Danny Vilela

By The Numbers:
Preparing for Tomorrow's Terrorism — Today

DANNY VILELA
New York University
April 27, 2016

Introduction

September 11th, 2001 marks one of the most devastating tragedies in recent American history. While the cultural, political, and overall social effects of the terrorist attacks have been studied and examined from numerous angles, we naturally look to examine the major casualties. As a rising global threat and constant hot topic for major news broadcasters, we hope to provide some transparency tomorrow by building a model that can predict the expected casualties of any given terrorist incident.

Using this data, domestic security researchers might be interested in focusing their time towards different high-profile targets and improving security at certain domestic checkpoints. Furthermore, the results of this prediction model would give emergency response units a better understanding of how to best prepare for future incidents of terrorism.

Problem

With the Global Terrorism Database, we can easily view individual terrorist incidents and numerous details per incident – including property damage, event date, and – most importantly – the number of people killed. We opt to filter our entire dataset on two conditions:

1. **We only look at the top 4 countries where terrorism is most frequent**. We do so because otherwise we would have a categorical variable with over 100 levels.
2. **We only look at the top 4 targets within said countries where terrorism has occured**. Again, we do so to limit the scope and complexity of our regression model.

With that, we begin by defining our problem: **Can we predict the number of deaths in a terrorist attack?**

We initially approach the question with hesitation: how can we possibly go about predicting the number of deaths in the chaotic aftermath of terrorism? We decide to utilize a two-way analysis

of variance model in order to define our prediction model, with the belief that the following effects will give us insight into predicting the number of deaths `n_killed` for any given terrorist attack:

1. the target of our terrorist attack: `targeted`
2. the country where our attack took place: `country`

We concede that our model is simple and that the results are likely to be limited in applications – and this would be a fair criticism. We choose to filter our dataset not only to allow for reasonable results, but also to better understand the historical climates and relationship with terrorism. Naturally, we also admit that there are numerous external factors (e.g. varying political landscapes, socio-cultural developments, etc.) that we won't be able to quantify and, therefore, include in our model. That said, we reason that our model will provide an interesting first-look into the goal of predicting the casualties inflicted by terrorism.

Model Definition

We can represent our earlier assertion through the regression model:

$$\texttt{n\_killed} = \mu + \texttt{targeted}_i + \texttt{country}_j + (\texttt{targeted} \cdot \texttt{country})_{ij} + \epsilon_{ijk}$$

Where `targeted` will act as our α effect, `country` will act as our β effect, and their product will represent our interaction effect $\alpha\beta$.

Data Sourcing and Cleaning

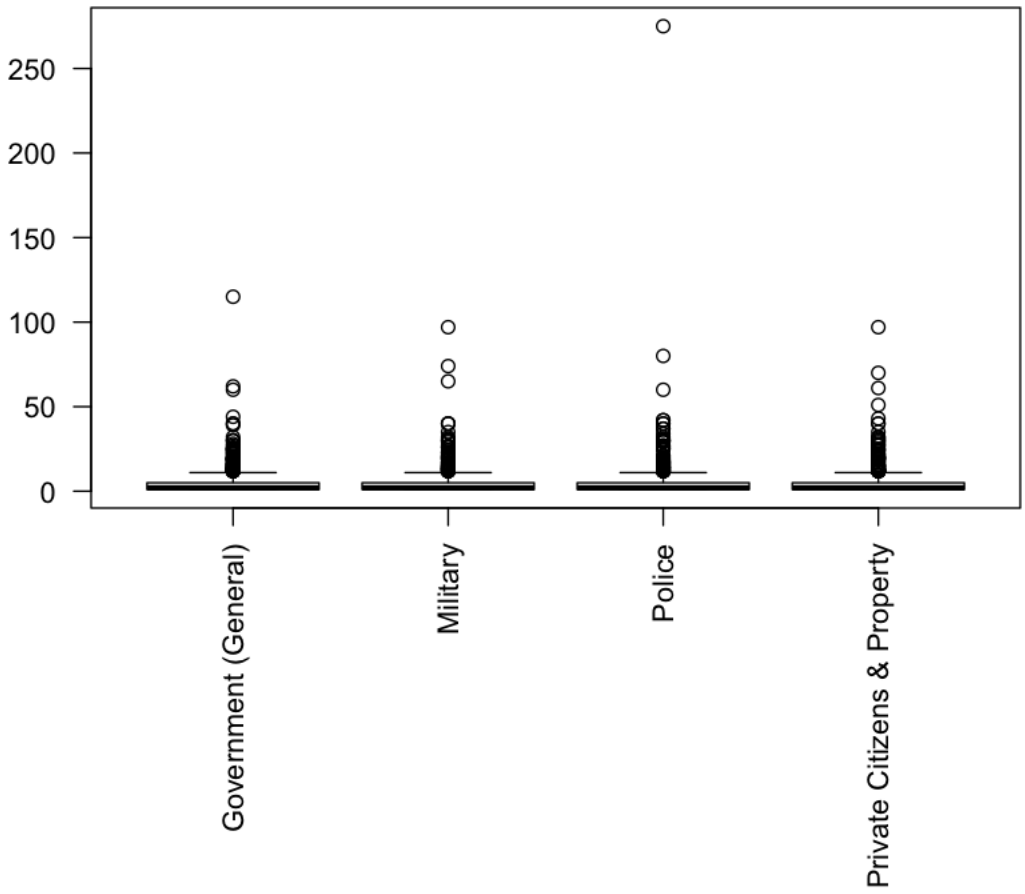
I sourced my data by using Enigma's dataset finder, which allowed me to find the Global Terrorism Database, maintained by the University of Maryland. Our raw dataset contained almost 100,000 observations with 125 variables per observation, so our first task was to reduce the complexity and choose what predictors would ultimately be included in our model.

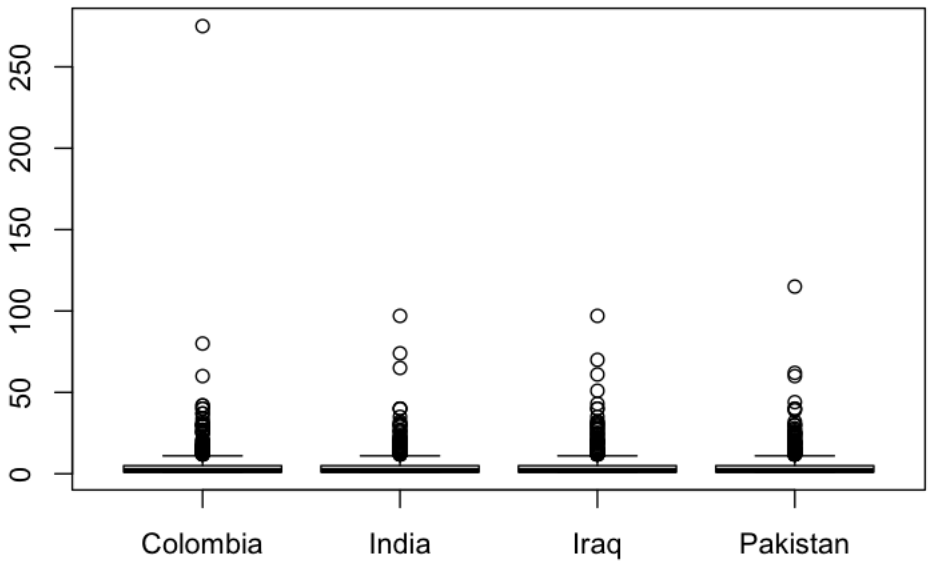
I first filtered my data by ensuring that our response variable `n_killed` was not `NA` and greater than 0. Then, I further filtered the dataset to only contain the top 4 `country` observations, resulting in Iraq, India, Afghanistan, and Pakistan remaining in our dataset. Lastly, we further filter our dataset to only contain the top 4 `targeted` locations, resulting in "Private Citizens & Property", "Police", "Government (General)", and "Military" remaining in our dataset.

Lastly, since our filtered-down dataset had over 9000 observations, I decided to split the dataset in half such that the first half was my training set, and the second half was my testing set. Unless otherwise explicitly stated, I will be using the first half (training set) for the data analysis.

