A Suicide Rate Model Based on Economic, Violence and Health Indicators

Hongyuan Zhang and Yuxi Deng

Abstract

Much research suggests that suicide is related to social factors. In this paper, we explore the factors that best fit a country's suicide rate. We considered three types of country statistics, economic statistics, health statistics and violence statistics. We used the best subset technique to select a model with the least multicollinearity and high adjusted R-squared. We finally have a model with the adjusted R-squared value of 0.3409. This model may help provide new research directions and suicide prevention.

Background and Significance

Suicide has been a leading cause of death. According to World Health Organization (WHO), approximately 800,000 people commit suicide every year, which means one person commits suicide every 40 seconds. We decided to explore the relationship between suicide rate and the social factors in a country. Ample research has been done on exploring how social factors affect suicide rate in a country. One prominent study is from Durkheim where he suggests "the victim's acts which at first seem to express only his personal temperament are really the supplement and prolongation of a social condition which they express externally" (p.299). Additionally, Ceccherini-Nelli and Priebe (2010) found a long-run association between economic factors (unemployment rate, RGDP, and CPI) and suicide rates for four countries. However, there is no statistical model to try to fit the suicide rates by country using these possible social factors and indicators. We are interested in developing such a model which may help suicide prevention across the world.

Methods

Data Collection and Variable Selection

Our dataset contains data from 183 different countries. Age-standardized suicide rates (per 100 000 population) is the response variable. The dataset has 15 explanatory variables from World Health Organization (WHO), World Bank, Central Intelligence Agency(CIA) and United Nations Office on Drugs and Crime(UNODC). A table of data sources is attached in the Appendix A. We used data from 2015 due to comprehensive data coverage. For all the missing values from 2015 in our datasets, we substituted with the data from the most recent year (no earlier than 2000) from the organizations we collected data from. If there is still no available value, we marked "NA" for that cell.

The 15 explanatory variables are mostly from 3 categories: economic indicators, violence indicators, and health indicators. Economic indicators demonstrate the affluence of material life of citizens, violence data indicates the safety and the lawful order in the country, and health data indicates the physical well-being of citizens. Thus, we believe these data are representative of the comprehensive sociological scene in a country and are reasonably related to the response variable. Besides, for each country, we include data on a developed country or no (0 for developing and I for developed countries), population density and the region the country is in.

Table 1. List of Explanatory Variables

Economic Indicators	Health Indicators	Violence Indicators	Other
GDP per capita *(GDP per capita)* CPI *(CPI)* % of Tax revenue in GDP *(TaxRevenue%GDP)* Unemployment rate *(UnemploymentRate)* GINI Coefficient *(GINI)*	Life expectancy at birth *(LifeExpectancy)* Age-standardized mortality rate per 100,000 population *(MortalityRate)* Age-standardized DALYs for alcohol use disorders *(AlcoholDALY)*	Homicide rate per 100,000 population *(HomicideRate)* Sexual violence rate per 100,000 population *(SexualViolenceRate)*	Population Density per Sq. Km *(PopulationDensity)* Developed *(Developed)* Region *(Region)* % of Urban population *(UrbanPopulation%)*

	Age-standardized DALYs for drug use disorders *(DrugDALY)*		

For the variable *SexualViolenceRate*, there are 68 missing values, but we believe it's highly related to the response variable, we decided to keep it and test its significance later. We also found that *LifeExpectancy* and *MortalityRate* are correlated with each other, so we decided to keep both and test which one is the better indicator later. Lastly, we do not expect other variables to be correlated with each other.

Analytic Methods

We developed a multiple linear regression model for suicide rates using the best subset technique. Adjusted R-square and standard errors are used to assess the model.

Results

According to Appendix B Figure 1, the average suicide rate is almost the same for developed and developing countries. According to Appendix B Figure 2, Arab States countries have the lowest average, Latin American and the Caribbean countries have the second lowest, and the other three regions have similar higher averages.

