Title: Building and evaluation of a linear regression model

Description of the dataset

This compiled dataset pulled from four other datasets linked by time and place, and was built to find signals correlated to increased suicide rates among different cohorts globally, across the socio-economic spectrum.

Number of instances: 27,820
Number of attributes: 11

Note: Several of the attributes may be correlated, thus it makes sense to apply some sort of feature selection.

	country	year	sex	avg_age	suicides	population	suicides100k	country-year	HDI for year	gdp_for_ye ($)
1	Albania	1987	male	19.5	21	312900	6.71	Albania1987	NA	2156624
2	Albania	1987	male	44.5	16	308000	5.19	Albania1987	NA	2156624
3	Albania	1987	female	19.5	14	289700	4.83	Albania1987	NA	2156624
4	Albania	1987	male	75.0	1	21800	4.59	Albania1987	NA	2156624
5	Albania	1987	male	29.5	9	274300	3.28	Albania1987	NA	2156624
6	Albania	1987	female	75.0	1	35600	2.81	Albania1987	NA	2156624
7	Albania	1987	female	44.5	6	278800	2.15	Albania1987	NA	2156624
8	Albania	1987	female	29.5	4	257200	1.56	Albania1987	NA	2156624
9	Albania	1987	male	64.5	1	137500	0.73	Albania1987	NA	2156624

In this project, we try to find a correlation between the number of suicides per 100 k of the population of many countries and factors like the age group, the net population and GDP of the country.

We have considered 3 predictor variables:

1. avg_age
2. population
3. gdp per capita

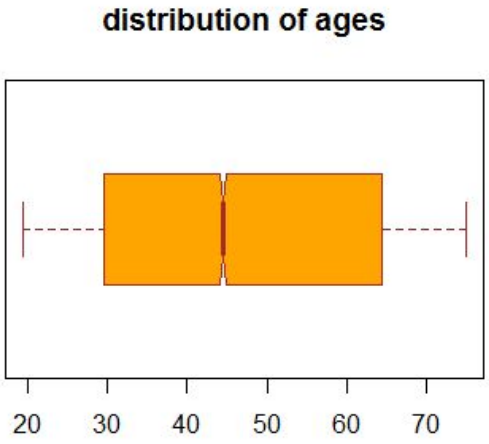
The response variable is: suicides100k

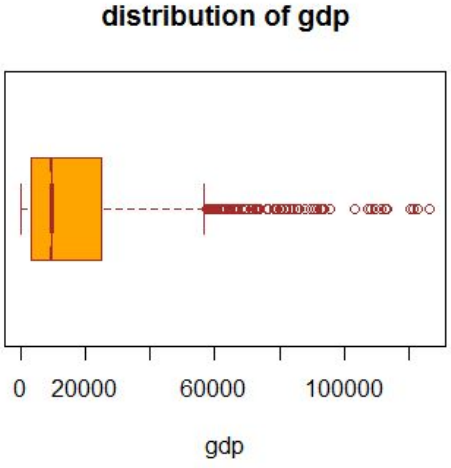
```
> summary(suicide_newer)
    avg_age         population            gdp           suicides100k
 Min.   :19.50   Min.   :     278   Min.   :   251   Min.   :  0.00
 1st Qu.:29.50   1st Qu.:   97498   1st Qu.:  3447   1st Qu.:  0.92
 Median :44.50   Median :  430150   Median :  9372   Median :  5.99
 Mean   :46.25   Mean   : 1844794   Mean   : 16866   Mean   : 12.82
 3rd Qu.:64.50   3rd Qu.: 1486143   3rd Qu.: 24874   3rd Qu.: 16.62
 Max.   :75.00   Max.   :43805214   Max.   :126352   Max.   :224.97
```

There are some missing values in our dataset. These missing values have been replaced while coding.