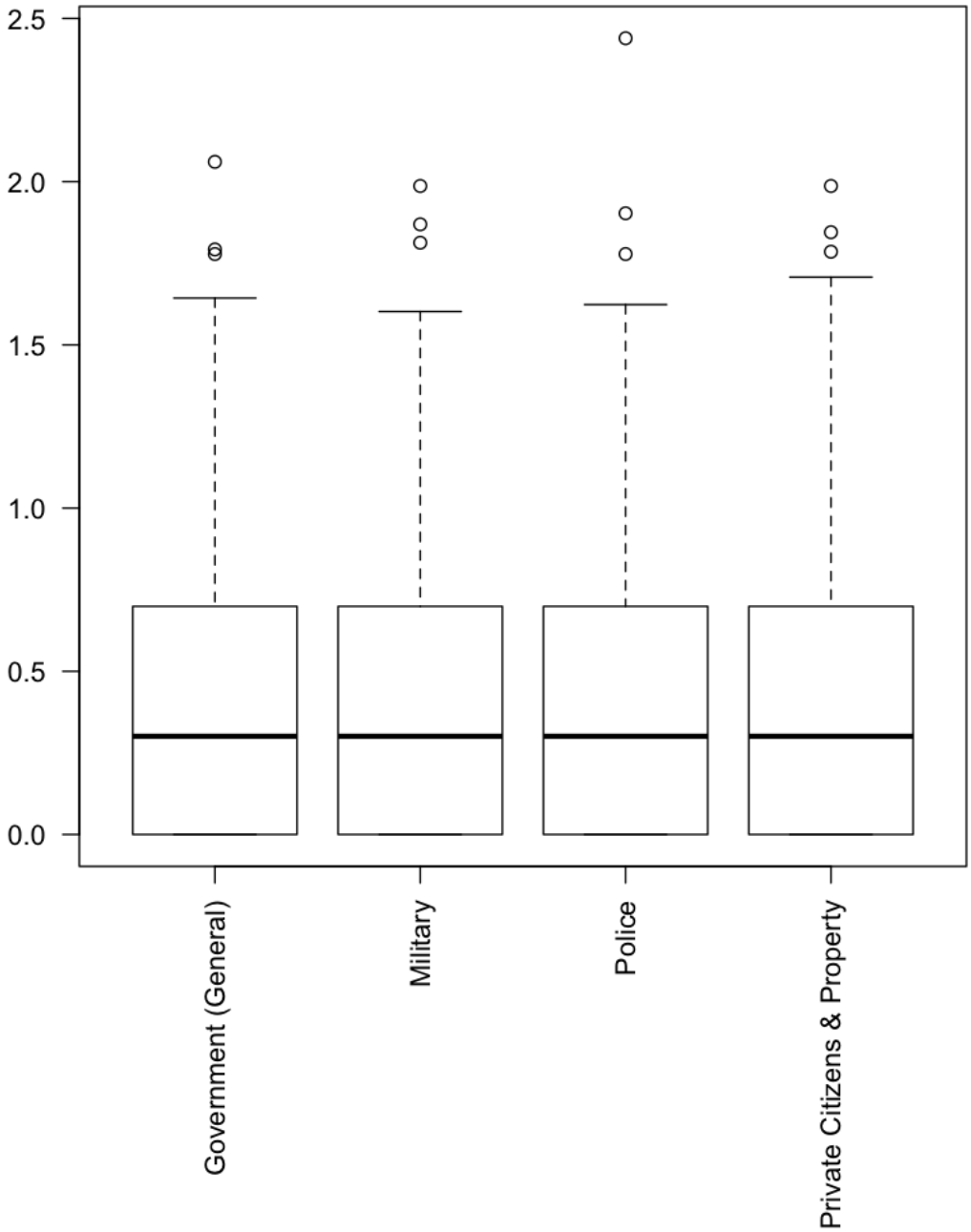
Exploratory Data Analysis

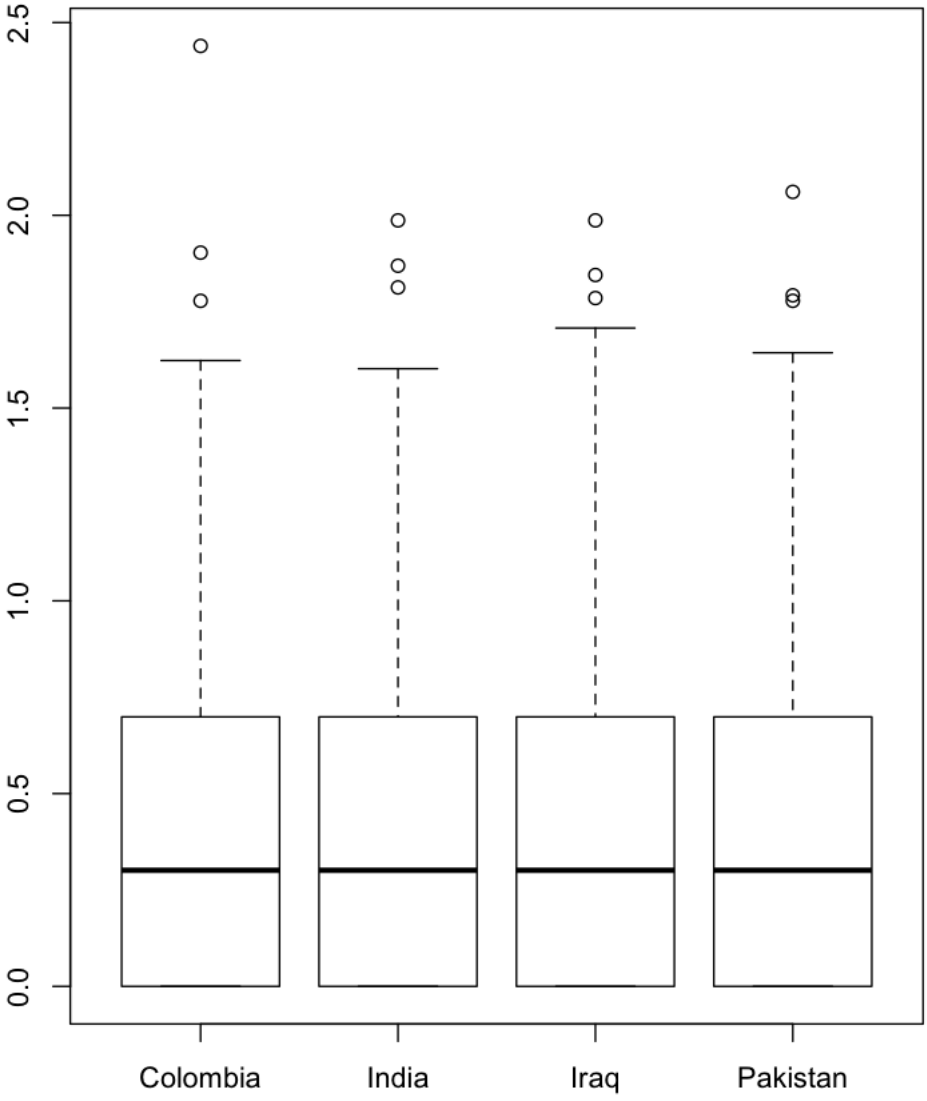
Let's view a side-by-side boxplot to see if there are `targeted` and `country` effects:





We see immediate evidence of nonconstant variance – it's evident that most terrorist attacks claim less lives than one outlier might lead us to believe. Exploring the dataset, we see that our outlier actually belongs to an incident in 1998 where the Revolutionary Armed Forces of Colombia (FARC) attacked government facilities and infrastructure. Given the dramatic variance, we naturally take the $\log_{10}$ of our `n_killed` response and note the results below:


2.5
2.0
1.5
1.0
0.5
0.0
Government (General)
Military
Police
Private Citizens & Property



Much nicer! Taking logs actually seems to have dramatically reduced the variance between both targets and countries. We update our model to represent the logged number of deaths per incident of terrorism:

$$\texttt{log.n\_killed} = \mu + \texttt{targeted}_i + \texttt{country}_j + (\texttt{targeted} \cdot \texttt{country})_{ij} + \epsilon_{ijk}$$

Data Analysis

First we check to see if we have at least one instance of each interaction:

```
> table(target, country)
                            Colombia India Iraq Pakistan
  Government (General)           451   296   23       81
  Military                       601   212   26       26
  Police                         564   610    6      153
  Private Citizens & Property    774   747   15      363
```

Success! We have enough data within our testing dataset such that we can account for our interaction effect between all levels of both of our categorical predictors. We do note that we're working with an unbalanced design since the number of observations at each factor-level combination are unequal.

Let's take a look at the output for our model object `log.deaths`:

```
> anova(log.deaths)

Analysis of Variance Table
Response: log.n_killed

                 +------+----------+-----------+-----------+------------+
                 |  Df  |  Sum Sq  |  Mean Sq  |  F value  |   Pr(>F)   |
+----------------+------+----------+-----------+-----------+------------+
| target         |    3 |   74.28  |  24.7587  |  178.949  | < 2.2e-16  |
| country        |    3 |   20.90  |   6.9669  |   50.355  | < 2.2e-16  |
| target:country |    9 |   12.54  |   1.3930  |   10.068  |  1.74e-15  |
| Residuals      | 4932 |  682.37  |   0.1384  +-----------+------------+
+----------------+------+----------+-----------+

> summary(log.deaths)

Call:
lm(formula = log.n_killed ~ target + country + target * country)

Residuals:
     Min       1Q   Median       3Q      Max
```

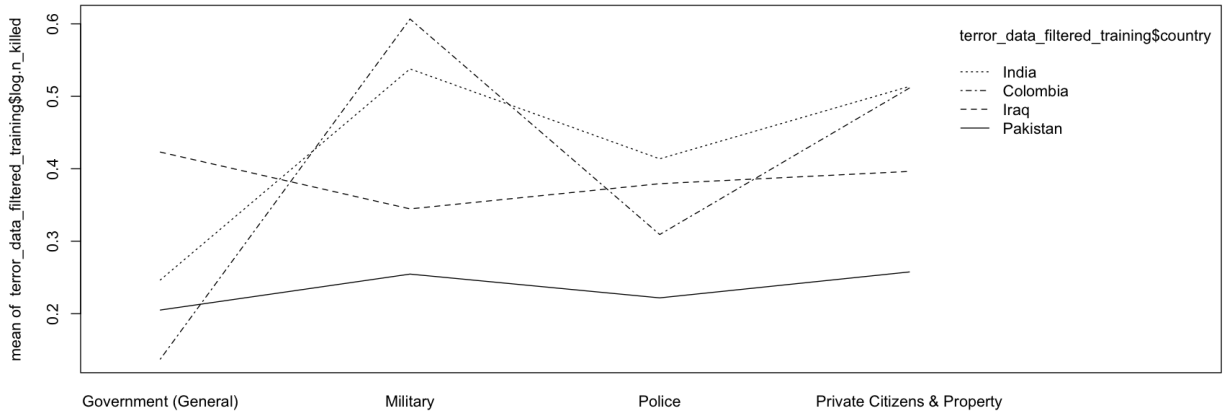
```
-0.60674 -0.25762 -0.06076  0.23836  2.30206

Coefficients:
                                    Estimate   Std. Error   t value   Pr(>|t|)
(Intercept)                        |  0.137278 |  0.017515  |  7.838  | 5.58e-15 |
Target
  Military                         |  0.469465 |  0.023173  | 20.259  |  < 2e-16 |
  Police                           |  0.171941 |  0.023497  |  7.318  | 2.93e-13 |
  Private Citizens & Property      |  0.374126 |  0.022035  | 16.979  |  < 2e-16 |
Country
  India                            |  0.109266 |  0.027824  |  3.927  | 8.72e-05 |
  Iraq                             |  0.285756 |  0.079513  |  3.594  | 0.000329 |
  Pakistan                         |  0.067746 |  0.044887  |  1.509  | 0.131297 |
Target * Country
  Military:India                   | -0.178132 |  0.040707  | -4.376  | 1.23e-05 |
  Police:India                     | -0.004754 |  0.035303  | -0.135  | 0.892876 |
  Private Citizens & Property:India | -0.106984 |  0.033737  | -3.171  | 0.001528 |
  Military:Iraq                    | -0.547876 |  0.108967  | -5.028  | 5.13e-07 |
  Police:Iraq                      | -0.215564 |  0.172125  | -1.252  | 0.210494 |
  Private Citizens & Property:Iraq | -0.400818 |  0.125398  | -3.196  | 0.001400 |
  Military:Pakistan                | -0.419839 |  0.086985  | -4.827  | 1.43e-06 |
  Police:Pakistan                  | -0.155150 |  0.056254  | -2.758  | 0.005836 |
  Private Citizens & Property:Pakistan | -0.321530 |  0.050742  | -6.337  | 2.56e-10 |

Residual standard error: 0.372 on 4932 degrees of freedom
Multiple R-squared: 0.1363,    Adjusted R-squared:  0.1337
F-statistic:  51.9 on 15 and 4932 DF,  p-value: < 2.2e-16

Regression Equation:

log.n_killed = 0.137278 + 0.469464552 Military + 0.171940807 Police
                          + 0.374126067 Private Citizens & Property
                       + 0.109266490 India + 0.285755658 Iraq + 0.067746488 Pakistan
                       - 0.178132102 Military:India - 0.004754371 Police:India
                          - 0.106983908 Private Citizens & Property:India
                       - 0.547875545 Military:Iraq - 0.215563716 Police:Iraq
                          - 0.400818350 Private Citizens & Property:Iraq
                       - 0.419839351 Military:Pakistan - 0.155149860 Police:Pakistan
```

```
- 0.321530420 Private Citizens & Property:Pakistan
```

How do we go about interpreting our regression model? We note that main effects don't hold much weight in a model with an interaction effect. Here, a `target` and `country` interaction effect would be saying that there are a set of effects that differentiate `country` subgroups and `target`subgroups, meaning they *are* fundamentally different from each other. We see that the interaction effect *is* statistically highly statistically significant with a p-value much lower than 0.001.

