Regression Analysis of Terrorist Attacks

By: Pranav Vinod

University of Massachusetts Dartmouth

MTH 522: Advanced Mathematical Statistics

Supervisor: Professor Donghui Yan

03/31/2022

Contents

Abstract

The aim of this project is to analyze terrorist attacks from 1970 till 2015 throughout the world and gain insights from the data. Initially, we shall explore the dataset to understand the structure of data by employing various analysis techniques. The bulk of this project will be in implementing linear regression on 3 subsets of the dataset based on the number of casualties and evaluating each model. Finally, we will summarize the results of each model.

Description of Dataset

This dataset is the Global Terrorism Database, an entry level database containing more than 20,000 records of terrorist attacks that took place around the world from 1970 to 2015. It is maintained by the National Consortium for the Study of Terrorism and Responses to Terrorism at the University of Maryland. The dataset can be found here.

The dataset is a large set containing 135 attributes and 181691 observations. Attributes include iyear, imonth and iday which are the year, month and day of the specific attack respectively. The set also has information on number of casualties in each attack (nkills), location (country, region, city, vicinity), the type of attack (attacktype1), the types of weapons used in attack (weapontype1), the target of the attack (targtype1, natlty1), the perpetrators of the attacks (gname, nperps) among others.

Data Exploration

To begin with we find a summary of the data. We find that there are 77 numeric variables and 58 categorical variables in the dataset.

```
> # Data exploration
> length(select_if(data,is.numeric))
[1] 77
> # Number of categorical variables
> length(select_if(data, is.character))
[1] 58
```

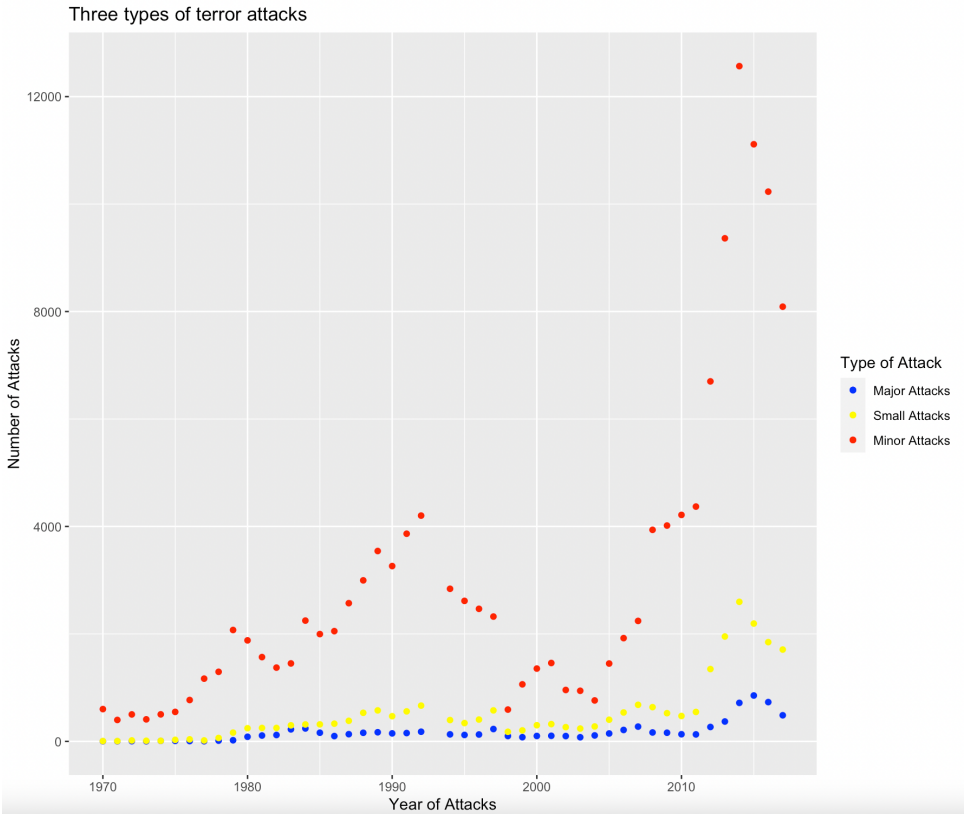
To explore the data further, we will try to find the number/percentage of missing values in the data. We find that 27% of the values in the dataset are missing.