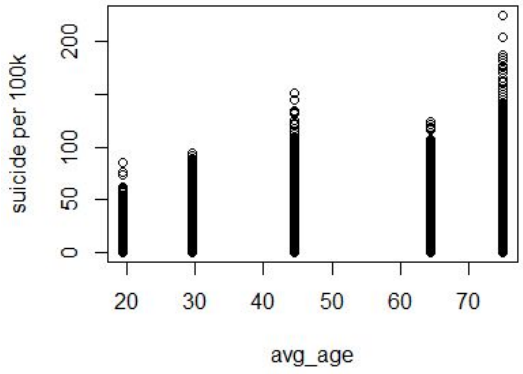
Data Analysis

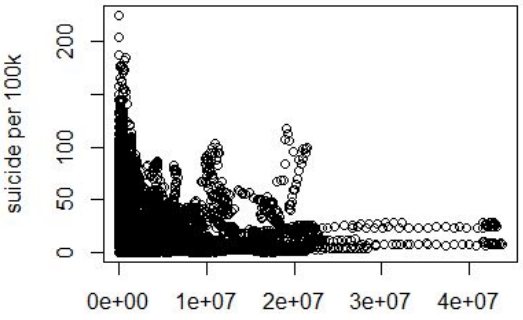
1. BOX PLOTS

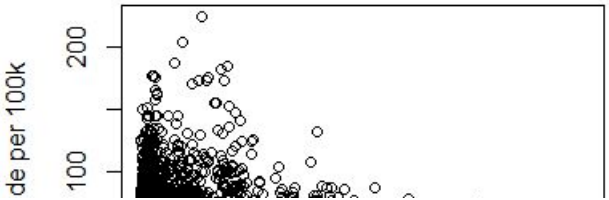




2. SCATTER PLOTS







From the plots we can draw the following inferences relevant to our analysis:

1. 75% of the people who commit suicide are 30 years or older.
2. The number of suicides per 100 k of the population decreases as the population increases.
3. An increase in the per capita GDP leads to a decrease in the no. of suicides per 100 k of the population.
4. The most number of suicides have been reported in the 70+ age group.
5. The least number of suicides have been reported among people less than 20 years of age.
6. A considerable number of suicides have also been reported in the age group of 40-50 years.

Building a regression model

We have split the data into train and test variables. 70% of the data has been used as the training data and the remaining 30% as the test data.
We have constructed a linear model between the three predictor variables and the response variable. The summary of the linear model gives us values for several parameters like intercept, the R-squared value, the standard error, different quartiles, etc.

The model above is achieved by using the lm() function in R and the output is called using the summary() function on the model.

Presentation and interpretation of model parameters

As mentioned before, we have considered average age, population and gdp per capita as the predictor variables. After building the regression model, we seek to know the dependencies of the number of suicides per 100 k of the population (i.e. our dependent variable) with respect to all of the three predictor variables. For that, we evaluate the coefficients of linear regression.

Using avg_age as the predictor variable

```
> summary(linmod1)

Call:
lm(formula = suicide_newer$suicides100k ~ suicide_newer$avg_age,
    suicide_newer = train)

Residuals:
    Min      1Q  Median      3Q     Max
-19.634 -11.771  -4.932   4.417 205.336

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           1.847074   0.290716   6.354 2.14e-10 ***
suicide_newer$avg_age 0.237158   0.005814  40.788  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 18.42 on 27818 degrees of freedom
Multiple R-squared:  0.05643,   Adjusted R-squared:  0.0564
F-statistic:  1664 on 1 and 27818 DF,  p-value: < 2.2e-16
```

Using gdp as the predictor variable

```
> summary(linmod2)

Call:
lm(formula = suicide_newer$suicides100k ~ suicide_newer$gdp,
    suicide_newer = train)

Residuals:
    Min      1Q  Median      3Q     Max
-13.012 -11.894  -6.827   3.802 212.152

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       1.279e+01  1.524e-01  83.888   <2e-16 ***
suicide_newer$gdp 1.792e-06  6.019e-06   0.298    0.766
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 18.96 on 27818 degrees of freedom
Multiple R-squared:  3.187e-06, Adjusted R-squared:  -3.276e-05
F-statistic: 0.08865 on 1 and 27818 DF,  p-value: 0.7659
```