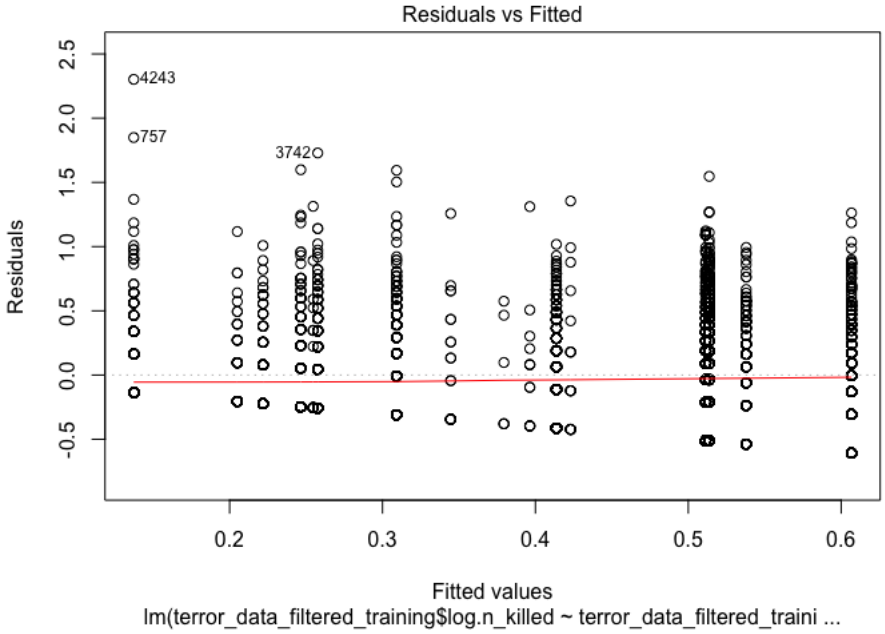
We also note the low R^2 of 13.64% and R^2_a of 13.37%, meaning that our model's predictors are only accounting for 13.64% of the variance in `log.n_killed`. This is less than ideal (but meaningless in context), so we move on to exploring shortcomings with our current model to understand and build our next model.

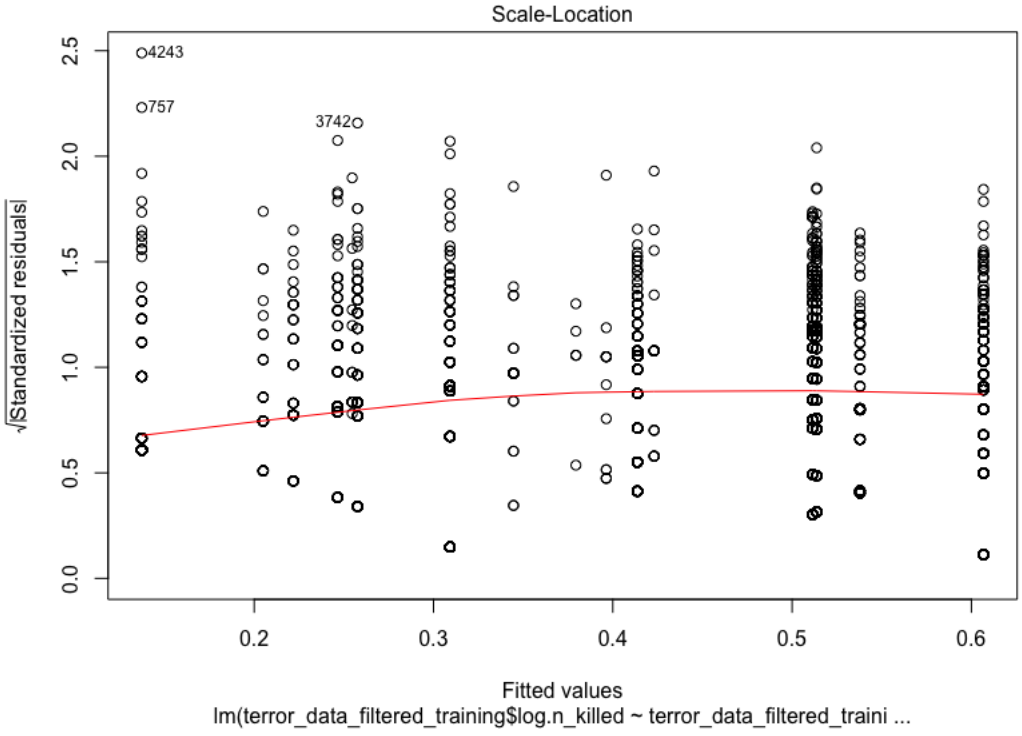
Let's look at our interaction plot to get a preliminary look into how our interaction effect might be at work:

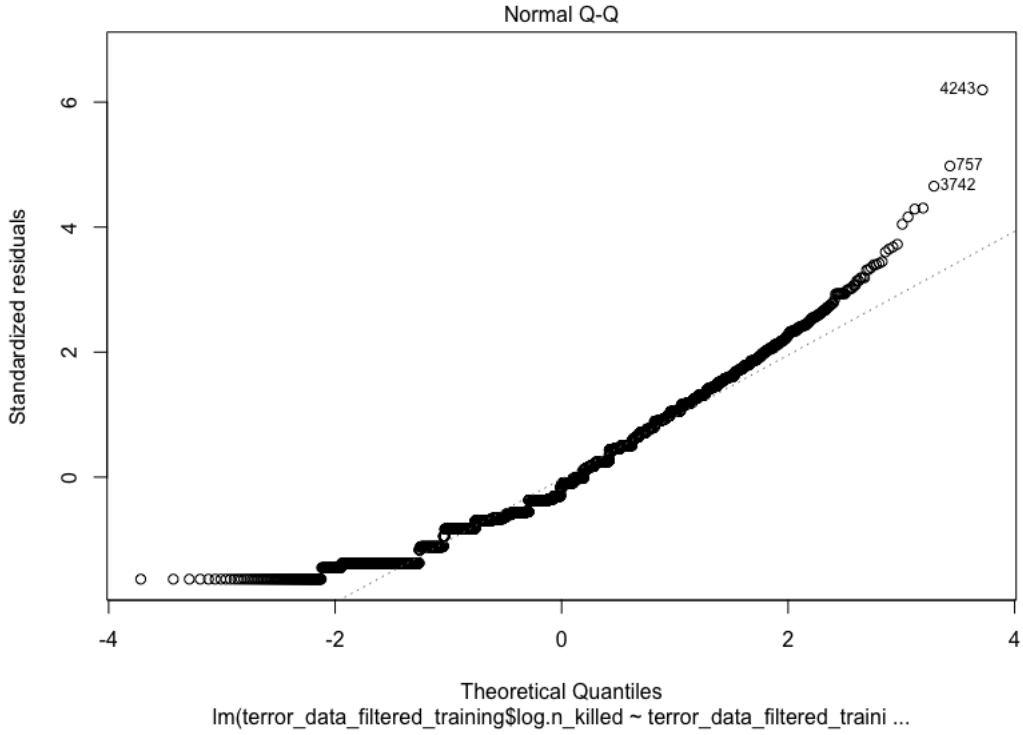


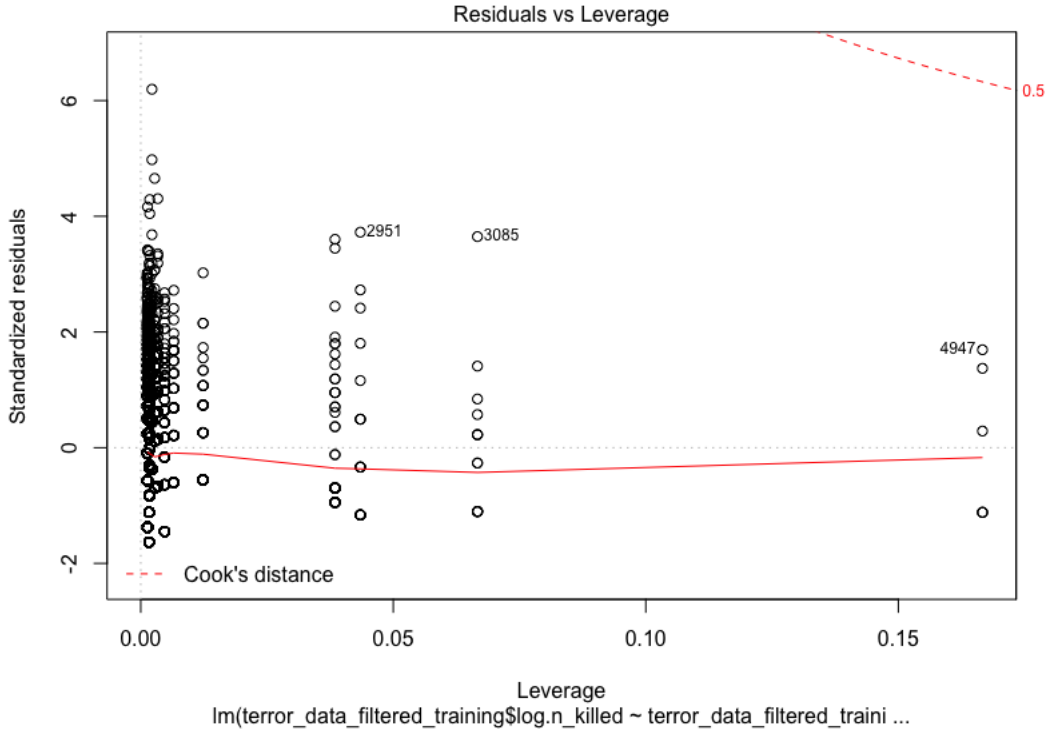
We see some parallel between India and Colombia's interaction lines, with both generally increasing, decreasing, then increasing again. Pakistan follows a similar pattern, but to a much lighter magnitude. We see that Iraq is unique in that its interaction effects seem to noticeably, albeit lightly, invert the general interaction effect. From this plot, we would be able to conclude that Military locations were at the most risk of terrorist deaths within these countries.

