Analyzing Global Terrorism Using Bayesian Hierarchical Models & Logistic Regression

Author: Will Noone
George Washington University: DATS 6450 – Bayesian Methods for Data Science

Introduction

Terrorism has become increasingly prevalent throughout the world in the past fifty years. With the emergence of a global marketplace & 24 hour news-cycle in developed economies, terrorism has proven a cost efficient platform for fringe groups to promote an ideological end. Analyzing the underlying conditions of nation-states producing terror groups or actors may yield insights as to mitigating successful attacks. The aim of this study though is to probe relationships among the Regions in which attacks occur, then to focus our lens upon economic and demographic factors common to nationalities of terror groups most active.

Data Overview

Source: Global Terrorism Database (GTD). The Global Terrorism Database (https://www.start.umd.edu/gtd/) is an open-source database managed by the University of Maryland offering information on terror attacks around the world from 1970 through 2017. Measurements are made for each incident, including the date and location, the weapons used, nature of the target, number of casualties, and identifying characteristics of the terror group responsible.

Data Definitions:

SUCCESS: (dichotomous; 1 = "Yes" or 0 = "No") This is defined according to the tangible effects of the attack. Success is not judged in terms of the larger goals of the perpetrators. The key question is whether or not the attack materialized or was thwarted via preventative action or chance.

REGION: (categorical) The nine geographic divisions in which terror incidents occurred.

YEAR: (datatime) The year in which a terror incident occurred

NATIONALITY: (categorical) The country from which a terror group is based.

GNAME: (categorical) Group name of a given terror organization.

Source: World Bank Development Indicators Database (WBI). The World Bank Development Indicators Database (https://data.worldbank.org/indicator) is the primary World Bank collection of development indicators, compiled from officially recognized international sources. Available are current and accurate national, regional and global measures across domains: education, economics, labor, health, infrastructure, environment, etc.

Data Definitions:

X.SH.XPD.GHED.GD.ZS: (percent 0-100) Domestic general government health expenditure (% of GDP)

X. SE.XPD.TOTL.GD.ZS: (percent 0-100) Government expenditure on education, total (% of GDP)

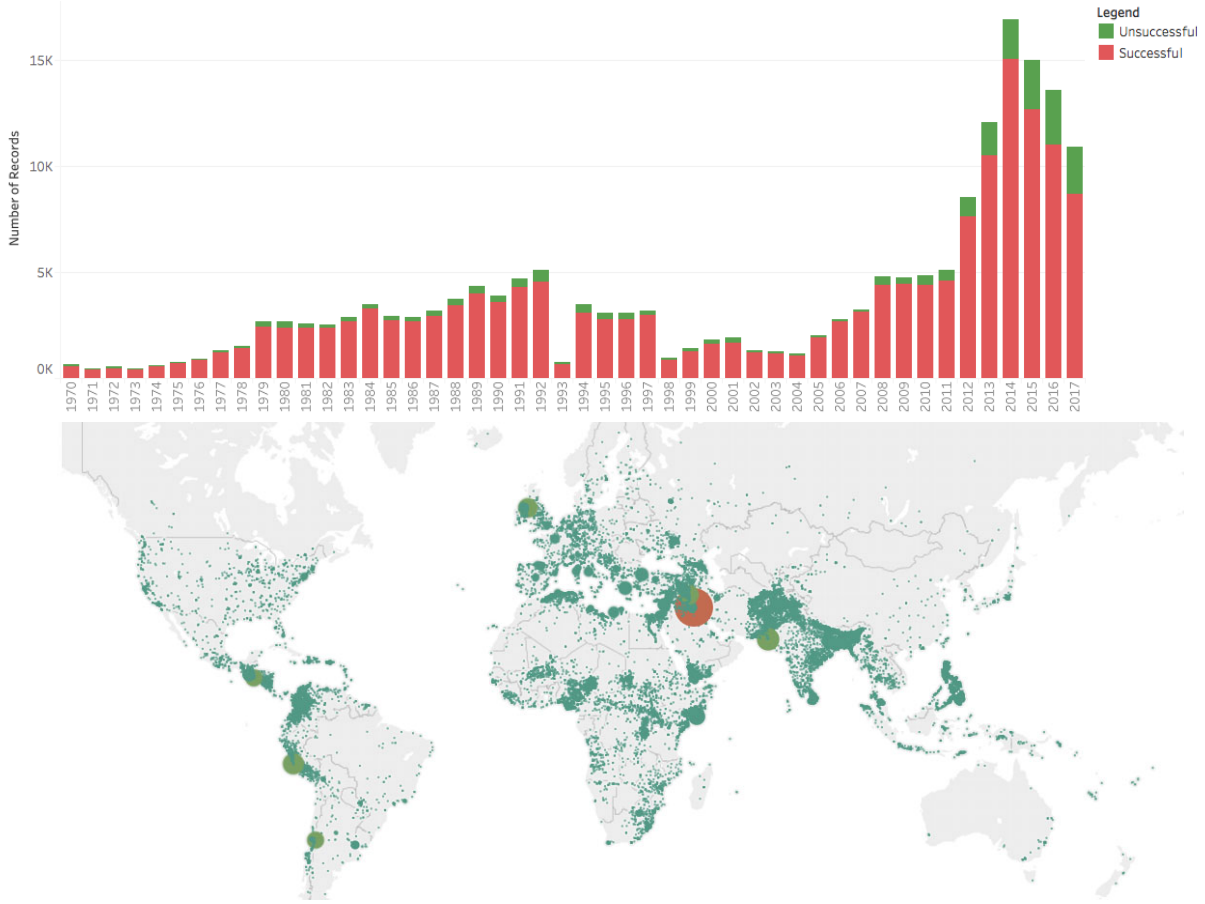
X. IT.NET.USER.ZS: (percent 0-100) Individuals using the Internet (% of population)

