Sogang University

Multiple Linear Regression Model for Suicide Rates in Germany

What factors impact this suicide rate and analyzing the model in depth using R

Name: Esrat Maria

Student ID: 120190185

Time Series Data Analysis and Forecasting

Professor Myung Suk Kim

Due Date: April 30th, 2020

Problem Motivation:

Suicide rates are defined as the deaths deliberately initiated and performed by a person in the full knowledge or expectation of its fatal outcome. Comparability of suicidal data between countries is affected by a number of reporting criteria, including how a person's intention of killing themselves is ascertained, who is responsible for completing the death certificate, whether a forensic investigation is carried out, and the provisions for confidentiality of the cause of death.

When the bubble of the ***Third Reich*** [1] collapsed around ***1944–45***, it became clear that anyone who supported the Nazis were in danger. Not only the Nazis [2] and their henchmen were in troubles, also the common man or woman was. They were frightened by the idea that the Russian communists would be victorious over the Germans and this was something that had been one of the many propaganda illusions set up by the ***NSDAP*** to control the mind of the German people and made many of them side up with the Nazi's. Due the many war crimes executed by Nazi officers or officials, they didn't see any other option to not commit suicide. Because they knew they would be sentenced (in most cases to death) by either the Russians, Americans, British or French[3].

On the other hand more and more doctors are committing suicide in Germany. Doctors bound to save lives. But more and more doctors in Germany are choosing to ignore the Hippocratic Oath when it comes to themselves. Now the question becomes why so many doctors are choosing to end their lives? The reason - High job stress plays a role, experts say[4], especially since the average German doctor puts in a 54 hour work week. One estimate says almost one in three doctors uses alcohol or drugs—sometimes both—in order to withstand the rigors of long hours and chronic stress. One third of the nation's doctors are unsatisfied with their quality of life. Financial difficulties are the reason behind some suicides as well as their private lives are neglected due to job demands.

In Germany, German suicide rates had a clear decreasing tendency between 1991 and 2006, they increased from 2007 to 2017. Deeper analyses of suicide data might help to understand better this change. The aim of this study is to analyze

1) What factors have impacted the spread of such suicidal rate across German;
2) Whether the decrease of suicide rates before 2007 as well as the increase from 2007 to 2017 are driven by the same suicide method.

[1] https://en.wikipedia.org/wiki/Nazi_Germany

[2] https://en.wikipedia.org/wiki/Battle_of_Berlin

[3] https://en.wikipedia.org/wiki/List_of_suicides_in_Nazi_Germany

[4] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6907772/

Literature Review:

The current study on this topic examines the extent to which the availability, attraction and lethality of particular methods of suicide affect suicide rates. A close relationship existed in the period from 1949 and 1989 between the peaks of the suicide rate and the percentage of low threshold suicide methods, whereby suicide in this context is understood to mean suicide by shooting or intake of solid, liquid or gaseous poisons.

A report on "***The impact of the reunification of Germany on the suicide rate***"[5] proposed- unemployment rates in East Germany have increased from near zero to 21% for women and 11% for men. It is generally assumed that these political and economic changes are associated with an increase in stress levels and possibly suicide rates.

Germany's death rate has been higher than its birth rates since the 1970s. Currently, the birth rate is 9.397 births per 1,000 people, which has been declining every year. The death rate is 11.392 deaths per 1,000 people. Additionally, the fertility rate in Germany is 1.59 births per woman[6].

The ***Journal of Death and Dying*** [7]ran a simple ***Suicide Opinion Questionnaire*** (SOQ) on 172 Germans and obtained statistically significant gender differences in attitudes toward suicide among Germans.

The reports on this topic were found on the internet when searched extensively. Most of them reported very elaborately that Stress and trauma are among the most common factors of suicide. Other possible causes may include abuse, accidents, bullying, injuries, and being close to someone who also committed suicide. Suicidal ideation may occur when a person is experiencing limited support, is exhausted, or feels there is no better solution.

