Analysis of Influencing Factors Leading to Suicidal Actions via Linear Regression and Regularization Methods

by

Anna Franziska Bothe

(576309)

Humboldt-Universität zu Berlin

School of Business and Economics

Ladislaus von Bortkiewicz Chair of Statistics

Dr. rer. nat. Sigbert Klinke

in fulfillment of the requirements

for Data analysis II

March 29, 2019

Contents

List of Figures

List of Tables

1 Introduction

Every year approximately 800.000 people commit suicide – without including the number of attempted suicides and not or falsely reported cases (World Health Organization, 2018).

In order to be able to take countermeasures and develop prevention strategies, the underlying and influencing factors have to be examined.

For the analysis financial (e.g. GDP), political (e.g. terrorism), cultural/social (e.g. HDI/family) as well as socio-economic (e.g. gender) factors influencing suicidal actions are considered.

Nowadays, most data sets are large-scaled which leads to interpretation problems because a model is explained by many variables. Moreover, explanatory variables in empirical data correlate always a little which causes multicollinearity (Auer and Rottmann, 2011).

The paper deals with shrinkage and selection methods in order to reduce the complexity of the data set. Due to the combination of several data sets and the attempt to record as many influencing factors as possible, the data to be analyzed tend to be multicollinear. Besides the regularizing method, the paper focuses on methods of detecting multicollinearity and provides approaches to act on it.

The structure of the paper is organized as follows. The second chapter describes the data sets, their collection and processing as well as their advantages and disadvantages, followed by the data cleaning and preparation process for the analysis. In the third section, the methods are briefly described, implemented, interpreted and lastly, compared. Finally yet importantly, a conclusion regarding the results, the influencing factors and the applied methods is drawn.

2 Data

The following subsection presents general information about the data to be analyzed and describes the five different sub-data sets that have been merged in order to construct the analyzed data set. Moreover, the data quality is discussed and the data cleaning as well as the preparation process is described in detail.

2.1 Data Sets

The data sets Suicide Rates Overview 1985-2016, the World Happiness Report 2015 and parts of the Global Terrorism Database are merged in order to construct a data set with as much details about the socio-economic circumstances as possible, potentially influencing suicidal actions. This includes, among other variables, the general development of the countries, the gross domestic products (GDP), terrorism, corruption but also individual related variables such as age and gender. Only the years 2014 and 2015 respectively is analyzed.

2.1.1 Suicide Rates Overview 1985-2016

The data set regarding suicidal action between 1985 to 2016 has been constructed by merging three different indices/data sources: The Human Development Index (HDI), World Development Indicators – GDP (current US$) by country and WHO Suicide Statistics (Rusty, 2018).

The Human Development Index (HDI) is measured by the following three dimensions regarding human development: "a long and healthy life, access to knowledge and a decent standard of living", United Nations (2019). Data sources for the index are e.g. the life expectancy at birth, expected years of education and the GNI per capita. The index is calculated in two steps. First, the dimension indices are created by setting a lower and upper bound for each indicator, e.g. life expectancy is min. 20 and max. 85 because no country has a life expectancy below 20 and above 85. After determining the minimum and maximum values, the dimension index (I) is calculated as follows:

$$I = \frac{\text{actual value} - \text{minimum value}}{\text{maximum value} - \text{minimum value}} \tag{1}$$

Secondly, the dimensional indices are aggregated in order to compute the HDI. It is the geometric mean of the indices:

$$\text{HDI} = (I_{Health} * I_{Education} * I_{Income})^{\frac{1}{3}} \tag{2}$$

The HDI is computed for 101 countries annually. Missing values are estimated via cross-country regression models (United Nations, 2018).

The World Development Indicators (WDI) data regarding the GDP (current US$) by country has been retrieved from the primary World Bank collection of development indicators which is considered to be the most current and accurate global development database. The data is collected via mainly external sources such as reports by the UN but also via survey conducted by the World Bank Group itself. The GDP data is based on external data but the given indicator is produced and compiled by the World Bank. The World Bank collects it in local currency by using the published information by the national authorities. The GDP (current US$) is one of the most popular WDI indicators and therefore, a lot of emphasis is put on a timely, correct and complete collection (World Bank, 2019).

WHO Suicide Statistics data set includes the number of suicides each year, the country, the year, age groups, the sex and the total amount of population as well the rate of suicides per 100k population, also called ”crude rate” (Szamil, 2017). The data is aggregated from the World Health Organization (WHO) Mortality Database. The source of the data are deaths registered in national civil registration databases. Suicide rates for 172 countries are estimated. The data is provided annually by the WHO (World Health Organization, 2018).

2.1.2 World Happiness Report 2015

The World Happiness Report is a survey that aims to measure global happiness of 156 countries’ citizens. The annual report is produced and published by the United Nations Sustainable Development Solution Network in cooperation with the Ernesto Illy Foundation. The six factors that are measured by the World Happiness Report are the GDP, the life expectancy, the generosity, the social support/family, freedom, and the degree of corruption of the country. Additionally, people are asked to assess their happiness on a scale from 1 to 10, with 10 being the highest (very happy) (United Nations Sustainable Development Solution Network, 2015).

2.1.3 Global Terrorism Database

The Global Terrorism Database (GTD) is the most comprehensive, unclassified open-source database recording terrorist incidents from 1970-2017. The data is retrieved from publicly available primary and secondary as well as national and international data sources such as archives, media and other databases. Sources are just included if they are proven to be credible (National Consortium for the Study of Terrorism and Responses to Terrorism (START), 2018). Only the variables regarding the amount of terrorist incidents per country in 2014 are taken into account for the analysis.

2.2 Data Quality

Data inconsistencies and possible bias, resulting from the collection process of the data, are described in this section.

The world happiness data set is the only random sample. The data sample is collected by the Gallup World Poll ensuring that the survey data represent 95% of the world's adult population (Gallup, 2019). Primary data collection introduces some bias depending on the interviewer and the surroundings during the collection among other things.