X. SP.RUR.TOTL.ZS: (percent 0-100) Rural population (% of total population)

Our analysis required construction of (2) data sets. The Analysis1_data used the GTD with measurements of Successful_Attacks out of Total_Incidents grouped by GNAME and REGION. The Analysis2_data required blending the GTD & WBI using a composite key; a concatenation of common attributes YEAR & COUNTRY_CODE (with respect to terror group nationality). This allowed for a LEFT OUTER JOIN to merge the Global Terror Database and World Bank Indicator data sets to align World Bank indicators with Global Terror data. To address occasional NULL values in the WBI attributes, we imputed the median.

Exploratory Data Analysis

Through brief exploratory analyses, we make a few observations. Viewing Total_Incidents as a function of YEAR, one can see the number of incidents increased precipitously in the early 2000s, peaking in 2014 with 16,903 attempted or successful attacks. Over this timeframe, total incidents occurred in the conflict laden Middle East & North Africa (41,078) and South Asia (37,212) regions respectively. The concentration of attacks between 2000 – 2017 local to these regions begs a number of questions. What is the probability of a successful attack in a particular region vs overall? Furthermore, can economic or demographic indicators (with respect to a terror group's nationality) explain the probability of a successful attack? We seek to apply Bayesian modeling of Hierarchies with Subjects within Categories and Logistic Regression to these cases.



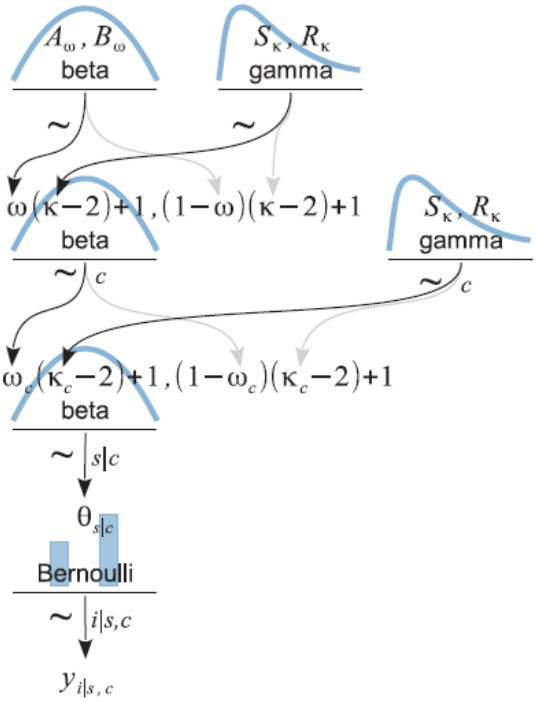
Geographic Concentration of Terror Attacks: 2000 - 2017		
Region	Successful_Attacks	Total_Incidents
Middle East & North Africa	36,008	41,078
South Asia	32,076	37,212
Sub-Saharan Africa	12,011	13,044
Southeast Asia	8,413	9,562
Eastern Europe	3,461	4,013
South America	2,262	2,508
Western Europe	2,424	3,200
North America	546	648
Australasia & Oceania	58	75