Now since we are trying to implement regression on years vs the number of attacks, we will split the dataset and keep only the needed rows.

We have split the dataset into 3 parts based on the number of casualties in each instance.
If there were more than 10 casualties in an attack, we have termed those as major attacks.If there were between 3-10 casualties, we have termed them as small attacks and if there were fewer than 3 attacks we have termed them as minor attacks.

Now, we have summarized the data to find out the number of attacks in each year in the dataset.

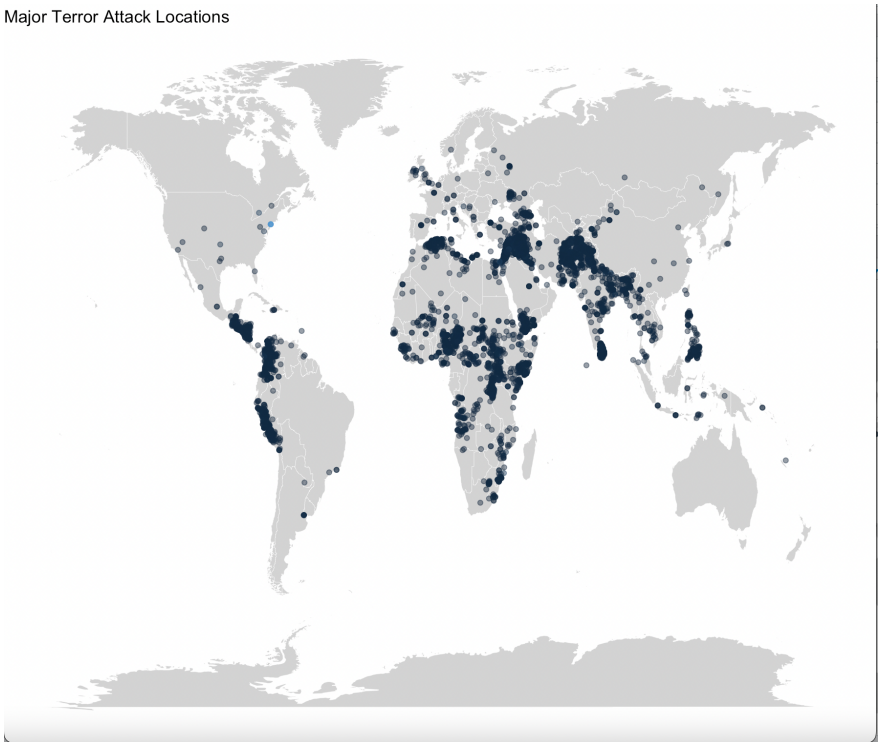
Once the split is complete we shall draw a scatter plot of the number of attacks per year for all the three categories of attacks that we have specified.



Based on this plot, we can see an interesting trend. While all three kinds of attacks experienced a rise in frequency around 2010, the number of minor attacks shows unusual behavior. There was a rise in the number of minor attacks around 1990 and 2000. Such a rise was not accompanied by a rise in major or small attacks.

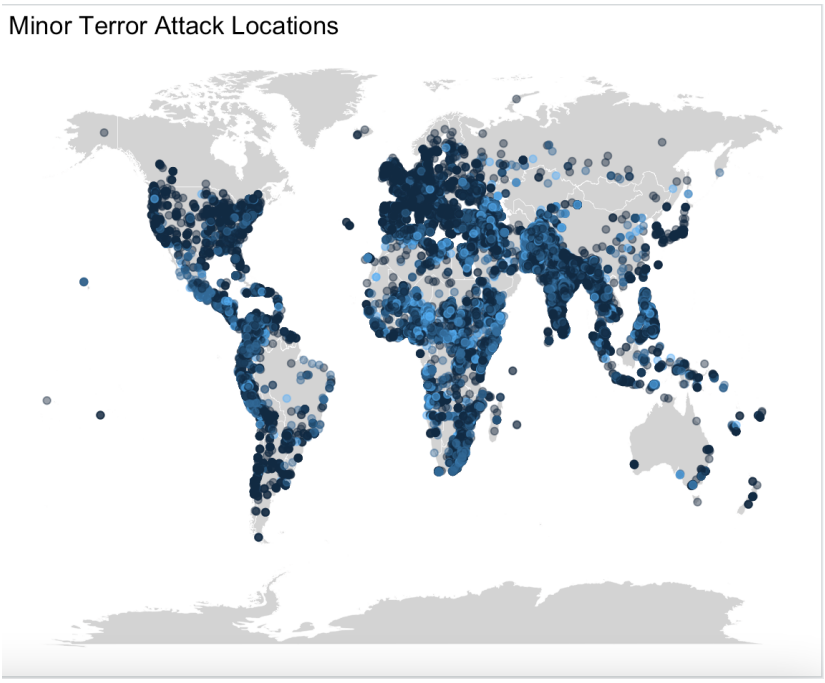
Additionally, to get an idea of where these attacks have been taking place for major and minor attacks respectively, we have plotted a geospatial plot:

For Major Attacks



We can see that there are clusters of regions where major terrorist attacks are concentrated. Most of them occur in East Asia and some parts of Africa.

For Minor Attacks



Unlike major attacks, minor attacks are spread throughout the world. From Europe, the US to most of southeast Asia and Africa, minor terror attacks take place everywhere.

Regression Analysis

For each of the three types of attacks, regression analysis was carried out on years vs number of attacks. The results were found to be as follows:

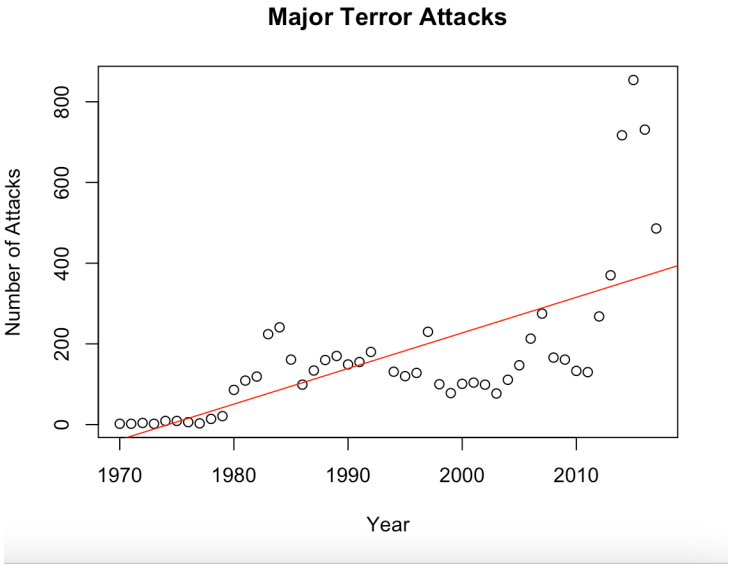
For Major attacks

```
Call:
lm(formula = count$Count ~ count$iyear, data = count)

Residuals:
    Min      1Q  Median      3Q     Max
-194.34  -88.40    2.39   36.96  494.36

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -17424.454   2893.416  -6.022 2.90e-07 ***
count$iyear      8.826      1.451   6.081 2.37e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 139.3 on 45 degrees of freedom
Multiple R-squared:  0.4511,    Adjusted R-squared:  0.4389
F-statistic: 36.98 on 1 and 45 DF,  p-value: 2.367e-07
```

There is a positive correlation between the number of major attacks and years. The final equation is as follows:

$$Attacks = -17424.454 + 8.826 * years$$

We can also see that the model has a R-squared value of 0.455 which indicates that the model only captures 45% of the total variance present in the data and it has a low score on the F-statistic as well.



This scatter plot shows us the regression line along with the count of attacks vs years.

To check the validity of the model, we will now run some model evaluation tests.