The other databases try to collect as much information as possible and consciously (non-randomly) exclude countries where e.g. the data quality is not reliable. The data from the suicide overview data set as well as the GTD data are mainly retrieved from secondary sources which is cheap and fast but can also be less trustable. Moreover, data from developed as well as developing countries are collected which might introduce bias due to the following reasons:

- Technology and communication boundaries: the main source of the GTD is the internet (GTD Codebook, 2017). Countries with poor technological development are reporting less online than e.g. highly developed countries.

- Culture: the WHO collects data of suicides, in some countries suicide is illegal or very dishonorable. Hence, suicides are misclassified or not reported at all (World Health Organization, 2018).

- Lack of transparency: some countries have a poor standard of bureaucracy and anti-corruption policy (World Health Organization, 2018). It leads to similar results as the cultural bias.

Counteractions are taken. The quality of the data that are retrieved by the World Bank Group is checked before including it in the database (World Bank, 2019). The WHO uses only data from 60 countries directly for the estimation of suicide rate due to data quality issues. The other 112 countries' suicide rates are based on modeling methods – even though, this introduces some systematical bias because the good-quality data normally comes from high-income countries. Furthermore, monitoring and surveillance of suicidal actions are constantly improved in order to enhance the comprehensiveness, timeliness and over-all quality of the data (World Health Organization, 2018). The happiness survey is standardized and follows strict rules in order to make it representative (Gallup, 2019).

In general, the global availability and quality of data regarding suicides is poor. But the data sets are retrieved from reliable sources that attempt to reduce disruptive factors as good as possible. Overall, the data quality is suitable for the analysis. Nevertheless, when discussing the results, the natural difficulties of the data collection should not be forgotten.

2.3 Data Cleaning and Preparation

In order to observe the effect of the given variables on the suicide rate per 100k population without having also a time effect, only data from one year is chosen by the following criteria: recent year, enough observations within the suicide database and an existing world happiness report referring to this year.

The World Happiness Report 2015 was published on 23^{th} April 2015. I assume that the queried happiness of the report in 2015 is a result of the socio-economic and individual circumstances of 2014. Hence, the data of 2014 are matched with the happiness report of 2015.

All incidents that have been recorded in the GTD are commented whether a doubt about the terrorism property exists or not. Additionally, three criteria regarding different severe definitions of the term terrorism are applied (GTD Codebook, 2017). In order to ensure that just actual terrorist incidents are analyzed, only observations, that fulfill all three criteria and that are without a doubt regarding their terrorism property, are included (cp. among others Bothe (2019), Stern and McBride (2013)).

The suicide data set shows missing values for Puerto Rico, Russia and South Korea. The HDI missing data for Russia and South Korea are missing at random (MAR) and the data are imputed by the HDI score that is published online for every year. The values for the HDI are missing not at random (MNAR) in case of Puerto Rico. Since Puerto Rico is an United States territory, it is not independent (Cardona et al., 2019). Estimating a HDI introduces more likely bias to the analysis than removing the observations which makes up less than 1.3% of the data. Thus, all observations regarding Puerto Rico are removed.

The happiness report is not conducted for all countries that are listed in the suicide data set. Since there is no information about the country selection for the happiness report, it is difficult to tell whether the data are missing at random or not. The countries with missing values are presented in table 1. Other variables are added to estimate if an imputation of the regional mean or median is justifiable.

Following abbreviations are used: Latin America/Caribbean = LA/C, Sub-Saharan Africa = SSA, Eastern Asia = EA.

Country	HDI	GDP/Capita	Suicides 100k/Pop	Region
Antigua & Barbuda	0.783	14,093	0	LA/C
Belize	0.715	5,448	8	LA/C
Cuba	0.769	7,459	15	LA/C
Grenada	0.750	9,456	0	LA/C
St. Lucia	0.729	9,372	8	LA/C
St. Vincent & Grenadines	0.720	7,496	7	LA/C
Average of LA/C	**0.735**	**9,962**	**18**	**LA/C**
Republic of Korea	0.896	29,120	32	EA
Average of EA	**0.891**	**40,328**	**9**	**EA**
Seychelles	0.772	16,018	4	SSA
Average of SSA	**0.722**	**9,147**	**5**	**SSA**

Table 1: Missing countries from the happiness report and their underlying data.

As observed in table 1, the missing data are from different regions, even though most unavailable data are from Central America/Caribbean. Some countries' data are close to the regional average and some are not, a missingness at random (MAR) is assumed. Since all countries vary strongly around their regional mean in at least one of the three variables, no imputation is performed and the observations are removed from the data set.

The different country naming, such as "Russia" or "Russian Federation", are adjusted for all data frames.

After the data cleaning and preparation process, the data set to be analyzed has 828 observations described by 26 variables. The gender, age and region variables are coded as dummies; the other 12 variables are scaled as numeric.

3 Analysis

The following chapter deals with the analysis of the underlying data set. Methods are described, outputs are interpreted and difficulties that occurred during the implementation of the models are discussed.

3.1 Multiple Linear Regression

The objective of multiple linear regression is to model the relationship between two or more independent variables and a response variable by fitting a linear equation to the observed data:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + ... + \hat{\beta}_k x_{ik} + u_i \tag{3}$$

with y_i as the numeric dependent variable for the ith sample, β_0 as the estimated intercept, β_j representing the estimated coefficients for the j^{th} prediction and x_{ij} as the value of the j^{th} predictor for the i^{th} sample. u_i represents the error term that describes only random effects that cannot be systematically captured by the model (Kuhn and Johnson, 2013). The Ordinary Least Squares (OLS) method attempts to minimize the sum-of-squared errors (SSE) between the observed and predicted dependent variable:

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{4}$$

with y_i being the observed value and $\hat{y}_i$ being the prediction (Auer and Rottmann, 2011).

In order to prevent the linear model from overfitting, a training and testing set is created with replacement and a partitioning of 70%/30%. Hence, the training set contains 588 observations and the testing set has 240 observations. The linear regression analysis is conducted with help of the R function lm[1]. The results of the regression fitted on the training set are shown in table 2. The table shows the independent variables plus the intercept, the estimates which represent the slope of each coefficient, the standard error, t-value and the p-value. Below statics regarding the analysis evaluation are presented.