Analyses & Experimental Results

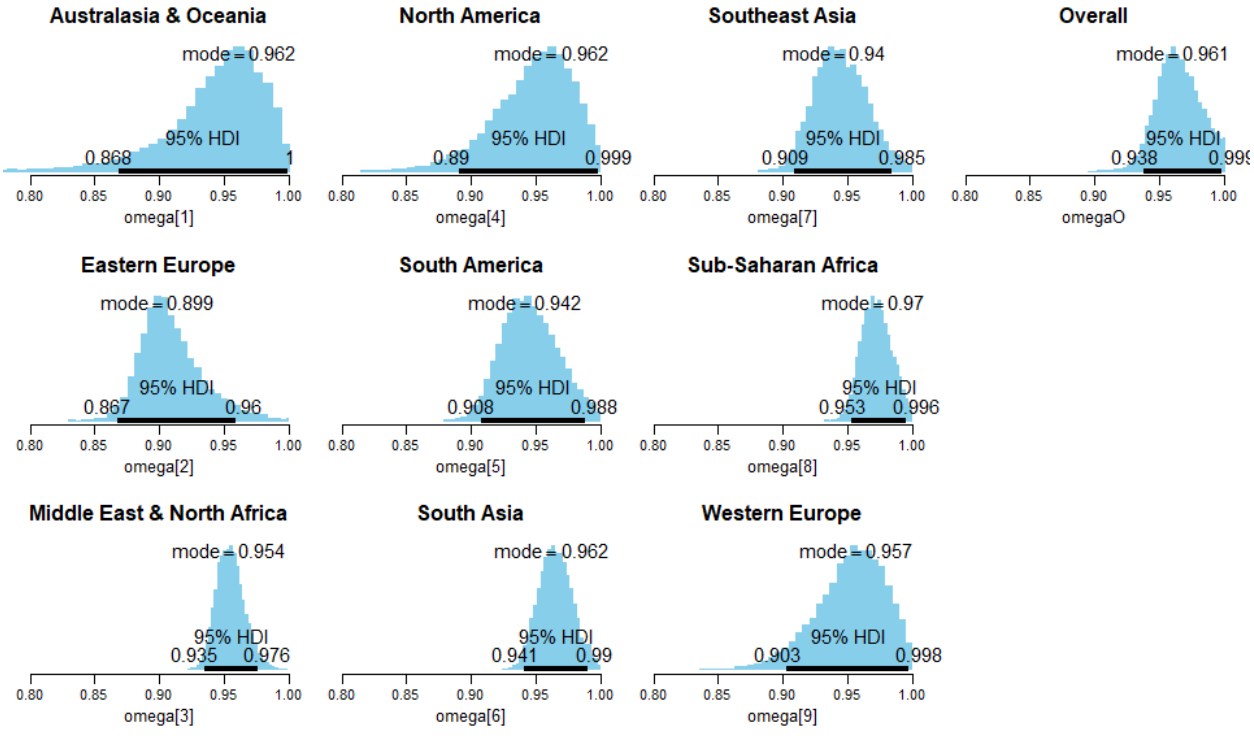
Part (a): Using Hierarchical Models to examine success rate among Regions & Terror Groups

The data structure (particularly Analysis1_data) invites hierarchical descriptions across multiple levels. With GNAME (Terror Group) as subjects, and Region as categories, we estimate the probability of a successful attack by Region. We then extend our analysis to horizontal comparison among select Terror_Groups and vertical comparison of Terror_Groups to the respective Region mode.

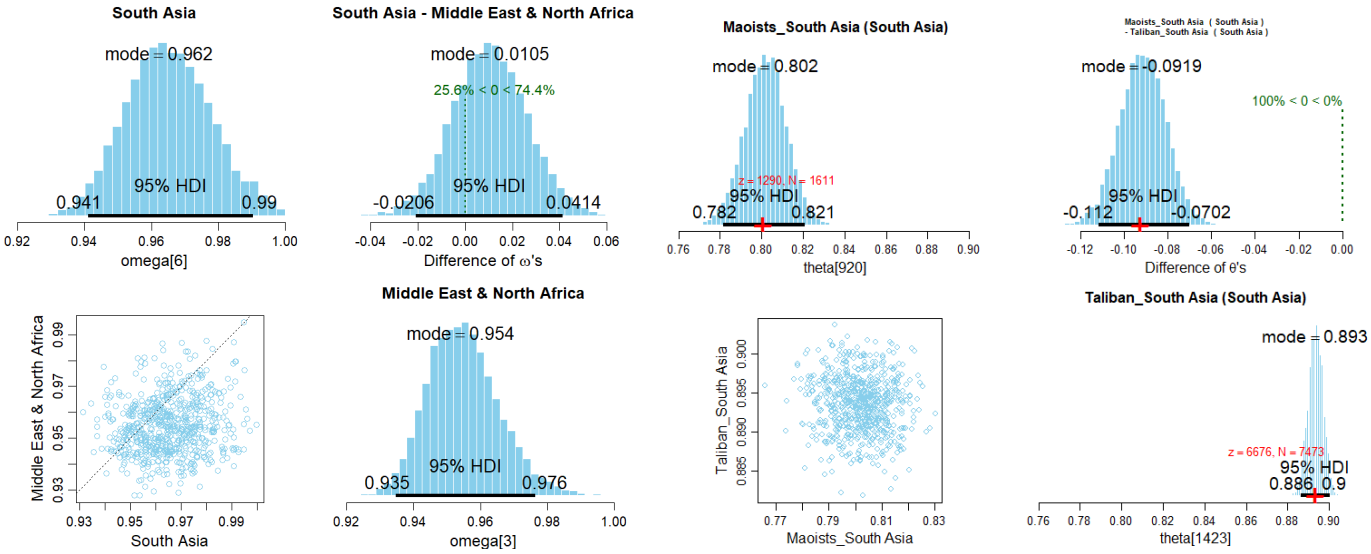
The hierarchy model to the right (**Kruschke 252**) provides a guide to implementation. An initial caveat, our model's likelihood uses a binomial instead of Bernoulli distribution. Each incident is denoted as $y_i|_{s,c}$ to indicate attacks by Terror_Goups (subjects s) within Regions (categories c). The underlying bias of a Terror_Group within a Region is denoted $\theta_{s|c}$. The biases of Terror_Groups within Regions are assumed to be distributed as a beta density with mode ω_c and concentration κ_c. Each Region has its own modal bias ω_c, from which all Terror_Group biases in the Region are assumed to be drawn. We assume that Region modes come from a higher-level beta distribution that describes the variation across Regions. The modal bias across Regions is denoted ω and the concentration of the Region biases is denoted κ.