Let's look at the diagnostic plots:


Residuals vs Fitted
Residuals
Fitted values
lm(terror_data_filtered_training$log.n_killed ~ terror_data_filtered_traini ...
4243
757
3742
-0.5
0.0
0.5
1.0
1.5
2.0
2.5
0.2
0.3
0.4
0.5
0.6


Scale-Location
√|Standardized residuals|
Fitted values
lm(terror_data_filtered_training$log.n_killed ~ terror_data_filtered_traini ...
4243
757
3742
0.0
0.5
1.0
1.5
2.0
2.5
0.2
0.3
0.4
0.5
0.6


Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(terror_data_filtered_training$log.n_killed ~ terror_data_filtered_traini ...
4243
757
3742


Residuals vs Leverage
Standardized residuals
Leverage
lm(terror_data_filtered_training$log.n_killed ~ terror_data_filtered_traini ...
Cook's distance
0.5
2951
3085
4947

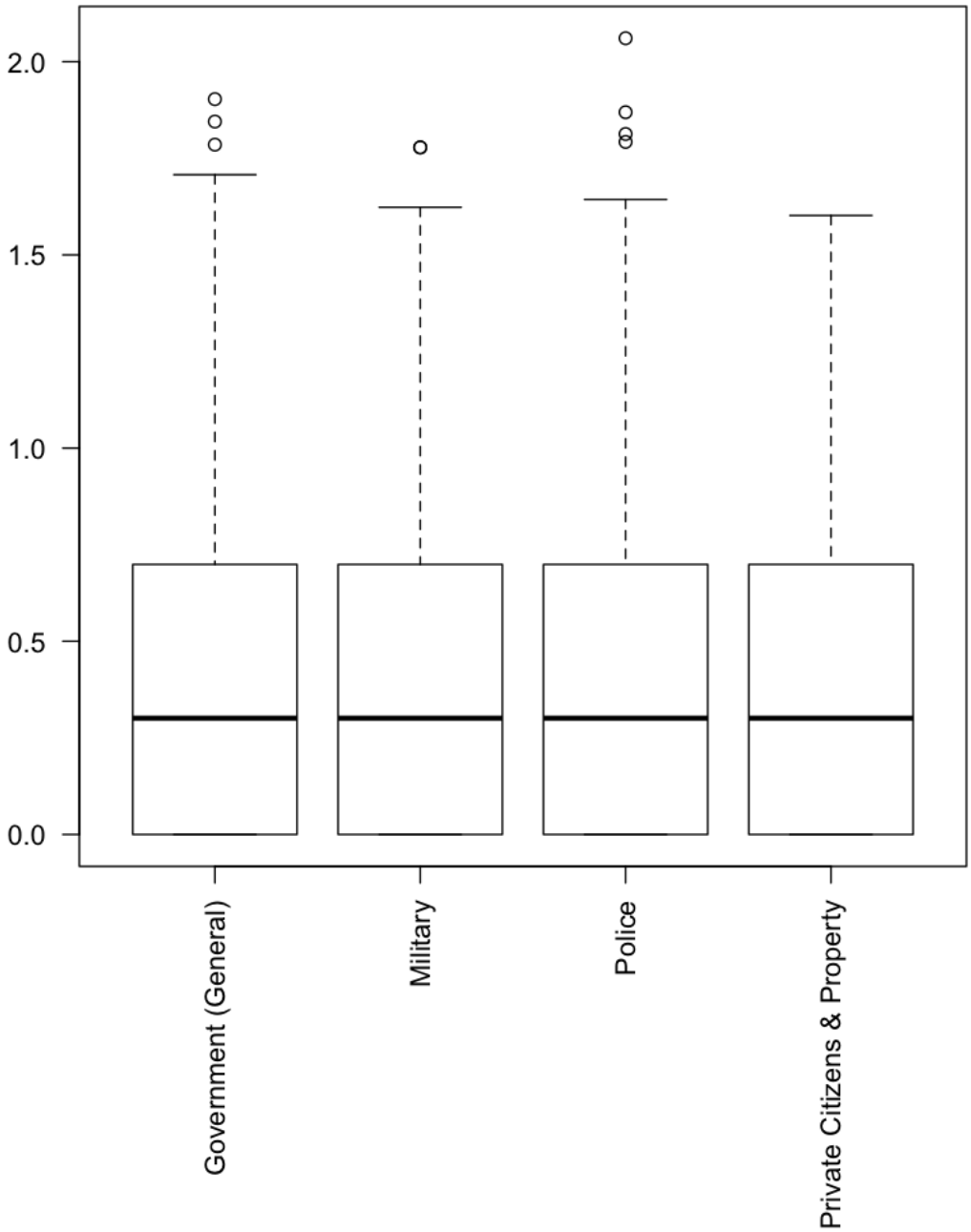
Through our standardized residual plot, we can see three distinct outliers at rows 757, 3742, and 4243. These rows map to the following observations:

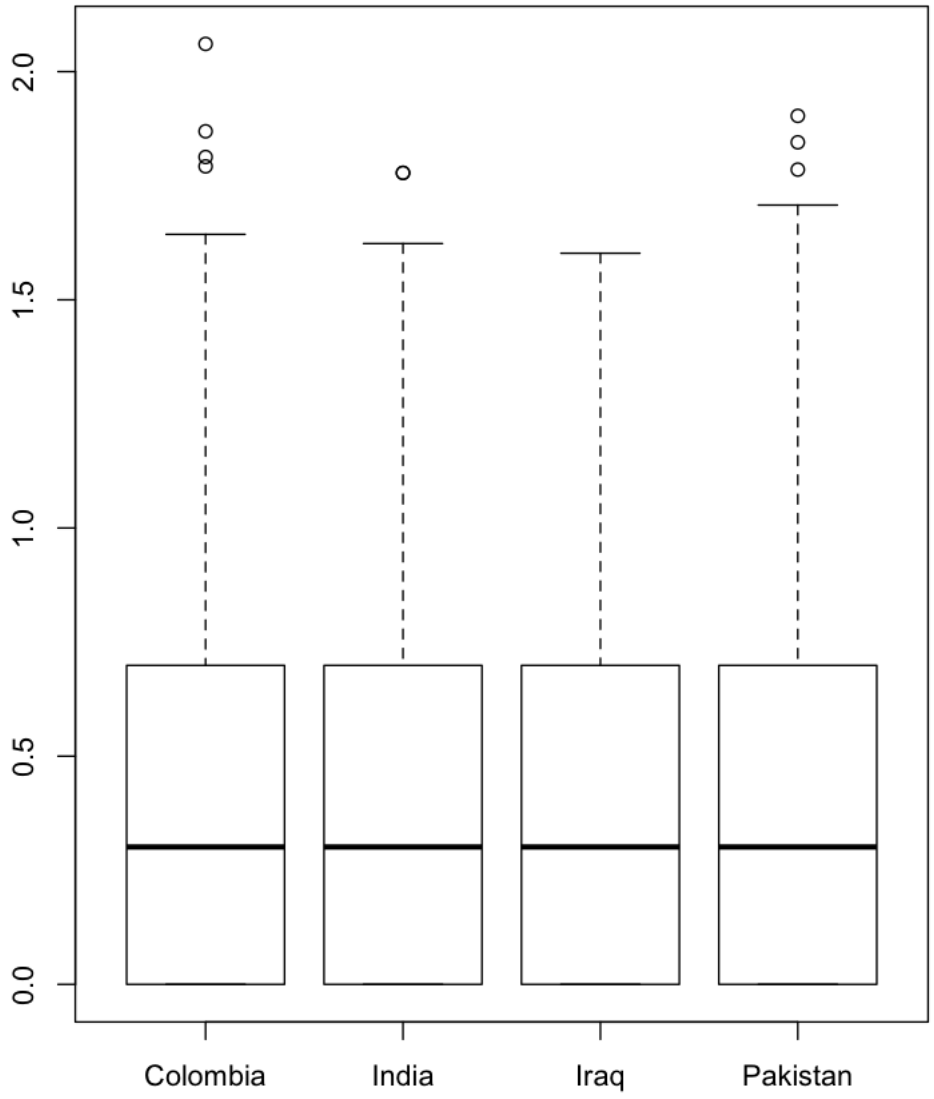
	year	target	country	n_killed	log.n_killed
[757]	1985	Government (General)	Colombia	97	1.986772
[3742]	1996	Private Citizens & Property	Pakistan	97	1.986772
[4243]	1998	Government (General)	Colombia	275	2.439333

Evidently, these outliers are among the top-leading instances of terrorism when ranked by number of casualties *for what they were – members in their particular groups* (coincidentally, however, these observation are leading the overall number of deaths count – regardless of groupings). These incidents might be having a crucial effect on our data, so we opt to remove said outlier observations to understand how they influence our model.

Data Analysis: Removed Outliers

Let's plot our side-by-side boxplots yet again, but after removing the outliers noted in the `log.n_killed` residual plots:


2.0
1.5
1.0
0.5
0.0
Government (General)
Military
Police
Private Citizens & Property



Not much of a change! As we noted earlier, taking the log of `n_killed` did us a big favor and got us most of the way there. Turning to the output for our model object `log.deaths.nout`:

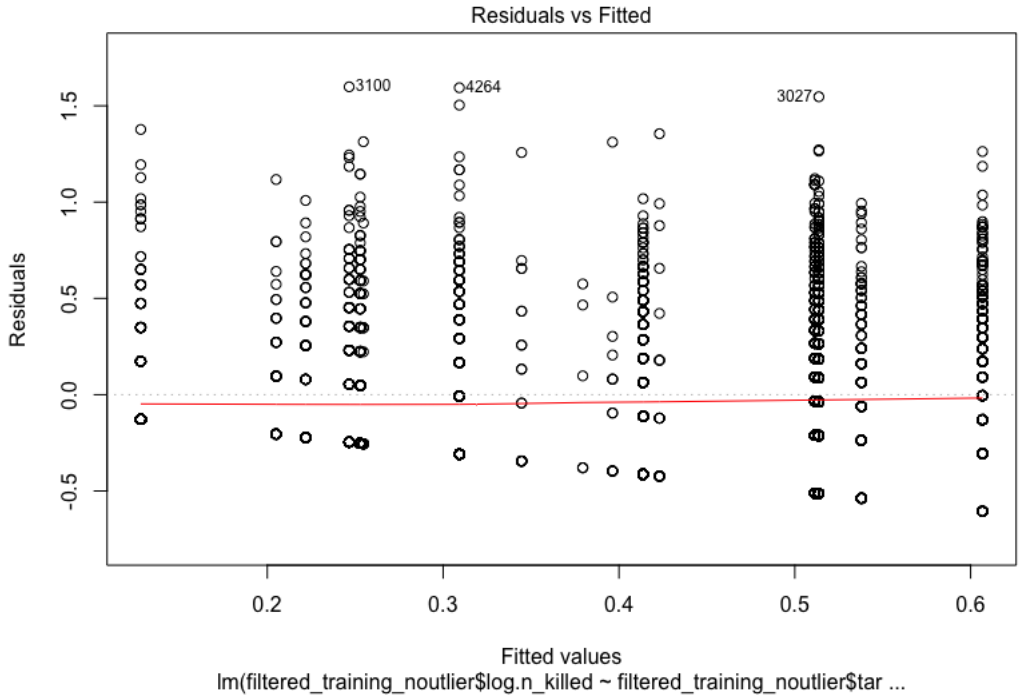
```
> anova(log.deaths.nout)

Analysis of Variance Table
Response: log.n_killed
```