[1] https://www.rdocumentation.org/packages/stats/versions/3.5.3/topics/lm

The null hypothesis that is tested for each coefficient implies that the coefficient is equal to zero which means that it has no effect on the response variable. A low p-value indicates that the null hypothesis can be rejected; thus, the coefficient has an effect on the number of suicides per 100k population. Seven of the coefficients show a p-value below 0.01 which means that they are statistically significant. The null hypothesis can also be rejected for six other variables who have a significance level of 0.05 and 0.1. A low p-value leads to a higher t-value because both indicators are related.

The dummy variable *Male* seems to have a significant positive effect on the response variable. If it is 1 (= male), *Suicides per 100k population* increases on average by 13.206. In contrast to *Generosity* which has a significant decreasing effect on the number of suicides. If it rises by 1 unit, the number of *Suicides per 100k population* decreases by 12.459 on average.

The R^2 as well as the standard error give an indication about the goodness-of-fit of the model on the sample data. The R^2 indicates what share of the variance of the model can be explained with the given independent variables. By adding variables, the R^2 slowly increases. Therefore, an adjusted R^2 is computed. It adapts the R^2 to the number of variables (Kuhn and Johnson, 2013). The R^2 as well as the adjusted R^2 are around 0.5. Thus, about 50% of the model's variance can be explained by the predictor variables.

The F-statistic shows that the overall model is also statistically significant on a level of $\alpha = 1\%$.

Table 2: Linear regression results

Independent Variable	Estimate	Std. Error	t-value	$Pr(>\lvert t\rvert)$
Intercept	-30.958^{***}	(8.538)	-3.626	0.0004
HDI 2014	33.530	(21.523)	1.558	0.120
GDP 2014	-0.000^{*}	(0.000)	-1.819	0.070
GDP/Capita	-0.00001	(0.00004)	-0.190	0.850
Male	13.206^{***}	(0.831)	15.895	0.000
25-34 yearolds	3.081^{**}	(1.471)	2.094	0.037
35-54 yearolds	5.855^{***}	(1.411)	4.150	0.00004
5-14 yearolds	-7.030^{***}	(1.428)	-4.925	0.00001
55-74 yearolds	6.144^{***}	(1.431)	4.295	0.00003
>75 yearolds	12.622^{***}	(1.461)	8.640	0.000
Happiness Score	0.982	(1.143)	0.859	0.391
Economy GDP/Capita	8.448	(6.645)	1.271	0.205
Family	-2.383	(3.706)	-0.643	0.521
Health-Life-Expectancy	-14.353	(9.701)	-1.479	0.140
Freedom	5.991	(5.794)	1.034	0.302
Trust-Government-Corruption	-6.012	(5.877)	-1.023	0.307
Generosity	-12.459^{**}	(5.469)	-2.278	0.024
Australia/New Zealand	5.666	(5.348)	1.060	0.290
Central/Eastern Europe	8.868^{***}	(3.172)	2.796	0.006
Eastern Asia	14.326^{**}	(6.143)	2.332	0.021
Latin America/Caribbean	5.579	(3.782)	1.475	0.141
Middle East/Northern Africa	-7.422^{*}	(4.174)	-1.778	0.076
North America	21.445^{*}	(11.861)	1.808	0.072
Southeastern Asia	1.727	(5.026)	0.344	0.732
Western Europe	3.920	(3.905)	1.004	0.316
Terrorist Incidents	0.014^{***}	(0.005)	3.049	0.003
Observations	588			
R^2	0.514			
Adjusted R^2	0.493			
Residual Std. Error	10.017 (df = 562)			
F Statistic	23.796^{***} (df = 25; 562) (0.000)			

Note: *p<0.1; **p<0.05; ***p<0.01

3.1.1 Residual Diagnostics

One assumption of the linear regression model is that a linear relationship between the dependent and independent variables is given. Residual analysis allows to test this assumption. Residuals are defined as the difference between the observed and the predicted value of the response variable:

$$u_i = y_i - \hat{y}_i \tag{5}$$

with $E[u \big| x_1, x_2, ..., x_k] = 0$ and $E[u] = 0$.

The linear regression that has been described and interpreted in 3.1, provides the following output regarding the relationship between the residuals and the fitted variables:

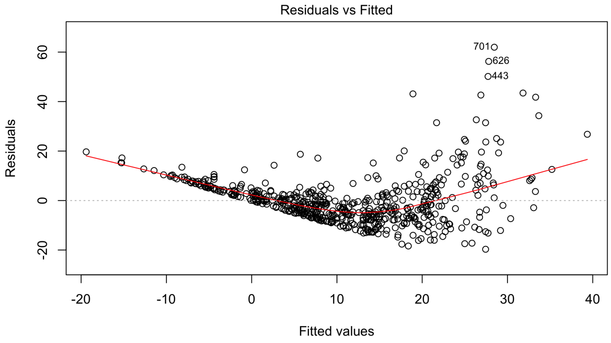

Figure 1: Residual diagnostics of the residuals and the fitted values

Figure 1 shows a non-random, u-shaped pattern which implies a non-linear relationship between the response variable and the independent variables. Hence, the performed linear regression is not an appropriate model for the underlying data yet. Consequently, a non-linear transformation is applied in order to achieve linearity for the regression model. After testing quadratic, reciprocal and exponential model transformation on the response variable, the exponential model provides the best results: the highest R^2, the lowest residual standard error and a random pattern of the distribution of the residuals.

