Project 1 – Global Terrorism Data

By – Ishan Bhargava (02017165)

# Abstract

The objective of this undertaking is to investigate acts of terrorism and gather useful insights from the available information. Initially, we will scrutinize the data using various analytical techniques to comprehend its structure. The primary emphasis of this project will be to utilize linear regression on three distinct segments of the data, grouped according to the number of casualties, and analyze the effectiveness of each model. Ultimately, we will summarize the discoveries from each model.

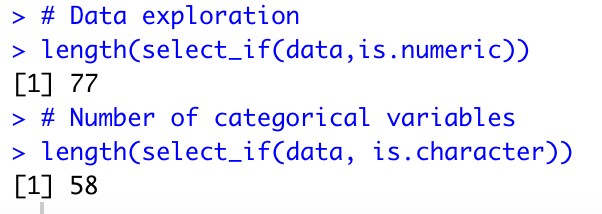
# Description of Dataset

The Global Terrorism Database (GTD) is a publicly available collection of data on acts of terrorism worldwide that took place between 1970 and 2017. It comprises detailed and organized information on both local and global terrorist incidents, and currently covers over 180,000 attacks. The database is managed by a team of researchers from the National Consortium for the Study of Terrorism and Responses to Terrorism (START), who are based at the University of Maryland. The dataset can be found [here](https://www.kaggle.com/datasets/START-UMD/gtd?resource=download).

The dataset is an extensive collection that consists of 181,691 observations and 135 attributes. Some of the attributes include iyear, imonth, and iday, which represent the year, month, and day of the particular attack. The dataset also contains information about the number of casualties (nkills) in each attack, the location of the attack (country, region, city, vicinity), the type of attack (attacktype1), the weapons used in the attack (weapontype1), the target of the attack (targtype1, natlty1), and the individuals or groups responsible for the attacks (gname, nperps), among others.

# Data Exploration

* First, we find the summary of the data. There are 77 numeric variables and 58 categorical variables in the dataset.



* Adding a Scatter plot of year vs number of attacks.

Chart, scatter chart

Description automatically generated

* Creating a missingness map to explore the data further and find the number/percentage of missing values in the data. We see that 27% of the values in the dataset are missing.

A picture containing graphical user interface

Description automatically generated

We have split the dataset into 3 parts based on the number of casualties in each instance.

1. If there were more than 10 casualties in an attack, they are classified as major attacks.
2. If there were between 3-10 casualties, they are classified as small attacks.
3. If there were fewer than 3 casualties, they are classified as minor attacks.

* After the partitioning process is finished, we generated a scatter plot that displays the frequency of attacks per year for each of the three attack categories we designated.

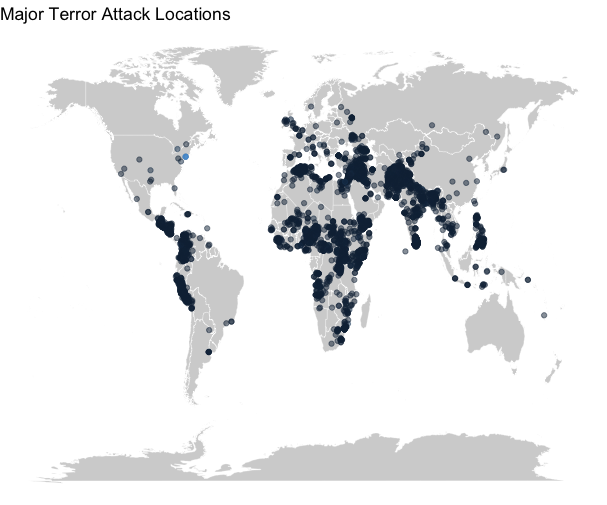
Chart

Description automatically generated

Using this plot, we can observe an intriguing pattern. Despite all three types of attacks displaying an increase in frequency around 2010, the number of minor attacks exhibited peculiar behavior. It saw an upsurge in frequency in both 1990 and 2000, which was not mirrored in the frequency of major or minor attacks.

