Model on Suicide Analysis and Prediction

“Ming Chu Cheng(Miranda), Wanying Li, Tal Jacobi, Jayant Bishnoi”

5/5/2021

STA 9750 Final Project

Introduction

Nowadays, people are intelligent and continuously seeking a high quality of life. At the same time, people are also faced with an enormous amount of pressure such as the cost of living, workload, relationships, even freedom…. etc. When people feel helpless, depressed and hopeless, they would contemplate ending their life in order to ease their unhappiness. The topic of suicide is always a complicated subject within society. Therefore, we try to use different features to explore the topic of suicide by using multiple linear regression and modeling analysis. We wish many people would contemplate the consequence of suicide and hope the result of suicide analysis could help some Governors’ awareness and decision making through various means of advertising.

There are in total of five data sources for our project analysis and prediction, including master.csv, gdp.csv, CPI_total.csv, Unemployment_total_data.csv and Happiness.csv. All the resource comes from World Bank and Kaggle. The largest data sets of Suicides (master) has 27820 rows and 10 columns which contains historical data from 1985 to 2016. As we hope our data analysis has more diversification and accuracy, we find more component data sets and join them together to become our new data sets, encompassing historical data from 1985 to 2020. As the data in earlier years and some countries are relatively scarce and drop all NA, we select an analysis time frame from 2009 to 2015 which has 2940 observations and 28 columns which is called complete_join for our main data set. The complete_join data set comprises many global market index and rates. Consequently, we select different types of components to perform deep analysis and prediction, such as Suicides per 100k, our dependent variable. Population, GDP, CPI, Unemployment Rate, Social support, Freedom to make life choice, Perceptions of corruption, Confidence in national government, Democratic quality are our predictors. Order data cleaning procedures includes conversion of column types from the csv file, removal of unnecessary symbol, rename column name and pivot_longer the column and value and then join them together.

Summary of the suicides rate of the countries from 2009 to 2015

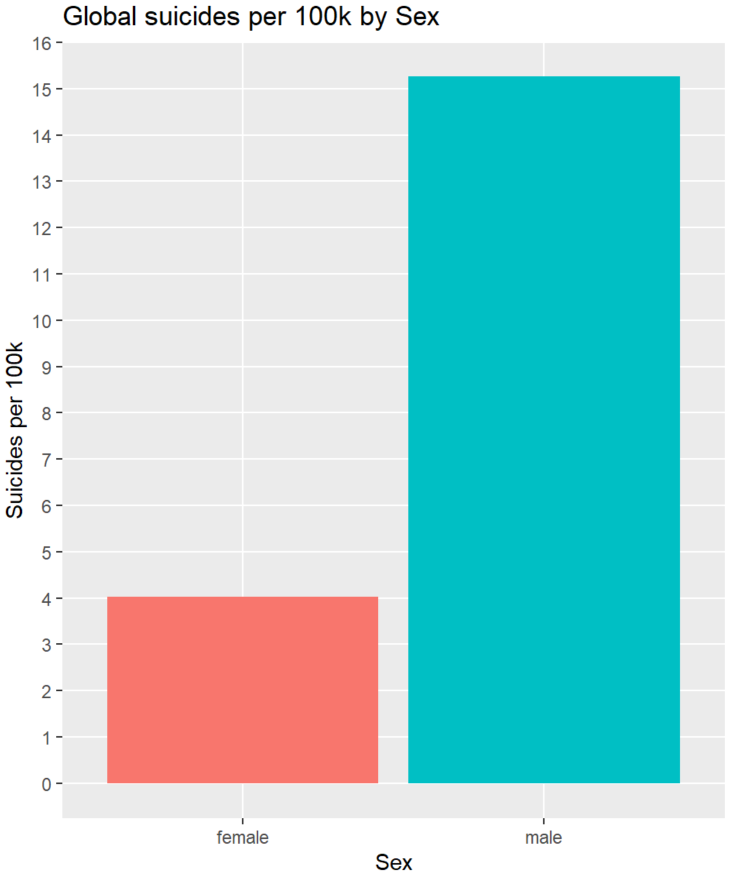
The summary chart provides some summary statistics of our analysis data set. It has contained 9 predictors and the dependent variables (suicide 100k pop) average, median and standard deviation. The table starts with descending by the suicide 100k pop. We can see that most of the top ten suicide rate countries is in Europe or near Europe. The highest average rate of suicides per 100K people is the Lithuania country. Comparing with the United States, Lithuania has 2.5 times of average of per 100k people. But its population is less than the United States 6 times. Meanwhile, comparing with others top nine countries, they also have 1.3 – 1.85 times higher comparing with the United States‘ average of per 100k people. It seems that most people in these counties, are feeling helpless and hopeless every day.