Model	R^2	Std. Error
No transformation	0.5142	10.02
Quadratic	0.6690	1.118
Reciprocal	0.3555	1.005
Exponential	0.7360	0.744

Table 3: Results of implementing different non-linear transformation techniques on the response variable

After the exponential transformation of the response variable, the new residual diagnostics can be seen in figure 2. A slight tendency regarding an u-shaped pattern can still be observed. However, it is only due to a few observations on the left-upper corner. All marked observations are shown in the Cook's distance plot which is an indicator for influential outliers. Outliers may influence the interpretation of the model caused by an increase of the error term (Kassambara, 2018). The rule-of-thumb is that a variable has a high influence if the Cook's distance exceeds the following equation:

$$C_i = \frac{4}{(n-p-1)} \tag{6}$$

with n being the number of observations and p the number of independent variables (Fox, 1997). The Cook's distance of the analyzed data is 0.00712. 33 observations exceed this value. According to Cook and Weisberg (1982), only observations with $D_i > 0.5$ should be explored in detail. All calculated Cook's distances are below this threshold. The residuals vs. leverage plot in figure 2 gives also an indication about the existence of outliers. If a standardized residual is greater than 3, it is a potential outlier. If the value exceeds the following equation, it indicates an observation with a high leverage value.

$$L_i = \frac{2(p+1)}{n} \tag{7}$$

Observations with a high leverage value should be examined in detail (James et al., 2013). The leverage value that should not be exceeded by the data is 0.0884. As observed in figure 2, several observations are above the value and therefore examined in detail because those observations might influence the regression results.

The outliers' values of the variables of the investigated observations are randomly distributed. The extreme observations make up 5.5% of all data. When removing those 33 observations from the training set, the R^2 increases by about 7% and the residual standard error is reduces by approximately 8%. But by removing the outliers on the training set, bias are introduced that lead to an overfitting of the data. When predicting the model, the training MSE is improved in contrast to the MSE of the testing set. Therefore, the outliers are not removed from the data. As a further analysis approach, one could examine every outlier observation closely to find out what variable causes the extreme value and decide one a case based level whether to impute a mean or median for this variable or not. This specific analysis exceeds the scope of this paper though.

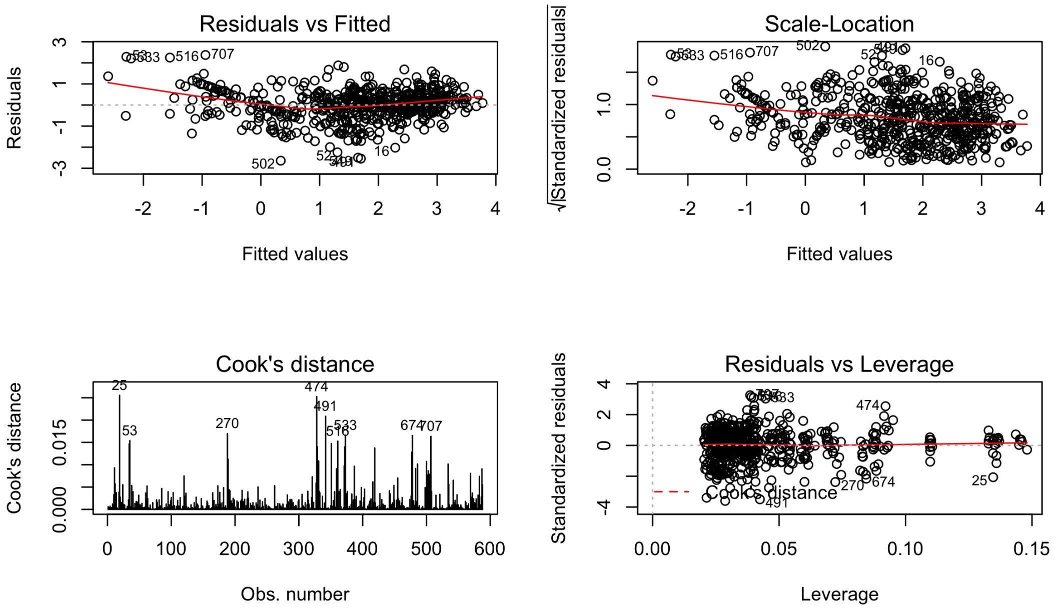

Figure 2: Residual diagnostics of the residuals and the fitted values, the scale-location plot for testing the homogeneity of variance, the Cook's distance to inspect potential outliers and the residuals vs. leverage plot to check on the influences of outliers

Finally yet importantly, the scale-location plot is shown in figure 2. It gives an indication about the homogeneity of square root standardized residual variances. The variance is homogeneous if the residuals spread equally around a horizontal line (Kassambara, 2018). The red line is not horizontal but rather descends from about 1.15 to 0.75 with the increase of the fitted values. This means that the left-sided residuals of those predictors are more spread out. As discussed before, the spread is primarily due to the outliers on the top left. Thus, the data are somewhat heteroscedastic.

To put it in a nutshell, the distribution of the residuals is not optimal but good enough for the application of a linear model.

3.1.2 Multicollinearity

An additional assumption of the Classical Linear Regression Model (CLRM) is that the explanatory variables are not (perfectly) correlated; thus, an explanatory variable must not be a linear function of another explanatory variable – X_1 must not be perfectly described by a linear function of X_2 (Auer and Rottmann, 2011). If multiple predictors correlate with each other, it is called multicollinearity (Kuhn and Johnson, 2013). Consequently, large variances and covariances of the estimators, wide confidence intervals, large standard errors, that are not robust to small changes, arise. These results in e.g. statistically insignificant coefficients which makes an accurate prediction/estimation difficult (Ghosh, 2017).

Multicollinearity leads to overfitting. In order to handle multicollinearity different techniques and methods are applied. At first, indicators for multicollinearity are checked based on the linear regression analysis and secondly, the variance inflation factor is conducted. Additionally, two penalized models are presented and applied (see 3.2).

For the underlying analysis, five different data frames are merged that contribute all explanatory factors for suicidal actions but also partially similar explanatory values. Hence, the data is very likely to have multicollinearity issues.

One way of getting an indication if multicollinearity exists, is to have a look at the correlation matrix of the variables (see figure 3). Correlation can be classified as presented in table 4.

High correlation	0.5 - 1.0
Moderate correlation	0.3 - 0.5
Low correlation	0.1 - 0.3

Table 4: Classification of correlation

Correlation can also be negative; the classification is analogous to the positive classification presented in table 4.

