Predicting Outcome of Terrorist Attacks

STAT 471/571/701, Fall 2018

Andrew Huang and Eric Zeng

12/12/2018

Contents

Executive Summary **1**

Data Summary / EDA **3**
Data Origins . . . 3
Goal of the study . . . 3
EDA . . . 3

Findings / Analysis **10**
Entire Dataset Analysis . . . 10
Data by Country . . . 10
Data by Target . . . 12
Data by Type . . . 14

Future Work **15**
Conclusion . . . 15

Appendix **16**
Works Cited . . . 16
Removed variables . . . 16
Summary of cleaned data . . . 16
Code to Collapse Factor Levels . . . 18
Data by Country . . . 19
Data by Target . . . 24
Data by Type . . . 26

Executive Summary

Acts of intentional violence at a sub-national level have occurred since the development of civilizations. From religious acts of terror to violence with political intent, terrorism has manifested itself in various forms throughout history. The term itself was developed in the 1790s to describe Maximilien Robespierre's Jacobin regime as the "Reign of Terror", but it was popularized following the 1983 Beirut barracks bombings and the 2001 World Trade Center attacks. These acts of violence, regardless of motive, if successful, often claim the lives of the innocent and bystanders.[1]

This leads to the question - how can people use historical data to predict the outcome of terrorist activity? The answer to this lies in statistical analysis of terrorist incidents. Using location data, attack type, group names, target data, and other past terrorism data, we can apply statistical models to best predict the outcome of events - success, number of killed/wounded, total property damage, etc. With a better understanding of what factors can lead to a foiled or successful attempt, people can better try to prevent such attacks in the future.

We found that the predictors for terrorism varied across different regions (countries), targets, and types. Regions with more political and economic tension tend to have more violent, successful terrorism plots involving many casualties than developed areas. Terrorist incidents targeting local law enforcement tend to occur in large urban centers (cities, towns) in southern regions of the world. Militaries engaging terrorist should be wary of the level of resources available to their enemy; the more similar a terrorist groups is to a militia or PMC (private military contractor), the more successful the terrorists tend to be. Ransom incidents tend to end with more released hostages as the years progress, a positive outlook for the future of combating terrorism.

Data Summary / EDA

Data Origins

The origins of the data is the Global Terrorism Database (GTD). The GTD was developed by the National Consortium for the Study of Terrorism and Responses to Terrorism (START) at the University of Maryland, College Park, in Maryland, USA. The database contains incidents of terrorism from 1970 to 2017, and is still under development. There are over 181,000 incidents in the database and 135 factors, including a few main factors listed below:

- `iyear`, `imonth`, `iday`: incident year, month, and day
- `country_txt`, `region_txt`, `provstate`, `city`: country, region, providence/state, city names
- `crit1`, `crit2`, `crit3`: which of the three criterion the incident satisfies (see below)
- `attacktype1_txt`: a text descriptor for the attack type; there are other variables regarding the type of attack
- `targtype1_txt`, `targsubtype1_txt`, `natlty1_txt`: a text descriptor for the target type, subtype, and nationality; there are other variables regarding the type of targets
- `gname`, `gnucertain1`, `individual`: group name, and indicator variables for presence of guns and if individual attack
- `weaptype1_txt`, `weapsubtype1_txt`: type of weapon used in attack
- `success`, `nkill`, `nwound`, `propextent_txt`: indicates if the incident was successful, the number of killed and wounded, and the extent of property damage (respectively)

To be included in the study, an incident must qualify with three fields:

- The incident must be intentional – the result of a conscious calculation on the part of a perpetrator.
- The incident must entail some level of violence or immediate threat of violence -including property violence, as well as violence against people.
- The perpetrators of the incidents must be sub-national actors. The database does not include acts of state terrorism.

Additionally, it must satisfy two of the following three criterion:

- Criterion 1: The act must be aimed at attaining a political, economic, religious, or social goal.
- Criterion 2: There must be evidence of an intention to coerce, intimidate, or convey some other message to a larger audience (or audiences) than the immediate victims.
- Criterion 3: The action must be outside the context of legitimate warfare activities.

In general, the GTD does not include plots that are not enacted or attempted. For an incident to be considered, the attackers must be "out the door", or en route to execute the attack. This means, according to their handbook, "in general if a bomb is planted but fails to detonate; if an arsonist is intercepted by authorities before igniting a fire; or, if an assassin attempts and fails to kill his or her intended target, the attack is considered for inclusion in the GTD, and marked success=0."[2]

Goal of the study

The goal of the study is to utilize data on terrorist attacks and identify which factors can be best used to predict the outcome of such attacks. In this study, we will be analyzing various outcomes, from success, number of killed, and random outcomes.

EDA

First, we read in the data given in csv format. There are 181,691 observations and 135 total variables. However, this must be further cleaned. There were three main steps in the data cleaning process for eliminating variables:

1) There were many variables for multiple groups; for instance, there are 3 groups for target (target1, targettype1, targetsubtype1, corp1, target1, natlty1, etc.), 3 groups for attack type, 3 groups for claim, and 3 groups for weapon types. These are present in the case that multiple groups stage an attack, or multiple targets are targeted. However, for the most parts of the dataset, the second and third group for most predictors were NA, and thus were dropped.
2) We filtered variables that were just encodings of other variables. For instance, there were two variables `country` and `country_txt`. The former is a number encoding for a country, while the latter is the name of the country. For purposes of easier readability, we kept the text description.
3) We finally dropped variables that contained too many NA's. This included number of killed US citizens, group that claimed the incident, etc. Considering the number of US wounded/kill/perp is quite specific, it makes sense that many incident do not report this. Since these caused our models to fail to run, we ended up removing this from the overall data set for the rest of the study.

In general, this final cleaned dataset was used as a baseline for each of our subset analysis. See the appendix for the full list of removed variables. Unless otherwise noted, each subset will be derived from these 31 remaining variables (see below). Additionally, the -9 and -99s that were encoded for missing variables were coded into NA in R. We omitted these NA's from the cleaned dataset due to the large number of examples we had from the database already.

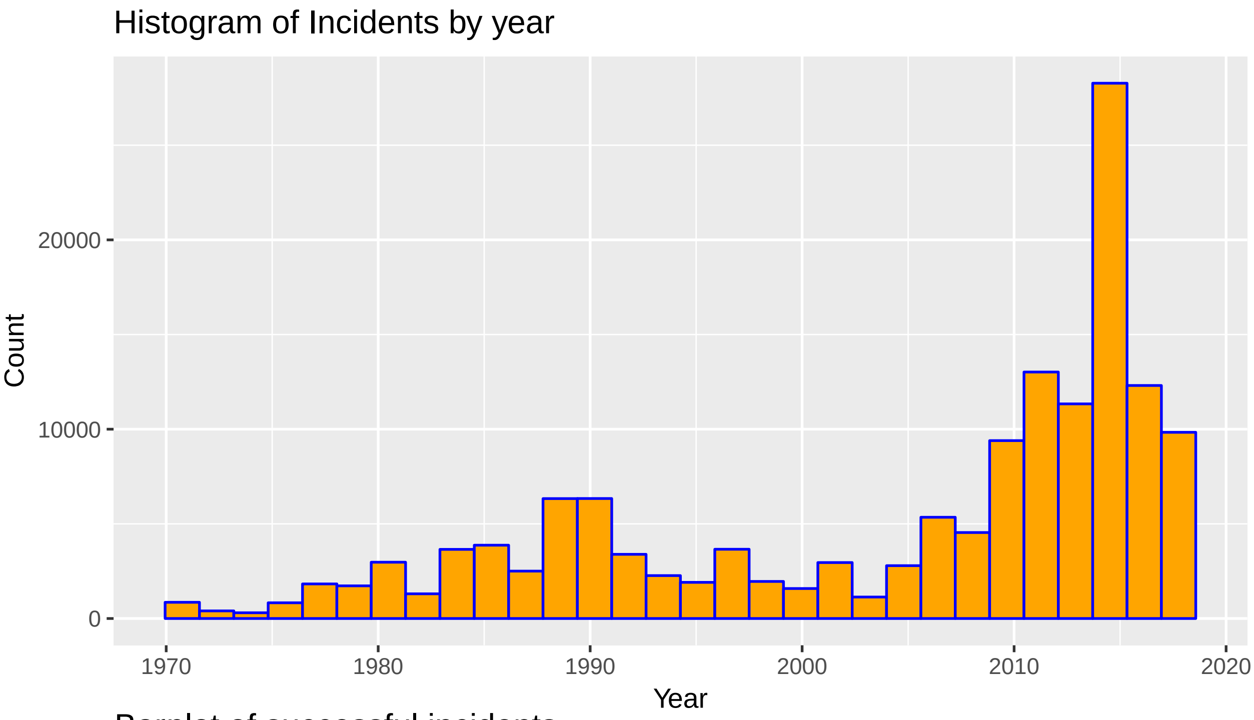
```
##  [1] "iyear"            "imonth"           "iday"            
##  [4] "extended"         "country_txt"      "region_txt"      
##  [7] "provstate"        "city"             "latitude"        
## [10] "longitude"        "specificity"      "vicinity"        
## [13] "crit1"            "crit2"            "crit3"           
## [16] "doubtterr"        "multiple"         "success"         
## [19] "suicide"          "attacktype1_txt"  "targtype1_txt"   
## [22] "targsubtype1_txt" "natlty1_txt"      "gname"           
## [25] "guncertain1"      "individual"       "weaptype1_txt"   
## [28] "weapsubtype1_txt" "nkill"            "nwound"          
## [31] "propextent_txt"
```

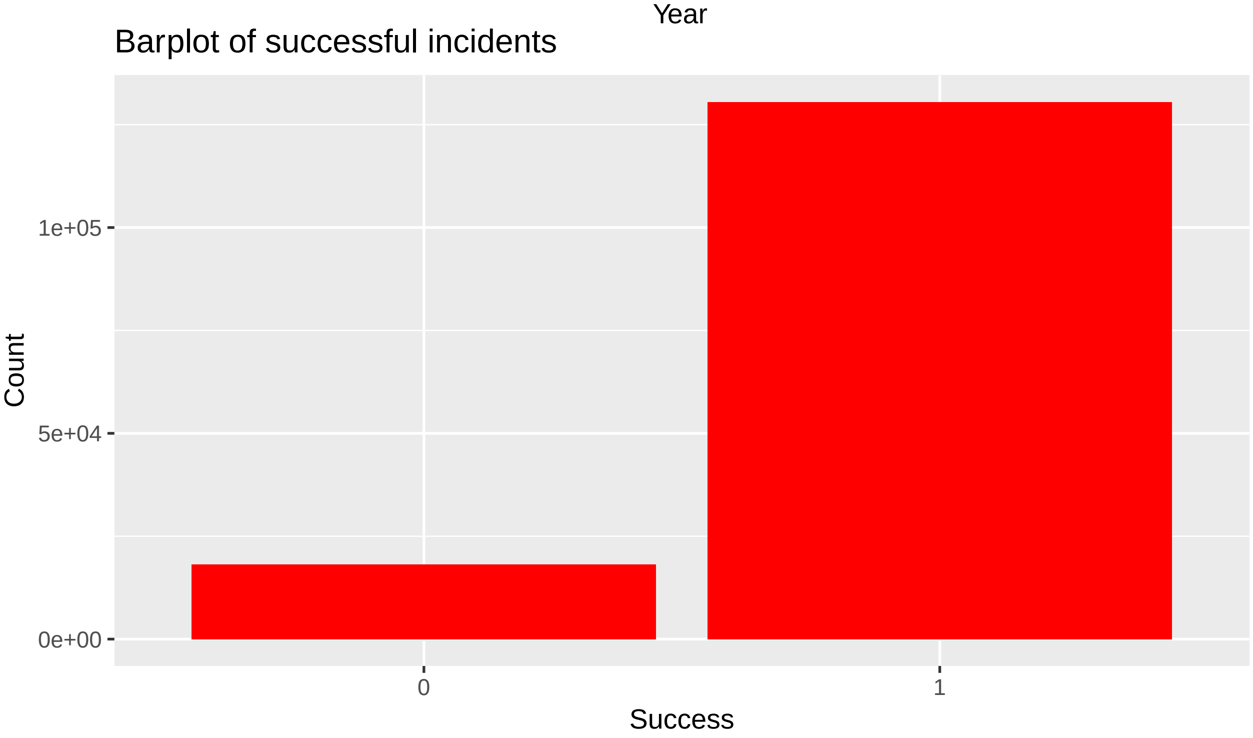
Let's first get a sense of the (cleaned) dataset as a whole. Through the summary of the dataset (see Appendix for full summary), we can elucidate a few key insights from the data. Dropping the NA's yielded 148,627 observations on 31 variables.

```
## [1] 148627     31
```

The years range from 1970 to 2017, with a huge left skew in data, meaning there are a lot more reported incidents in the recent years, which makes sense given the development of the Internet. Additionally, there are significantly more successful than unsuccessful incidents, likely due to the fact that only incidents where the perpetrators were "out of the door" were recorded, as aforementioned.