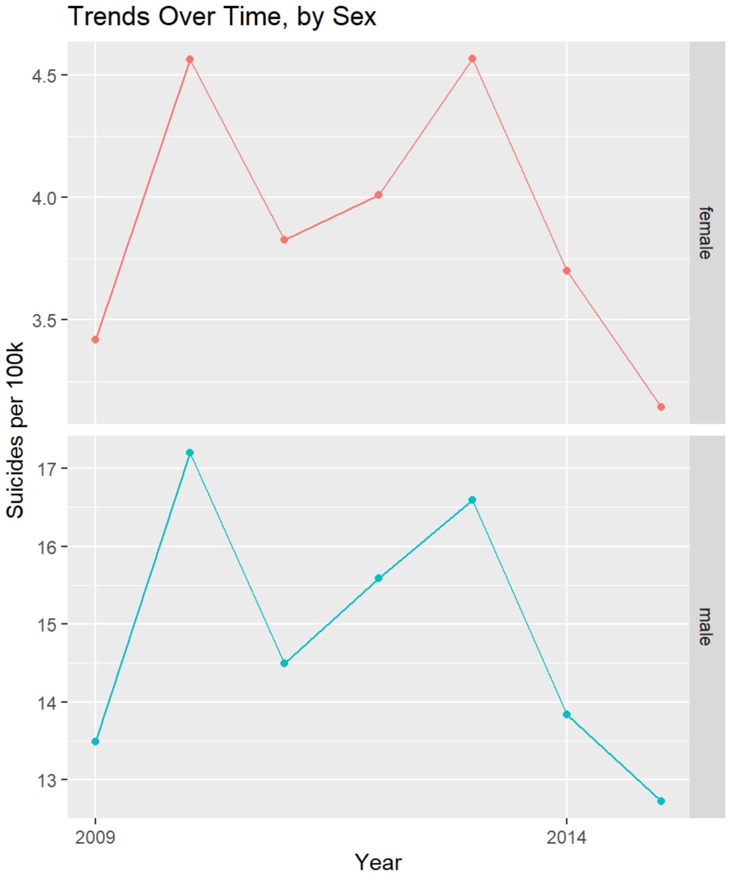
	Sample Size	Suicides 100k pop			Population			CPI Index			Unemployment Rate			GDP R	
Country	n	Avg_S	Median_S	SD_S	Avg_P	Median_P	SD_P	Avg_CPI	Median_CPI	SD_CPI	Avg_U	Median_U	SD_U	Avg_GDP	Median_
Lithuania	72	34.17	19.06	32.22	239889	197766	113892	2.36	2.20	1.64	13.80	13.57	2.34	0.63	
Belarus	36	25.05	12.29	25.75	749644	732032	372172	16.46	18.12	2.52	2.37	0.50	2.68	0.98	
Hungary	60	23.96	13.74	27.27	788480	629552	386399	3.19	3.93	2.17	10.22	11.00	1.30	1.55	
Latvia	60	22.59	10.58	22.74	162033	138115	78661	2.15	2.26	1.68	14.30	15.05	2.56	-0.03	
Slovenia	72	22.49	14.64	27.18	162186	130976	82972	1.50	1.79	0.78	8.31	8.51	1.45	-1.04	
Ukraine	72	21.81	11.33	20.93	3522987	3419905	1751853	15.76	10.72	15.55	8.46	8.47	0.67	-3.59	
Uruguay	72	21.21	10.32	21.82	262870	255262	93221	8.00	8.34	0.84	6.97	6.86	0.52	3.97	
Estonia	60	20.17	12.57	20.90	104084	91189	48228	2.30	2.78	2.09	10.38	10.02	2.31	0.09	
Croatia	36	18.65	13.55	20.95	336922	267022	158852	1.80	2.22	1.53	16.82	17.25	0.64	-1.06	
Serbia	12	18.14	10.62	23.67	569403	446709	288581	7.69	7.69	0.00	22.15	22.15	0.00	2.89	
Belgium	60	17.79	14.13	15.20	868566	687164	382621	2.00	2.19	1.16	7.98	8.29	0.55	1.47	
Austria	60	17.46	12.00	20.99	670260	529401	335369	2.24	2.00	0.60	5.04	4.87	0.38	1.23	
France	60	16.55	11.16	17.78	4970423	3922543	2041935	1.39	1.53	0.62	9.46	9.40	0.58	1.20	
Finland	60	16.33	10.62	12.89	425922	335123	188896	1.99	1.48	0.96	8.14	8.19	0.37	0.61	
Poland	60	15.55	6.15	15.82	3002213	2841410	1359698	2.29	2.58	1.57	9.74	9.64	0.46	2.87	
Czech Republic	60	15.32	7.92	15.66	827910	688456	423175	1.69	1.47	0.96	6.81	6.95	0.40	1.13	
Iceland	24	15.28	12.88	14.02	24904	22969	10096	4.53	4.53	0.67	5.69	5.69	0.32	2.72	
Switzerland	24	14.40	10.48	14.54	623340	495970	313550	-0.59	-0.59	0.11	4.30	4.30	0.18	-0.61	

Proportion of Global Sex

Global suicides per 100k bar chart presents the proportion of gender of global suicides per 100k people. It obviously shows that a man is higher pone of ending their life than female. Most importantly, there is a big ratio occurrence. Approximately, the ratio is (1:4). It means for every 1 female contemplating suicide globally, it has 4 males that are thinking suicide, from 2009 to 2015. What's the thought on why a male is contemplating negative thinking than a female in the world? What factors causes many males to end their precious life? We will explore the data in depth; we will discover the reasons.

Besides, we also did an analysis of gender by year. Both male and female suicide rates peaked in 2010. Surprisingly, male suicide rates sharply increased two years from 2011 to 2013. Instead, female suicide rates increased slowly for two years. It might relate to the global economic crisis in 2011. Many people's wealth suddenly decreased significantly after August 8, 2011. After 2013, some factors might have changed their mindset and encouraged them not to give up their life. Or some factors might give them hope in order to build up their courage to live in the world.