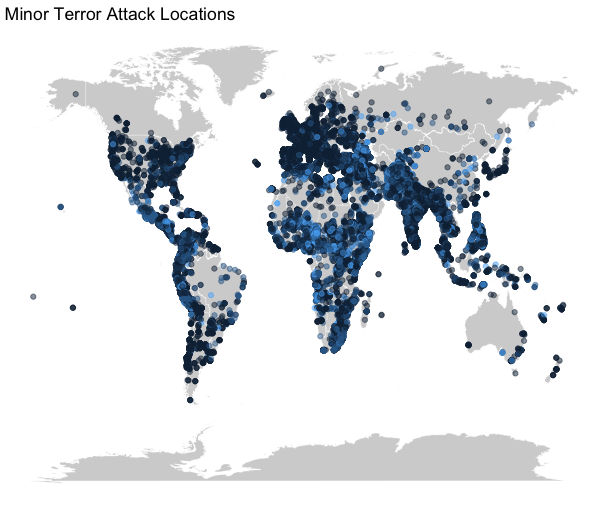
* Plotting geo-spatial plots for all three kinds of attacks.

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



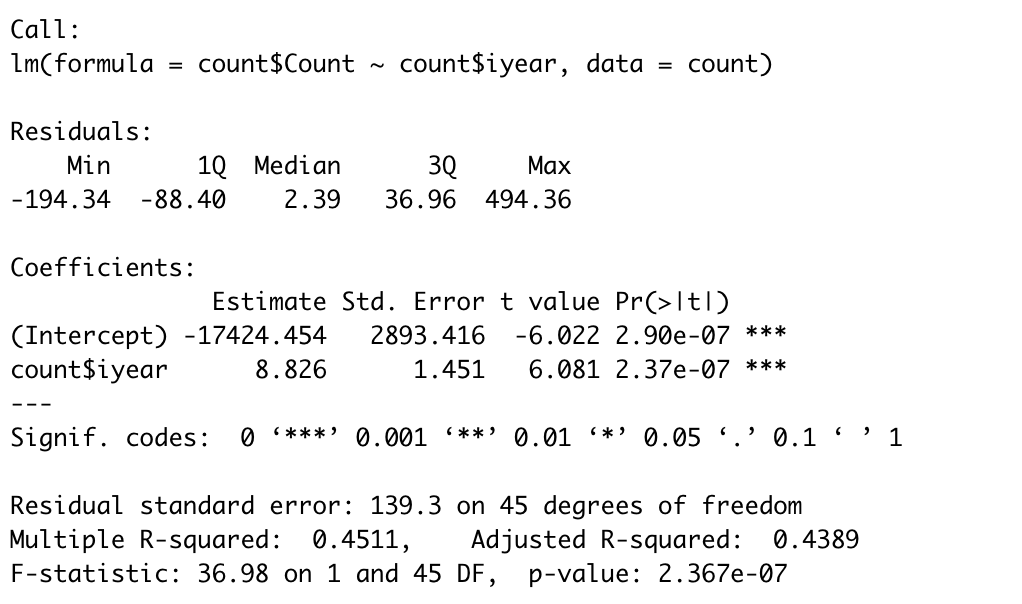
Unlike major attacks, minor attacks are spread throughout the world. From Europe, the US to most of southeast Asia and Africa, minor terror attacks take place everywhere.

# 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:

We carried out regression analysis on year vs number of attacks for each of the three classes of attacks.