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```





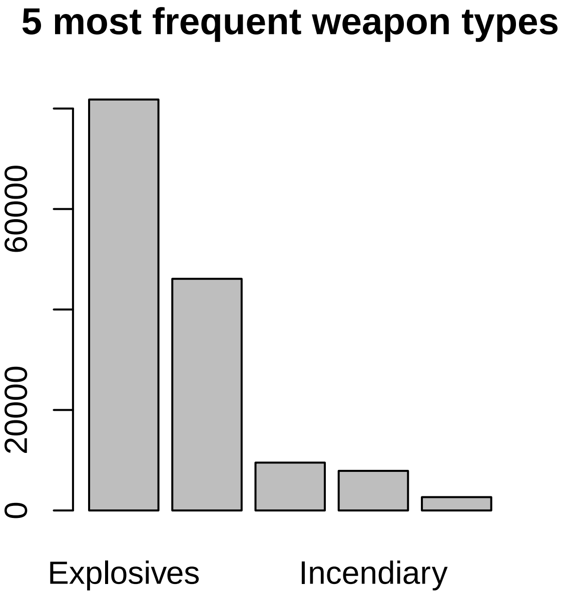
From the histogram of incidents by year, it is clear that there is an influx of recent events, so it makes sense to generate a subset of data for recent events as well as those from past events. Thus, we created two subset splits, one for data from 2017, and one for data from 1997-2016. We choose the range 1997-2016 because before 1997, our GTD is quite space for some variables. This ensures we still have comparable datasets with similar sparsity.

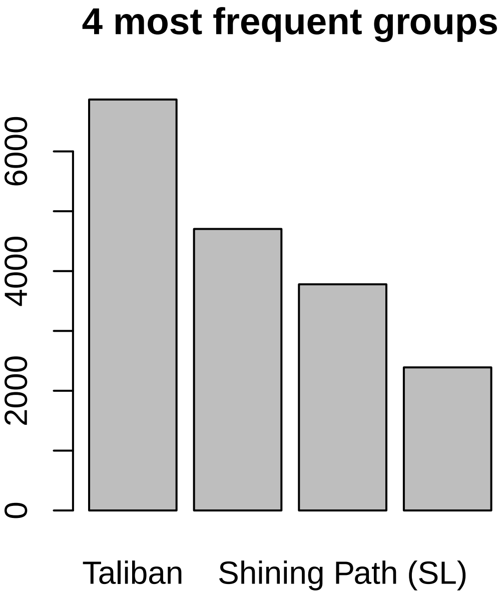
Looking at the summary of the number of number of killed and wounded (respectively), we see that on average there are 2.192 killed and 3.408 wounded, but the medians are both 0. The maximum of these incidents were both the tragic attack on the World Trade Center on 9/11/2001.

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##    0.000    0.000    0.000    2.192    2.000 1384.000
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##    0.000    0.000    0.000    3.408    2.000 8191.000
```

Looking at the types of attacks, the top 5 weapons of choice range from explosives and firearms down to melee incidents. The most frequent 4 groups (the top factor was "Unknown") were Taliban, ISIL, Shining Path, and New People's Army.





To view a geographical distribution of the incidents, we plot the total killed as a function of country location on a global map. We see that there is a large number of incidents in Iraq, as well as the South Asian region of Afghanistan, Pakistan, and India. Something worthy of noting is also the small, but still significant number of incidents in the USA, Western Europe, and South America.

```
## 183 codes from your data successfully matched countries in the map
## 17 codes from your data failed to match with a country code in the map
## 60 codes from the map weren't represented in your data
```

A similar map to look at, the aggregate number of people killed by country, displays similar information. However, it is notable that in this chloropleth, the shade of some South Asian countries (notably India), as well as many countries in Western Europe and the USA, drop off significantly. This signifies that while there are a large number of incidents in these countries, either they do not end up being successful, or are stopped at the source quickly. This makes these countries worthy of further examination.