Many of the variables show a high correlation which can be observed in figure 3. As shown in the correlation matrix e.g. the Human Development Index (*HDI 2014*) is highly correlated (0.8) with the health-life-expectancy (*Health-Life-Expectancy*) which can be explained by the Human Development Index measuring the health as one of the three key components of the HDI (United Nations, 2019). It is an indication that multicollinearity exists but since the correlation matrix only shows the correlation between two explanatory variables and no multicollinear relationships between groups of variables, further examination is necessary.

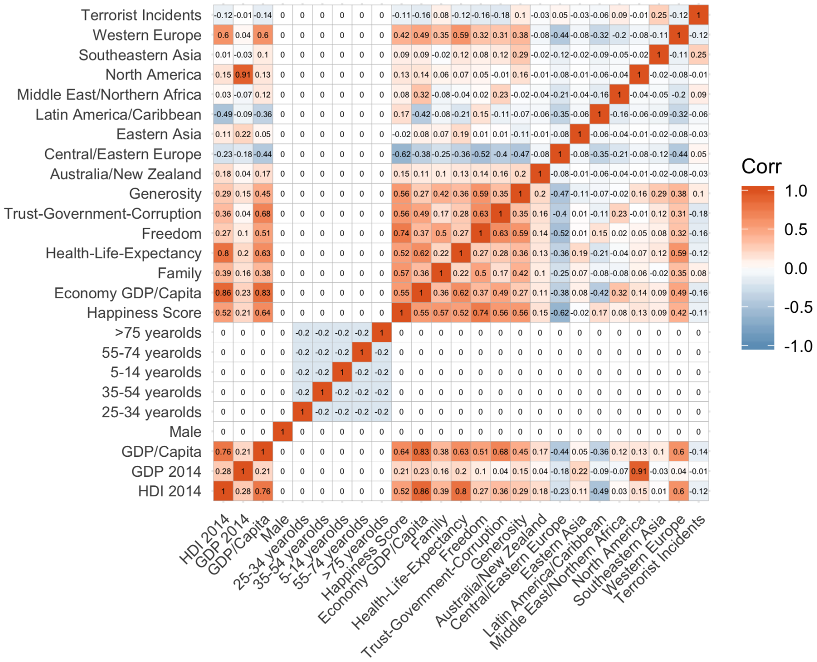

Figure 3: Correlation Matrix compiled of the independent variables

Another method for the detection of multicollinearity is called variance inflation factor (VIF). The VIF measures how strongly every single explanatory variable is influenced by multicollinearity by indicating how much the estimated variance of the i^{th} coefficient increases due to its linear dependence on other explanatory variables. The VIF only quantifies (multi-)collinearity but it does not tell which explanatory variable causes the collinearity. It is computed for each explanatory variable (Obrien, 2007). The lowest possible value of VIF is 1; the highest value is not defined (Allison, 2012). No fixed threshold whether a VIF value is (too) large or not has been determined. Due to Craney and Surles (2002) a legitimate cutoff value of the VIF_i is between 5-10.

The VIF score shown in table 7 in combination with the values of the correlation matrix shown in figure 3 indicates which variable(s) causes multicollinearity.

For the analysis, I consider a VIF_i of 10 as an indicator for severe multicollinearity. Table 7 shows that the variables *HDI 2014*, *GDP 2014*, *Economy GDP/Capita* and the regional dummy variables *Central/Eastern Europe*, *Latin America/Caribbean*, *North America* as well as *Western Europe* are above the threshold.

As shown in figure 3, *HDI 2014*, *GDP/Capita*, *Happiness Score*, *Health-Life-Expectancy* and *Freedom* have each at least five pairwise correlations that are considered as highly correlated (> 0.5). Excluding the variables with a high VIF score and many pairwise correlations, such as *HDI 2014*, could one approach of reducing multicollinearity but could also lead to misleading results. Another method that is more efficient, are penalized models.

3.2 Penalized Regression

Penalized regression is a regularization method that introduces a term to the regression that penalizes the model for having too many predictor variables or inflated parameter estimates that occur in case of an overfitted model or due to multicollinearity issues. The consequence of applying a regularization method is either a shrinkage of coefficient values towards zero (Ridge) or even a shrinkage to zero and thereby a selection of coefficients (Lasso) by minimizing the following equation:

$$min \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda\left(\alpha \sum_{j=1}^{p}|\hat{\beta}_j|+(1-\alpha)\sum_{j=1}^{p}\hat{\beta}_j^2\right) \tag{8}$$

with $\alpha = 0$ for Ridge regression and $\alpha = 1$ for Lasso. The amount of shrinkage is determined by the tuning parameter λ. An increasing λ shrinks the coefficients (James et al., 2013). Before conducting the analysis, the data need to be prepared. The data are split into a training and testing set as described in 3.1 and a seed is set for reproducibility. Furthermore, cross-validation with 10-folds and five repetitions is used to find the best λ and to prevent overfitting.

3.2.1 Ridge Regression

To find an optimal λ, the sequence is set between 0.0001 to 0.1 with a length of 20. After cross-validation, the optimal λ is returned to be 0.0842. Setting the tuning parameter to this value shrinks the coefficient as much as possible without increasing the RMSE.

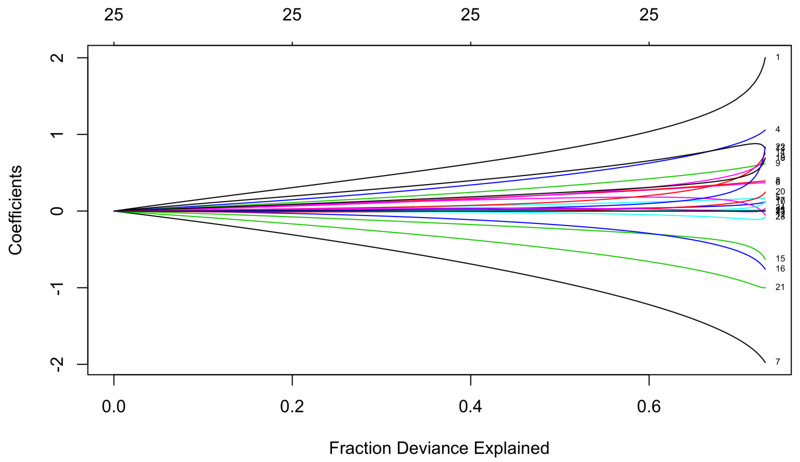

Figure 4: Relationship between the size of the coefficient and the explained fraction deviance

The Fraction Deviance Explained states how much variability of the model can be explained by the magnitude of the coefficients. A jump or inflation of the coefficients indicates overfitting. In the given figure 4, no such jump can be observed. Therefore, overfitting is very unlikely.