Statement of Research Objectives:

Which factors are affecting suicide across nation? What age groups are more likely to be consumed by the thought of suicide? In Germany the common factors for suicidal thoughts are depression, family issues, love affairs or divorce. Depression/stress has been one of the most widely found reasons for suicide across the nation. The objective of this research is to find out what age groups are more prone to commit suicide and what factors are mostly causing it.

In this research paper I make a multiple regression model to find out or predict what age is more sensitive to committing suicide. Such research can promote caution across nation and take necessary measures to spread help among people who are in need of it. Mental health is important and talking about it is even more important. Through many ups and downs in life, suicide is never supposed to bring a solution to it. This research is hoping to spread awareness across nation by analyzing suicide rate and reasons so that people feel at ease to talk about whatever is troubling them in life and changes their perspective of viewing the problem that is concerning them.

---

[5] https://www.tandfonline.com/doi/abs/10.1080/13811119908258332

[6] https://worldpopulationreview.com/countries/germany-population/

[7] https://journals.sagepub.com/doi/abs/10.2190/H1CB-YFJD-W51B-P741

Description of Data:

Every data is from data.oecd[8] [9] and is a slightly modified version of some data obtained from github[10].

Dependent variable:

1. age_group

In the dataset the age has been divided into 5 groups. The ranges are**: 0-14, 15-29, 30-44, 45-59 and 60+**. The age groups gives us a clear insight on what age group is more likely to commit suicide by analyzing the data that are dependent on this variable. When we figure out which group has a higher rate of committing suicide then the government can take necessary measures to come up with solutions to tackle that certain age group.

Independent variable:

2. **year**

In this dataset the year range is **1985-2018**. Throughout the range the suicide rate across Germany has been calculated along with the reason why that may resonate the suicides. It will be awakening to see which year had a higher rate of suicide and what were the causes that led to suicide are.

3. **sex**

The suicide data has been calculated between male and female. By analyzing the dataset we can get an insight on which sex is more likely to commit suicide. We can also analyze what kind of suicide is more triggered by which sex.

4. **toxic gas**

People living with air pollution have higher rates of depression and suicide, a systematic review of global data [11] has found this information. Cutting air pollution/toxic gas exploitation around the world could prevent millions of people becoming depressed, the research suggests. This assumes that exposure to toxic air is causing cases of depression. Scientists believe this is likely but is difficult to prove beyond doubt. Pollution or being open to toxic gas can harm mental health.

5. **hang**

Suicide committed by hanging themselves has been quite common among males and females throughout the year. From a rough analyzation this rate has been seen to be quite common among females. Hanging is

[8] https://data.oecd.org/healthstat/suicide-rates.htm

[9] https://www.macrotrends.net/countries/DEU/germany/suicide-rate

[10] https://vincentarelbundock.github.io/Rdatasets/doc/vcd/Suicide.html

[11] https://www.theguardian.com/environment/2019/dec/18/depression-and-suicide-linked-to-air-pollution-in-new-global-study

often considered to be a simple suicide method that does not require complicated techniques and it has a high mortality rate[12].

6. **drown**

When considering suicide in the age group of 50 years and older suicide committed by drowning represents 25% of all suicidal deaths, and within females in this group represents the most common form of suicidal death.

7. **gun**

The majority of gun deaths are self-inflicted. The easy availability of firearms to those in distress makes suicide attempts far more likely to result in death. This suicide method has been really common all across Germany throughout the year.

8. **knife**

Knife initiated suicide rates are one of the many methods of suicide attempts in Germany. However from many existing reports this attempt is not likely to result in death.

9. **jump**

Survivors of falls from hazardous heights are often left with major injuries and permanent disabilities from the impact-related injuries. A frequent scenario is that the jumper will sit on an elevated highway or building-ledge as police attempt to talk them down. Almost all falls from beyond about 10 stories are fatal.

10. **poison**

Suicide attempts using poison have surged among young people, particularly girls. More young people than ever are trying to kill themselves using poison. The rate of attempted suicide by poison has more than doubled among people under 19.