```
## 183 codes from your data successfully matched countries in the map
## 17 codes from your data failed to match with a country code in the map
## 60 codes from the map weren't represented in your data
```

Based on these maps, it makes sense to also subset out specific countries. Thus, we developed a subset for 1) Iraq, the country with by far the most incidents and total number killed, 2) the USA, given the fact that we are based in the USA and for the large number of unsuccessful or small incidents, 3) Japan, for being a country with a small number of incidents and number killed, and 4) Syria, for the recent developments of the Syrian Civil War. For larger context of countries that score high on the Human Development Index (HDI), we also subset off a developed countries dataset, consisting of those located in North America, Western Europe, and East Asia.

Some final subsets that we created were for the target type, and the type of attack. Based on the target types, we created a subset for 1) attacks on police, 2) attacks on military, and 3) attacks on police agencies.

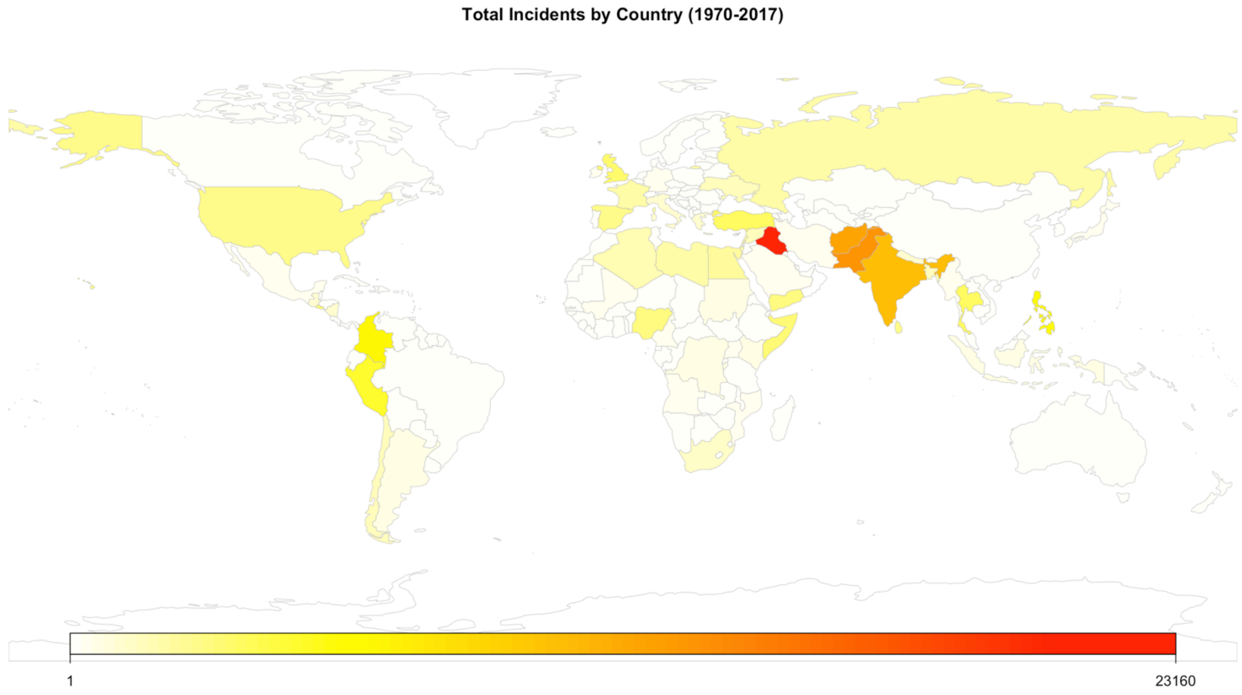


Figure 1: ‘Total Incidents by Country (1970-2017)’

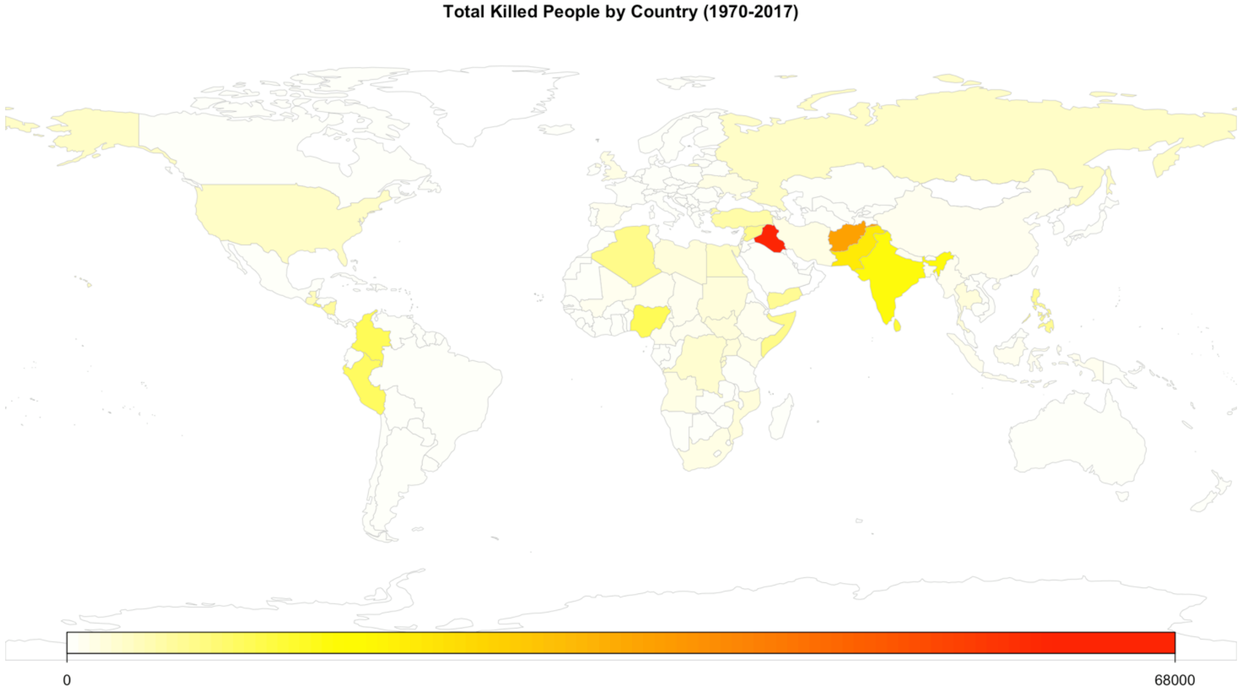
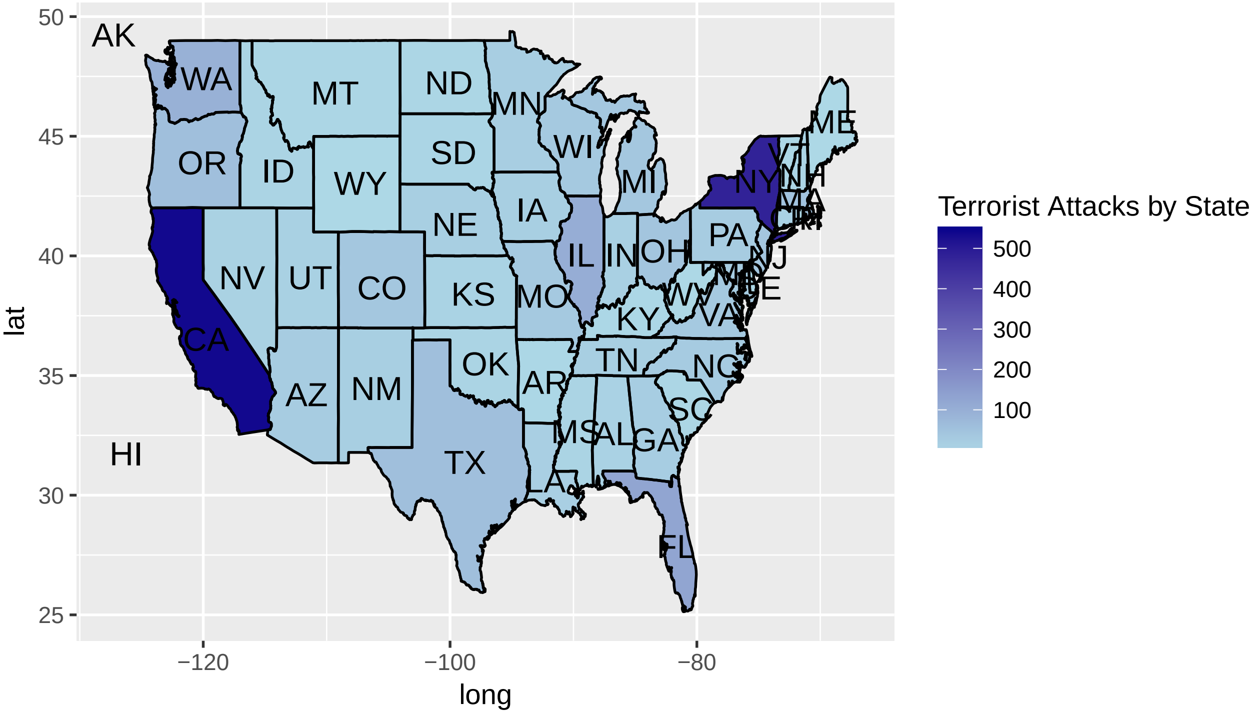


Figure 2: ‘Total Killed People by Country (1970-2017)’

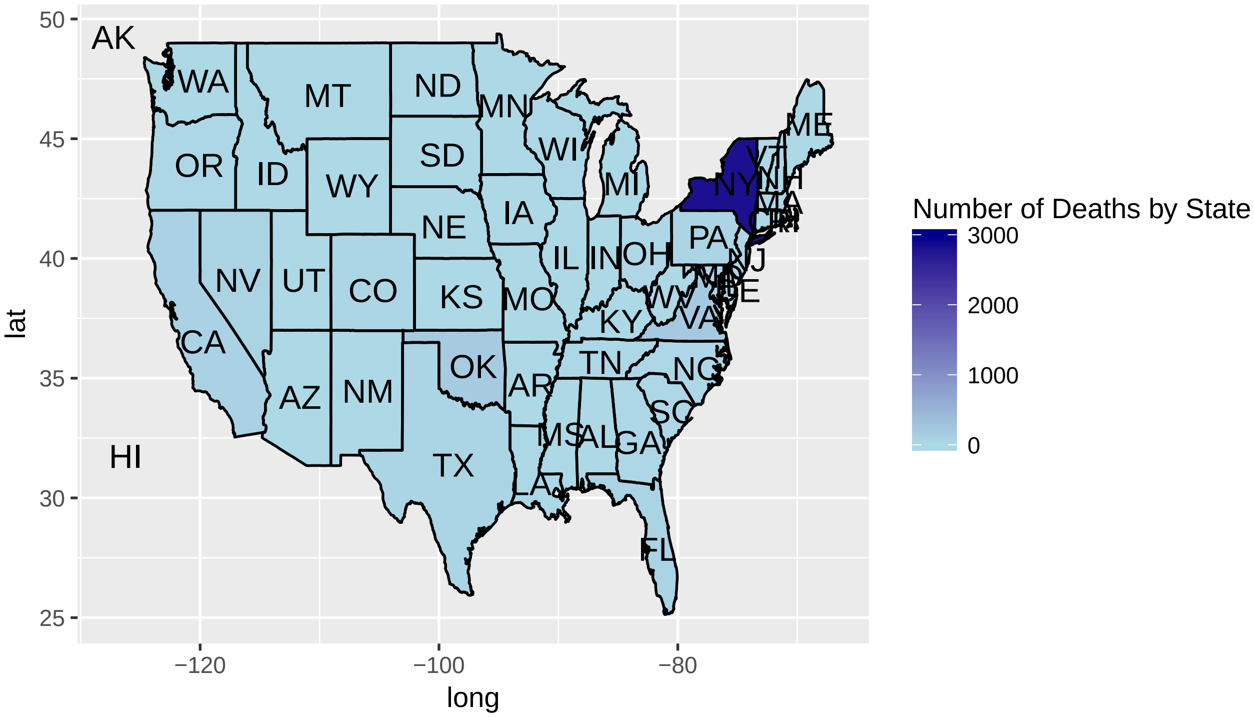
Due to the armed nature of these bodies, it makes sense to branch these off to separate subsets to analyze if it results in a change in the number of successful attacks and killed people. Additionally, with a large number of predictors based on property damage and also for kidnappings / ransoms, we subset both of these into their own datasets for analysis.

Out of curiosity, we make a plot of total terrorist attacks by state in the USA. Note that there are a significant number of attacks in New York and California, but even populous states like Texas and Florida have a significantly less number of attacks.

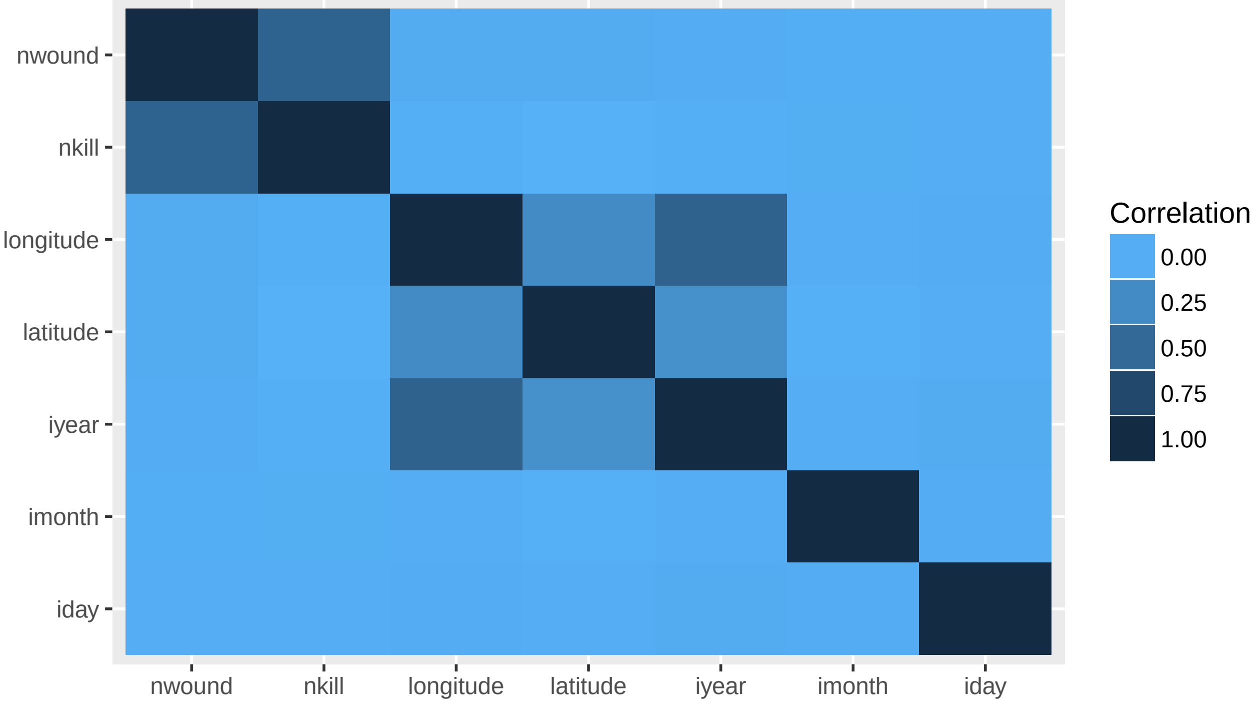
```
## Warning: package 'maps' was built under R version 3.4.4
```



It is extremely interesting when plotting this in comparison to the number of deaths by state, as seen below. The number of deaths in California, as well as of any other state, are dwarfed by the number of deaths in the 9/11 attacks in New York. Looking at the actual dataframe, while California has 534 attacks, the most in the nation, it is surpassed by New York, Virginia, and Oklahoma in terms of number of deaths. This could be because many of the attempts are unsuccessful, or are stopped before they can worsen.



Let's do a correlation heatmap for all the numeric values in the dataset. While all of these pairings do not make the most logical sense to correlate, it is good reaffirmation to see some of the values. There does not seem to be a correlation between the dates of the attacks, and latitude and longitude are very mildly, but significantly correlated. It is also good reaffirmation to see that the number of wounded and number of killed are highly correlated.



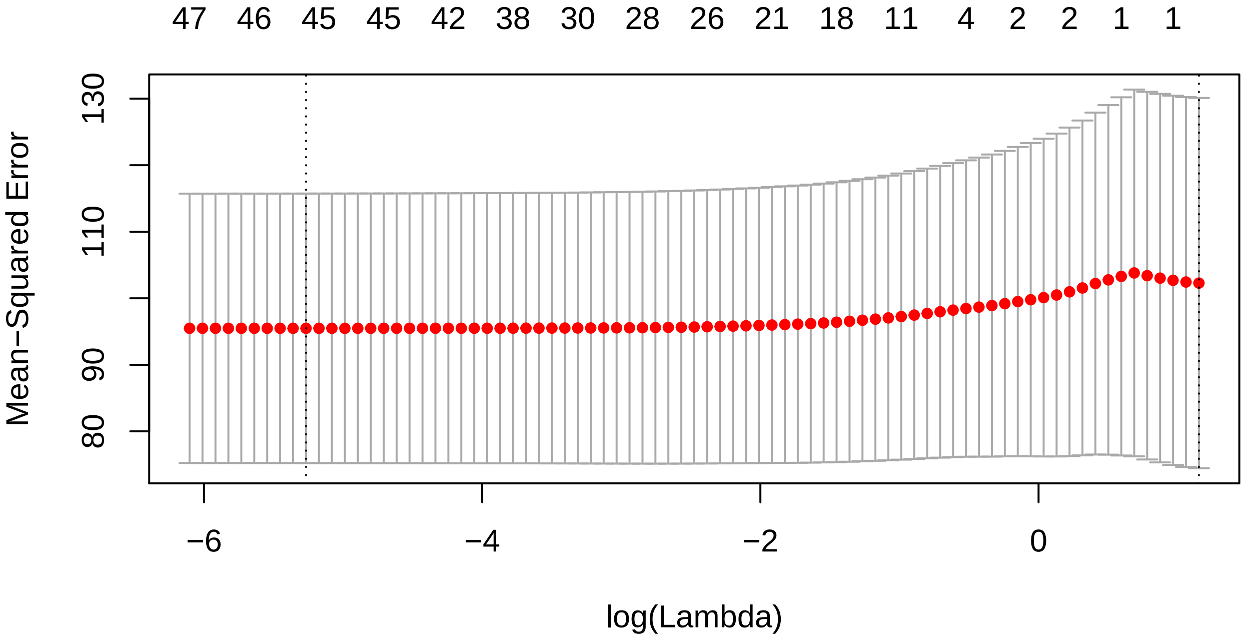
Before we go onto the analysis, we will collapse all factor levels into the top 5 + an other category to improve the runtime of our code for dataset we are predicting on `success`. If we were to include all 3537 levels of variables like `gname`, then our classification models would take too long to train. See the appendix for the code to collapse factors with many levels.

Findings / Analysis

Entire Dataset Analysis

We will start with the whole dataset `data_clean` and by training a classifier on `nkill` with a logistic classifier. We will run cv.glm with $\alpha = 1$ to run LASSO regression for model selection. Once we take the smallest model at most 1 std deviation of cross-validated mean errors away from the minimum cvm, we will run Anova() iteratively to remove the most insignificant grouping until all predictors are significant at the $\alpha = 0.01$ level.

Note that for this specific subset of the data, we will use `lambda.min` instead of `lambda.1se` as `lambda.1se` implies a model with 0 nonzero coefficeints (essentially, just an intercept). Thus for feasibility, we will use `lambda.min`.



Adjusted R-Squared for Entire Dataset (LASSO): 0.1540607

RSE for Entire Dataset (LASSO): 9.3059462

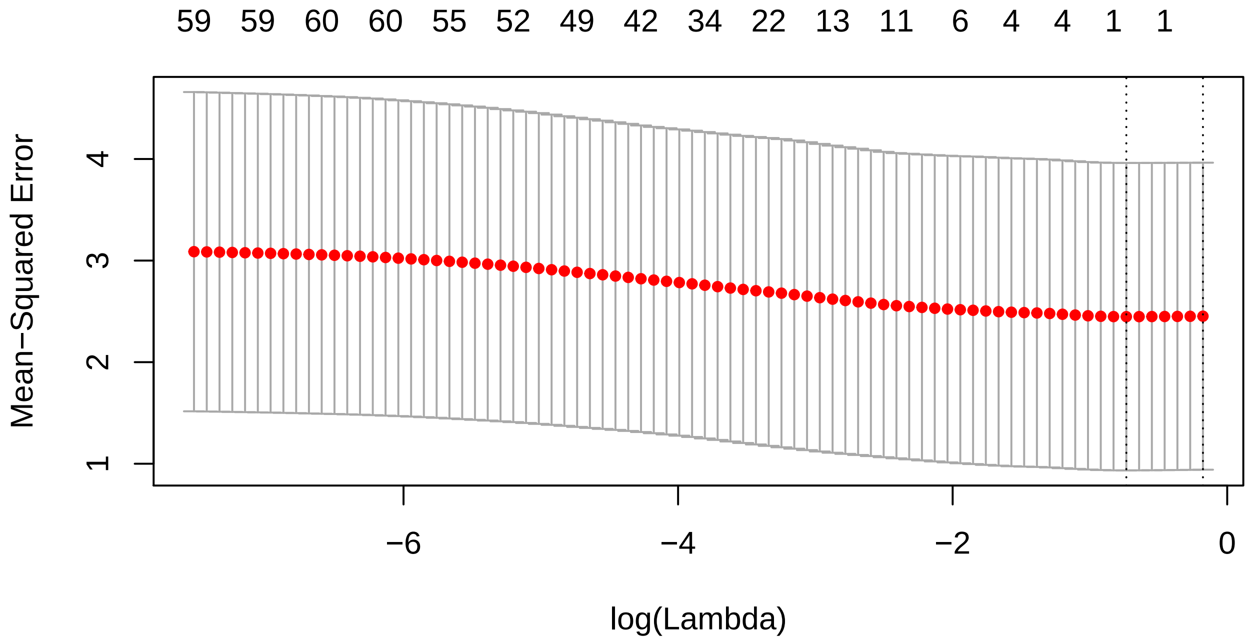
From these R-Squared values, it seems that the model is not able to explain a lot of variance in the nkill dataset, but with a pvalue of 2.2e-16, the model does seem significant in that regard. Due to the poor performance of the model on the overall dataset, we subset the data into smaller datasets to use for further analysis.

Data by Country

Let us take a look at the differences between countries, namely Iraq, Japan, USA, and Syria. We will build a linear regression model through LASSO (cvglm) and run Anova() to remove (backward select out) categories that are not significant at the $\alpha = 0.01$ level.

Starting with Japan, we first remove `nwound` before modeling building as it would be highly correlated with `nikill`. In reality, we can't really use `nwound` to predict `nkill` as they occur at the same time (it makes sense that more wounded implies more killed, and vice versa). We are interested in deeper relations, if they exist. The summaries can be found in the appendix.

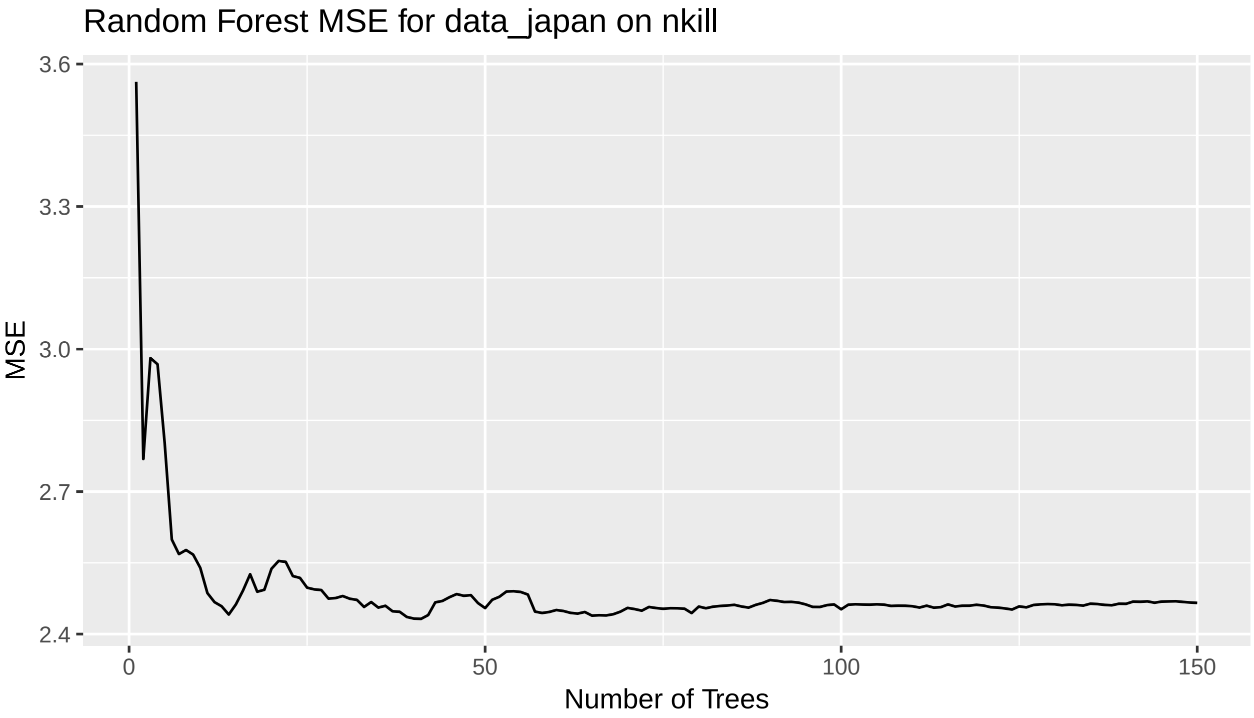
Below is our graph of cross-validated mean errors.



Adjusted R-Squared for Japan (LASSO): 0.3373682

RSE for Japan (LASSO): 1.2710812

Let us also try running random forest model on this dataset. We must remove `provstate`, `city`, `targsubtype_txt`, `natlty1_txt`, and `gname` as they have over 53 levels. We tuned our random forest appropriately with `mtry` and `ntree`. Displayed is the MSE of our RF:



MSE at 150 trees: 2.4656854

The same approaches were applied the USA, Syria, and Iraq. Their outputs are shown in the appendix and the overall summaries are:

	Adjusted.R.Squared..LASSO.	RSE..LASSO.	MSE.of.Random.Forest
Japan	0.337368159125459	1.27108122909616	2.46568544197762
USA	0.579077190945226	25.0891313741301	921.011397597494
Syria	0.139196432809236	9.61117862561429	88.8768283594253
Iraq	0.108303563422869	9.42474183236123	80.4923681087204

Taking a look at the Anova() and full summary outputs in the Appendix, we see some interesting findings:

For Japan, our most significant variable is `attacktype1_txtHostage Taking (Barricade Incident)`. Examining the japan dataset we see two data points where this categorical level is true:

	provstate	city	attacktype1_txt	nkill	nwound
179	Tokyo	Sumida	Hostage Taking (Barricade Incident)	0	0
246	Kanagawa	Sagamihara	Hostage Taking (Barricade Incident)	19	26

We see that only the incident in Kanagawa/Sagamihara actually involved any target deaths (with a decent number of 26 people wounded). It seems that this model has overfit to this single datapoint.

For the USA, we find a much simpler model with only 5 predictors (vicinity and the 4 levels of `propextent_txt`). `propextent_txtCatastrophic (likely >= $1 billion)` is the most significant predictor holding all other constant, followed by `vicinity1`. The former maybe due to the tragic 9/11 incident which is clearly in the Catastrophic level for property damage; this maybe another case of overfitting to a single point of extreme noise. The latter (`vicinity1`) suggests that large centers of population have more targets killed. This may simply be due to population density and also does not serve as a good predictor.

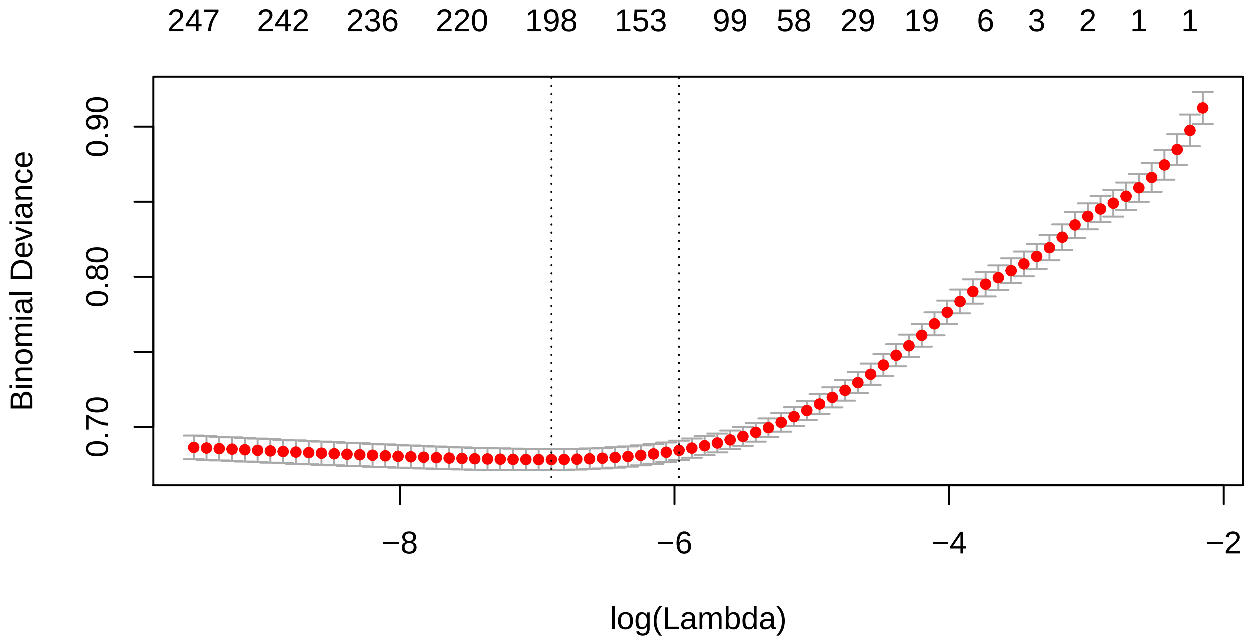