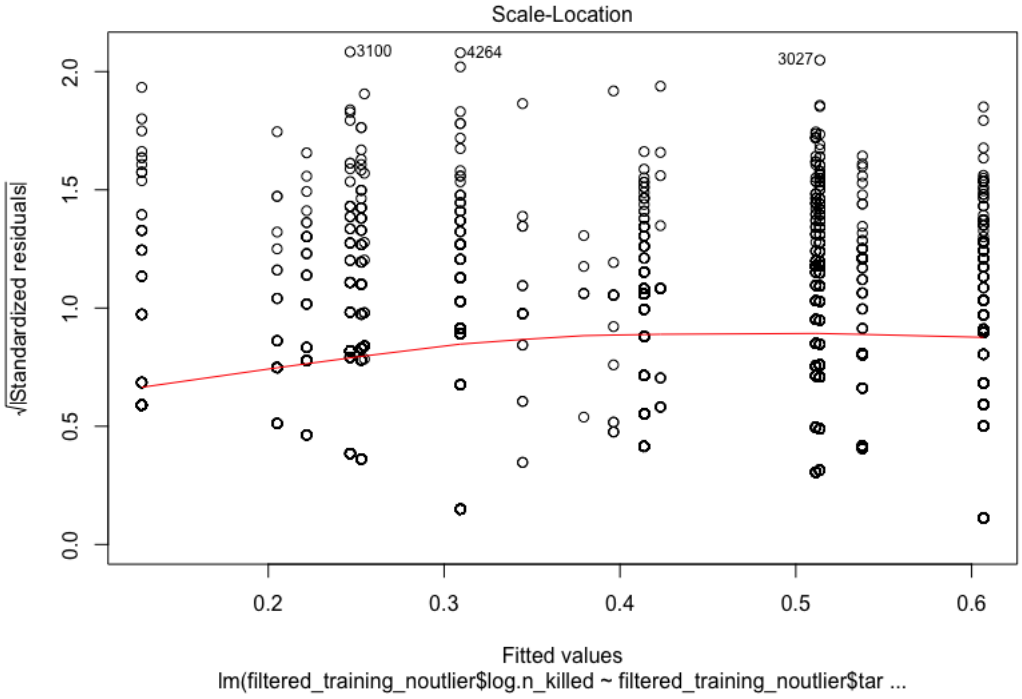
```
                +-------+---------+-----------+-----------+------------+
                |  Df   | Sum Sq  |  Mean Sq  |  F value  |   Pr(>F)   |
+---------------+-------+---------+-----------+-----------+------------+
| target        |    3  |  75.75  |  25.2499  |  185.586  |  < 2.2e-16 |
| country       |    3  |  21.38  |   7.1252  |   52.370  |  < 2.2e-16 |
| target:country |    9  |  13.18  |   1.4649  |   10.767  |  < 2.2e-16 |
| Residuals     | 4929  | 670.61  |   0.1361  +-----------+------------+
+---------------+-------+---------+-----------+

> summary(log.deaths.nout)

Call:
lm(formula = log.n_killed ~ target + country + target * country)

Residuals:
     Min       1Q   Median       3Q      Max
-0.60674 -0.25284 -0.06076  0.23836  1.59855

Coefficients:
                                    Estimate   Std. Error  t value   Pr(>|t|)
(Intercept)                        |  0.12803 |  0.01741  |  7.355  | 2.22e-13 ***
Target
  Military                         |  0.47871 |  0.02301  | 20.806  | < 2e-16  ***
  Police                           |  0.18119 |  0.02333  |  7.767  | 9.74e-15 ***
  Private Citizens & Property      |  0.38337 |  0.02188  | 17.520  | < 2e-16  ***
Country
  India                            |  0.11851 |  0.02762  |  4.291  | 1.81e-05 ***
  Iraq                             |  0.29500 |  0.07886  |  3.741  | 0.000185 ***
  Pakistan                         |  0.07699 |  0.04453  |  1.729  | 0.083854
Target * Country
  Military:India                   | -0.18738 |  0.04038  | -4.640  | 3.57e-06 ***
  Police:India                     | -0.01400 |  0.03503  | -0.400  | 0.689394
  Private Citizens & Property:India | -0.11623 |  0.03348  | -3.472  | 0.000521 ***
  Military:Iraq                    | -0.55712 |  0.10806  | -5.156  | 2.63e-07 ***
  Police:Iraq                      | -0.22481 |  0.17069  | -1.317  | 0.187881
  Private Citizens & Property:Iraq | -0.41006 |  0.12436  | -3.297  | 0.000982 ***
  Military:Pakistan                | -0.42909 |  0.08627  | -4.974  | 6.78e-07 ***
  Police:Pakistan                  | -0.16440 |  0.05580  | -2.946  | 0.003230  **
```

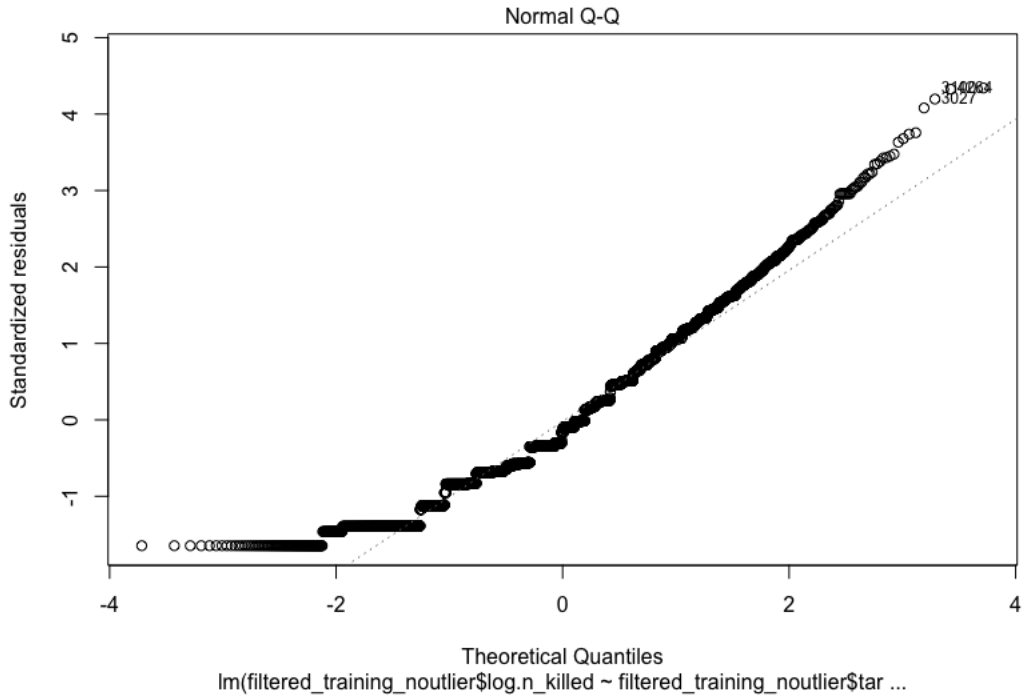
```
  Private Citizens & Property:Pakistan | -0.33555 |  0.05034  | -6.665  | 2.93e-11 ***

Residual standard error: 0.3689 on 4929 degrees of freedom
Multiple R-squared: 0.1413, Adjusted R-squared: 0.1386
F-statistic: 54.05 on 15 and 4929 DF,  p-value: < 2.2e-16
```

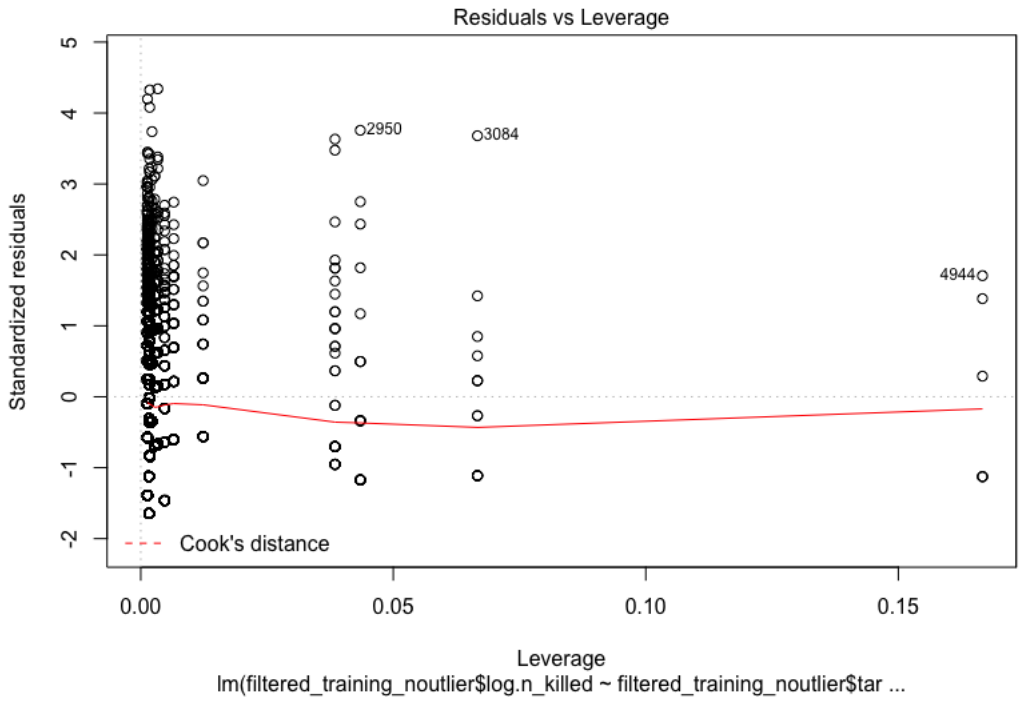
Looking at our output, we see that the interaction effects between the `target` Police and the all listed countries are the only interaction effects with a non-highly significant ("***") p-value. Otherwise, however, our output is similar to that of our previous regression model with the exception of a slight increase in our model's R^2 – from 13.63% to 14.13% (meaningless in context). Both models were highly statistically significant at the 0.001 level.

We look to the diagnostic plots for the residuals:



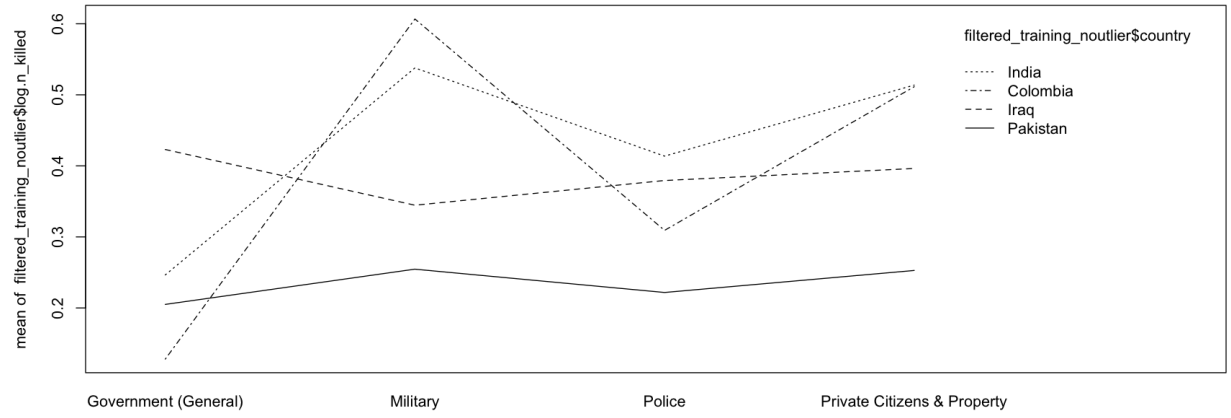

Scale-Location
3100
4264
3027
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
0.2
0.3
0.4
0.5
0.6
Fitted values
lm(filtered_training_noutlier$log.n_killed ~ filtered_training_noutlier$tar ...


Normal Q-Q
3027
Standardized residuals
-1
0
1
2
3
4
5
-4
-2
0
2
4
Theoretical Quantiles
lm(filtered_training_noutlier$log.n_killed ~ filtered_training_noutlier$tar ...