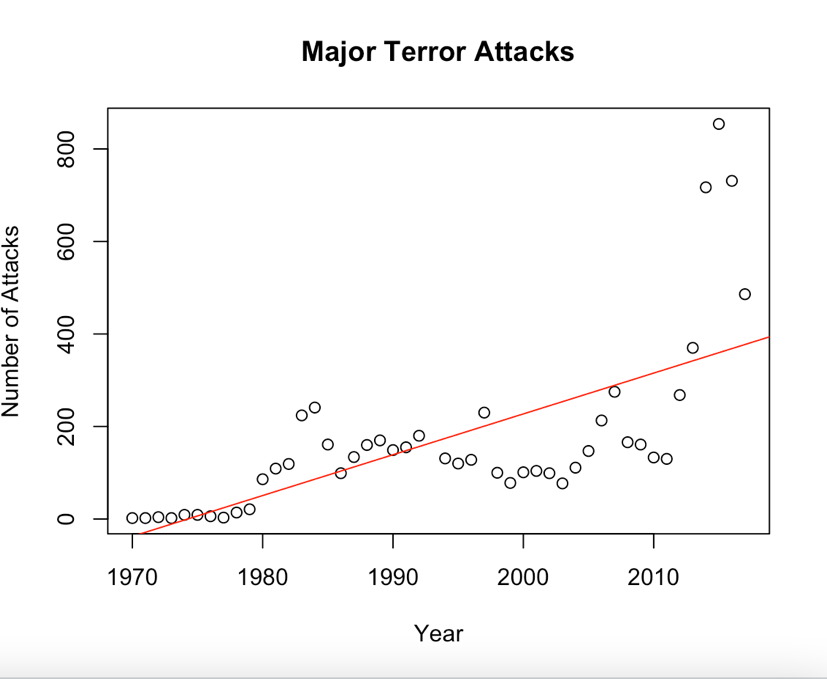
For Major attacks



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

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.



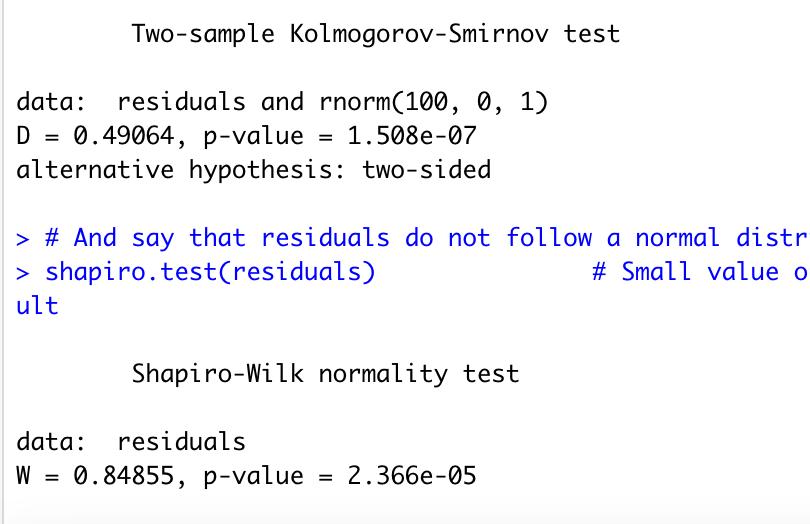
This scatter plot shows us the regression line along with the count of attacks vs years.

# Running Tests

1. **Normality Test**

Conducting KS Test and Shapiro Wilk test.