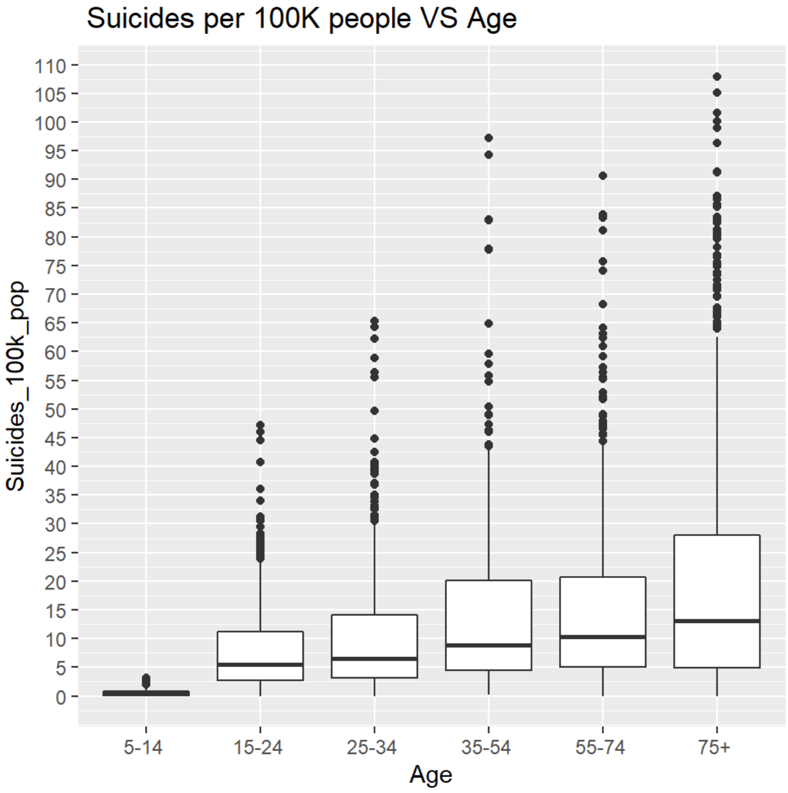




Distribution of Global Age

We use a boxplot to analyze age. From the box plots, we can see the age distribution for different groups of age. The thick line in the middle of the box indicates the median suicides per 100K people (Suicides_100k-pop) for the age group, the bottom of the box is the lower quartile, and the top of the box is the upper quartile. The endpoints represent the smallest and largest suicides per 100K people values, excluding outliers, which are represented by the dots.

In the boxplot, we can see that group 35–54 and group 55-74 have very close 50% (median) of all respondent suicides per 100 people rate lower than 8 rates. Also, they also have similar 75% (upper quartile) of all respondent rate lower than 20 rates. For the group over 75 age, we can identify that the median of over 75 age is approximately 12.5 rates, the upper quartile is 26.5 rates. Overall, in the different group of age it appears we have many potential outliers in different group of ages.



Multiple linear Regression

We use a statistical technique - multiple linear regression to do several explanatory predictors to predict the outcome of a dependent variable which is Suicide_100k-pop. This multiple linear regression is regressed on the following nine predictors: xi1: Population, xi2: CPI_Index, xi3: GDP_Rate, xi4: Unemployment_Rate, xi5: Social_support, xi6: Freedom_to_make_life_choices, xi7: Perceptions_of_corruption, xi8: Confidence_in_national_government, xi9: Democratic_Quality

The multiple regression model:

$$yi = \beta 0 + \beta 1xi1 + \beta 2xi2 + \ldots + \beta pxip + \epsilon$$

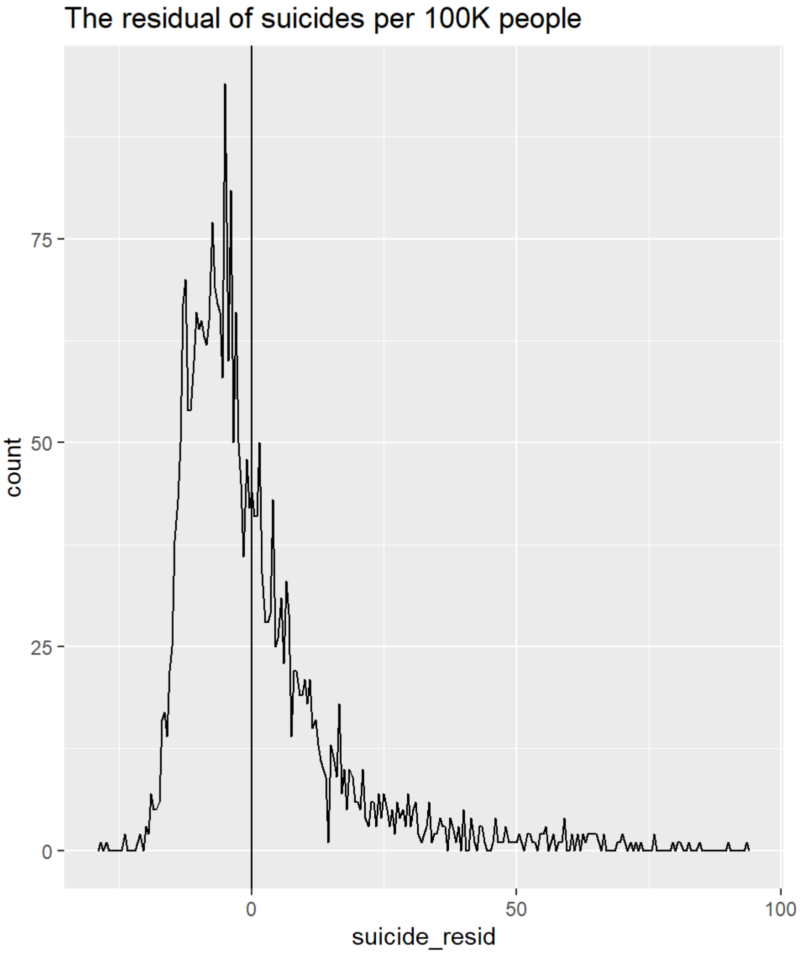
Where, for i=n observations:

- yi = dependent variable
- xi = explanatory predictors
- β0 = y-intercept
- βp = slope coefficients for each explanatory predictor (Estimate)
- ϵ = the model's error term (also known as the residuals)

Multiple Linear Regression model VS residual

As the population of average suicides per 100k people is unknown, so we assume that the population of average suicides per 100k people follows a normal distribution N (0, 1). The epsilon (ϵ) is a normally distributed variable centered at zero. We always hope our sample average in population will meet the population average of the suicides population in order to have high accuracy estimated regression line that is close to true line. Not exactly the same, but we can use null hypothesis test to know whether our dataset is unbiased.