Out normal q-q plot is better, but not great. We see a long tail at the low end, reflecting an inherent asymmetry in our data: many instances of terrorism do not claim any lives, and most do not claim more than 2 or 3. While this gives us some peace of mind, it does not (yet) give us a spectacular model for representing our data. We still see evidence of non-constant variance, and face the question of how to account for such variance in our regression model.

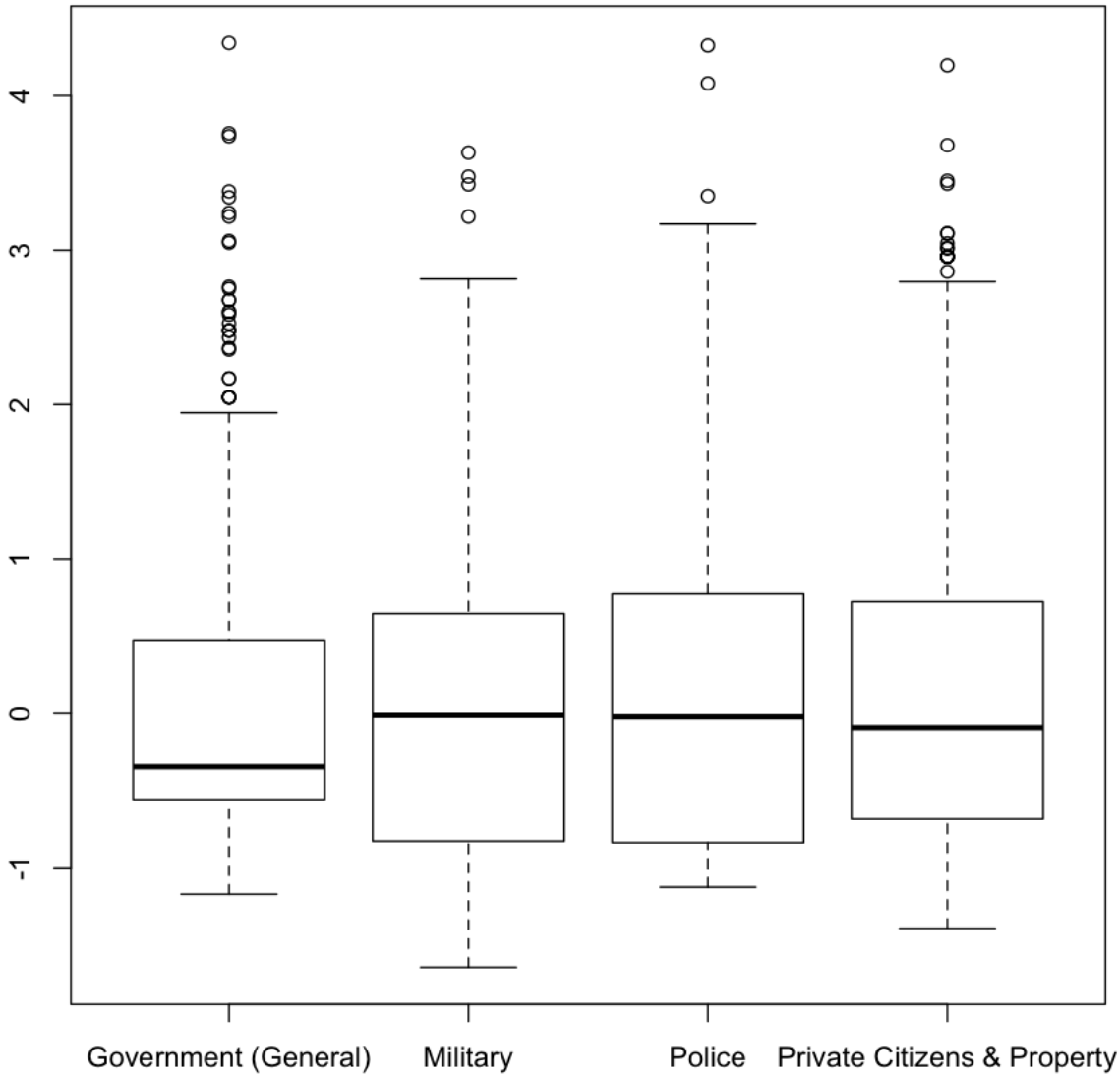
We look to our interaction plot to see if our interaction effect might tell us anything about our subgroups:

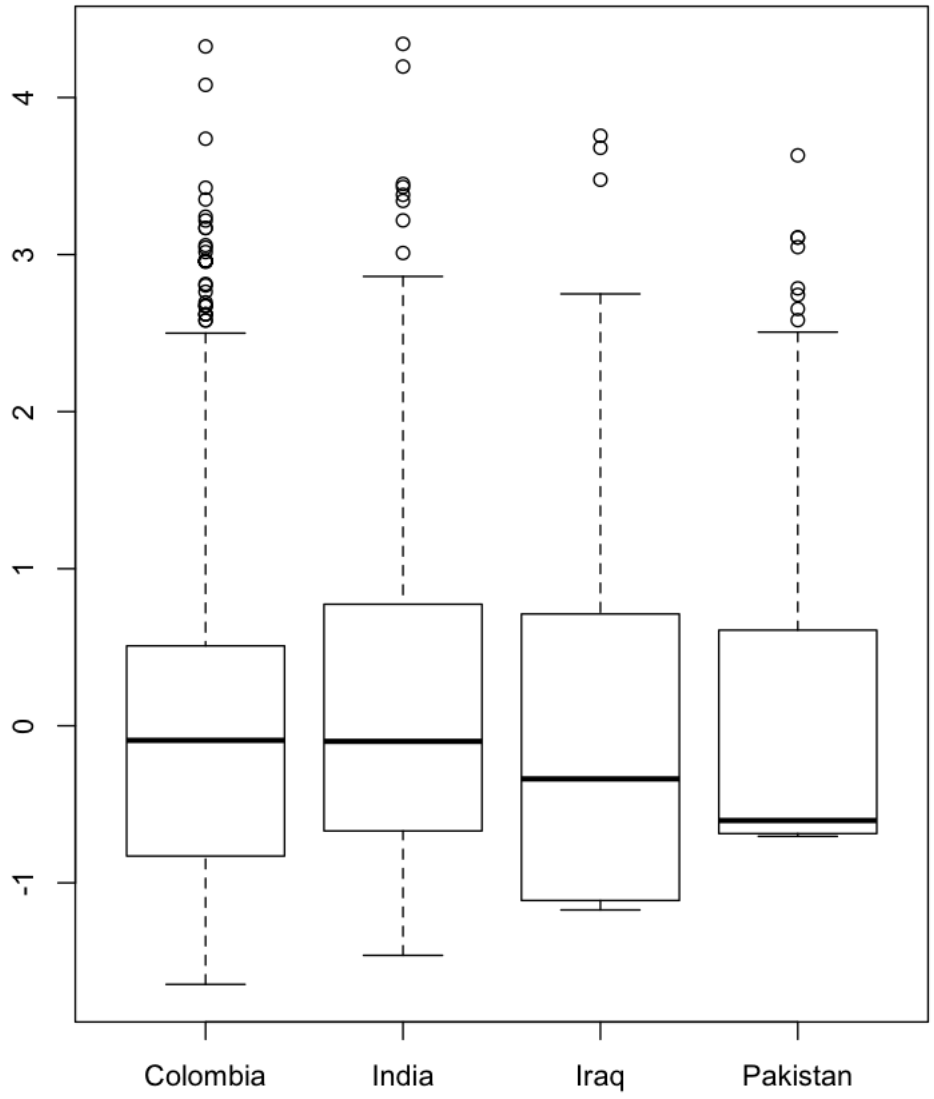


For the most part – no surprise – our interaction plot with removed outliers is the same. India

and Colombia share the same general slopes, with Colombia's `n_killed` varying greatly between each `target` group whereas Iraq and Pakistan's slopes are much closer to flattening.

Looking to the residual plots:





we see that they look much better! There's still a noticeable right tail, and some persisting evidence of nonconstant variance. We look to Levene's test to see if nonconstant variance is indicated:

```
> Anova(levene, type = 3)

Anova Table (Type III tests)
Response: abs(std.resd)
```

```
                +---------+------+----------+-----------+
                |  Sum Sq |  Df  | F value  |  Pr(>F)   |
+---------------+---------+------+----------+-----------+
| (Intercept)   |  122.17 |    1 | 379.4796 | < 2.2e-16 |
| target        |   39.79 |    3 |  41.1990 | < 2.2e-16 |
| country       |   17.30 |    3 |  17.9124 | 1.469e-11 |
| target:country |    9.57 |    9 |   3.3015 | 0.0005041 |
| Residuals     | 1586.83 | 4929 +----------+-----------+
+---------------+---------+------+
```

Interesting! We see that the interaction effect in the Levene's test is still highly statistically significant, meaning we should not omit the interaction effect from our model. Therefore, everything we've had so far has been in support of the notion that we *still* have nonconstant variance despite taking the log of our response variable, `n_killed`. We face the question: how do we account for the heteroscedasticity?

Enter: Weighted Least Squares

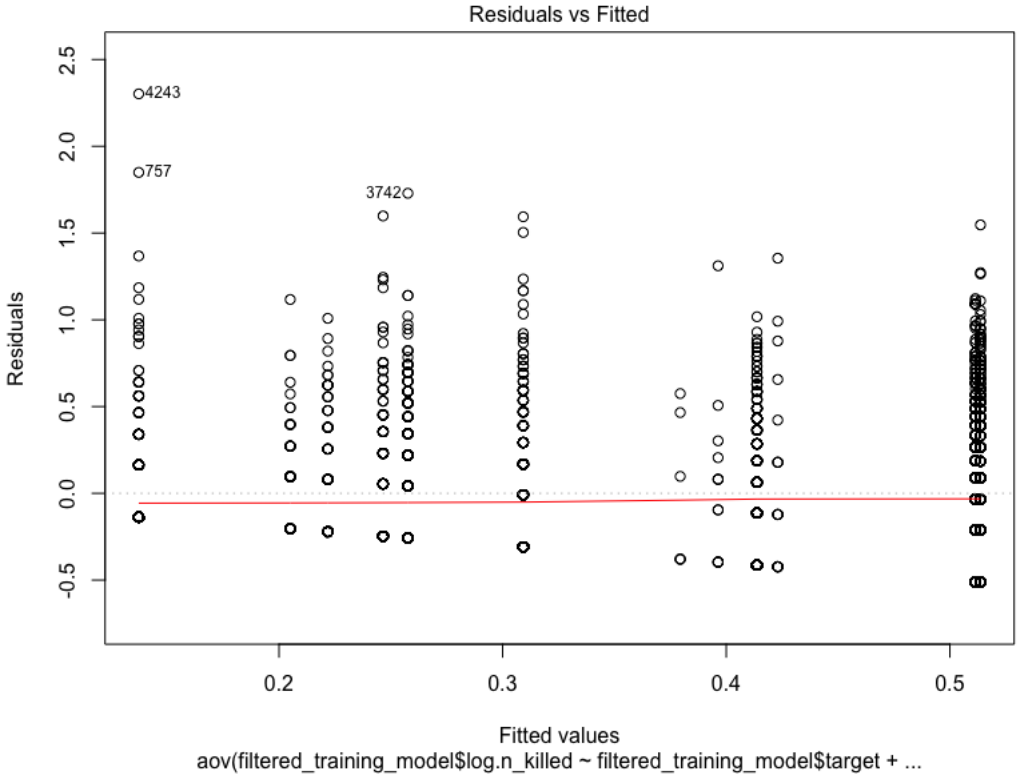
We decide to utilize weighted least squares in place of ordinary least squares. We go back to our original logged dataset (outliers **not** removed) and construct the weights for each subgroup combination. After determining the weights for each subgroup, we get the following output:

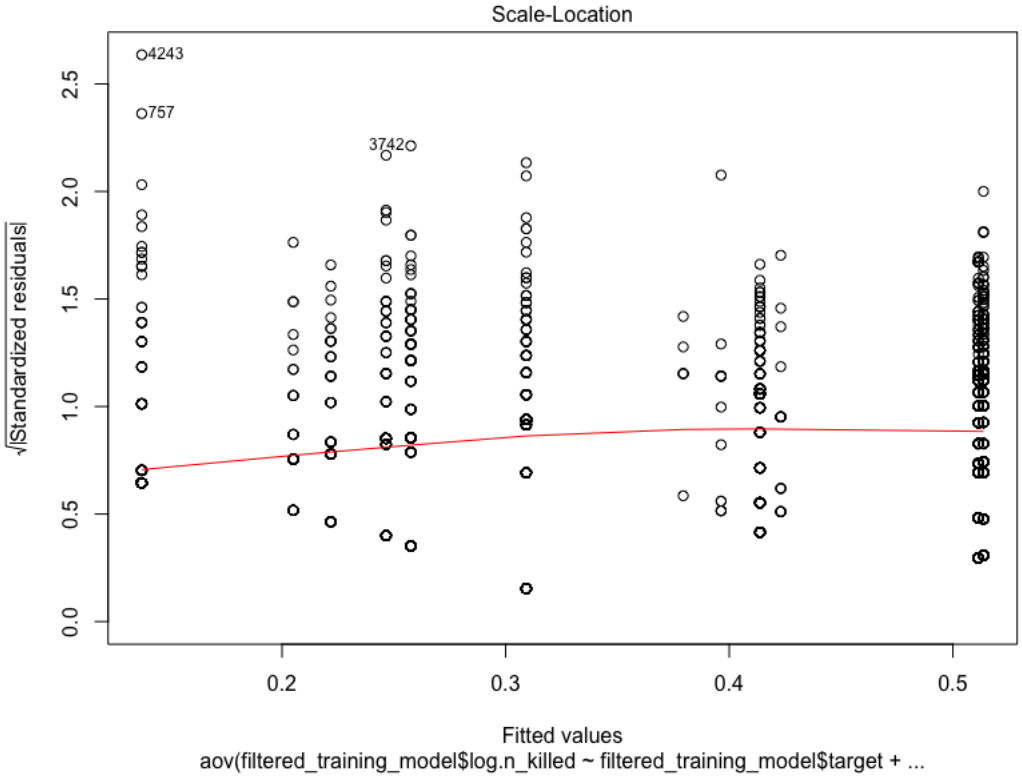
```
> Anova(filtered_training_w, type = 3)

Anova Table (Type III tests)
Response: log.n_killed

                +---------+------+----------+-----------+
                |  Sum Sq |  Df  |  F value |  Pr(>F)   |
+---------------+---------+------+----------+-----------+
| (Intercept)   |  8.30   |    1 |  59.9233 | 1.236e-14 |
| target        |  42.08  |    2 | 151.8726 | < 2.2e-16 |
| country       |   3.48  |    3 |   8.3747 | 1.507e-05 |
| target:country |   7.75  |    6 |   9.3205 | 3.556e-10 |
| Residuals     | 563.92  | 4071 +----------+-----------+
+---------------+---------+------+

Call:
```

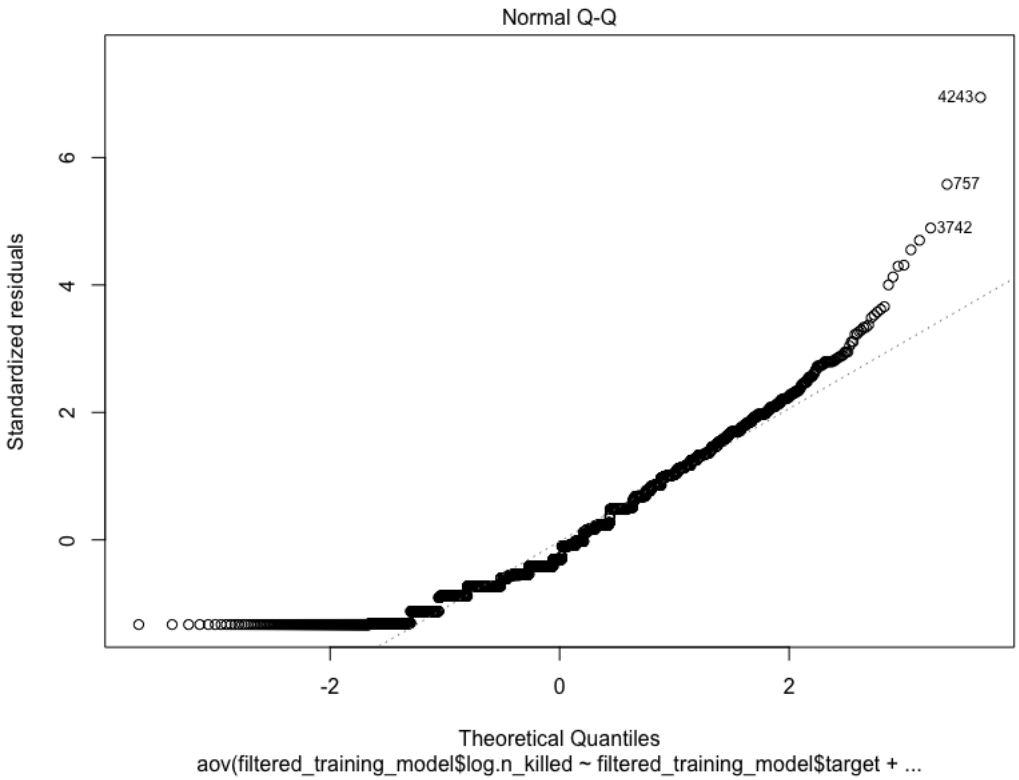
```
lm(formula = log.n_killed ~ target + country + target * country, weights = weight)

Weighted Residuals:
    Min      1Q  Median      3Q     Max
-0.5250 -0.2594 -0.1214  0.2426  2.2750

Coefficients:
                                        Estimate    Std. Error  t value    Pr(>|t|)
(Intercept)                           | 0.137278  |  0.017734  | 7.741   | 1.24e-14 ***
Target
  Police                              | 0.171941  |  0.023364  |  7.359  | 2.22e-13 ***
  Private Citizens & Property         | 0.374126  |  0.022008  | 17.000  |  < 2e-16 ***
Country
  India                               | 0.109266  |  0.027154  |  4.024  | 5.83e-05 ***
  Iraq                                | 0.285756  |  0.081348  |  3.513  | 0.000448  **
  Pakistan                            | 0.067746  |  0.045136  |  1.501  | 0.133448
Target * Country
  Police:India                        | -0.004754 |  0.034122  | -0.139  | 0.889194
  Private Citizens & Property:India   | -0.106984 |  0.033083  | -3.234  | 0.001231 ***
  Police:Iraq                         | -0.215564 |  0.194830  | -1.106  | 0.268609
  Private Citizens & Property:Iraq    | -0.400818 |  0.124137  | -3.229  | 0.001253  **
  Police:Pakistan                     | -0.155150 |  0.059422  | -2.611  | 0.009061  **
  Private Citizens & Property:Pakistan | -0.321530 |  0.051790  -| 6.208  | 5.89e-10 ***

Residual standard error: 0.3702 on 4071 degrees of freedom
  (865 observations deleted due to missingness)
Multiple R-squared:  0.1212,Adjusted R-squared:  0.1188
F-statistic: 51.04 on 11 and 4071 DF,  p-value: < 2.2e-16
```

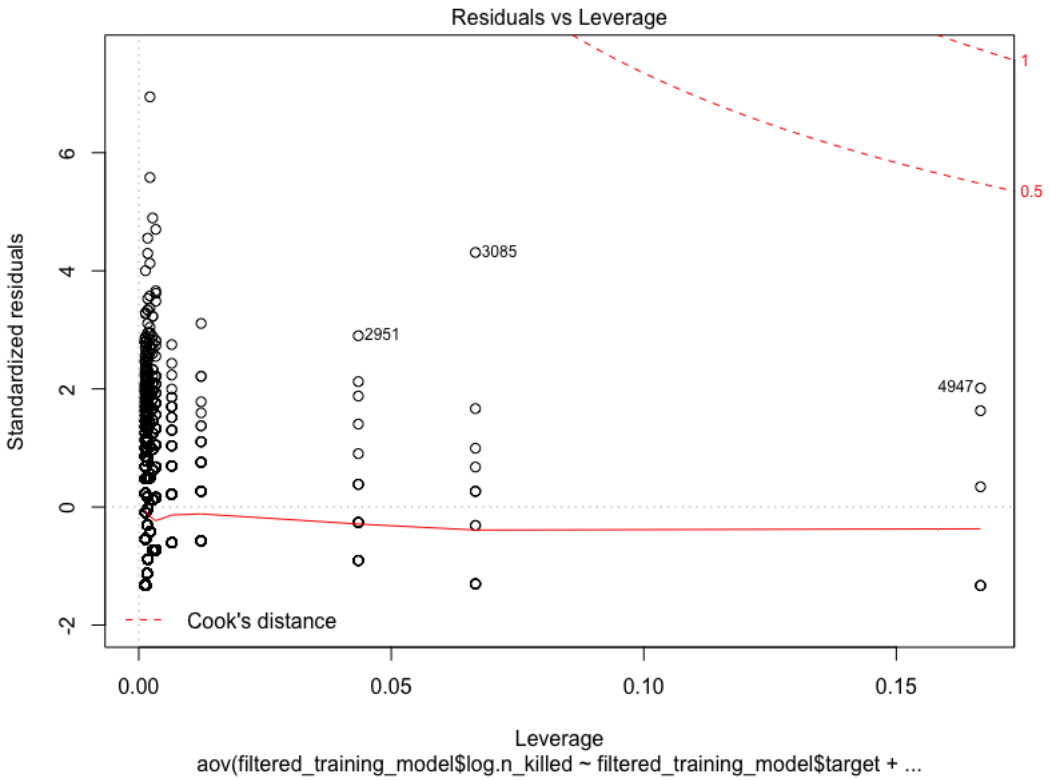
Similar findings as before. We note that some subgroups have been removed from the regression not because of an accident, but because their values were omitted due to missing data. Our regression output hasn't changed much at all: both the R^2 and adjusted R^2 remain in the low tens (representing the log number of deaths), with a good amount of our subgroups marking as highly statistically significant. Our interaction effect is still incredibly statistically significant (and thus there is very strong evidence of an interaction effect), so we won't need to do multiple comparisons on main effects. Our residual standard error of 0.3702, which doesn't mean much either given we're dealing with weighted least squares (since there's no single standard deviation of the errors).