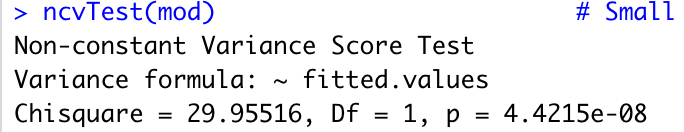
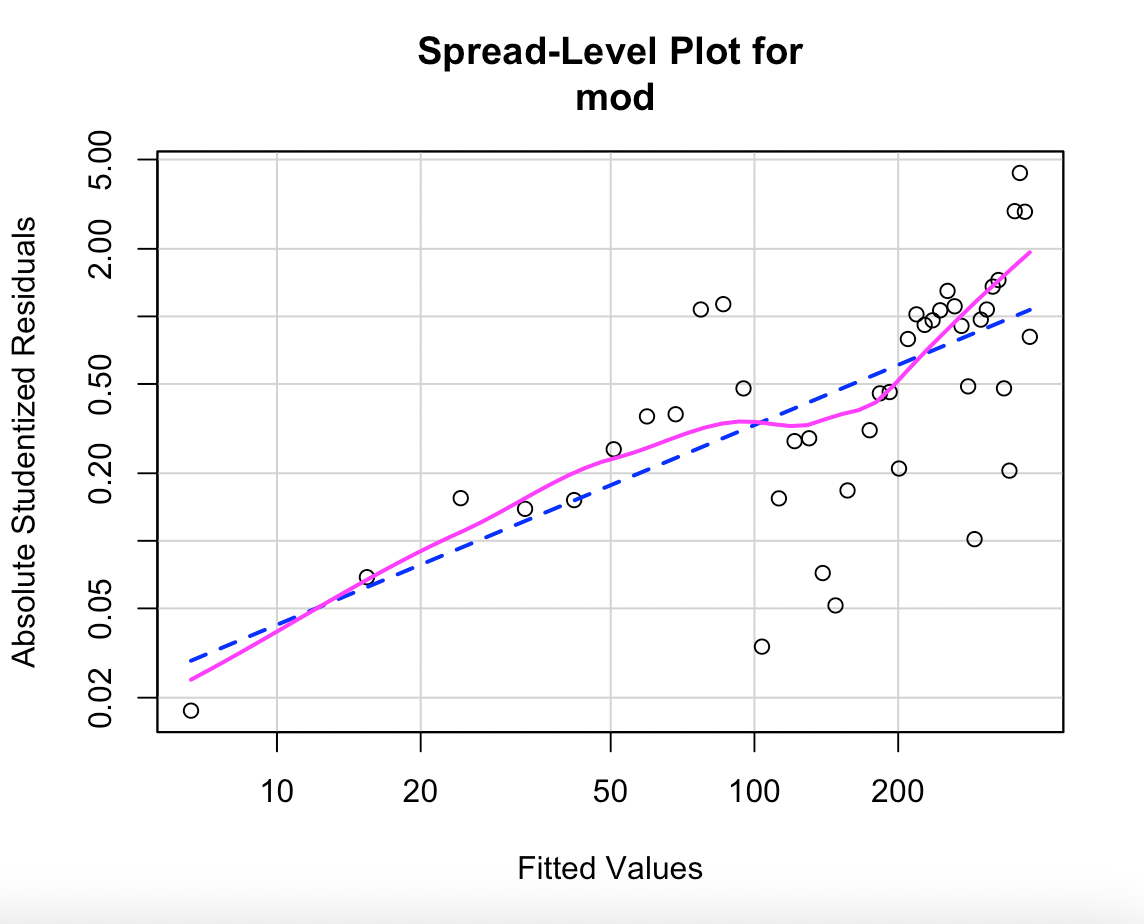
A small p value would indicate that we can reject our hypothesis while a large one would help us not reject our hypothesis that the residuals follow a normal distribution.



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

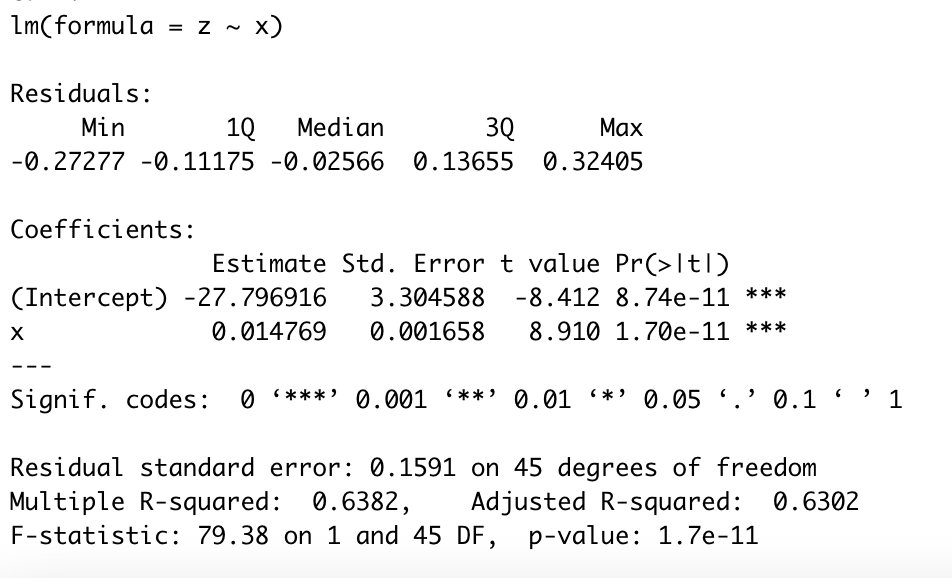
1. **Constant Variance Test**

We employed the ncvTest() function to verify our hypothesis that the residuals have a consistent variance.



The assumption appears to be invalid, as the calculated value is considerably low. As a solution, we will conduct power transformations on our outcome variable y to reestablish a uniform variance. The recommended power transformation value is 0.1104.