For Syria, there are a lot more predictors than in the USA or Japan dataset (suggesting a complexity of terrorism characterizations in this area). More notably, we see that `crit3` is highly significant. `crit3` incidents are described as "Outside International Humanitarian Law", which aligns with the modern day rising refugee crisis. Additionally, `suicide`, `weapsubtype1_txtAutomatic or Semi-Automatic Rifle`, and `weapsubtype1_txtVehicle` are quite good predictors of `nkill`. This seems to differ from the tools used by terrorists in developed countries such as Japan and the US.

For Iraq, we see that some noteworthy significant predictors are `iyear`, `extended`, `crit1`, `suicide`, as well as many `attacktype1_txt` and `weapsubtype1_txt` levels such as Assassination, Bombing, Infrastructure, Hostages, Arson, and Automatic/semi rifle/guns. Given that most of our data is recent (with post-1997 incidents better reported than those before), it makes sense that Iraq appears to be the most violent region. The early 2000s War on Terror may play a role in influencing so many military-ques incidents, though such a thought is only a conjecture; we would need more complex analyses to test that hypothesis. Regardless, it seems deadly terrorism incidents in Iraq are mainly influenced by extended, political/economic/religious, military aggression.

Data by Target

For our datasets by target of `police`, `military`, and `police`, we will train a classification model to predict on `success`. This is because we expected a lower number of deaths if these terror incidents are more specifically targeted. We are instead interested in seeing how different target entities resist terrorist attacks, or if the profiles of terrorists differ across targets.

We will use a logistic regression selected through cross-validation, then backwards select out using Anova(). Here is a sample cross-validated deviance graph by lambda value for government. The remaining relevant outputs are in the Appendix.



	Residual.Deviance..LASSO.	Cross.validated.Accuracy
Government	14554.0176266645	0.8771443103528
Military	12243.384732204	0.933481442760063
Police	10398.9245686216	0.965631609305903

Taking a look at the government dataset, we have no variables of real surprise. Assassination, Bombing, Infrastructure, Hostage seem to be strong predictors. Otherwise, we find it somewhat surprising that `crit1` was not chosen as it relates to politically charged terror attacks.

In the military dataset, we see nearly all standard military equipment in weapsubtype are significant; this includes levels such as rifles, explosives, handguns, knives, pipe bombs, time bombs, projectiles, and even vehicles. The scale of terrorism (maybe even warfare in this category) is clearly much higher as successful terrorist must be well equipped to target a military entity(s).

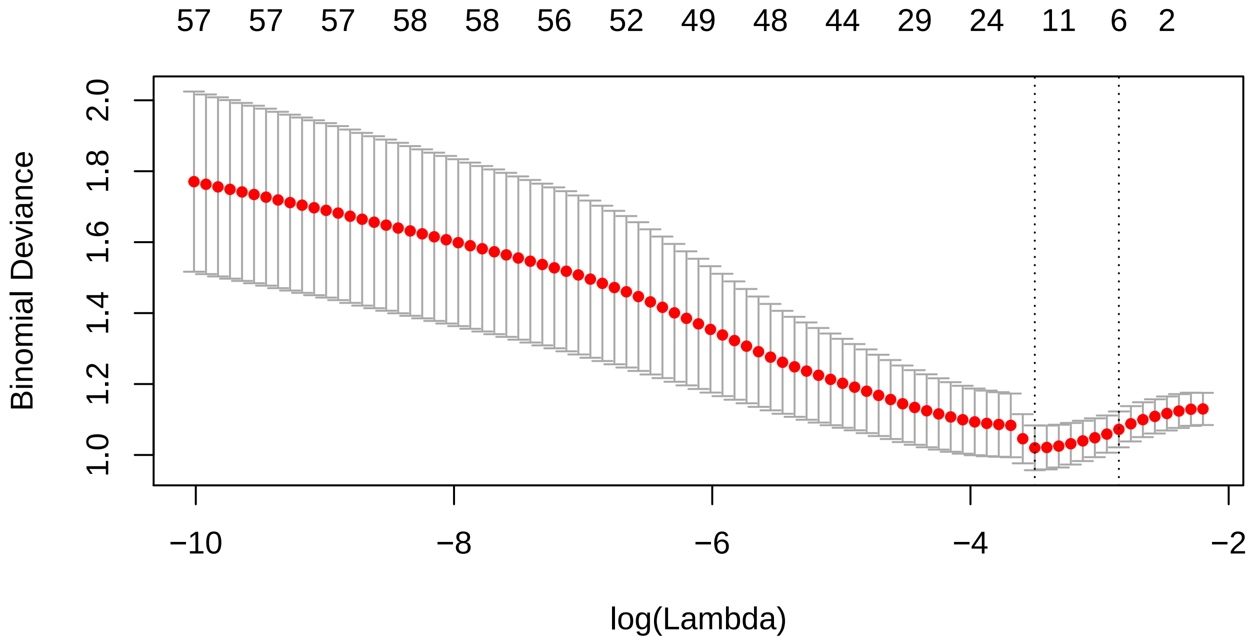
As for the police dataset, we see some interesting predictors that relate to geographical attributes. `latitude`, `specificity2`, and `specificity3` are very significant (holding each other and all other predictors constant) at predicting success:

	Estimate	Pr...z..
latitude	−0.0132080753617859	1.59447897484529e−15
specificity2	−0.239770036712441	0.0404771396127338
specificity3	−0.351718433398738	1.18905392328452e−05

More specifically, latitude decreases the log-odds ration of success by -0.0132081 per 1 degree increase in latitude; the more north we go, the less likely a police-targeted terror attack succeeds. `specificity` records how granular the location data is; it discretely ranges from 1 (meaning lat/lang of the city/town in known) to 4 (meaning no 1st order administrative region is known). The presence of `specificity2`, and `specificity3` also decrease the log-odds ration by -0.23977 and -0.3517184 of success respectively. In general, southern large cities are more likely to be hit by successful terror attacks targeting the police. These may be areas of political unrest where anti-government rebel groups are common.

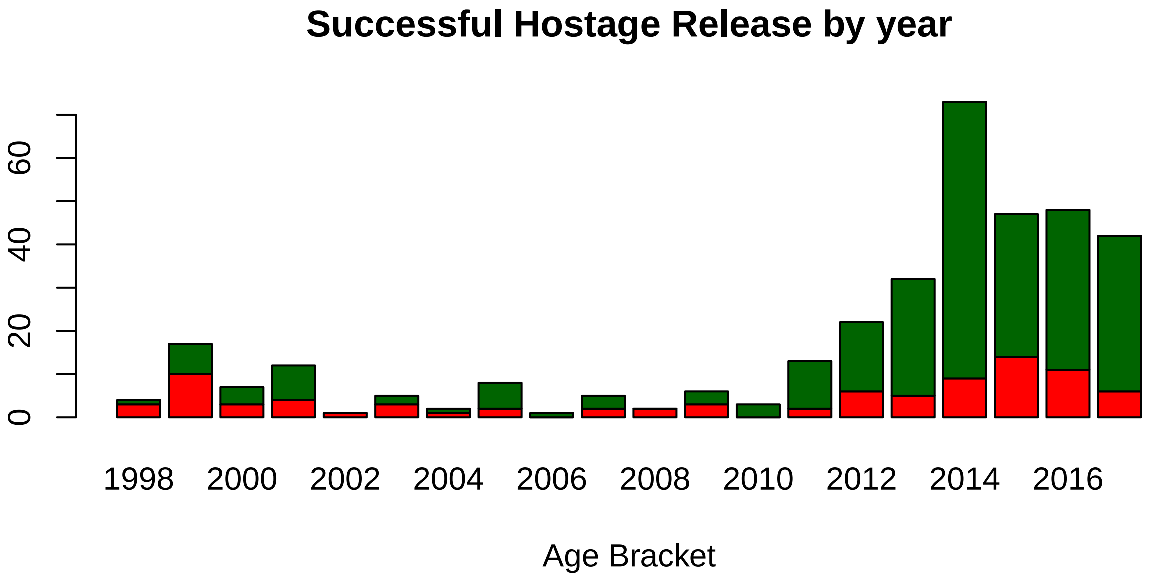
Data by Type

We tested a subset of data on only attacks where hostages were taken and some form of a ransom was demanded. These were interesting because these attacks have all been successful - some form of kidnapping has taken place - but the event is not over. We take the other predictors in the dataset and predict on if the kidnapping will be resolved successfully (release of the hostages or successful rescue), or not.



We fit a binomial classification regression on “successful” or “unsuccessful”. From the ANOVA and LASSO analysis, the most significant predictors for predicting the outcome of the kidnapping were “iyear”, “attacktype1_txt”, “weapsubtype1_txt”, “propextent_txt”, and “ransomamt”. While the latter four variables are explainable - the type of attack, the type of weapons used, the extent of property damage, and the overall ransom demanded - the year seems to be out of place as a significant predictor.

After plotting the year vs. successful hostage releases (successful in green, unsuccessful in red), it is clear that the reasoning for this is because many kidnappings in recent years have been able to be resolved successfully. This is a positive trend that we hope keeps happening!



Future Work

There are many aspects that can be focused on for future works. From the raw data, there was a column of the number of perpetrators. This seems logically correlated with the success of the attack, but due to the sparsity of the reporting, we were unable to use the variable. There were other issues related to this as well.

We would to explore the additional hypothesis mentioned throughout our analysis. This includes studying if or how the US's political stances have shaped terrorism, and whether large southern cites tend to have high crime rates. These question cannot be answered with the GTD, but we can use these results as a basis of what to search for or expect with the future work.

Many of our comparison models, such as the four for the countries, vary drastically in sample size. This led to a wide range of errors/accuracies between the four models trained. In the future, we may want to use more size-comparable choices of countries or partitions such that size of the sub-dataset is not a factor in modeling. Additionally, the skewness of some predictors suggest that we should have filtered even more out during the preprocessing step. Our large number of potential predictors to train on raises concerns about p-hacking, a situation where we may falsely find significance simply due to so many variables and not a true underlying relationship. [3]

Conclusion

Our country datasets most notably showed the difference between developed and non-developed countries. Developed countries (Japan, USA) tended to have fewer terrorist incidents overall, but also ones that involve less militaristic conflict. There seems to be no strong predictor of the severity of casualties in these developed countries, which may have led our model to overfit onto random noise. On the other hand, non-developed countries were tied strongly with military conflict, targets, and weapons. This suggest an important, but somewhat expected, result that more war-prone regions will tend to have terror attacks with higher casualties. For kidnappings, our analysis revealed an increasing trend of recent kidnappings ending with the successful release of the hostages.

As terrorist evolve in the modern day, local authorities continue to stay a step ahead to prevent their attacks. Knowing what factors, such as location, weapons available, socioeconomic/political tensions, and potential targets can let authorities better predict where the next big attack may be.

Appendix

Works Cited

[1] https://books.google.com/books?id=6qSjk2C9x6wC&pg=PA161#v=onepage&q&f=false

[2] https://www.start.umd.edu/gtd/downloads/Codebook.pdf

[3] https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1002106

Removed variables