11. **others**

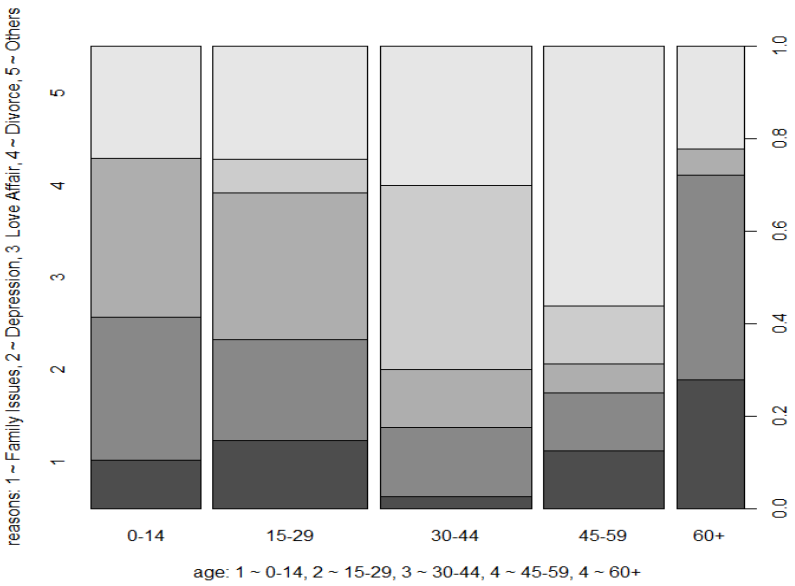
This category can be any other suicidal attempt other than the one mentioned above. In the dataset this variable has not been seen vey widely.

12. **reason**

In the dataset this variable has been divided into 5 range. Like, **family issues, depression/stress, love affairs, divorce or others**. Needless to mention depression has been the most common reason for committing suicide among people living in Germany.

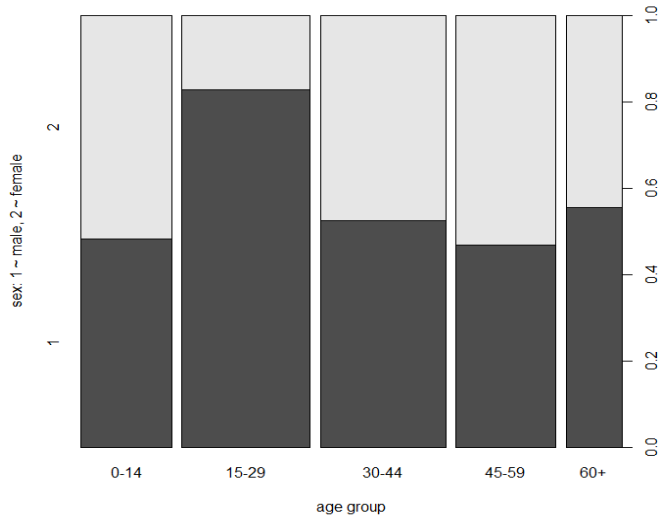
Now we will analyze our dataset by plotting different diagrams to know clearly what kind of data we are working with.

[12] https://en.wikipedia.org/wiki/Suicide_by_hanging



The above graph shows which age groups are more likely to commit suicide for what specific reason. For example the graph shows that the age group **30-44** is more likely to commit suicide because of divorce related reasons.

In Germany it has been seen that males are more prone to suicide than females. This could be due to inefficient work-life balance. The results are very clear from the graph below:



Applied Methodology:

We will analyze the dataset and make our multiple linear regression model using R studio. Since the dataset has non-numerical data, in R we transform them into categorical data by using defined labels. The commands are given below:

$$\boldsymbol{dataset\$sex} = \boldsymbol{factor}(\boldsymbol{dataset\$sex}, \boldsymbol{levels} = \boldsymbol{c}('\boldsymbol{male}', '\boldsymbol{female}'), \boldsymbol{labels} = \boldsymbol{c}(\mathbf{1}, \mathbf{2}))$$

$$dataset\$reason = factor(dataset\$reason, levels = c('family\ issues', 'depression', 'love\ affair', 'divorce', 'others'), labels = c(1, 2, 3, 4, 5))$$