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



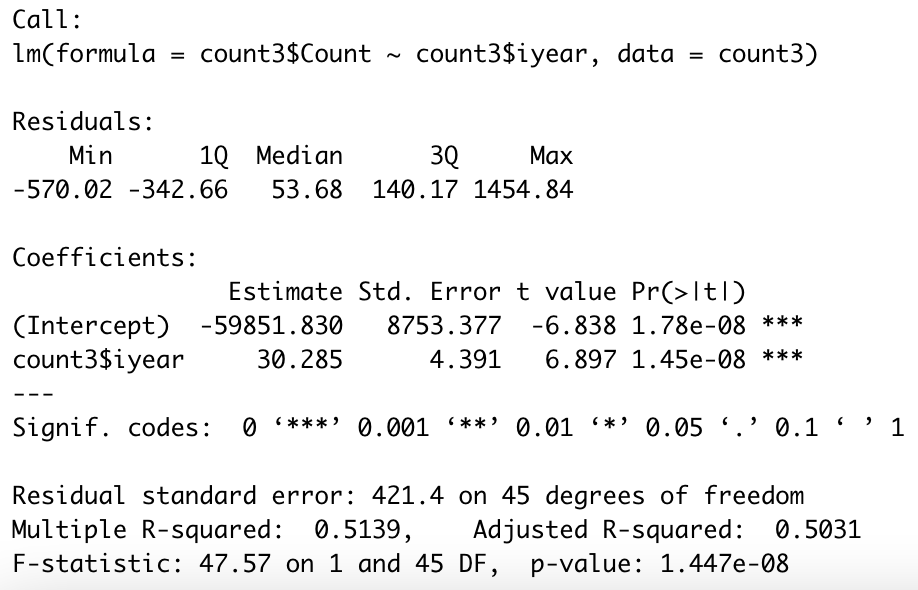
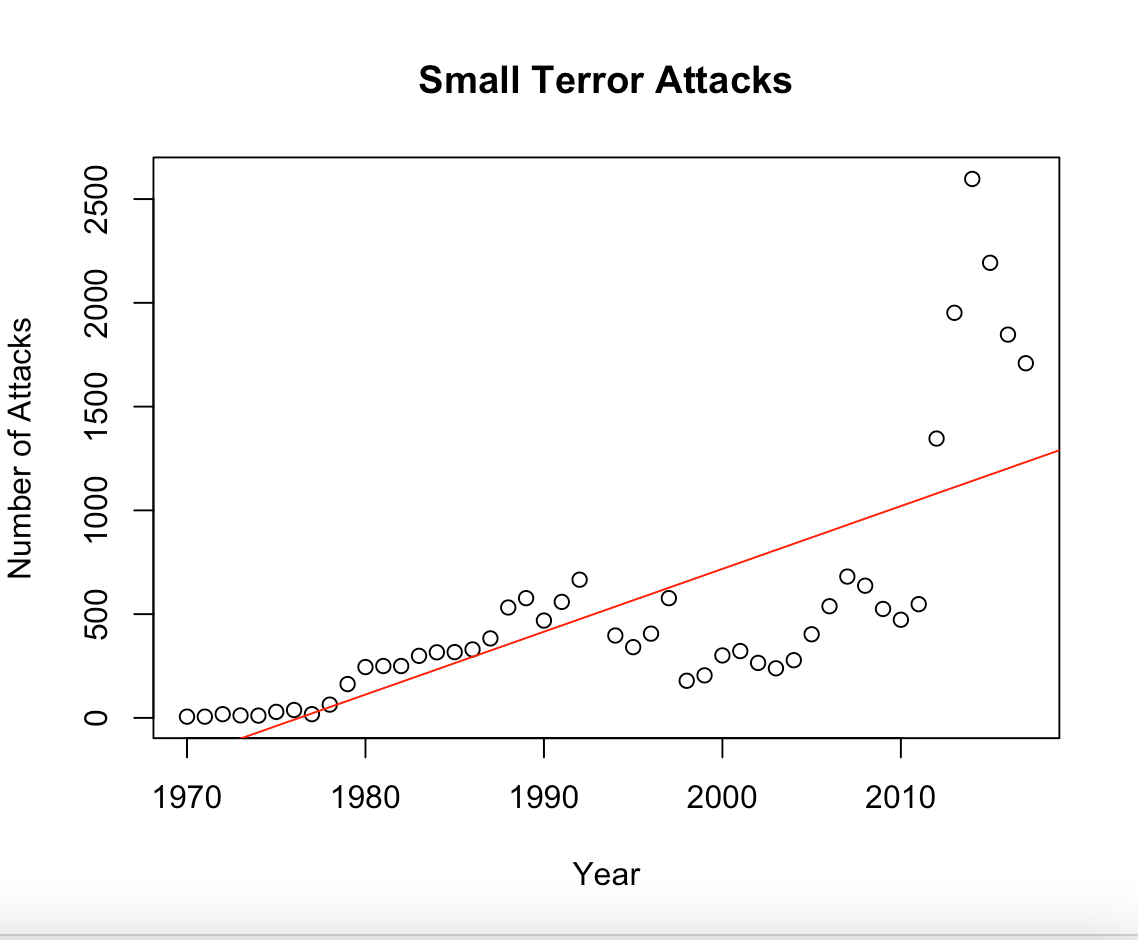
We can see the rise in R-squared statistics, now it captures 63.8% of the total variance.