```
# filter data
data_clean <- data %>% select(-c(approxdate, resolution, location, summary, alternative,
                                 alternative_txt, attacktype2, attacktype2_txt,
                                 attacktype3, attacktype3_txt, targtype2, targtype2_txt,
                                 targsubtype2, targsubtype2_txt, corp2, target2, natlty2,
                                 natlty2_txt, targtype3, targtype3_txt, targsubtype3,
                                 targsubtype3_txt, corp3, target3, natlty3, natlty3_txt,
                                 gsubname, gname2, gsubname2, gname3, gsubname3, motive,
                                 guncertain2, guncertain3, claim2, claimmode2, claimmode2_txt,
                                 claim3, claimmode3, claimmode3_txt, compclaim,
                                 weaptype1, weapsubtype1, weaptype2, weaptype2_txt,
                                 weapsubtype2, weapsubtype2_txt, weaptype3, weaptype3_txt,
                                 weapsubtype3, weapsubtype3_txt, weaptype4, weaptype4_txt,
                                 weapsubtype4, weapsubtype4_txt, nhostkid, nhostkidus,
                                 nhours, ndays, divert, kidhijcountry, ransom, ransomamt,
                                 ransomamtus, ransompaid, ransompaidus, ransomnote,
                                 hostkidoutcome, hostkidoutcome_txt, nreleased, addnotes,
                                 scite1, scite2, scite3, dbsource, INT_LOG, INT_IDEO,
                                 INT_MISC, INT_ANY, related, ishostkid))