Figure 4 shows that all 25 predictor variables are kept during the shrinkage, none are removed. Every line represents one variable.

Other graphics regarding the relationship between the coefficients and log λ, the variables importance and relationship between RMSE and the regularization parameter λ are attached to the appendix.

3.2.2 Lasso

The advantages of the Least Absolute Shrinkage and Selection Operator (Lasso) is that it does not only shrink the coefficients but also select features. If there is a group of highly correlated data which cause multicollinearity, Lasso tends to select the feature from the group that is most strongly correlated with the response variables and ignores the other variables (Tibshirani, 1994).

The sequence for finding the optimal λ is set between 0.0001 to 0.01 with the length of 20. The resulting optimal λ is 0.00114. A small value for λ implies that the used Lasso is very similar to the OLS model. The effect of the regularization term increases with an increasing λ.

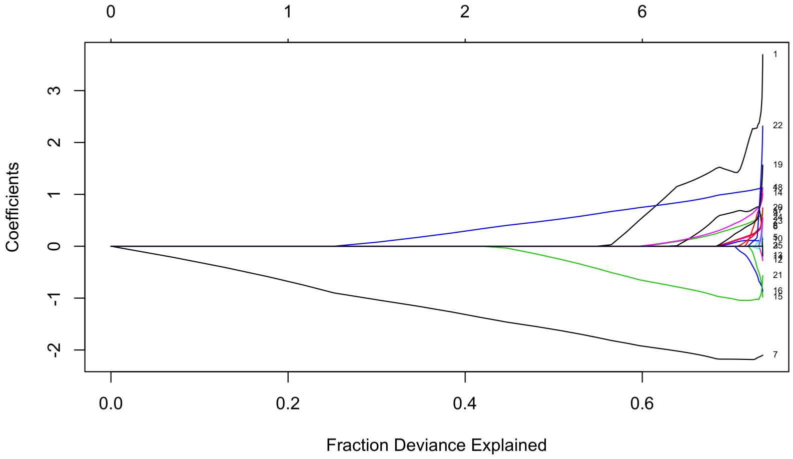

Figure 5: Relationship between the size of the coefficient and the explained fraction deviance

As observed in figure 5, about 40% of the variance of the model can be explained by its coefficients; with only six variables about 60% of the variability is explainable.

Variables 1 (*HDI 2014*), 4 (*Male*), 7 (*5-14 yearolds*) and 21 (*Middle East/Northern Africa*) are high performing variables for Lasso as well as Ridge regression because they are "shrinked" at last as shown in figure 6 and 9 – attached to the appendix.

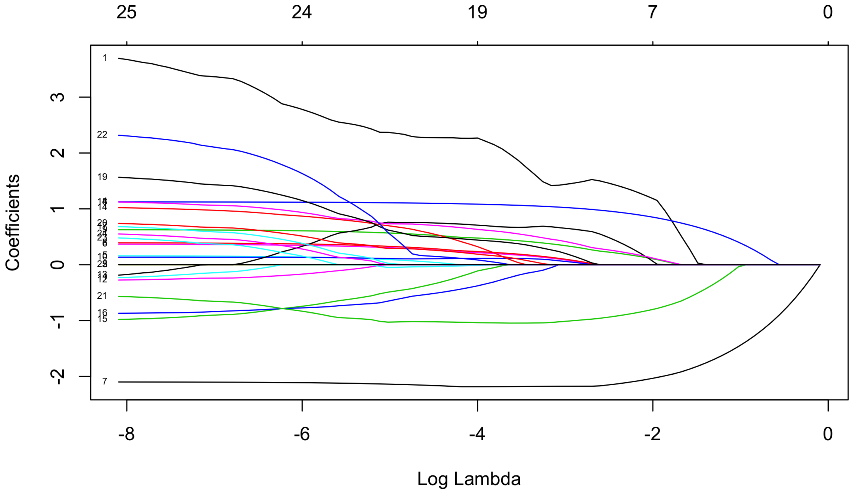

Figure 6: Relationship between coefficients and the value of log lambda

3.3 Evaluation of Models

Comparing the results of the linear regression, Ridge regression and Lasso by their average RMSE and their average R^2 gives the following output:

Model	RMSE	R^2
Linear	0.7564	0.7218
Lasso	0.7560	0.7219
Ridge	0.7584	0.7217

Table 5: Comparison of regression model by the models mean of RMSE and the mean of R^2

Lasso has the lowest average RMSE and the highest average R^2. Therefore, it is the best model. Unfortunately, all results are very close. Figure 7 shows the comparison of the performance of Lasso compared to the performance of the linear model. The points represent the different variables. All are very close to the line. Points above the line indicate a higher RMSE for the linear model; thus, the Lasso model performs better because the RMSE is lower.