After running the linear regression and residual plot, we discover that our model is not close to the normal distribution. We can see as below residual plot shows that our model has skewness. Thus, we have to deal with it before doing prediction in order to make our prediction to be more accurate.



Model Comparision

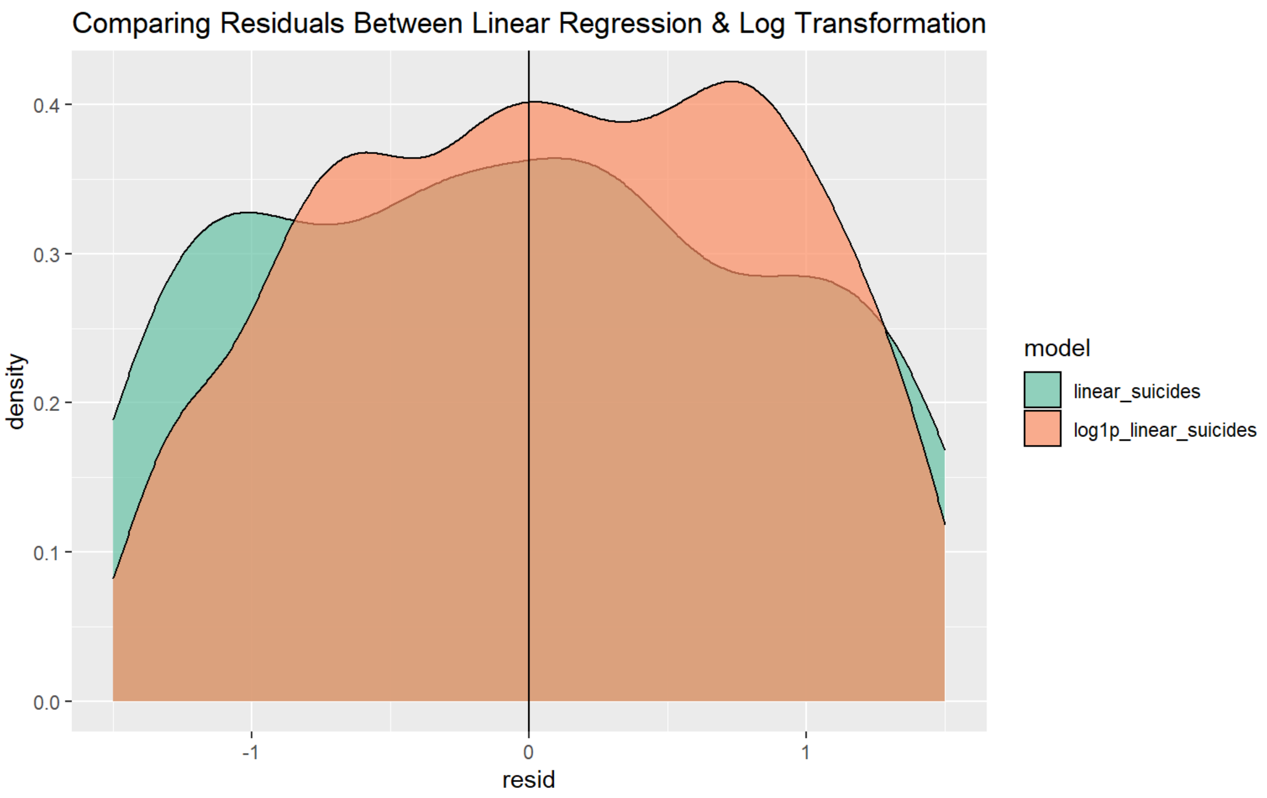
We try to use base log transformation, polynomial regression and log1p transformation to adjust our model. Please see the following plot. We can see that the residual decreases frequently and the distribution of my fitted model moves much closer to the normal distribution N(0,1). Our prediction will be much better now.

linear_suicides:

y_hat(suicides_100k_pop) = 7.30 + (-0.000000244)Population + 0.387CPI_Index + 0.0126GDP_Rate +(-0.201)Unemployment_Rate +19.5Social_support + (-17.90)Freedom_to_make_life_choices + 2.14Perceptions_of_corruption + (-6.02)confidence_in_national_government) +5.88Democratic_Quality)

log1p_linear_suicides:

y_hat(suicides_100k_pop) = 1.797 + (-0.02)Population + (-0.031)CPI_Index +(-0.043)GDP_Rate +(-0.164)Unemployment_Rate + (4.025)Social_support + (-1.814)Freedom_to_make_life_choices + (-0.65)Perceptions_of_corruption + (-0.938)confidence_in_national_government) +(0.189)Democratic_Quality)



Log Transformation linear Regression

After comparing with residual, we decide to use log1p transformation to adjust our model. The log1p model as below:

y_hat(suicides_100k_pop) = 1.797 + (-0.02)Population + (-0.031)CPI_Index +(-0.043)GDP_Rate +(-0.164)Unemployment_Rate + (4.025)Social_suport + (-1.814)Freedom_to_make_life_choices + (-0.65)Perceptions_of_corruption + (-0.938)confidence_in_national_government) +(0.189)Democratic_Quality)