# remove text or IDs
data_clean <- data_clean %>% select(-c(country, region, attacktype1, targtype1,
                                       targsubtype1, natlty1, claimmode, propextent,
                                       propcomment))

# re-code -9 or -99 to NA
data_clean <- data_clean %>% select(-c(nkillus, nkillter, nwoundus, nwoundte, nperps,
                                       nperpcap, claimed, claimmode_txt, property,
                                       propvalue, weapdetail, eventid, corp1, target1))

data_clean$vicinity[data_clean$vicinity < 0] <- NA
data_clean$doubtterr[data_clean$doubtterr < 0] <- NA
data_clean <- na.omit(data_clean)
```

Summary of cleaned data

```
##      iyear          imonth            iday          extended
##  Min.   :1970   Min.   : 0.000   Min.   : 0.00   Min.   :0.00000
##  1st Qu.:1994   1st Qu.: 4.000   1st Qu.: 8.00   1st Qu.:0.00000
##  Median :2011   Median : 6.000   Median :15.00   Median :0.00000
##  Mean   :2005   Mean   : 6.468   Mean   :15.54   Mean   :0.03228  
##  3rd Qu.:2014   3rd Qu.: 9.000   3rd Qu.:23.00   3rd Qu.:0.00000  
##  Max.   :2017   Max.   :12.000   Max.   :31.00   Max.   :1.00000  
##                                                                    
##       country_txt                         region_txt   
##  Iraq       :23159   Middle East & North Africa:43663  
##  Pakistan   :13113   South Asia                :40537  
##  Afghanistan:11844   South America             :14080  
##  India      :10414   Sub-Saharan Africa        :12777  
##  Colombia   : 6282   Western Europe            :11971  
##  Philippines: 6001   Southeast Asia            :11021  
##  (Other)    :77814   (Other)                   :14578  
##              provstate             city           latitude     
##  Baghdad           :  7426   Baghdad  :  7370   Min.   :-53.15  
##  Balochistan       :  3563   Unknown  :  6273   1st Qu.: 12.10  
##  Saladin           :  3203   Mosul    :  2112   Median : 31.66  
##  Al Anbar          :  3029   Karachi  :  2028   Mean   : 24.04  
##  Khyber Pakhtunkhwa:  3018   Lima     :  1883   3rd Qu.: 34.53  
##  Nineveh           :  2969   Mogadishu:  1333   Max.   : 74.63  
##  (Other)           :125419   (Other)  :127628                   
##    longitude        specificity       vicinity           crit1       
##  Min.   :-157.86   Min.   :1.000   Min.   :0.00000   Min.   :0.0000  
##  1st Qu.:  11.26   1st Qu.:1.000   1st Qu.:0.00000   1st Qu.:1.0000  
##  Median :  44.19   Median :1.000   Median :0.00000   Median :1.0000  
##  Mean   :  32.18   Mean   :1.373   Mean   :0.07357   Mean   :0.9872  
##  3rd Qu.:  69.42   3rd Qu.:1.000   3rd Qu.:0.00000   3rd Qu.:1.0000  
##  Max.   : 179.37   Max.   :4.000   Max.   :1.00000   Max.   :1.0000  
##                                                                       
##      crit2            crit3          doubtterr         multiple     
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000  
##  1st Qu.:1.0000   1st Qu.:1.0000   1st Qu.:0.0000   1st Qu.:0.0000  
##  Median :1.0000   Median :1.0000   Median :0.0000   Median :0.0000  
##  Mean   :0.9932   Mean   :0.8774   Mean   :0.1603   Mean   :0.1414  
##  3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:0.0000  
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000  
##                                                                     
##     success          suicide                                attacktype1_txt 
##  Min.   :0.0000   Min.   :0.00000   Bombing/Explosion            :77923  
##  1st Qu.:1.0000   1st Qu.:0.00000   Armed Assault                 :33845  
##  Median :1.0000   Median :0.00000   Assassination                 :15294  
##  Mean   :0.8778   Mean   :0.04106   Facility/Infrastructure Attack: 7865  
##  3rd Qu.:1.0000   3rd Qu.:0.00000   Hostage Taking (Kidnapping)   : 6754  
##  Max.   :1.0000   Max.   :1.00000   Unknown                       : 5029  
##                                       (Other)                       : 1917  
##                      targtype1_txt  
##  Private Citizens & Property:34273  
##  Military                   :22956  
##  Police                     :20804  
##  Government (General)       :18780  
##  Business                   :14893  
##  Transportation             : 5500  
##  (Other)                    :31421  
##                                                    targsubtype1_txt
##  Unnamed Civilian/Unspecified                          :  9886     
##  Police Security Forces/Officers                        :  9180
##                                                         :  8491
##  Military Personnel (soldiers, troops, officers, forces):  6614
##  Military Unit/Patrol/Convoy                            :  6378
##  Government Personnel (excluding police, military)      :  5945
##  (Other)                                                :102133
##       natlty1_txt                                                gname
##  Iraq       :22738   Unknown                                    :67999
##  Pakistan   :12723   Taliban                                    : 6867
##  India      :10549   Islamic State of Iraq and the Levant (ISIL): 4704
##  Afghanistan:10163   Shining Path (SL)                          : 3779
##  Colombia   : 5995   New People's Army (NPA)                    : 2391
##  Philippines: 5840   Al-Shabaab                                 : 2375
##  (Other)    :80619   (Other)                                    :60512
##   guncertain1       individual         weaptype1_txt
##  Min.   :0.0000   Min.   :0.000000   Explosives:81793
##  1st Qu.:0.0000   1st Qu.:0.000000   Firearms  :46106
##  Median :0.0000   Median :0.000000   Unknown   : 9523
##  Mean   :0.0901   Mean   :0.003398   Incendiary: 7882
##  3rd Qu.:0.0000   3rd Qu.:0.000000   Melee     : 2655
##  Max.   :1.0000   Max.   :1.000000   Chemical  :  276
##                                      (Other)   :  392
##                                   weapsubtype1_txt     nkill
##  Unknown Explosive Type                   :37776   Min.   :   0.000
##  Unknown Gun Type                         :27376   1st Qu.:   0.000
##                                           :12343   Median :   0.000
##  Automatic or Semi-Automatic Rifle        :12283   Mean   :   2.192
##  Vehicle                                  : 9150   3rd Qu.:   2.000
##  Projectile (rockets, mortars, RPGs, etc.): 8753   Max.   :1384.000
##  (Other)                                  :40946
##      nwound                                               propextent_txt
##  Min.   :   0.000                                                :94236
##  1st Qu.:   0.000   Catastrophic (likely >= $1 billion)          :    6
##  Median :   0.000   Major (likely >= $1 million but < $1 billion):  792
##  Mean   :   3.408   Minor (likely < $1 million)                  :38265
##  3rd Qu.:   2.000   Unknown                                      :15328
##  Max.   :8191.000
##
```

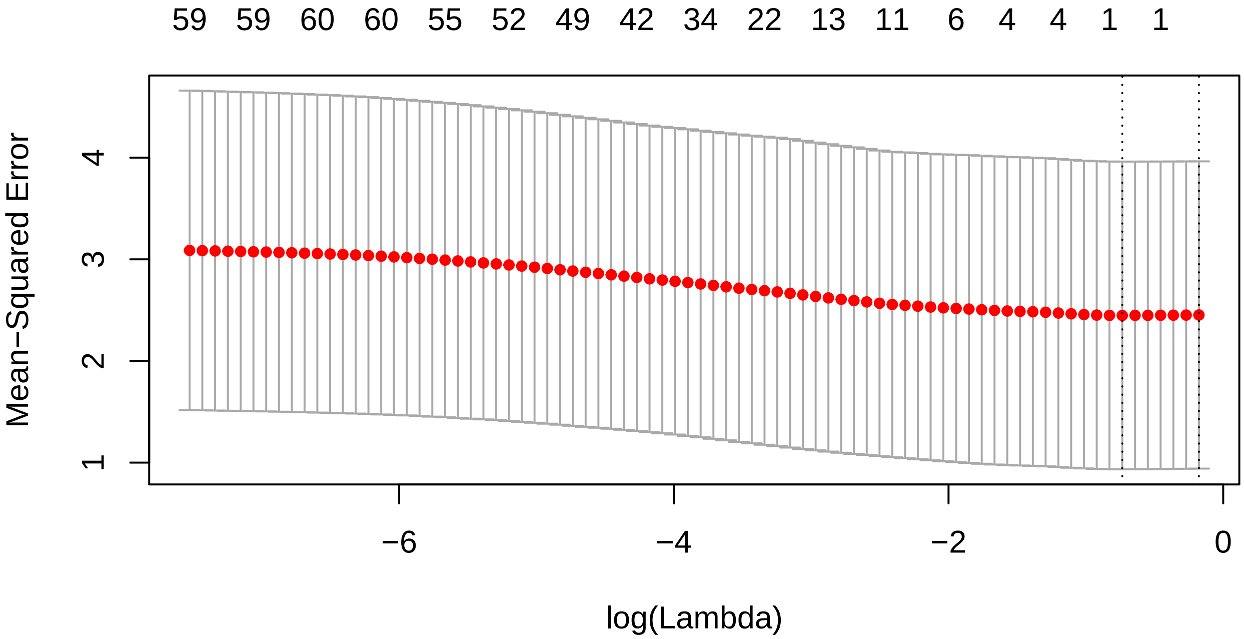
Code to Collapse Factor Levels

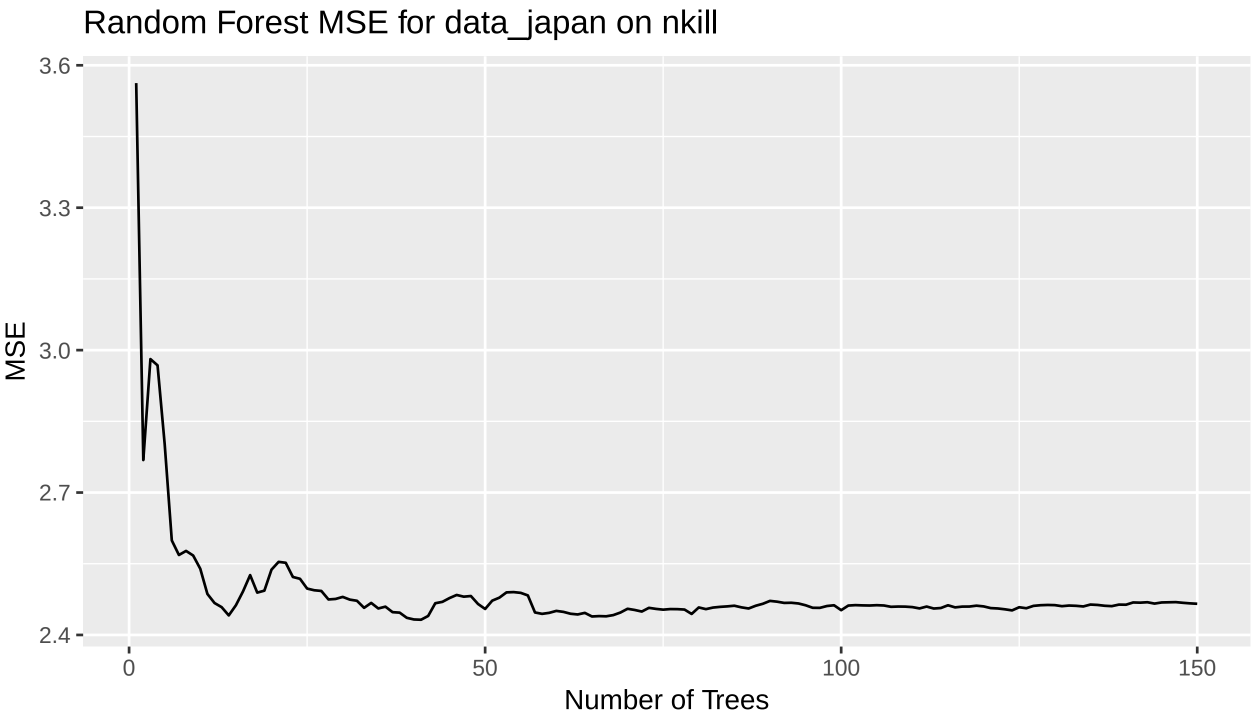
```r
collapse <- function(d, type=''){
  tgt_cols <- colnames(d)[sapply(d, function(x) length(levels(x))) > 6]

  for (col in tgt_cols) {
    if (col != type) {
      sorted <- sort(table(d[,col]), decreasing=TRUE)
      top5 <- sorted[1:5]
      if ('Unknown' %in% names(top5)) {
        top5[names(sorted[6])] <- sorted[[6]]
      }
      d[,col] <- fct_collapse(d[,col], 'Unknown' = levels(d[,col])[!(levels(d[,col])
                                                                     %in% names(top5))])
    }
  }
}
```

Data by Country

Japan Final Anova output, LASSO cvm-by-lambda, and RF mse-by-ntree

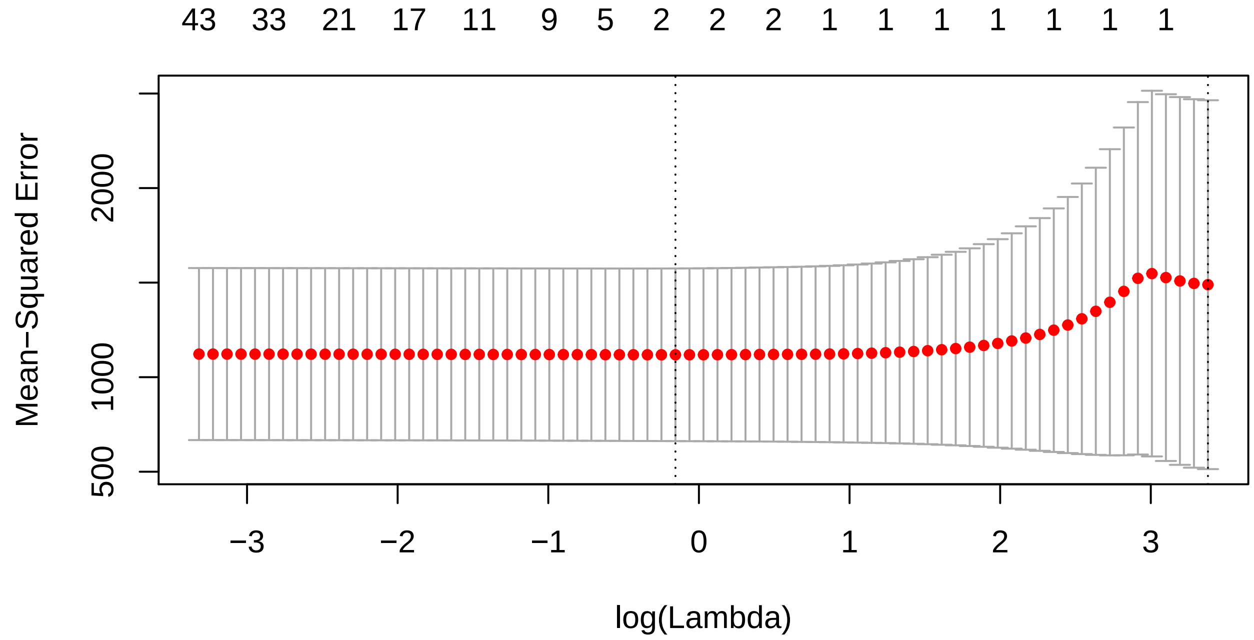
```
## Anova Table (Type II tests)
##
## Response: nkill
##                   Sum Sq Df F value   Pr(>F)
## attacktype1_txt 159.153  7 14.0725 3.89e-15 ***
## weaptype1_txt    43.388 11  2.4414 0.006767 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