Since there are 68 missing values in *SexualViolenceRate,* we first tested its significance to decide whether we need to keep it in the dataset. We used the best subset technique on all the explanatory variables except *SexualViolenceRate* and two categorical variables *Developed* and *Region*, with countries that have no missing values in all columns. Based on adjusted R^2 and VIF, we picked Model 1 (see Table 2) since it has highest R^2 among all models with low VIF. However, as shown in Appendix C Figure 5, the residuals do not follow a normal distribution for Model 1 and there are obvious outliers. We found including either categorical variable does not help satisfy the assumptions after testing the assumption. Thus, we transformed the data and found square root transformation yields the most normally distributed residuals for Model 1. We used the best subset technique again with all explanatory variables but with the transformed response variable. Using the same criteria, we picked Model 2. It turns out that Model 1 and Model 2 have the same 5 explanatory variables. All the assumptions are met in Model 2. Using Model 2 as the reduced model, the p-value for *SexualViolenceRate* is 0.838392, which suggests it is not significant, so we excluded *SexualViolenceRate* from our dataset.

We again used the best subset technique on the remaining explanatory variables except two categorical variables *Developed and Region* with countries that do not have missing values for any columns. Considering both high adjusted R^2 and low VIF, we picked Model 3. Again, we have the non-normality and outlier problem. Including either categorical variables does not help us meet the model assumption. Thus, we transformed our response variable using square root and performed the best subset technique with the 12 explanatory variables. All the model assumptions are met. Using the same criteria, we picked Model 4 as the final model.

Table 2. Multiple linear regression models

	UrbanPopulation%	HomicideRate	MortalityRate	GINI	AlcoholDALY	DrugDALY	UnemploymentRate	GDP per capita	CPI	Adjusted R^2
Model 1		-0.09198	0.035071	-0.10908	0.005856		-0.266628			0.2977
Model 1 VIF		1.276208	1.187708	1.317697	1.184781		1.049941			
Model 2		-0.0147885	0.0050664	-0.01936	0.000853		-0.048422			0.309
Model 2 VIF		1.276208	1.187708	1.317697	1.184781		1.049941			
Model 3	2.86E-02	-1.15E-01	4.53E-02		7.72E-03	-5.18E-03	-1.44E-01	4.20E-05		0.3414
Model 3 V	1.748224	1.16535	1.785947		1.27894	1.026745	1.051383	1.828378		
Model 4	6.86E-03	-1.61E-02	7.82E-03	-9.71E-03	1.06E-03	-7.23E-04	-2.51E-02	6.88E-06	-6.22E-03	0.3409
Model 4 V	1.81976	1.32395	2.01587	1.435831	1.289278	1.046816	1.083116	1.8766	1.184886	

Discussion

Our model has 9 explanatory variables with an adjusted R-square of 0.3409 and standard error of 0.7721 (Appendix D). This means that our model explains about 34% of the variability of suicide rates, and the data points have an average distance of 0.7721 from the regression line. Considering the range of the suicide rates, which is 34.60, this distance is relatively small, which indicates that our model is relatively precise.

Considering the coefficients of the terms, some of the relationships are what we expected. For example, the expected suicide rate will increase by 7.820e-03 per 100,000 population when mortality rate increases by 1 person after accounting for corresponding changes in all the other explanatory variables in the model. We also expected the positive sign for *AlcoholDALY*. However, the economic factors are not associated with suicide rates in the way as suggested in Ceccherini-Nelli and Priebe's study. We should notice that the coefficient of each term is not large compared to its standard error, which is the expected error in evaluating the coefficient. This indicates that we did not quite precisely estimate the coefficients and some coefficients can be in fact zero (no association). Also, there exists a moderate correlation between the variables. Thus, we may not interpret the coefficients further.

Our study has limitations. First, nonresponse bias exists since there are missing values in our dataset which are all from developing countries. Second, although we substituted the missing values, these values may not be accurate since different years may have largely different values. Third, besides the explanatory variables we chose, there are other important social factors such as culture and mental illness that do not have available data. Fourth, the data are from an observational study, not an experiment. Therefore, even though the model reveals relationships between our explanatory variables and the response variable, it does not imply a causal link between them. Fifth, all the data are estimates, and thus they may be inaccurate, which might lead to inaccuracy in our model.