Using population as the predictor variable

```
> summary(linmod3)

Call:
lm(formula = suicide_newer$suicides100k ~ suicide_newer$population,
    suicide_newer = train)

Residuals:
    Min      1Q  Median      3Q     Max
-13.284 -11.907  -6.854   3.783 212.228

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)             1.274e+01  1.257e-01 101.377   <2e-16 ***
suicide_newer$population 4.016e-08  2.906e-08   1.382    0.167
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 18.96 on 27818 degrees of freedom
Multiple R-squared:  6.864e-05, Adjusted R-squared:  3.27e-05
F-statistic:  1.91 on 1 and 27818 DF,  p-value: 0.167
```

1. Formula Call

The first item shown in the output is the formula R used to fit the data. It needs the predictors and the target/response variable, together with the data being used.

2. Residuals

The next item in the model output talks about the residuals. Residuals are essentially the difference between the actual observed response values and the response values that the model predicted. The Residuals section of the model output breaks it down into 5 summary points.

In this case, we can see that the distribution of the residuals do not appear to be strongly symmetrical. That means that the model predicts certain points that fall far away from the actual observed points.

3. Coefficients

The next section in the model output talks about the coefficients of the model. Theoretically, in simple linear regression, the coefficients are two unknown constants that represent the *intercept* and *slope* terms in the linear model.

Coefficient - Estimate

The coefficient Estimate contains two rows; the first one is the **intercept**. Let us take the predictor variable avg_age as an example. The intercept is essentially the expected value of the no. of suicides corresponding to the average of all the age groups in the dataset. The second row in the coefficients is the **slope**. The slope term in our model is saying that for a unit increase in the average age of an age group, the number of suicides per 100 k of the population goes up by approximately 0.24, likewise for all other predictor variables.

Coefficient - Standard Error

The coefficient Standard Error measures the average amount that the coefficient estimates vary from the actual average value of our response variable. The standard errors corresponding to the predictor variables average age, population and gdp per capita are 0.05, 6.19*10^-6 and 2.9*10^-8 respectively.

Coefficient - t value

The coefficient t-value is a measure of how many standard deviations our coefficient estimate is far away from 0. We want it to be far away from zero as this would indicate we could reject the null hypothesis - that is, we could declare a relationship between our parameters exists. In our regression model, all the t-values are significantly greater than zero, hence there exists a relationship between all our considered predictor variables and the dependent variable.

Coefficient - Pr(>t)

The Pr(>t) acronym found in the model output relates to the probability of observing any value equal to or larger than t. Typically, a p-value of 5% or less is a good cut-off point. In our model, the p-values are very close to zero.

Three stars (or asterisks) represent a highly significant p-value. Consequently, a small p-value for the intercept and the slope indicates that we can reject the null hypothesis.

4. Residual Standard Error

Residual Standard Error is a measure of the quality of a linear regression fit. Theoretically, every linear model is assumed to contain an error term E. The Residual Standard Error was calculated with 27816 degrees of freedom. Simplistically, degrees of freedom are the number of data points that went into the estimation of the parameters used after taking into account these parameters (restriction).

5. Multiple R-squared

The R-squared statistic provides a measure of how well the model is fitting the actual data. It takes the form of a proportion of variance. In our example, the value we get is 0.057. Or roughly 5.7% of the variance found in the response variable can be explained by the predictor variable.

6. F-Statistic

F-statistic is a good indicator of whether there is a relationship between our predictor and response variables. The further the F-statistic is from 1 the better it is. In our model, the F-statistic is 563, which is greater than one.

Predicted data

1	6.471650
2	12.400594
3	6.471650
4	19.633906
5	8.843228
6	19.633906
7	12.400594
8	8.843228
9	17.143749
10	12.400594
11	17.143749
12	12.400594
13	19.633906
14	6.471650
15	19.633906
16	12.400594

Using avg_age

1	12.78730
2	12.78730
3	12.78730
4	12.78730
5	12.78730
6	12.78730
7	12.78730
8	12.78730
9	12.78730
10	12.78730
11	12.78730
12	12.78730
13	12.78725
14	12.78725
15	12.78725
16	12.78725

Using gdp

1	12.75458
2	12.75438
3	12.75365
4	12.74289
5	12.75303
6	12.74344
7	12.75321
8	12.75234
9	12.74753
10	12.75450
11	12.74782
12	12.75559
13	12.74347
14	12.75483
15	12.74291