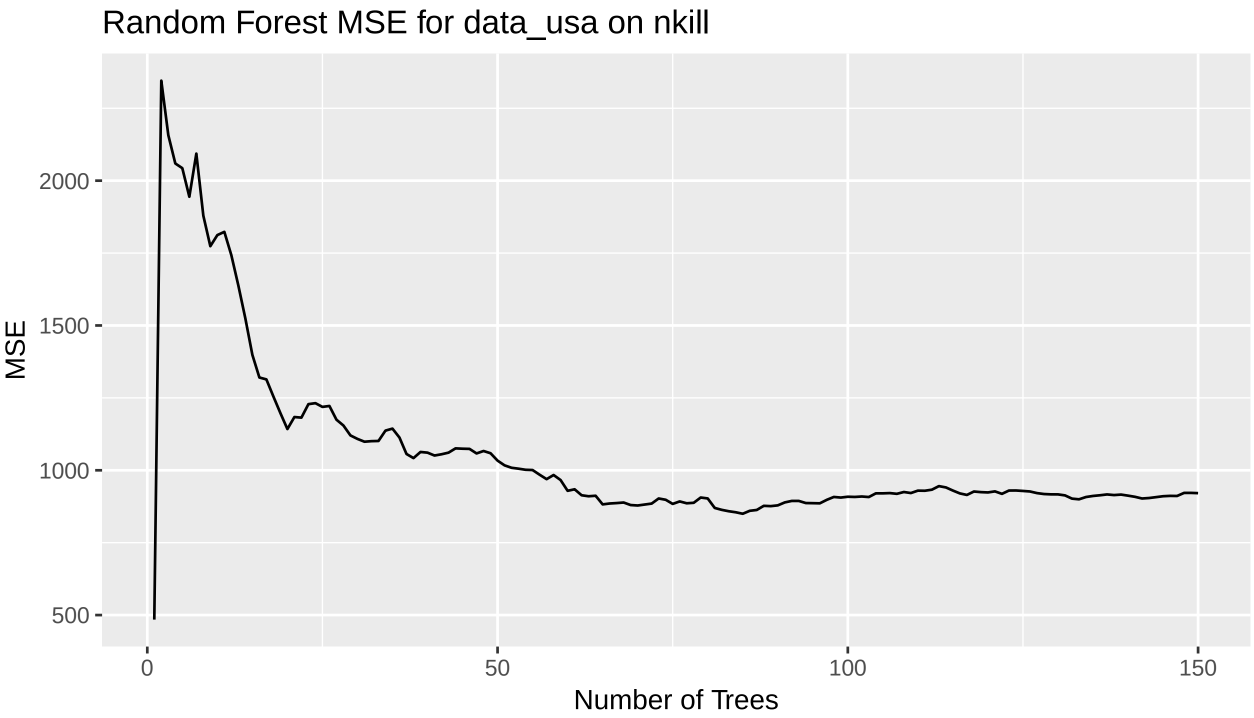




USA Final Anova output, LASSO cvm-by-lambda, and RF mse-by-ntree

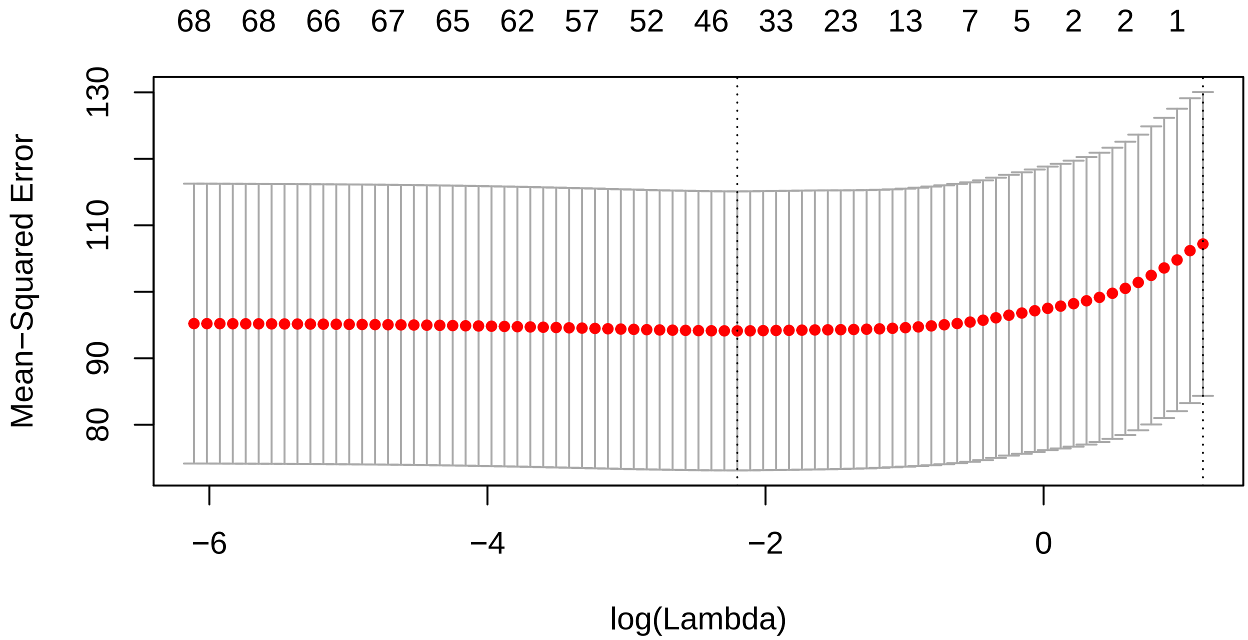
```
## Anova Table (Type II tests)
##
## Response: nkill
##                  Sum Sq Df F value    Pr(>F)
## vicinity          13018  1  20.681 5.674e-06 ***
## propextent_txt 2259873  4 897.538 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

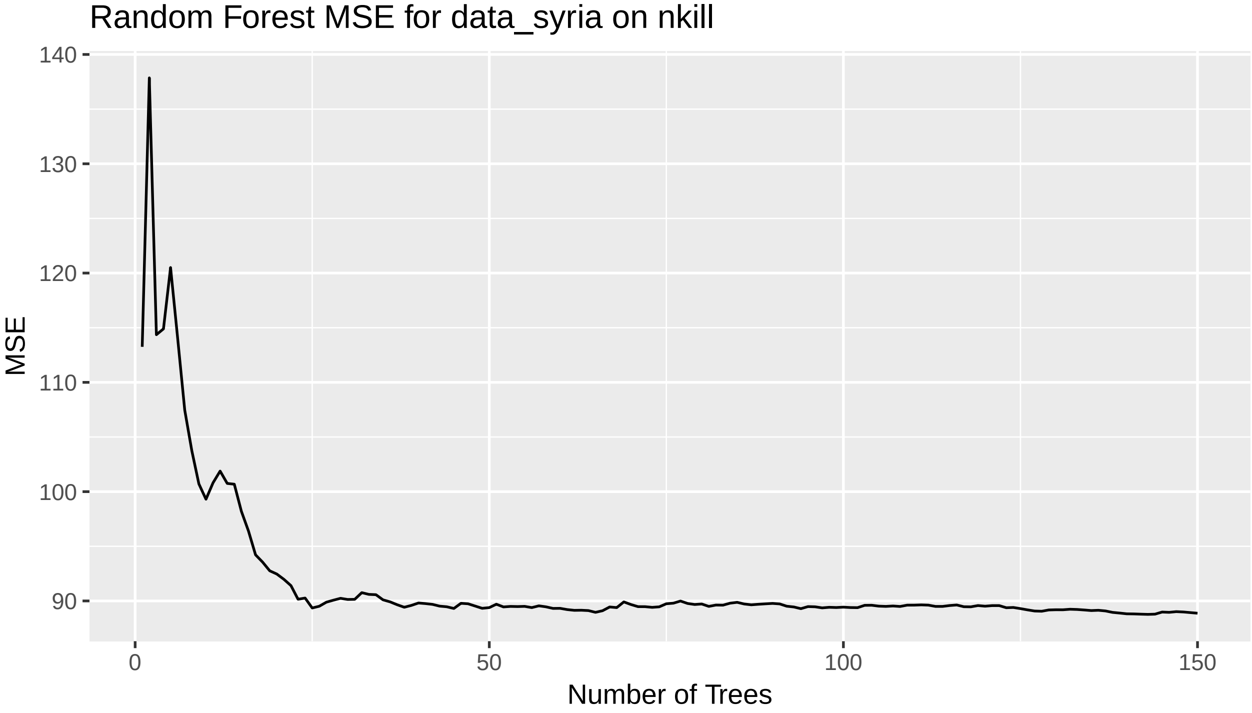




Syria Final Anova output, LASSO cvm-by-lambda, and RF mse-by-ntree

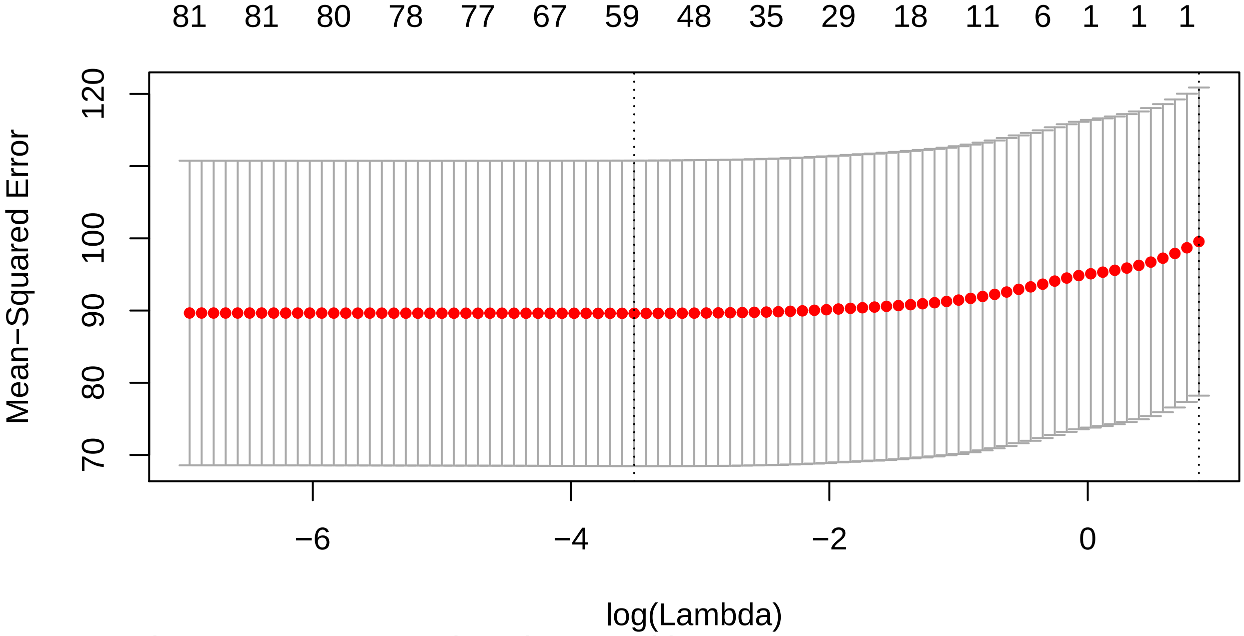
```
## Anova Table (Type II tests)
##
## Response: nkill
##                  Sum Sq Df F value    Pr(>F)
## crit3            1513.2  1 16.3806 5.439e-05 ***
## doubtterr        1372.7  1 14.8600 0.0001206 ***
## success          1639.8  1 17.7517 2.663e-05 ***
## suicide          2923.4  1 31.6470 2.194e-08 ***
## attacktype1_txt  3361.3  8  4.5485 1.731e-05 ***
## weapsubtype1_txt 5252.1 19  2.9925 1.554e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