Our final model helps predict a country's suicide rate, given the country's statistics. This study may also provide evidence for the weak relationship between sexual violence rate and the suicide rate for a country. However, due to the nonresponse bias, this evidence may only be used with caution. Further research on the significance of sexual violence rate is needed if there is more available data. This study also provides new directions for research on suicide, for example, the association between urbanization and suicide rate for a country and the similarity for suicide rates between developed and developing countries. This study may also help better allocate suicide prevention resources, as shown in Appendix B Figure 2 that African, Asian and the Pacific, Europe and North American countries have higher suicide rate averages.

References

Ceccherini-Nelli, A., & Priebe, S. (2010). Economic factors and suicide rates: associations over time in four countries. Social Psychiatry and Psychiatric Epidemiology, 46(10), 975-982. doi:10.1007/s00127-010-0275-2

Durkheim, E. (1951). Suicide, a study in sociology; a study in sociology. Glencoe, IL: Free Press.

Holmes, R. M., & Holmes, S. T. (2005). Suicide: Theory, Practice, and Investigation. Sage Publications.

Suicide. (n.d.). Retrieved December 13, 2017, from http://www.who.int/mediacentre/factsheets/fs398/en/

Appendix A. Variable Sources

Variable	Data source
Developed	https://www.cia.gov/library/publications/the-world-factbook/appendix/appendix-b.html#D
GDP per capita	https://data.worldbank.org/indicator/NY.GDP.PCAP.CD
TaxRevenue%GDP	https://data.worldbank.org/indicator/GC.TAX.TOTL.GD.ZS
CPI (Consumer Price Index)	https://data.worldbank.org/indicator/FP.CPI.TOTL.ZG
UnemploymentRate	https://data.worldbank.org/indicator/SL.UEM.TOTL.NE.ZS
GINI	https://www.cia.gov/library/publications/the-world-factbook/rankorder/2172rank.html
LifeExpectancy	http://apps.who.int/gho/data/node.main.688
MortalityRate	http://apps.who.int/gho/data/node.main.11
AlcoholDALY	http://apps.who.int/gho/data/node.main.A1218?lang=en
DrugDALY	http://apps.who.int/gho/data/node.main.A1218?lang=en
HomicideRate	https://data.unodc.org/
SexualViolenceRate	https://data.unodc.org/
UrbanPopulation% Population Density	https://www.socialexplorer.com/tables/WDI2015
SuicideBoth	http://apps.who.int/gho/data/node.main.MHSUICIDEASDR?lang=en
Region	https://www.cia.gov/library/publications/the-world-factbook/