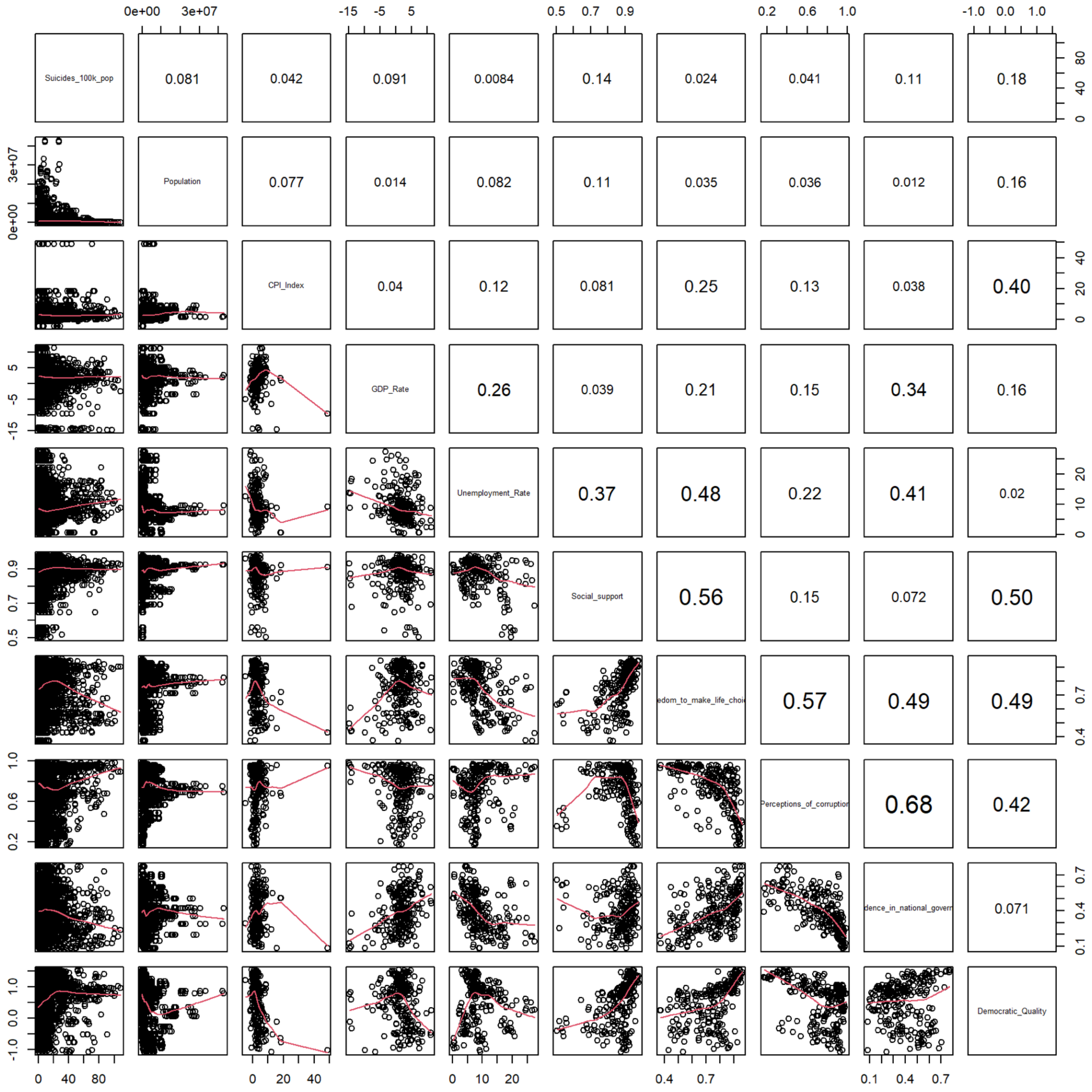
The outcome of our model as following table shows that β6 = -1.814 will influence the suicides_100k_pop for a unit change of Freedom to make life choices, holding other predictors constant. This indicates negative relationship between Freedom_to_make_life_choices and Suicide_100k_pop which means one unit increase of Freedom_to_make_life_choices, Suicides_100k_pop will decrease 1.814 unit holding other predictors constant. Also, we use null hypothesis test whether partial slope coefficient of "Freedom to make life choices" is statistically significant. The p-value is smaller than 0.05. The slope coefficient of "Freedom to make life choices" is statistically significant. We can reject the null hypothesis β6 = 0. It means that 95% confidence that the interval has contained the true values of population average. 5 % the average suicides_100k_pop in the rejected area. "Democratic Quality" and "Social Support" also have the same situation. Their p-value is also smaller than 0.05. The slope coefficient of "Democratic Quality" and "Social Support" are statistically significant. We can reject the null hypothesis β9 & β5 = 0. 95% confidence that the interval has contained the true values of population average. It has strongly evidence that these three predictors can control the suicides_100k_pop for a unit change. But the result has surprised us. "Social support" increases one unit, Suicides_100k-pop will also increase 4.025 unit and "Democratic Quality" increase one unit, Suicides_100_pop will increase 0.189 unit.

Term	Estimate	SE	T-statistic	P-value
(Intercept)	1.797	0.363	4.951	< 0.001
log1p(Population)	-0.02	0.011	-1.798	0.072
log1p(CPI_Index)	-0.031	0.021	-1.479	0.14
log1p(GDP_Rate)	-0.043	0.019	-2.274	0.023
log1p(Unemployment_Rate)	-0.164	0.039	-4.243	< 0.001

log1p(Social_support)	4.025	0.43	9.369	< 0.001
log1p(Freedom_to_make_life_choices)	-1.814	0.315	-5.765	< 0.001
log1p(Perceptions_of_corruption)	-0.65	0.195	-3.33	< 0.001
log1p(Confidence_in_national_government)	-0.938	0.231	-4.064	< 0.001
log1p(Democratic_Quality)	0.189	0.026	7.176	< 0.001

Correlation Plot

We check our model whether it contains multicollinearity and heteroscedasticity in our data matrix scatterplot. Look at the left-hand side of our data scatterplot. We can see that the scatterplot does not have a cone-like shape which means these predictors are not heteroscedasticity. But the scatterplot shows some predictors have influence leverage points that influenced our fitted model direction such as the one approximately located at (50, -14), is on CPI_Index predictor axis with GDP_Rate axis because this point is far from other observation points and has an x-coordinate. Also look at the right-hand side of our data using numbers to point out the relationship between our predictors. The highest correlation is between Perceptions_of_corruption and confidence_in_national_government which has 0.68. The second high correlation is between Perceptions_of_corruption and Freedom_to_make_life_choices which has 0.57. Both are not very high. Therefore, our model does not have multicollinearity. Checking correlation between each predictor can help our model prediction and analysis accurately.