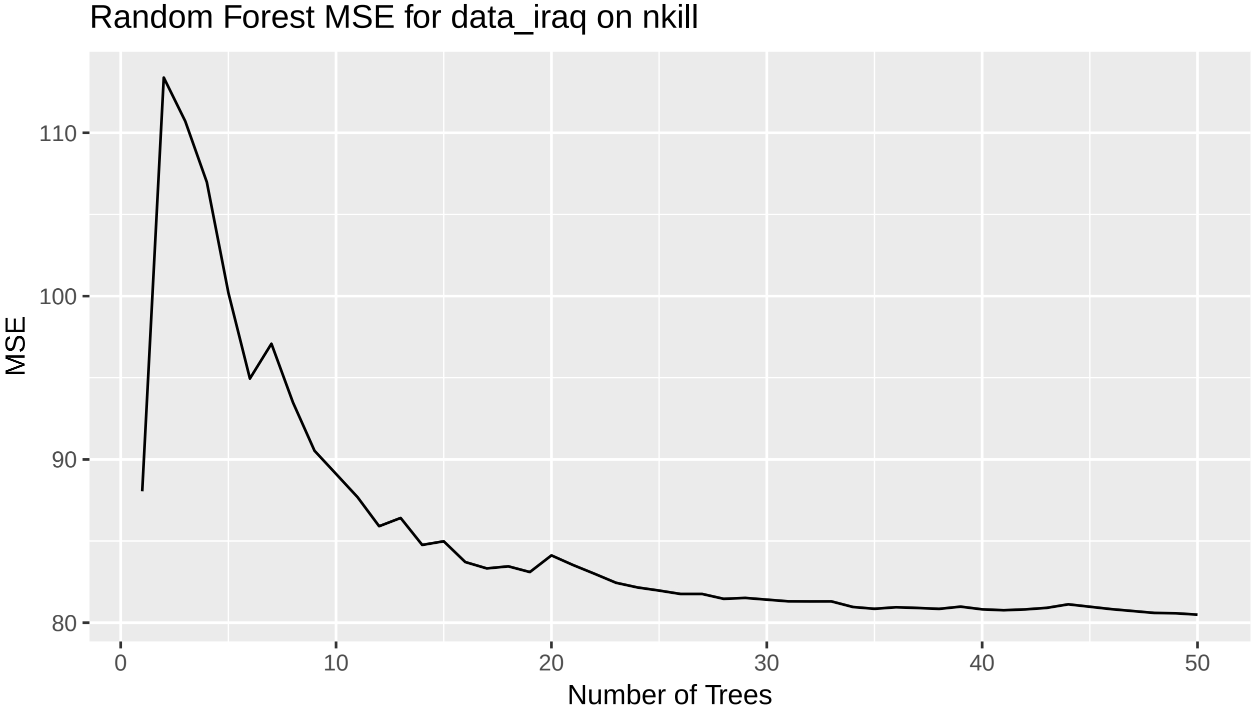




Iraq Final Anova output, LASSO cvm-by-lambda, and RF mse-by-ntree

```
## Anova Table (Type II tests)
##
## Response: nkill
##                    Sum Sq Df  F value    Pr(>F)
## iyear                1178  1  13.2655 0.0002709 ***
## extended             3299  1  37.1444 1.114e-09 ***
## longitude            2235  1  25.1611 5.313e-07 ***
## crit1                 630  1   7.0875 0.0077679 **
## success              2919  1  32.8626 1.001e-08 ***
## suicide             39879  1 448.9547 < 2.2e-16 ***
## attacktype1_txt   27355  8  38.4954 < 2.2e-16 ***
## targtype1_txt     15649 20   8.8089 < 2.2e-16 ***
## guncertain1        7480  1  84.2047 < 2.2e-16 ***
## weapsubtype1_txt  22108 28   8.8891 < 2.2e-16 ***
## propextent_txt    15622  3  58.6251 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

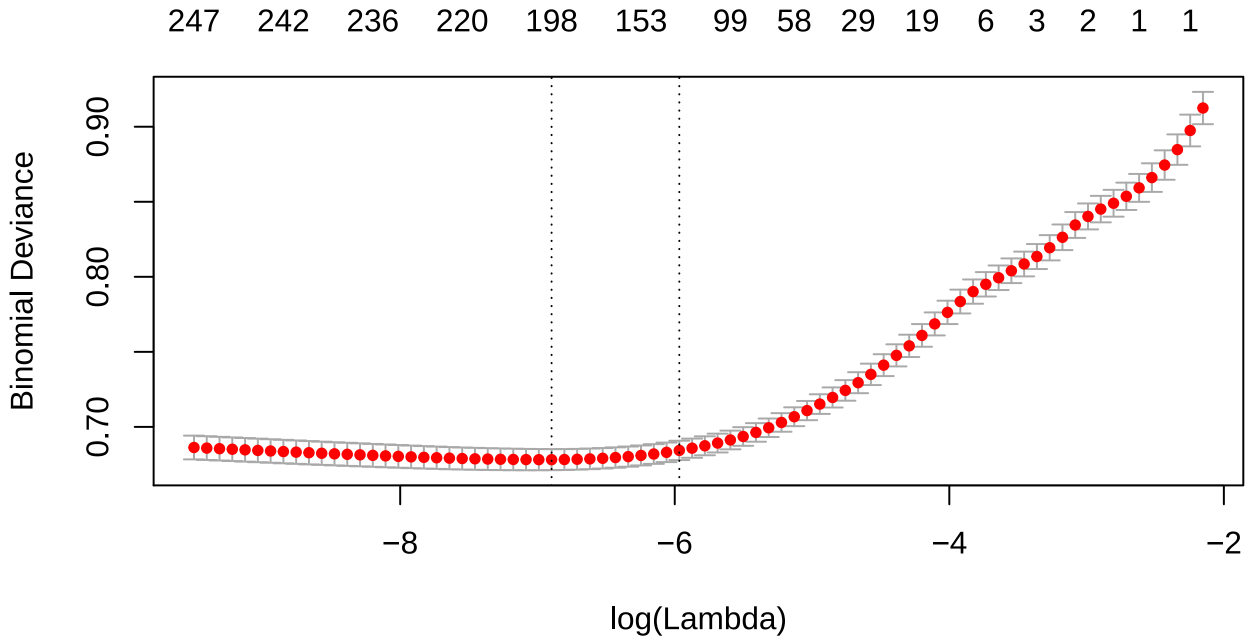




Data by Target

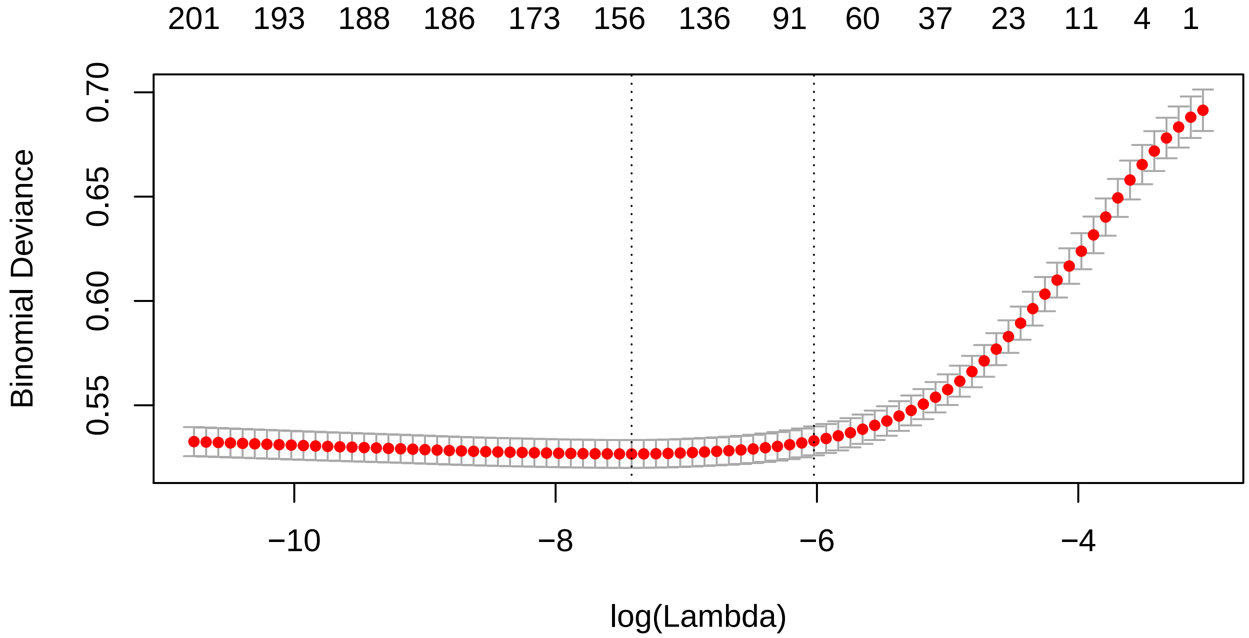
Government:

```
## Analysis of Deviance Table (Type II tests)
##
## Response: success
##                   LR Chisq Df Pr(>Chisq)
## iyear               183.97  1  < 2.2e-16 ***
## extended            147.85  1  < 2.2e-16 ***
## region_txt          167.59 11  < 2.2e-16 ***
## multiple             19.89  1  8.209e-06 ***
## suicide              26.61  1  2.492e-07 ***
## attacktype1_txt    2952.56  8  < 2.2e-16 ***
## targtype1_txt        61.82  1  3.772e-15 ***
## weaptype1_txt        57.53 11  2.654e-08 ***
## individual            8.00  1   0.004689 **
## weapsubtype1_txt    232.63 29  < 2.2e-16 ***
## nkill               674.31  1  < 2.2e-16 ***
## propextent_txt      437.96  4  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



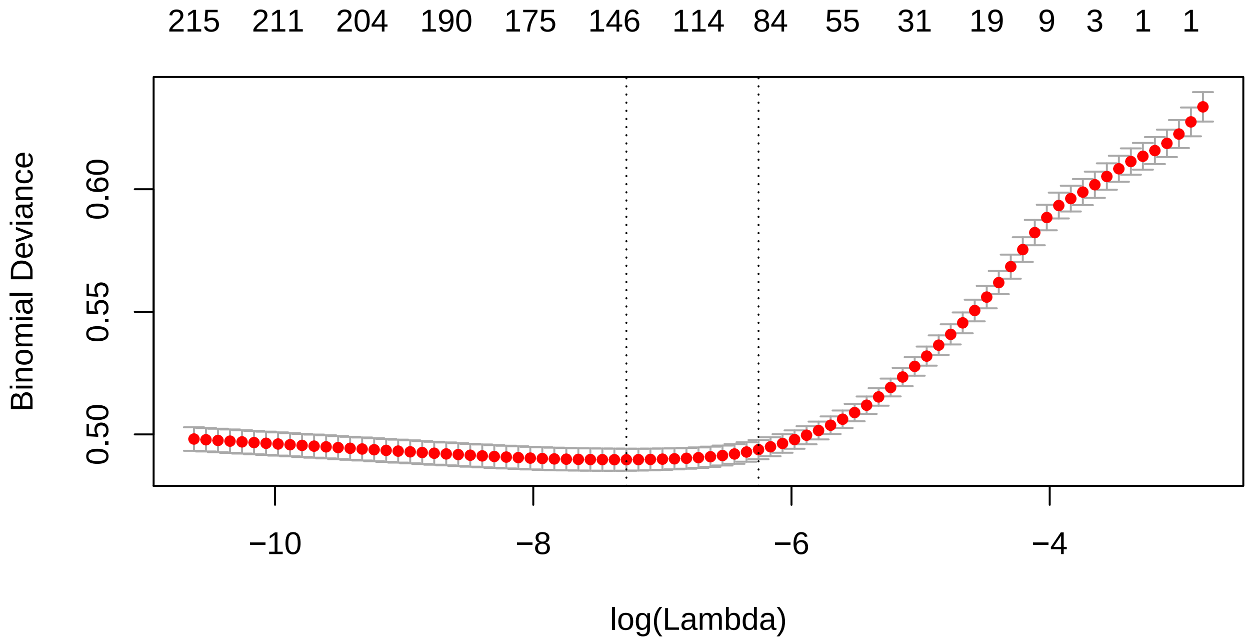
Military:

```
## Analysis of Deviance Table (Type II tests)
##
## Response: success
##                   LR Chisq Df Pr(>Chisq)
## iyear               137.03  1  < 2.2e-16 ***
## extended             16.20  1  5.698e-05 ***
## region_txt          183.94 11  < 2.2e-16 ***
## multiple             21.38  1  3.764e-06 ***
## attacktype1_txt     329.11  8  < 2.2e-16 ***
## individual           17.56  1  2.785e-05 ***
## weapsubtype1_txt   779.83 28  < 2.2e-16 ***
## nkill               25.00  1  5.741e-07 ***
## nwound             545.45  1  < 2.2e-16 ***
## propextent_txt     628.88  3  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



Police:

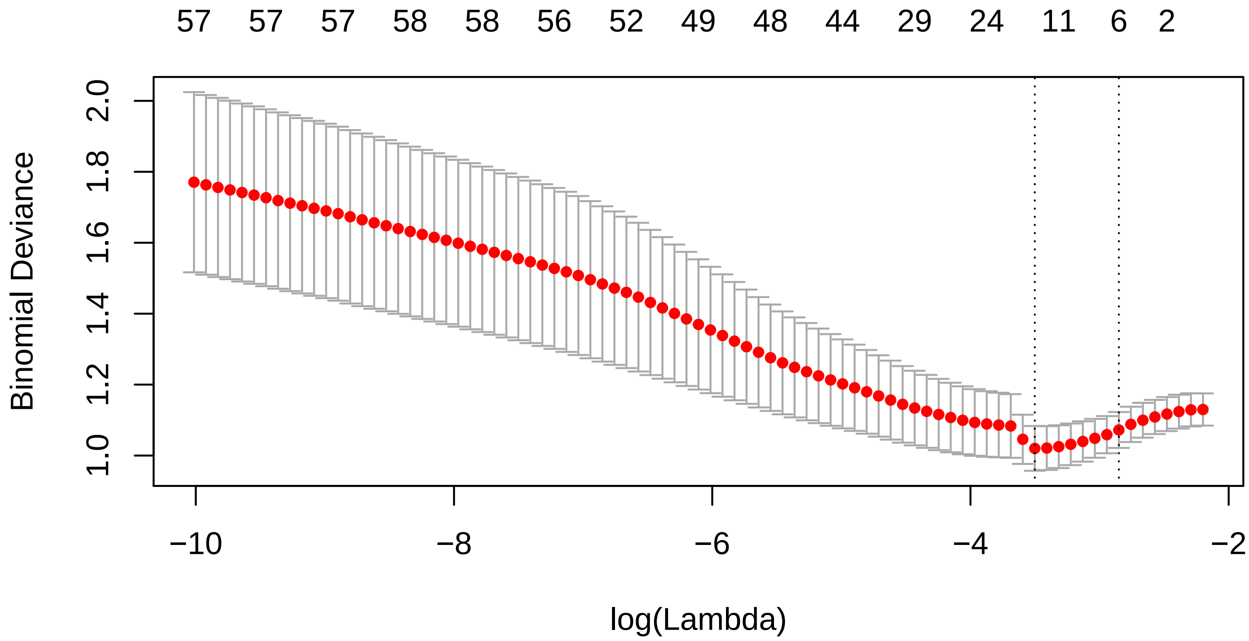
```
## Analysis of Deviance Table (Type II tests)
##
## Response: success
##                  LR Chisq Df Pr(>Chisq)
## latitude            68.02  1  < 2.2e-16 ***
## specificity         20.93  3  0.0001087 ***
## suicide             13.35  1  0.0002584 ***
## attacktype1_txt    992.08  8  < 2.2e-16 ***
## weapsubtype1_txt   663.04 30  < 2.2e-16 ***
## nkill              437.03  1  < 2.2e-16 ***
## nwound             149.77  1  < 2.2e-16 ***
## propextent_txt     539.98  3  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



Data by Type

Ransom:

```
## Analysis of Deviance Table (Type II tests)
##
## Response: hostkidoutcome_txt
##                  LR Chisq Df Pr(>Chisq)
## iyear              24.8993  1   6.04e-07 ***
## attacktype1_txt    13.6236  5   0.018186 *
## weapsubtype1_txt   29.4836 11   0.001909 **
## propextent_txt      9.3504  2   0.009324 **
## ransomamt          10.7086  1   0.001066 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```


57 57 57 58 58 56 52 49 48 44 29 24 11 6 2
Binomial Deviance
2.0
1.8
1.6
1.4
1.2
1.0
−10
−8
−6
−4
−2
log(Lambda)