a) Normality Test

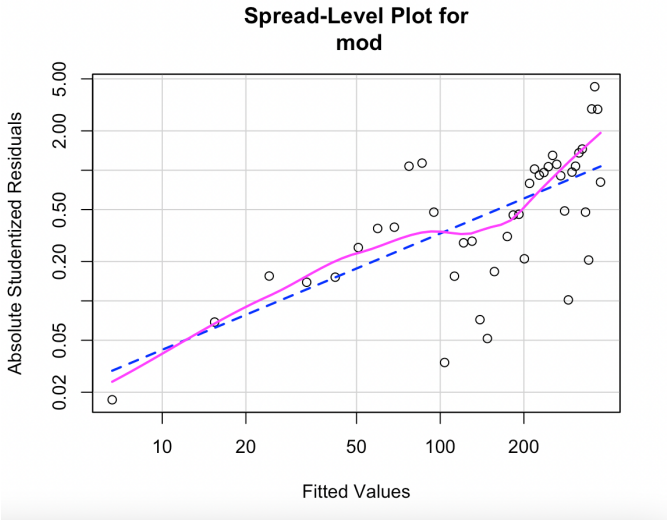
To check the normality assumption of our model, we will conduct the KS test along with the Shapiro Wilk test and look for the p value. A small p value would indicate that we can reject our hypothesis while a large one would help us not reject our hypothesis that the residuals follow a normal distribution.

```
        Two-sample Kolmogorov-Smirnov test

data:  residuals and rnorm(100, 0, 1)
D = 0.49064, p-value = 1.508e-07
alternative hypothesis: two-sided

> # And say that residuals do not follow a normal distr
> shapiro.test(residuals)               # Small value o
ult

        Shapiro-Wilk normality test

data:  residuals
W = 0.84855, p-value = 2.366e-05
```

A small value on both the tests indicate that our assumption of normality does not hold true for this model.

b) Constant Variance Test

Similarly we will test our assumption of constant variance of the residuals using the Cook Weinberg test. We are looking for a large enough value of p that would help us not reject the null hypothesis.




```
> ncvTest(mod)                              # Small
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 29.95516, Df = 1, p = 4.4215e-08
```

Such a small value indicates that the assumption is violated. Now we will perform power transformations on our outcome y to restore the constant variance.

The suggested power transformation is 0.1104

After performing the power transformation, we implement regression again and find the following results:

```
lm(formula = z ~ x)

Residuals:
     Min       1Q   Median       3Q      Max
-0.27277 -0.11175 -0.02566  0.13655  0.32405

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -27.796916   3.304588  -8.412 8.74e-11 ***
x             0.014769   0.001658   8.910 1.70e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1591 on 45 degrees of freedom
Multiple R-squared:  0.6382,    Adjusted R-squared:  0.6302
F-statistic: 79.38 on 1 and 45 DF,  p-value: 1.7e-11
```

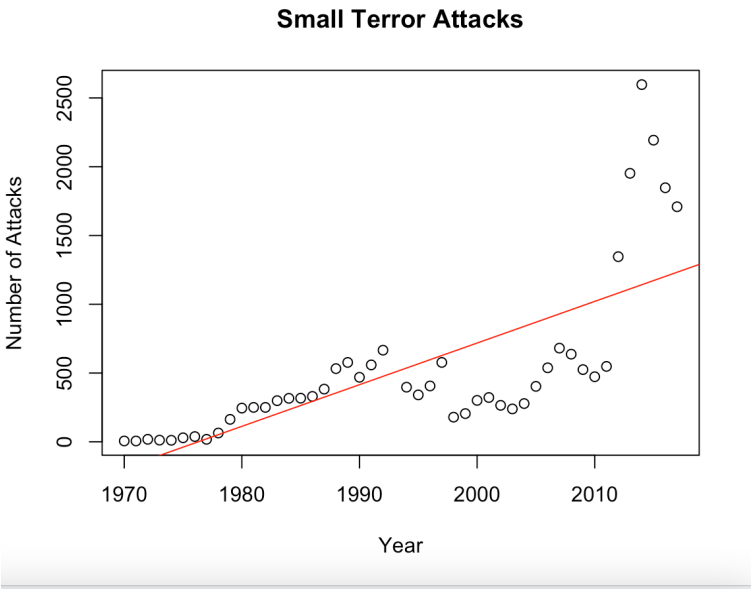
We can see the rise in R-squared statistics, now it captures 63.8% of the total variance.

Statistic	**Before transformation**	**After transformation**
R-squared	0.451	0.638
F-statistic	36.98	79.38
P value(Cook Weinberg test)	4.4e-8	0.01
P value(KS test)	1.7e-7	0.84

We can see that after performing power transformation, our model performs better on evaluation metrics.