Variance Inflation Factor (VIF)

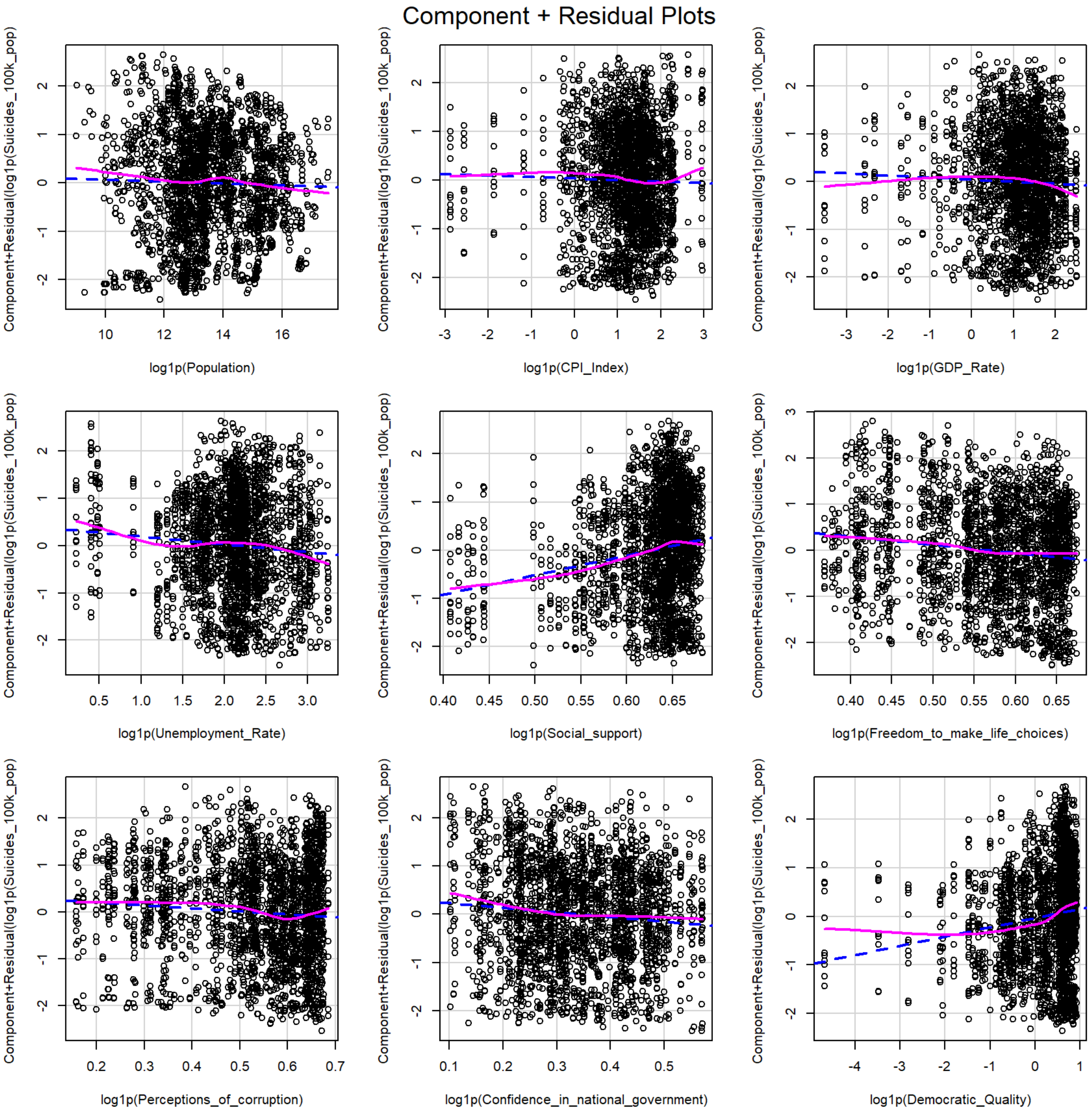
As per VIF for a regression model variable is equal to the ratio of the overall model variance. Thus we test Variance inflation factor to ensure our model does not have multicollinearity. The outcome of VIF as the following table, we can see the range of VIF in our model is between 1 to 2.7. It is a very low outcome. The highest one is “Perceptions of corruption”. “Freedom to make life choices” and “Confidence in national government” are the same as 2.5. This outcome is matching the outcome of the previous correlation plot. We can be sure that our fitted model does not have Multicollinearity.

Variance Inflation Factor(VIF) table

	x
log1p(Population)	1.1
log1p(CPI_Index)	1.1
log1p(GDP_Rate)	1.1
log1p(Unemployment_Rate)	1.7
log1p(Social_support)	1.9
log1p(Freedom_to_make_life_choices)	2.5
log1p(Perceptions_of_corruption)	2.7
log1p(Confidence_in_national_government)	2.5
log1p(Democratic_Quality)	1.9

Residual

In the following Component residual plots, we can see whether each predictor has a linear relationship to Suicides per 100k people. The blue dash line is the best fit line. The pink line is the residuals line. We can indicate that GDP_Rate and Democratic_Quality do not have a linear relationship with Suicides per 100k people even though we have used log1p transformation regression. Thus, we can indicate that both may influence our prediction the most.