Once we write the following command:

$$output = lm(age\_group \sim ., data = dataset)$$

We get the output like below:

```
Call:
lm(formula = age ~ ., data = dataset)

Residuals:
    Min      1Q  Median      3Q     Max
-9.3626 -2.4266 -0.2215  2.4979  7.3889

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 12.6222113 59.3981238   0.213   0.8320
year         0.0002899  0.0297567   0.010   0.9922
sex2        -0.9324932  0.6416665  -1.453   0.1484
toxicgas    -0.0348761  0.1748569  -0.199   0.8422
hang         0.2089594  0.2464941   0.848   0.3980
drown        0.0886625  0.0725224   1.223   0.2235
gun          0.0537472  0.0535801   1.003   0.3175
knife        0.0568835  0.0319232   1.782   0.0769 .
jump        -0.1172509  0.1494693  -0.784   0.4341
poison      -0.0695249  0.0437676  -1.589   0.1144
others       1.7648168  0.7015040   2.516   0.0130 *
age_group2   6.1836695  0.9270088   6.671 5.37e-10 ***
age_group3  24.2649801  0.9796893  24.768  < 2e-16 ***
age_group4  38.2070181  0.9943040  38.426  < 2e-16 ***
age_group5  59.4183789  1.1624515  51.115  < 2e-16 ***
reason2     -2.1244163  1.0781300  -1.970   0.0507 .
reason3      0.3340445  1.1064919   0.302   0.7632
reason4     -1.3203477  1.2688218  -1.041   0.2998
reason5     -0.2454343  1.0336549  -0.237   0.8127
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.61 on 141 degrees of freedom
Multiple R-squared:  0.9693,    Adjusted R-squared:  0.9654
F-statistic: 247.7 on 18 and 141 DF,  p-value: < 2.2e-16
```

From the above output we can see that the adjusted R^2 value is ***0.9654*** which is very close to **1**. So this proves our model to a better model. The ***p-value*** of our ***X*** variables are not significant because most of them are above the alpha value ***0.05***. This may occur from the relationship among variables. Now we first have to run a multicore linearity test.

Variation Inflation Factor Analysis:

If we have any variable with high VIF then that high VIF doesn't make any sense. So, now we will check the VIF values of our variables. The output console of R looks like below:

```
> vif(output)
toxicgas     hang    drown      gun    knife     jump   poison   others      sex
1.031321 1.088118 1.054040 1.051726 1.036346 1.084758 1.069878 1.075265 1.074137
> max(vif(output))
[1] 1.088118
```

We can see that all the values are relatively low. Some researchers use either ***10 or 5*** as a criteria for VIF. In that sense all the values are below the considered range.

Variable selection using partial F test:

To run partial ***F-test*** we run the below command:

$$result = step(output, data = dataset, direction = "backward")$$

The output is like below:

```
Step:  AIC=421.06
age ~ knife + poison + others + sex + age_group + reason

            Df Sum of Sq   RSS    AIC
<none>                    1890 421.06
- sex        1        25  1915 421.15
- knife      1        34  1924 421.89
- poison     1        50  1940 423.23
- others     1        67  1957 424.60
- reason     4       147  2036 425.01
- age_group  4     51746 53636 948.37

Call:
lm(formula = age ~ knife + poison + others + sex + age_group +
    reason)

Coefficients:
   (Intercept)           knife          poison          others            sex2
      14.29869         0.05059        -0.08330         1.54828        -0.84366
age_group15-29  age_group30-44  age_group45-59  age_group60+         reason2
       6.42968        24.15596        38.18442        59.70639        -2.33848
       reason3         reason4         reason5
       0.24345        -1.34578        -0.20987
```

The above image include the final step of the calculation. As the AIC (archaic information criteria) value decrease we will have a better model. After doing the partial ***F-test*** we now know what are our significant variables. The summary model output is like below:

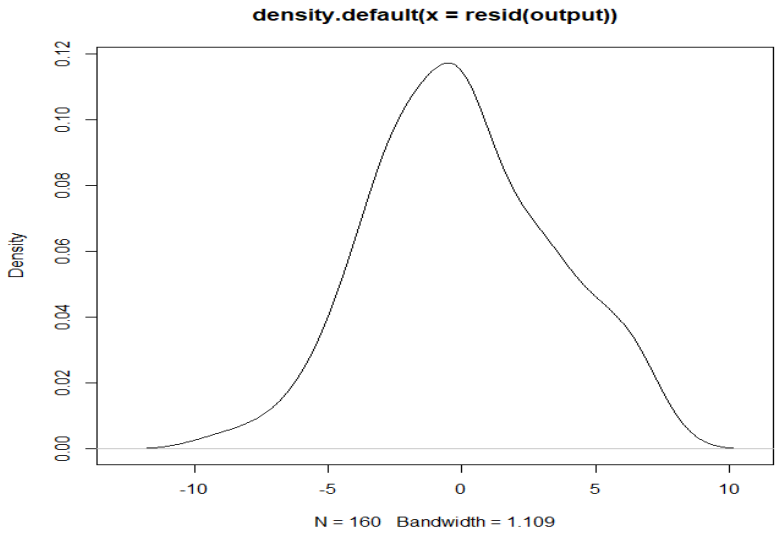
```
Call:
lm(formula = age ~ knife + poison + others + sex + age_group +
    reason)

Residuals:
    Min      1Q  Median      3Q     Max
-9.8755 -2.2158 -0.3864  2.4238  7.6487

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    14.29869    1.37118  10.428  < 2e-16 ***
knife           0.05059    0.03127   1.618   0.1078
poison         -0.08330    0.04231  -1.969   0.0508 .
others          1.54828    0.68039   2.276   0.0243 *
sex2           -0.84366    0.60732  -1.389   0.1669
age_group15-29  6.42968    0.90144   7.133 4.14e-11 ***
age_group30-44 24.15596    0.96494  25.034  < 2e-16 ***
age_group45-59 38.18442    0.97648  39.104  < 2e-16 ***
age_group60+   59.70639    1.12698  52.979  < 2e-16 ***
reason2        -2.33848    1.04520  -2.237   0.0268 *
reason3         0.24345    1.08527   0.224   0.8228
reason4        -1.34578    1.25107  -1.076   0.2838
reason5        -0.20987    1.01649  -0.206   0.8367
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.586 on 147 degrees of freedom
Multiple R-squared:  0.9685,    Adjusted R-squared:  0.9659
F-statistic: 376.2 on 12 and 147 DF,  p-value: < 2.2e-16
```

From the above image we can see that **p-value** is significant for variables and the significant ones have been marked with asterisk (*****) and it is below alpha value **0.05**.

Residual Plot:

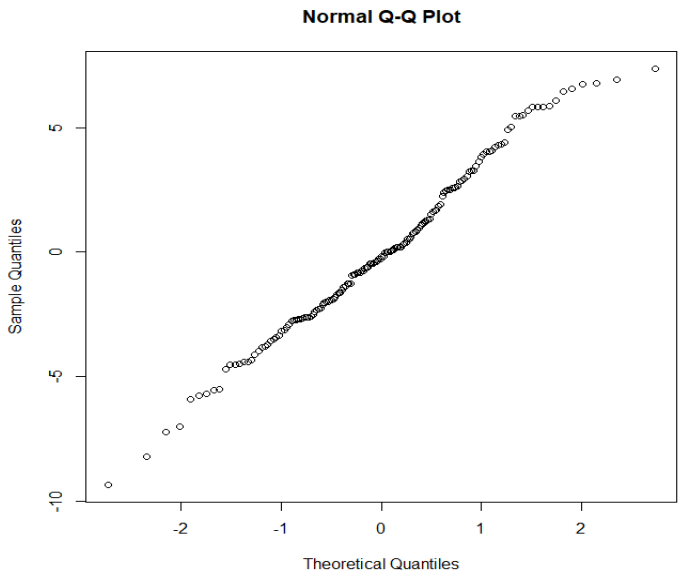
To make sure that our model's coefficients estimates are unbiased, and the estimation of the variance are reliable, we need to conduct a heteroscedasticity test.

From the below graph we can't intuitively come to a decision but however this shows that our data is not being biased.