Let's look at our residual diagnostic plots:


Residuals vs Fitted
Residuals
Fitted values
aov(filtered_training_model$log.n_killed ~ filtered_training_model$target + ...
4243
757
3742
2.5
2.0
1.5
1.0
0.5
0.0
-0.5
0.2
0.3
0.4
0.5

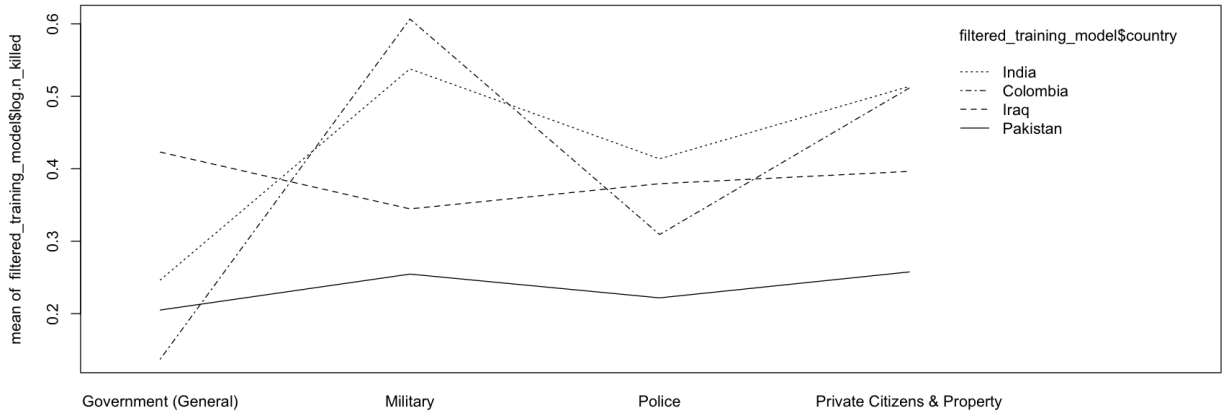

Scale-Location
4243
757
3742
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
2.5
0.2
0.3
0.4
0.5
Fitted values
aov(filtered_training_model$log.n_killed ~ filtered_training_model$target + ...


Normal Q-Q
Standardized residuals
Theoretical Quantiles
aov(filtered_training_model$log.n_killed ~ filtered_training_model$target + ...
4243
757
3742
0
2
4
6
-2
0
2

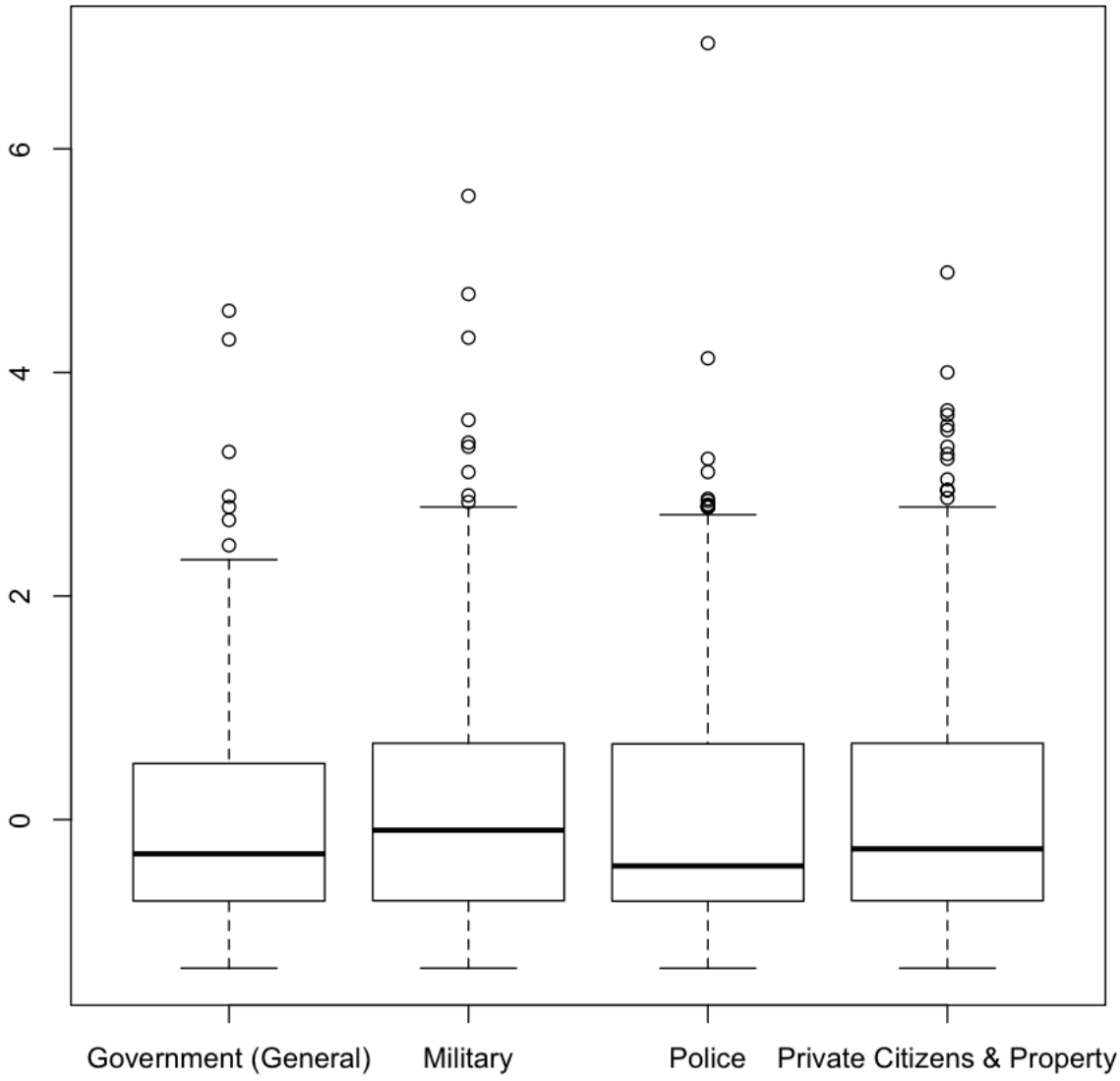


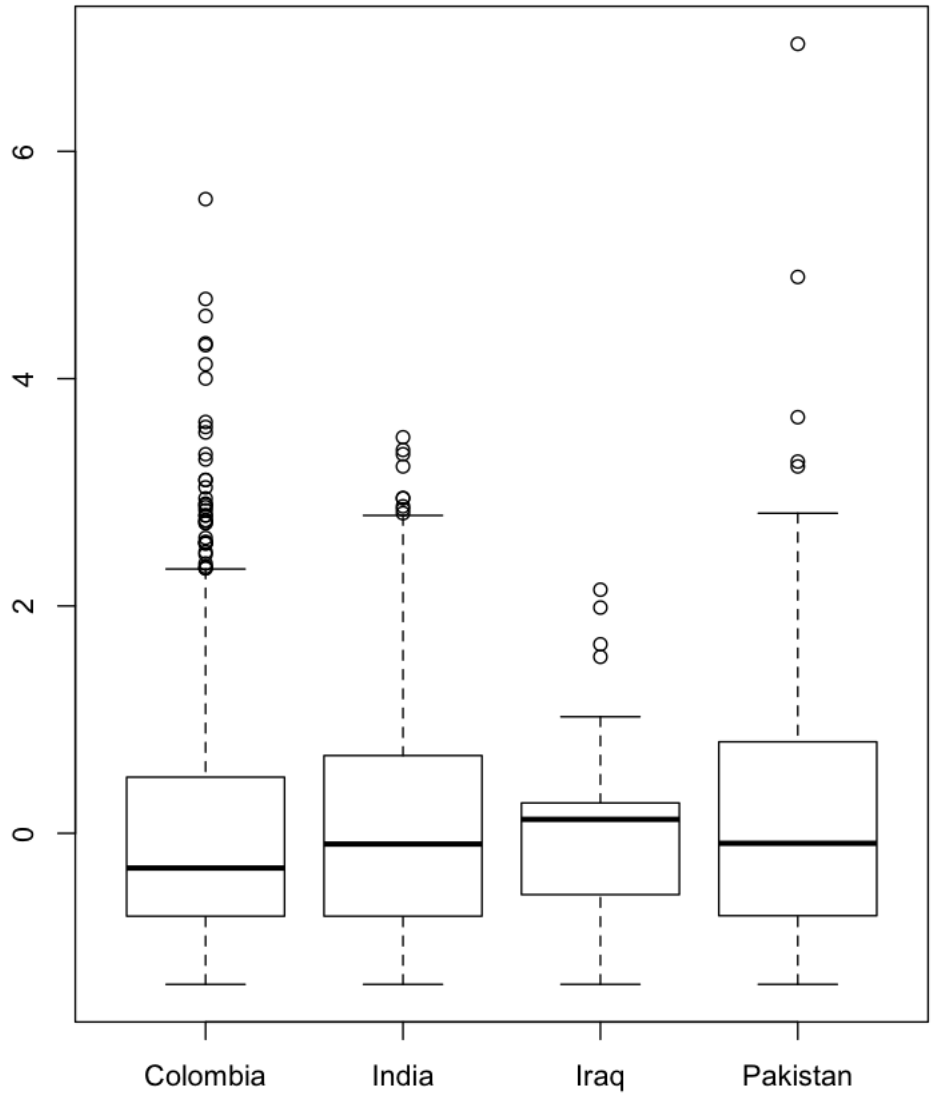
While our residuals vs fitted values plot seems to be mostly what we want, we cannot ignore the less-than-ideal q-q plot and its pronounced long tail at the low end of our data.

In context, we have our interaction plots for all subgroups:



And we note little overall difference between our previous interaction plot (on our logged, no-outlier OLS model) and our current plot. The same general relationships are holding up, and continue speaking to the lack of overwhelming parallelism (and thus, lack of similarity) between subgroups.


0
2
4
6
Government (General)
Military
Police
Private Citizens & Property



Our boxplots show that variance is pretty constant across all subgroups except for `country` Iraq with a slightly low variance relative to the other countries.

We're satisfied with our model's progression, and opt to report our regression model's equation as follows:

```
log.n_killed = 0.137278 + 0.171 Police + 0.374 Private Citizens & Property
                        + 0.109 India  + 0.285 Iraq + 0.067 Pakistan
                        - 0.004 Police:India
```

```
- 0.106 Private Citizens & Property:India
- 0.215 Police:Iraq
- 0.400 Private Citizens & Property:Iraq
- 0.155 Police:Pakistan
- 0.321 Private Citizens & Property:Pakistan
```

Resources

All files pertaining to this report, including the R data analysis script, dataset, plots, and this PDF are open to the public and hosted on GitHub (`github.com/dannyfig/multivariate/ANOVA_regression`).