Using population

Evaluation of Root Mean Square Error (RMSE)

The RMSE can be evaluated by using this formula.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (Predicted_i - Actual_i)^2}{N}}$$

The RMSE obtained while considering avg_age as the predictor variable:

```
> print(rmse1)
[1] 35.5837
```

The RMSE obtained while considering gdp as the predictor variable:

```
> print(rmse2)
[1] 34.65533
```

The RMSE obtained while considering population as the predictor variable:

```
> print(rmse3)
[1] 34.65033
```

Code

```
#for reading data
library(readr)
suicide_data <- read_csv("r-internship/suicidedata.csv")
View(suicidedata)
#for dropping columns that are not needed in our analysis
suicide_newer<- suicidedata[,c(4,6,11,7)]
View(suicide_newer)
#for finding the number of missing values
summary(suicide_newer)
#for cleaning the data by replacing the missing values by the median
of all the values in that column
suicide_newer$avg_age[is.na(suicide_newer$avg_age)]<-median(na.omit((
suicide_newer$avg_age)))



#for drawing the box plots
```

```
boxplot(suicide_newer$avg_age,main = "distribution of ages", xlab =
"age group", col = "orange", border = "brown", horizontal = TRUE,
notch = TRUE)
boxplot(suicide_newer$gdp,main = "distribution of gdp", xlab = "gdp",
col = "orange",border = "brown", horizontal = TRUE, notch = TRUE)
boxplot(suicide_newer$population,main = "distribution of population",
xlab = "population", col = "orange", border = "brown", horizontal =
TRUE, notch = TRUE)
#for drawing the scatter plots
plot(suicide_newer$suicides100k~suicide_newer$avg_age,
xlab="avg_age", ylab="suicide per 100k")
plot(suicide_newer$suicides100k~suicide_newer$population,
xlab="population", ylab="suicide per 100k")
plot(suicide_newer$suicides100k~suicide_newer$gdp, xlab="gdp",
ylab="suicide per 100k")
#for splitting the data for training and testing
n=nrow(suicide_newer)
trainIndex=sample(1:n,size = round(0.7*n),replace = FALSE)
train=suicide_newer[trainIndex, ]
test=suicide_newer[-trainIndex, ]

#making regression model

#Using avg_age as predictor variable
linmod1=lm(suicide_newer$suicides100k~suicide_newer$avg_age,suicide_n
ewer=train)
summary(linmod1)
#Using gdp as predictor variable
linmod2=lm(suicide_newer$suicides100k~suicide_newer$gdp,suicide_newer
=train)
summary(linmod2)
#Using population as predictor variable
linmod3=lm(suicide_newer$suicides100k~suicide_newer$population,suicid
e_newer=train)
summary(linmod3)




#for testing the linear regression model
```

```
#Using avg_age
new_data_test1=data.frame(suicideper100k=test$suicides100k,avg_age=te
st$avg_age)
pred1=predict(linmod1,new_data_test1)
View(pred1)
#Using gdp
new_data_test2=data.frame(suicideper100k=test$suicides100k,gdpcapita=
test$gdp)
pred2=predict(linmod2,new_data_test2)
View(pred2)
#Using population
new_data_test3=data.frame(suicideper100k=test$suicides100k,population
=test$population)
pred3=predict(linmod3,new_data_test3)
View(pred3)

#for finding the root mean squared error

#Using avg_age
rmse1=sqrt(sum((pred1-test$suicides100k)^2)/length(test$avg_age))
print(rmse1)
#Using gdp
rmse2=sqrt(sum((pred2-test$suicides100k)^2)/length(test$gdp))
print(rmse2)
#Using population
rmse3=sqrt(sum((pred3-test$suicides100k)^2)/length(test$population))
print(rmse3)
```

Conclusion

A regression model for predicting the number of suicides per 100 k of the population based on the age groups, GDP and net population has been successfully designed. Out of all these predictor variables, the age group seems to have the maximum effect on the response variable, compared to the GDP and net population of the country.