For Small Attacks

Similar analysis has been carried out for small attacks




```
Call:
lm(formula = count3$Count ~ count3$iyear, data = count3)

Residuals:
    Min      1Q  Median      3Q     Max
-570.02 -342.66   53.68  140.17 1454.84

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -59851.830   8753.377  -6.838 1.78e-08 ***
count3$iyear     30.285      4.391   6.897 1.45e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 421.4 on 45 degrees of freedom
Multiple R-squared:  0.5139,    Adjusted R-squared:  0.5031
F-statistic: 47.57 on 1 and 45 DF,  p-value: 1.447e-08
```

$$Attacks = -59851.8 + 30.285 * years$$

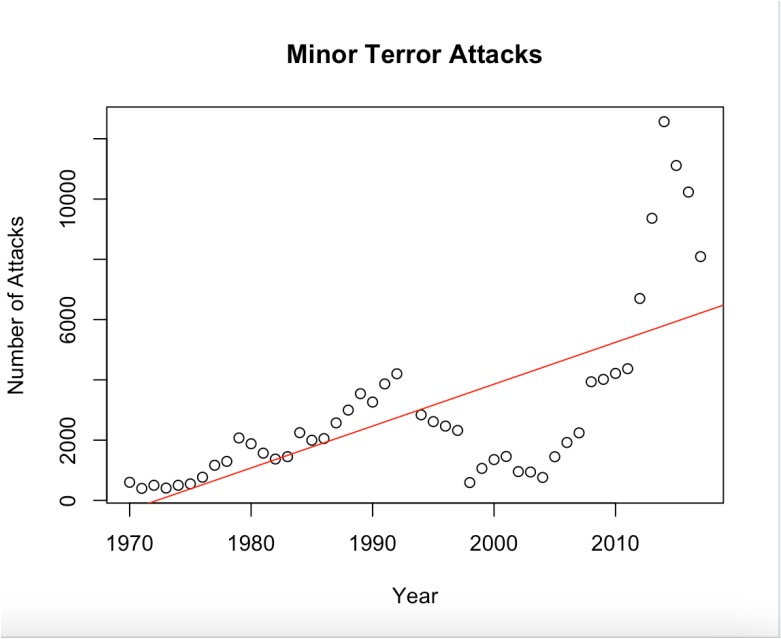
The model captures 51.4% of the total variance in the data and has a RSE of 421.4.

Further evaluation tests were conducted and the results have been reported in the following table.

Statistic	Before transformation	After transformation
R-squared	0.514	0.715
F-statistic	47.57	113.2
P value(Cook Weinberg test)	6.72e-8	0.787
P value(KS test)	3.87e-11	0.508

We can see that after performing power transformations, the model fit the data much better and performed better on evaluation metrics as well.

For Minor attacks



```
Call:
lm(formula = count2$Count ~ count2$iyear, data = count2)

Residuals:
    Min      1Q  Median      3Q     Max
-3650.0 -1060.2   248.1   805.1  6765.6

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -274009.3    45249.2  -6.056 2.58e-07 ***
count2$iyear     138.9       22.7   6.121 2.06e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2178 on 45 degrees of freedom
Multiple R-squared:  0.4543,    Adjusted R-squared:  0.4422
F-statistic: 37.47 on 1 and 45 DF,  p-value: 2.064e-07
```

$$Attacks = -274009.3 + 138.9 * years$$

The model captures 45.4% of the total variance in the data and has a RSE of 2178.

Further evaluation tests were conducted and the results have been reported in the following table.

Statistic	**Before transformation**	**After transformation**
R-squared	0.454	0.479
F-statistic	37.47	41.36
P value(Cook Weinberg test)	7.89e-7	0.89
P value(KS test)	3.87e-11	0.15

Again we can see that the model performs better after the power transformation has been completed.

Though the increase in performance is small as compared to the model for small attacks.

Result

The results of the models before power transformation have been presented below.

Statistic	Major Attacks	Small Attacks	Minor Attacks
RSE	139.3	421.4	2178
R^2	0.451	0.514	0.454
Adj. R^2	0.439	0.503	0.442
F-statistic	36.98	47.57	37.47

From the table above, we can see that regression performs best to find the relation between number of small attacks and years. It has the highest r-squared value of 0.514 and also the highest value for F-statistic with 47.57.

References

1. An Introduction to Statistical Learning - James, Witten, Hastie, Tibshirani