|  |  |  |
| --- | --- | --- |
| **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.

For Small Attacks

Similar analysis has been carried out for small attacks



𝐴𝑡𝑡𝑎𝑐𝑘𝑠 = − 59851. 8 + 30. 285 \* 𝑦𝑒𝑎𝑟𝑠

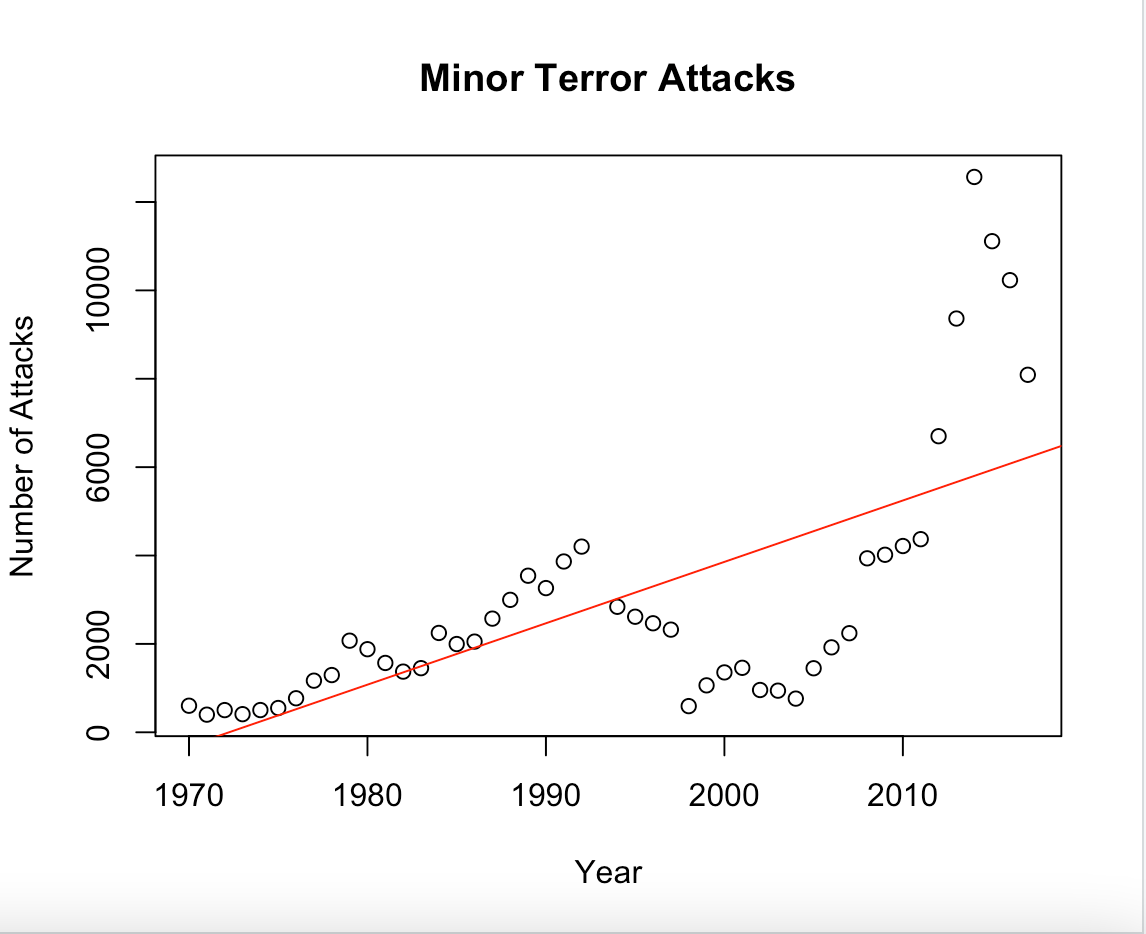
The model captures 51.4% of the total variance in the data and has a RSE of 421.4.