Normality Test:

A normality test is required to determine that our dataset is well modeled under a normal distribution. The normal QQ plot is given below for our dataset:



If we draw a nice linear line across the data points it should fit properly and there shouldn't be any major deviation.

Shapiro-Wilks test:

```
> shapiro.test(output$residuals)

        Shapiro-Wilk normality test

data:  output$residuals
W = 0.98897, p-value = 0.2441
```

We carry out the above test to check the normality of the residuals. From the output console of R in the above image we can see that the ***p-value*** is ***0.2441*** which is bigger than alpha value ***0.05***. This proves that the distribution of the residual is not significantly different from the normal distribution.

Breusch-Pagan Test:

To run this test we write the following command:

$$\boldsymbol{library(lmtest)}$$

$$\boldsymbol{bptest(output)}$$

The output console looks like below:

```
> bptest(output)

        studentized Breusch-Pagan test

data:  output
BP = 29.387, df = 18, p-value = 0.04385
```

We can see that the ***p-value*** is ***0.04385*** which is higher than significance level ***0.05***. This proves a strong pattern of heteroscedasticity among residuals.

So, finally our regression model,

$$\boldsymbol{Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \varepsilon}$$

$$\boldsymbol{Y = 14.299 + 0.1 X_1 - 0.08 X_2 + 1.54 X_3 - 0.84 X_4 + 6.42968 X_5 - 2.338 X_6 + \varepsilon}$$

Here,

$\boldsymbol{Y}$ = predicted suicide rate related to age group

$\boldsymbol{X_1}$= suicide committed using knife

$\boldsymbol{X_2}$= suicide committed using poison

$\boldsymbol{X_3}$= suicide committed using some other method

X_4= sex of the person (male/female)

X_5= age group that has committed the specified type of suicide

X_6= reasons why the suicide was committed

We can see that there is both positive and negative correlation among variables in the above regression model. Like we can say if we take a look at the reasons why the suicide is triggered among people and take necessary measures to address those reasons in a positive way then the total suicide rate can be minimized.

Expected original contribution:

In this research I analyzed the reasons why suicide is triggered among people in Germany and what age groups are more likely to be affected by this horrendous thought of suicide. Previous research works didn't focus much on the methods that are mostly used among people to commit suicide. However in this project the dataset included what are the most common ways of committing suicide among people and also the reasons why such thoughts are coming to people's mind.

A better work life balance can help prevent suicide in so many ways. If a suicidal thought appear to be imminent a local helpline should be made available in the country so that people in need can seek help. If the government is willing to take steps to prevent such thoughts the such analyzation can give us proper idea about what ways should be taken to minimize the suicide rate across nation.

Some limitations are witnessed during the whole research process. In the dataset there are data available from the year **1987-2018**. I only worked with 160 rows of data. To me it seemed like to do an in-depth analysis this amount of data is not enough and may not give a proper analyzation of the whole data.

For further research I think it would be better if we can work with some recent data. On the other hand some detailed focus can be paid on the reasons why people are committing suicide. In this dataset there are only few reasons of suicides mentioned. If we could do a further analysis on what kind of causes are initiating suicide then it would have been easier to take necessary steps to stop spreading such act across nation.

References:

1. https://en.wikipedia.org/wiki/Nazi_Germany
2. https://en.wikipedia.org/wiki/Battle_of_Berlin
3. https://en.wikipedia.org/wiki/List_of_suicides_in_Nazi_Germany
4. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6907772/
5. https://www.tandfonline.com/doi/abs/10.1080/13811119908258332
6. https://worldpopulationreview.com/countries/germany-population/
7. https://journals.sagepub.com/doi/abs/10.2190/H1CB-YFJD-W51B-P741

8. https://data.oecd.org/healthstat/suicide-rates.htm

9. https://www.macrotrends.net/countries/DEU/germany/suicide-rate

10. https://vincentarelbundock.github.io/Rdatasets/doc/vcd/Suicide.html

11. https://www.theguardian.com/environment/2019/dec/18/depression-and-suicide-linked-to-air-pollution-in-new-global-study
12. https://en.wikipedia.org/wiki/Suicide_by_hanging