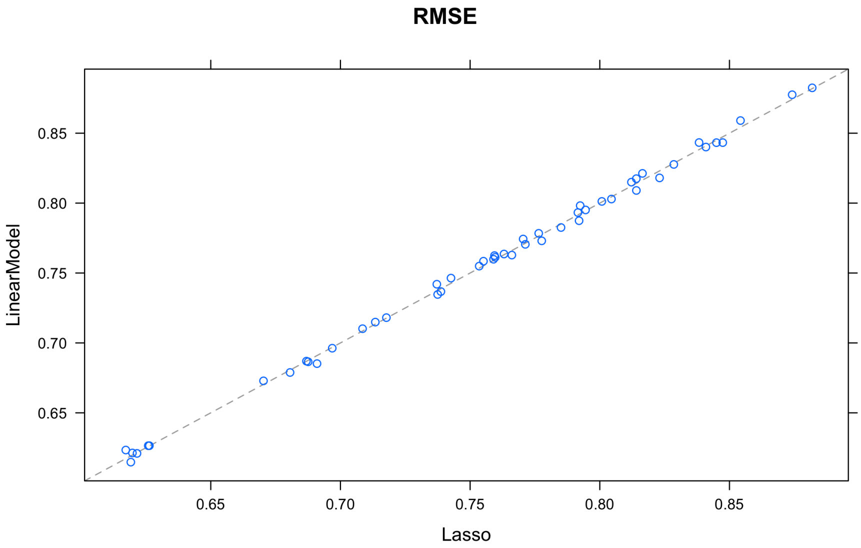

Figure 7: Impact of the RMSE on the performance of the linear model compared to Lasso

The variables importance differs slightly between Lasso and Ridge. The top six variables for each method are:

	Lasso	Ridge
1	HDI 2014	HDI 2014
2	5-14 yearolds	5-14 yearolds
3	North America	Male
4	Eastern Asia	Middle East/Northern Africa
5	Male	North America
6	Central/Eastern European	Health-Life-Expectancy

Table 6: Comparison of the variable importance

The full overview regarding the Ridge regression variable importance is presented in figure 10 and for Lasso in figure 13 – attached to the appendix.

When computing the VIF score, after conducting a linear regression analysis with the six most influencing factors resulting from Lasso, there is no multicollinearity anymore since all factor are below the value of 1.1. Nevertheless, the R^2 is still 0.6351. Even though the variable *North America* is very important due to Lasso, it is the only variable that is not statistically significant.

4 Conclusions

According to Lasso, the most important influencing factors are the HDI 2014 and socio-demographic variables such as age, gender and the place of residence. All of those variables are also highly significant due to the regression model with the logarithmic response variable. Interestingly, the increase of the Human Development Index leads to an increase of the number of suicides which could be explained with e.g. the constantly rising performance pressure of the western culture. Moreover, being a male has also a significantly positive effect on the number of suicides. Especially socio-demographic factors are difficult to influence in terms of prevention strategies. In contrast, the HDI can be influenced on a long term basis.

Surprisingly, happiness as well as the family support have no mentioning impact on suicidal actions. Also, terrorist incidents have no big influence on the number of suicides.

The model has to be treated with caution since the variance of the residuals shows a slight heteroscedasticity which influences the efficiency and the unbiasedness of the OLS estimator (cp. Kuhn and Johnson (2013)). As a next research approach, non linear models should be applied.

The approach of removing the variables with high VIF values and many pairwise correlation at the same time would have produced bad results since the variable with the highest VIF score and the most pairwise correlations is *HDI 2014* which is the most important variable according to Ridge regression and Lasso results. Therefore, this technique failed.

The multicollinearity problem of the data is solved via the shrinkage and selection method named Lasso. Lasso decreased the complexity from 25 to 6 variables without a major impact on the values of R^2 and the residual standard error for the model.

To put it in a nutshell, highly influential factors regarding suicidal action are chiefly socio-demographically based with the exception of the HDI. Penalizing methods such as Lasso are very efficient techniques to reduce the models complexity and to solve multicollinearity issues.

References

Allison, P. (2012): "When can you safely ignore multicollinearity," *Statistical Horizons*, 5.

Auer, B. and H. Rottmann (2011): "Statistik und Ökonometrie für Wirtschaftswissenschaftler–Eine anwendungsorientierte Einführung, 2," *Aufl., Wiesbaden.*

Bothe, A. F. (2019): "The Influence of Terrorist Hotbeds on the Global Transformation of Terrorist Tactics," B.S. thesis, Humboldt-Universität zu Berlin.

Cardona, R. J., Z. Rezaee, W. Rivera-Ortiz, and J. C. Vega-Vilca (2019): "Regulatory Enforcement of Accounting Ethics in Puerto Rico," *Journal of Business Ethics*, 1–14.

Cook, R. D. and S. Weisberg (1982): *Residuals and influence in regression*, New York: Chapman and Hall.

Craney, T. A. and J. G. Surles (2002): "Model-dependent variance inflation factor cutoff values," *Quality Engineering*, 14, 391–403.

Fox, J. (1997): *Applied regression analysis, linear models, and related methods.*, Sage Publications, Inc.

Gallup (2019): "How Does the Gallup World Poll Work?" *Gallup.*

Ghosh, B. (2017): "Multicollinearity in R," *datascience+.*

GTD Codebook (2017): "Global Terrorism Database. Codebook: Inclusion Criteria and Variables," `https://www.start.umd.edu/gtd/downloads/Codebook.pdf`, accessed: 2019-02-14.

James, G., D. Witten, T. Hastie, and R. Tibshirani (2013): *An introduction to statistical learning*, vol. 112, Springer.

KASSAMBARA, A. (2018): "Linear Regression Assumptions and Diagnostics in R: Essentials," *STHDA – Statistical tools for high-throughput data analysis.*

KUHN, M. AND K. JOHNSON (2013): *Applied predictive modeling*, vol. 26, Springer.

NATIONAL CONSORTIUM FOR THE STUDY OF TERRORISM AND RESPONSES TO TERRORISM (START) (2018): "Global Terrorism Database," *retrieved from https://www.start.umd.edu/gtd.*

OBRIEN, R. M. (2007): "A caution regarding rules of thumb for variance inflation factors," *Quality & quantity*, 41, 673–690.

RUSTY (2018): "Suicide Rates Overview 1985-2016," *Kaggle.*

STERN, J. AND M. MCBRIDE (2013): "Terrorism after the 2003 invasion of Iraq," *Group Brown University, Eisenhower Study Google Scholar.*

SZAMIL (2017): "Suicide in the Twenty-First Century," *Kaggle.*

TIBSHIRANI, R. (1994): "Regression Shrinkage and Selection Via the Lasso," *Journal of the Royal Statistical Society, Series B*, 58, 267–288.