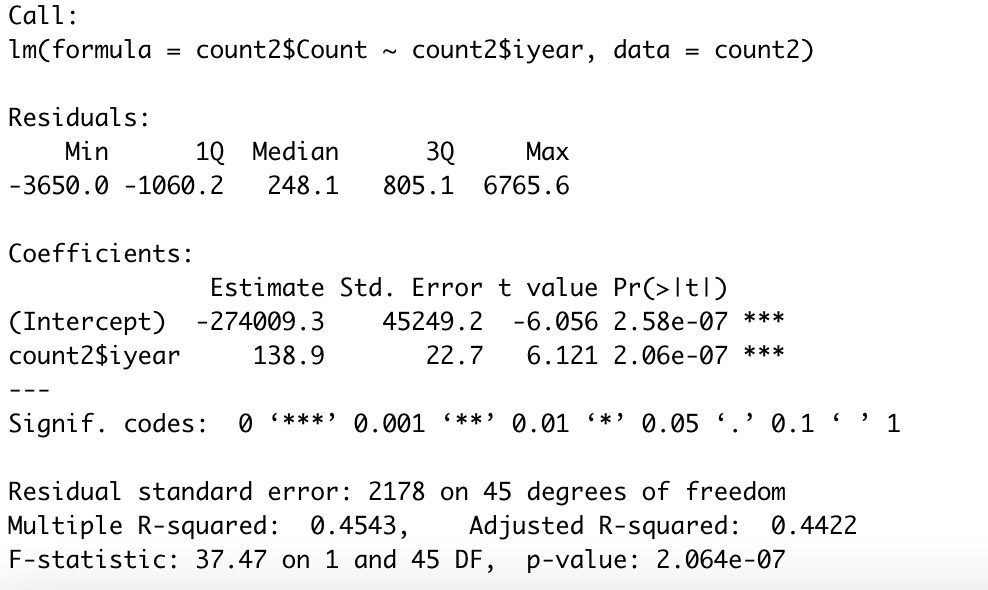
Further evaluation tests were conducted, and the results have been reported in the following table.

|  |  |  |
| --- | --- | --- |
| **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 |

After performing power transformations, the model fit the data much better and performed better as well.

For Minor attacks





𝐴𝑡𝑡𝑎𝑐𝑘𝑠 = − 274009. 3 + 138. 9 \* 𝑦𝑒𝑎𝑟𝑠

The model explains 45.4% of the total variation present in the dataset and has an RSE (Residual Standard Error) of 2178. Additional evaluation assessments were carried out, and their outcomes are presented in the table below.

|  |  |  |
| --- | --- | --- |
| **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 |

Upon conducting the power transformation on the data, we can observe that the model's overall performance has improved. However, the degree of enhancement is not as significant as it was for the model designed for small attacks.

# Result

Here are the statistics for the models constructed before power transformations were implemented:

- For Major Attacks: RSE = 139.3, R-squared = 0.451, Adjusted R-squared = 0.439, and F-statistic = 36.98.

- For Small Attacks: RSE = 421.4, R-squared = 0.514, Adjusted R-squared = 0.503, and F-statistic = 47.57.

- For Minor Attacks: RSE = 2178, R-squared = 0.454, Adjusted R-squared = 0.442, and F-statistic = 37.47.

From 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 the highest value for F-statistic with 47.57.