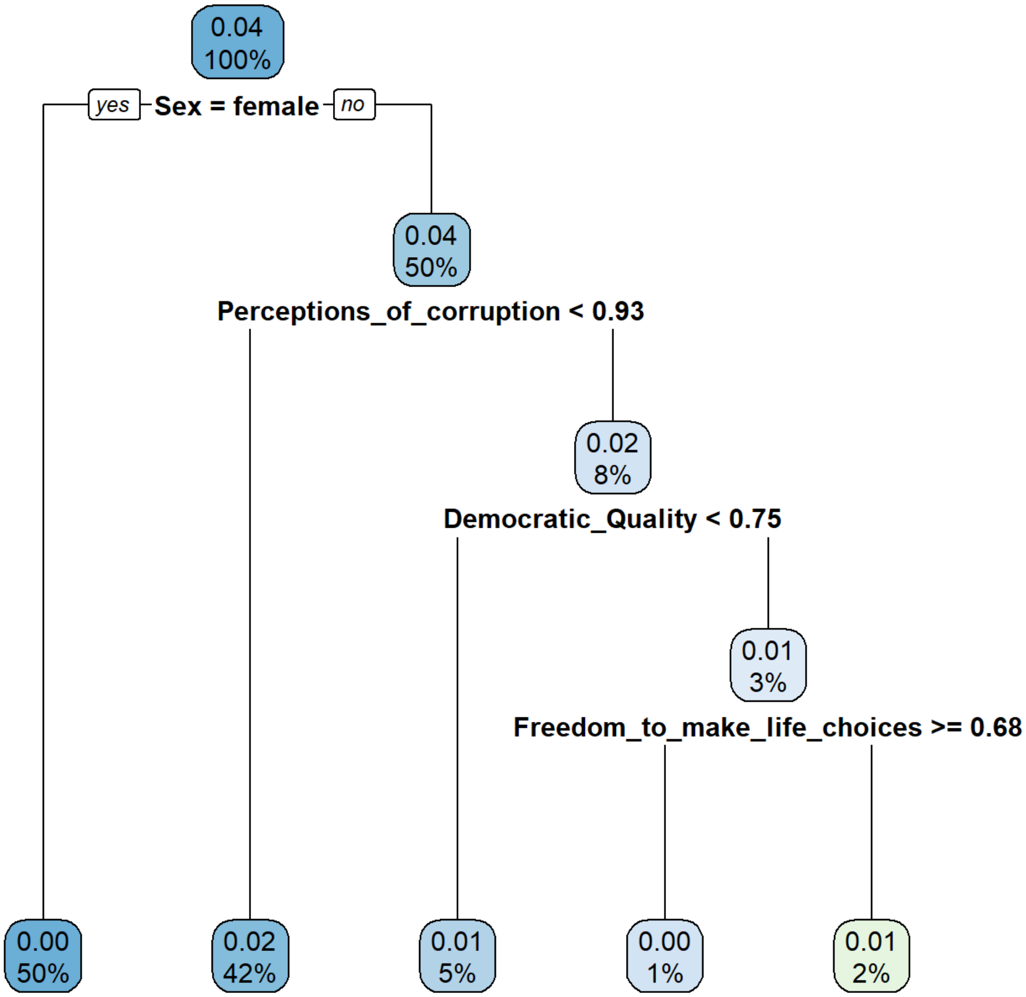
Decision Tree

We split the data 80/20, and use 80% of data to create a train set, and 20% to make predictions. According to the test data set, we can see the overall probability of suicide is 4%. If the sex is male, 50% of males will have a suicide probability of 4%. If Perception of Corruptions rate is smaller than 0.93, the chance of suicide is 2%. If Perception of Corruptions rate is larger than 0.93, the change of suicide is also 2%. If Democratic Quality rate is smaller than 0.75, the chance of suicide is 1%. If it is larger than 0.75, the chance of suicide is 1%. If Freedom to make life choice rate is larger than or equal to 0.68, the chance of suicide is 0. Instead, if it is smaller than 0.68, the chance of suicide is 1%.

The decision tree model we build helps us identify the most important factors that affect suicide rate. Through the model, we can easily find out that people are highly valued " perception of corruption", " democratic quality" and freedom to make life choices' The ultimate goal of our predicting model is to detect and visualize the relationship/patterns among all the variables, as well as determine whether an event (Suicide Y/N) will occur or not.

```
## Rows: 5,880
## Columns: 7
## $ Confidence_in_national_government <dbl> 0.37, 0.37, 0.37, 0.37, 0.37, 0.3...
## $ Perceptions_of_corruption         <dbl> 0.88, 0.88, 0.88, 0.88, 0.88, 0.8...
## $ Social_support                    <dbl> 0.68, 0.68, 0.68, 0.68, 0.68, 0.6...
## $ Democratic_Quality                <dbl> -0.32, -0.32, -0.32, -0.32, -0.32...
## $ Freedom_to_make_life_choices      <dbl> 0.44, 0.44, 0.44, 0.44, 0.44, 0.4...
## $ Sex                               <chr> "male", "male", "male", "male", "...
## $ Suicide                           <chr> "No", "No", "No", "No", "No", "No...
```

```
## [1] 4704   33
```