Appendix B. Boxplots

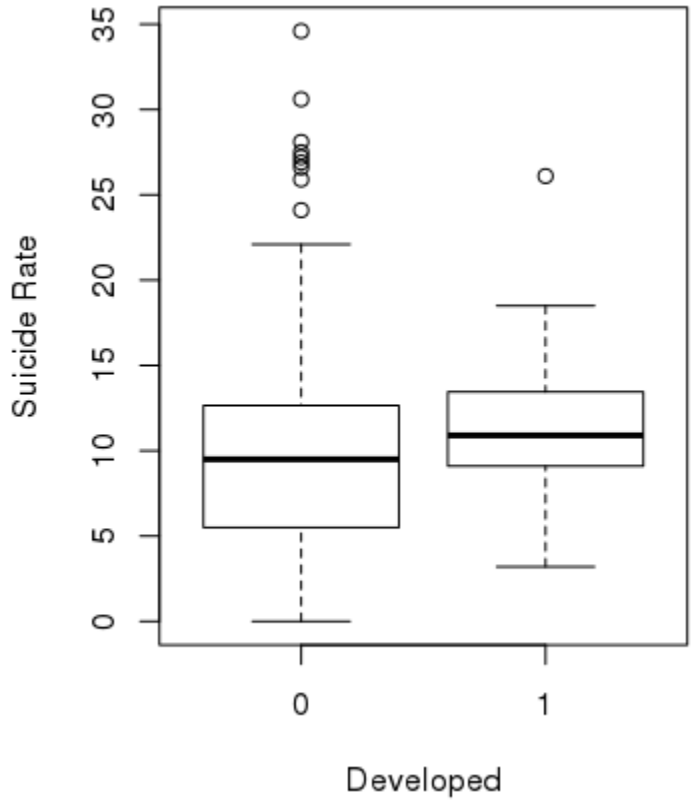


Figure 1. Boxplot for *Developed*

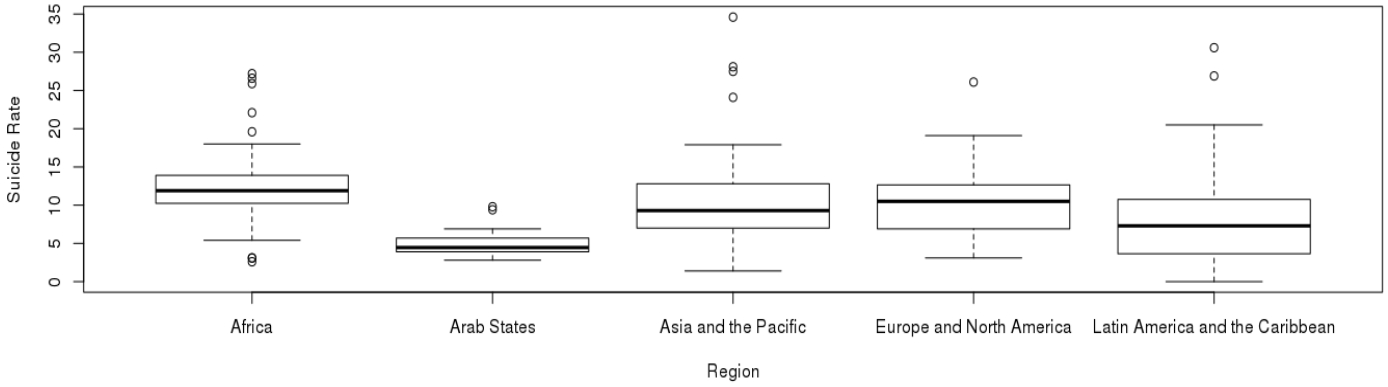


Figure 2. Boxplot for *Region*

Appendix C. Residual Plots

Model 1 residual plots:

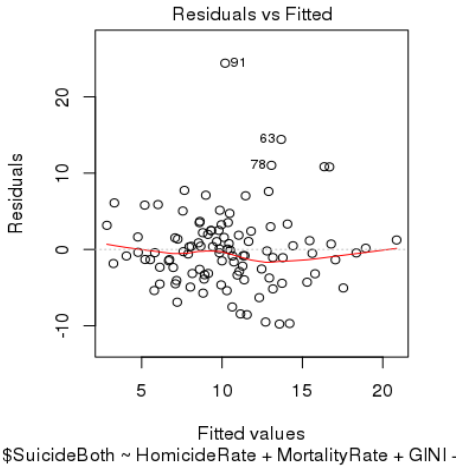


Figure 3. Residuals vs. Fitted

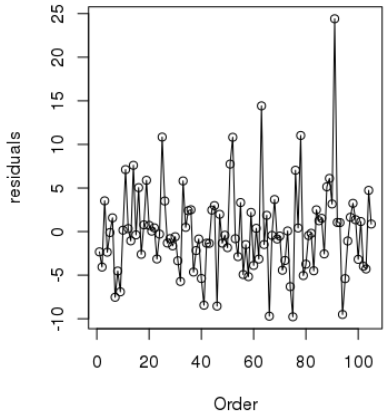


Figure 4. Residuals vs. Order

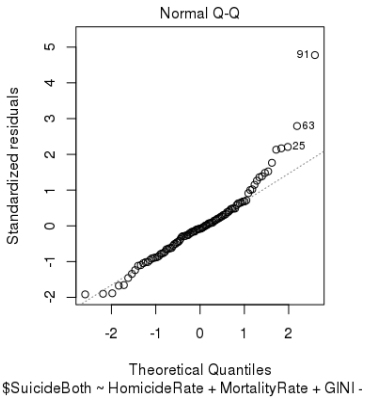


Figure 5. Normal Q-Q

Figure 3, 4, and 5 show obvious outliers. Figure 5 shows the non-normality of the residuals of Model 1.