UNITED NATIONS (2018): "Technical Notes - Calculating the human development indices," *United Nations Development Programme.*

——— (2019): "Human Development Reports," *United Nations Development Programme.*

UNITED NATIONS SUSTAINABLE DEVELOPMENT SOLUTION NETWORK (2015): "World Happiness Report 2015," *World Happiness Report.*

WORLD BANK (2019): "DataBank | World Development Indicators," *The World Bank.*

WORLD HEALTH ORGANIZATION (2018): "WHO Mortality Database," *WHO.*

Appendix

Variance Inflation Factor (VIF) Results

HDI 2014	18.264
GDP 2014	10.746
GDP/Capita	8.164
Male	1.011
25-34 yearolds	1.675
35-54 yearolds	1.711
5-14 yearolds	1.699
55-74 yearolds	1.693
>75 yearolds	1.651
Happiness Score	5.748
Economy GDP/Capita	14.122
Family	2.490
Health-Life-Expectancy	8.568
Freedom	3.884
Trust-Government-Corruption	3.967
Generosity	2.956
Australia/New Zealand	2.249
Central/Eastern Europe	12.285
Eastern Asia	3.697
Latin America/Caribbean	14.436
Middle East/Northern Africa	8.230
North America	11.064
Southeastern Asia	4.393
Western Europe	18.367
Terrorist Incidents	1.561

Table 7: Variance Inflation Factor

Ridge Regression Results

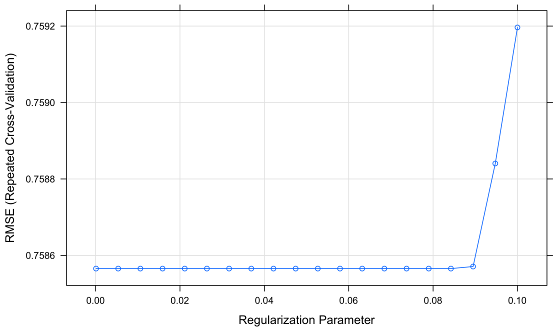

Figure 8: Relationship between RMSE and lambda

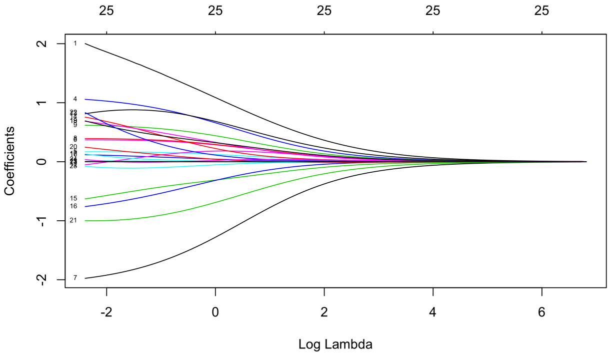

Figure 9: Relationship between coefficients and log lambda

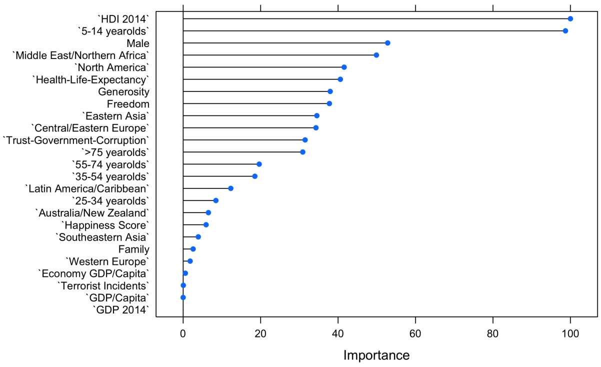

Figure 10: Variable importance resulting for ridge regression

Lasso Results

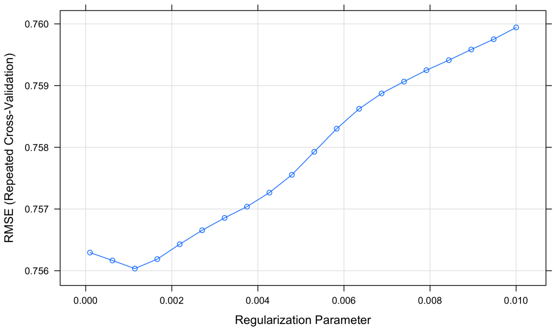

Figure 11: Relationship between RMSE and lambda

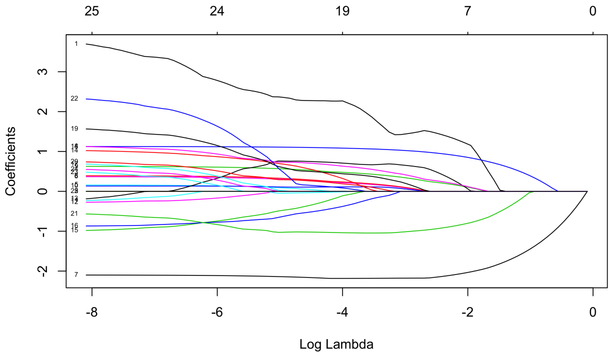

Figure 12: Relationship between coefficients and log lambda

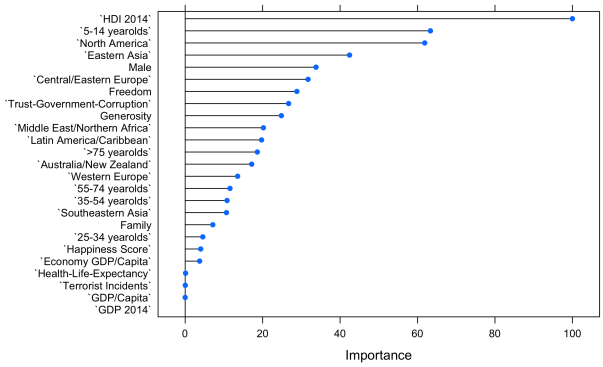

Figure 13: Variable importance resulting for lasso

Declaration of Authorship

I hereby confirm that I have authored this paper independently and without use of others than the indicated sources. All passages which are literally or in general matter taken out of publications or other sources are marked as such.

Berlin, March 29, 2019

Anna Franziska Bothe