0.04
100%
yes
Sex = female
no
0.04
50%
Perceptions_of_corruption < 0.93
0.02
8%
Democratic_Quality < 0.75
0.01
3%
Freedom_to_make_life_choices >= 0.68
0.00
50%
0.02
42%
0.01
5%
0.00
1%
0.01
2%

Step-Wise Model

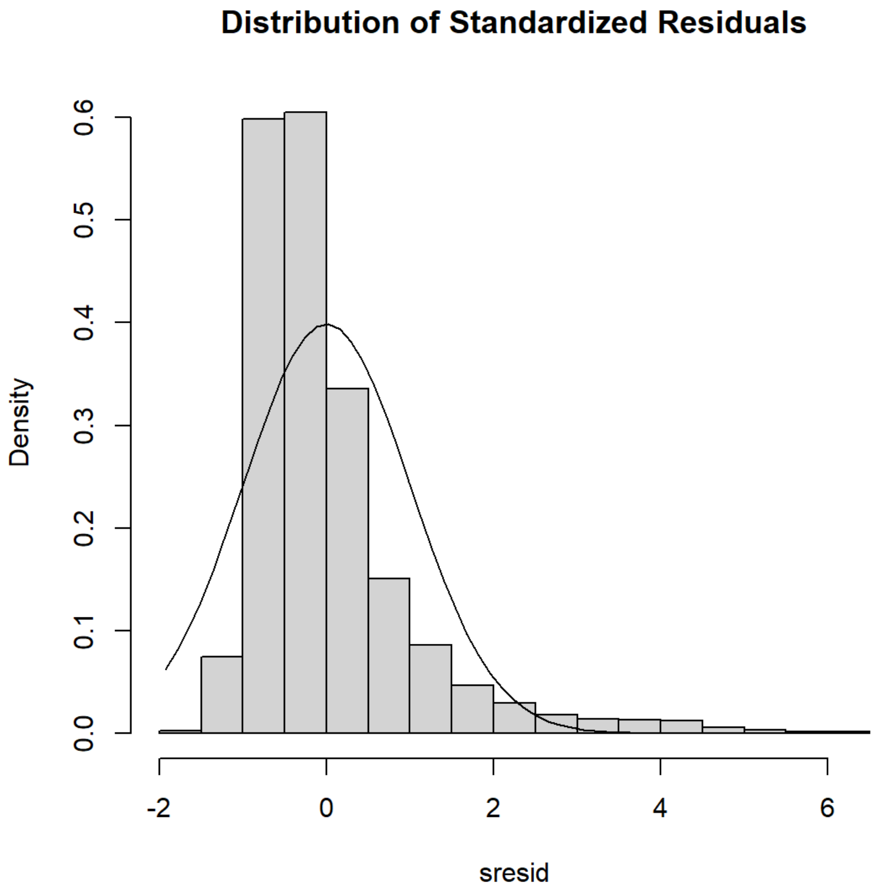
Step-wise model which is a combination of forward and backward selection, consists of iteratively adding and removing predictors in our fitted model in order to find the subset of predictors in our data set resulting in the best performing model with lower prediction error. After running the step-wise model, our best performing model is Population, CPI Index, Unemployment Rate, Social Support, Freedom to make life choices, Confidence in national government and Democratic Quality. R squared is 0.084, Adj.R squared is 0.0829. all partial coefficient and p-value as below table. The most influence is “Freedom to make life choice” to Suicides_100K_pop. When “Freedom to make life choice” increases one unit, it controls Suicide_100k_pop decreases 18.714 units, holding other predictors constant. This outcome is 2.46 times of “confidence in national government” decreasing. Also, it is 88.27 times of “Unemployment Rate”. The table also reveals the p-value of all predictors are smaller than 0.05 which means all rejects the null hypothesis = 0, all slope coefficient are statistically significant. It has strong evidence that these predictors can control the suicides_100k_pop for a unit change.

y_hat(suicides_100k_pop) = 9.558 + (0)Population + (0.374)CPI_Index +(-0.212)Unemployment_Rate + (20.46)Social_suport + (-18.714)Freedom_to_make_life_choices + (-7.592)confidence_in_national_government) +(5.597)Democratic_Quality)

Term	Estimate	SE	T-statistic	P-value
(Intercept)	9.558	2.821	3.389	< 0.001
Population	0	0	-4.755	< 0.001
CPI_Index	0.374	0.053	7.033	< 0.001
Unemployment_Rate	-0.212	0.048	-4.431	< 0.001
Social_support	20.46	3.095	6.611	< 0.001
Freedom_to_make_life_choices	-18.714	2.123	-8.813	< 0.001
Confidence_in_national_government	-7.592	1.536	-4.944	< 0.001
Democratic_Quality	5.597	0.399	14.041	< 0.001

Step-wise model density

In the distribution of a standardized Residual plot with Suicides as below, shows that the distribution is not exactly a normal distribution and it has a little skewness. But comparing with our model which didn’t do the log1p transformation, it is much closer to the population average N(0, 1). This prediction model would be more accurate.



Conclusion

In conclusion, our log1p regression model with low R square values can also be a good model. It is because some field of study have an inherently greater amount of unexplainable variation such as confidence in national government, Freedom to make life choices, etc. To explain human behavior generally is more difficult to predict than things that are physical. At least in our model analysis and prediction, we have tried our best to analyze and predict using diversification to make the outcome accurate.

We find that global suicides are prone to males instead of females. The age between 35-75 is the in most negative thinking. They maybe faced with an enormous amount of pressure. Our prediction reveals that freedom and confidence in national government are the most influenced subjects on people to have an contemplate suicide, instead of say related with economy. Thus, we hope this result of suicide analysis and prediction could help some Governors' awareness that the country's wealth are their people. If they live in the world feeling hopeless and helpless, no one will be willing to contribute their ability and knowledge to their country to build up a prosperous country.