Model 2 residual plots:

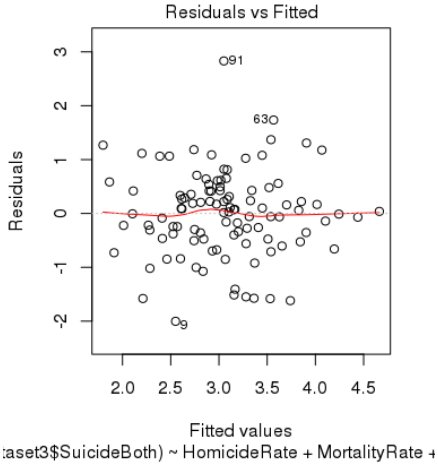


Figure 6. Residuals vs. Fitted

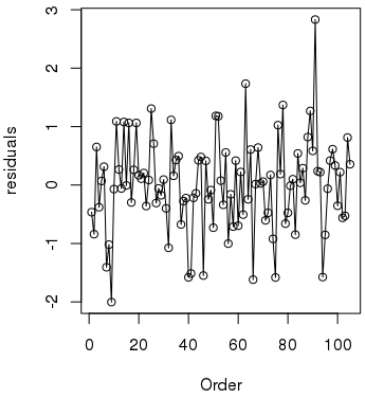


Figure 7. Residuals vs. Order

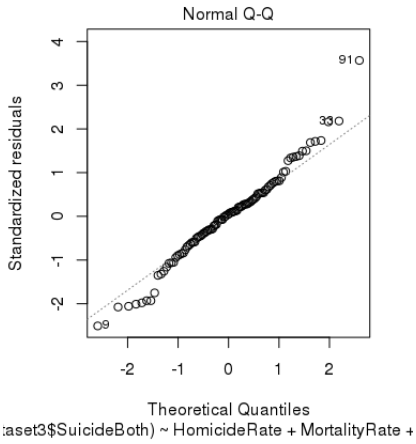
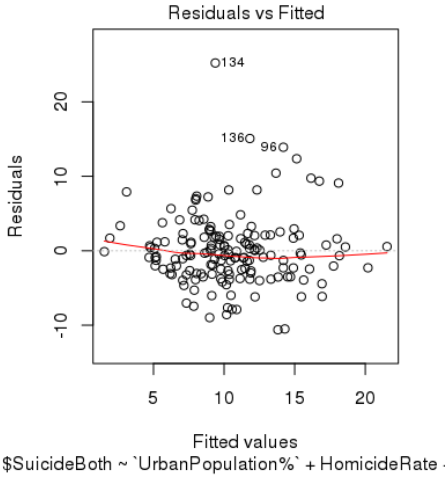
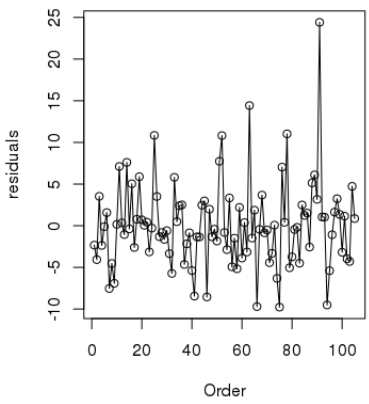


Figure 8. Normal Q-Q

After a square root transformation of the response variable, the outliers are modified (compare Figure 3, 4, 5 and Figure 6, 7, 8), the residuals are more normally distributed, and all the model assumptions are basically met.

