
## Model 1: restricted model
## Model 2: suicides.100k.pop^p ~ sex * age + gdp_per_capita
##
##   Res.Df     RSS Df Sum of Sq      F Pr(>F)
## 1    360 0.27627
## 2    359 0.27456  1 0.0017088 2.2343 0.1359
```

```
linearHypothesis(mlr4, c('age55-74 years + sexmale:age55-74 years = age75+ ye
ars + sexmale:age75+ years'))
```

```
## Linear hypothesis test
##
## Hypothesis:
## age55 - 74 years - age75  + years  + sexmale:age55 - 74 years - sexmale:ag
e75  + years = 0
##
## Model 1: restricted model
## Model 2: suicides.100k.pop^p ~ sex * age + gdp_per_capita
##
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1    360 0.52070
## 2    359 0.27456  1   0.24614 321.83 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# 95% CI
confint(mlr4, level = 0.95)
```

```
##                                2.5 %        97.5 %
## (Intercept)             8.854701e-01  9.120205e-01
```

```
## sexmale                  1.135636e-01  1.411918e-01
## age15-24 years           3.129261e-01  3.405543e-01
## age25-34 years           3.730021e-01  4.006303e-01
## age35-54 years           4.392157e-01  4.668439e-01
## age55-74 years           3.995161e-01  4.271443e-01
## age75+ years             3.517748e-01  3.794031e-01
## gdp_per_capita          -7.496436e-07 -2.917933e-07
## sexmale:age15-24 years   1.619672e-01  2.010394e-01
## sexmale:age25-34 years   1.456935e-01  1.847657e-01
## sexmale:age35-54 years   9.295158e-02  1.320238e-01
## sexmale:age55-74 years   1.431509e-01  1.822231e-01
## sexmale:age75+ years     3.169077e-01  3.559799e-01
```