Model 3 residual plots:





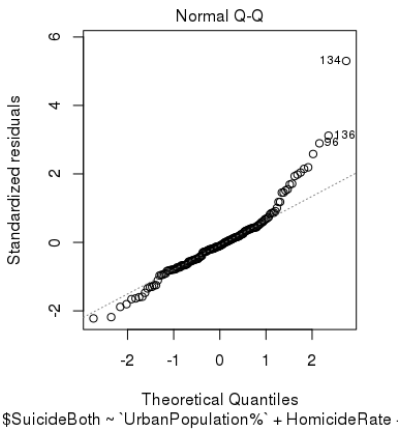
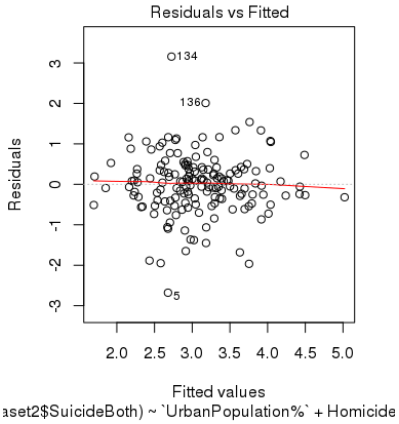
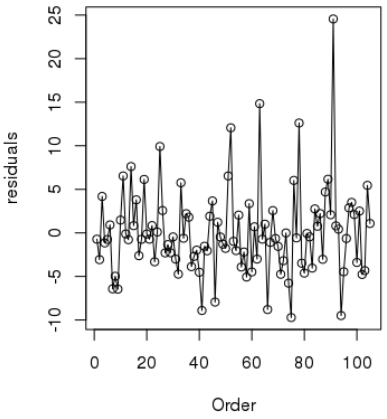


Figure 9. Residuals vs. Fitted Figure 10. Residuals vs. Order Figure 11. Normal Q-Q

Figure 9, 10, and 11 show obvious outliers. Figure 11 shows the non-normality of the residuals of Model 3.

Model 4 residual plots:





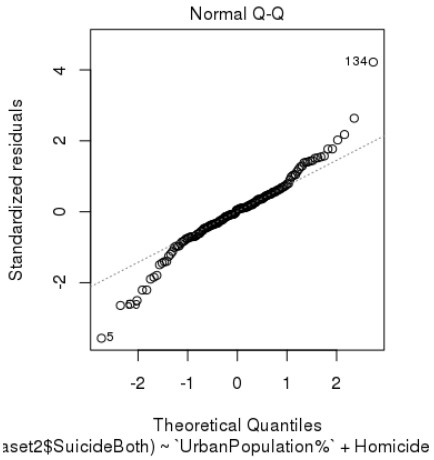


Figure 12. Residuals vs. Fitted Figure 13. Residuals vs. Order Figure 14. Normal Q-Q

After a square root transformation of the response variable, the outliers are modified (compare Figure 9, 10, 11 and Figure 12, 13, 14), the residuals are slightly more normally distributed, and the model assumptions are better met.

Appendix D. Model 4

	Estimate	Std. Error	Value	Pr(>|t|)
(Intercept)	1.821e+00	3.973e-01	4.585	9.42e-06
UrbanPopulation%	6.861e-03	3.674e-03	1.867	0.06380
HomicideRate	-1.610e-02	5.529e-03	-2.911	0.00414
MortalityRate	7.820e-03	9.751e-04	8.020	2.63e-13
AlcoholDALY	1.063e-03	2.489e-04	4.269	3.45e-05
DrugDALY	-7.230e-04	4.431e-04	-1.632	0.10482
UnemploymentRate	-2.509e-02	9.302e-03	-2.697	0.00778
GDP per capita	6.884e-06	4.647e-06	1.481	0.14056
CPI	-6.220e-03	5.946e-03	-1.046	0.29718
GINI	-9.710e-03	7.900e-03	-1.229	0.22091
Standard Error	0.7721			
Adjusted R^2	0.3409			