Marginal Posterior Distributions for Successful Attacks by Region



Above are the marginal posterior distributions for the Region (category) level modes with the overall omegaO realizing a probability of 0.961. Eastern Europe (omega2) with 0.889 and Sub-Sahara Africa (omega8) with 0.970 realized the min and max Region level modes respectively. To the middle left, we compare the Regions of focus (South Asia, Middle East & North Africa) per our EDA. The difference in modes is 0.0105 with South Asia having a higher probability of success than Middle East & North Africa. Because 0.00 is contained within the HDI though, we cannot assert that there is in fact a credible difference between the two Regions. When κ is large, the Region biases ω_c are generally more tightly concentrated (See Appendix for Kappa Posterior Distributions). The certainty of the estimate at a given level in the hierarchy depends on how many parameter values are contributing to that level. In the present application, there are nine Region parameters contributing to the overall mode, but thousands of Terror_Groups contributing to each Region. We therefore are more certain of the estimate within each Region compared to that of the overall level.



If we did not incorporate Region information into our hierarchical model, and instead put thousands of Terror_Groups under a single distribution, then the modal estimates for two Terror_Groups with the same "success rate" would be identical regardless of the Region which they attacked. To the middle right, we compare the top two Terror_Groups within the South East Region. The Taliban (theta 1423) had a greater probability of success than Maoists (theta 920) by 0.0919. This difference is credible as 0.00 does not fall within the HDI. Though shrinkage effect of the subject to category is small, we can measure it. For example, Maoists had success rates of (1290/1611 = 0.8001). The subject level mode is pulled toward the category mode for South Asia, which is 0.96. One shortfall of our analysis is that within each Region there are only a few well established Terror _Groups committing the majority of the attacks, which in execution realize lower success rates as a consequence of larger attempted sample sets. Said another way, our data is not well balanced in that there are many Terror_Groups with a low "N", nearly always equaled by "z", which one could argue makes the Mode for this analysis not ideal.

Part (b): Using Robust Multiple Logistic Regression to examine probability of a successful attack for a given set of conditions the native country of a terror group.

Our data included a dichotomous dependent variable "success", which lends quite well to building a Bayesian logistic regression model. The next part of our analysis aims explain the probability of a successful attack given economic or demographic indicators (with respect to a terror group's nationality). While the Region in the Hierarchical model was with respect to the location of the attack, we now delve into the conditions that produce a successful attack (ie the nationality of the Terror_Group). For this analysis, we chose to only use data from the Middle East & North Africa Region to make MCMC more tenable. While the menu of available variables from the WBI was extensive, we chose to perform two regressions. The first used Government Expenditure on Health and Education (as % GDP) as explanatory variables. The second used Internet Users and Rural population (as % of Total Population) as explanatory variables. Under the Bayesian logistic model, the linear combination of metric predictors maps to a probability value via the logistic function, and the predicted 0's and 1's are under a Bernoulli distribution. The Generalize Linear Model aims to express the combined influence of predictors as their weighted sum upon a dependent variable.

The GLM can be written as follows:

$$\mu = f\left(\text{lin}(x), [\text{parameters}]\right)$$
$$y \sim \text{pdf}\left(\mu, [\text{parameters}]\right)$$

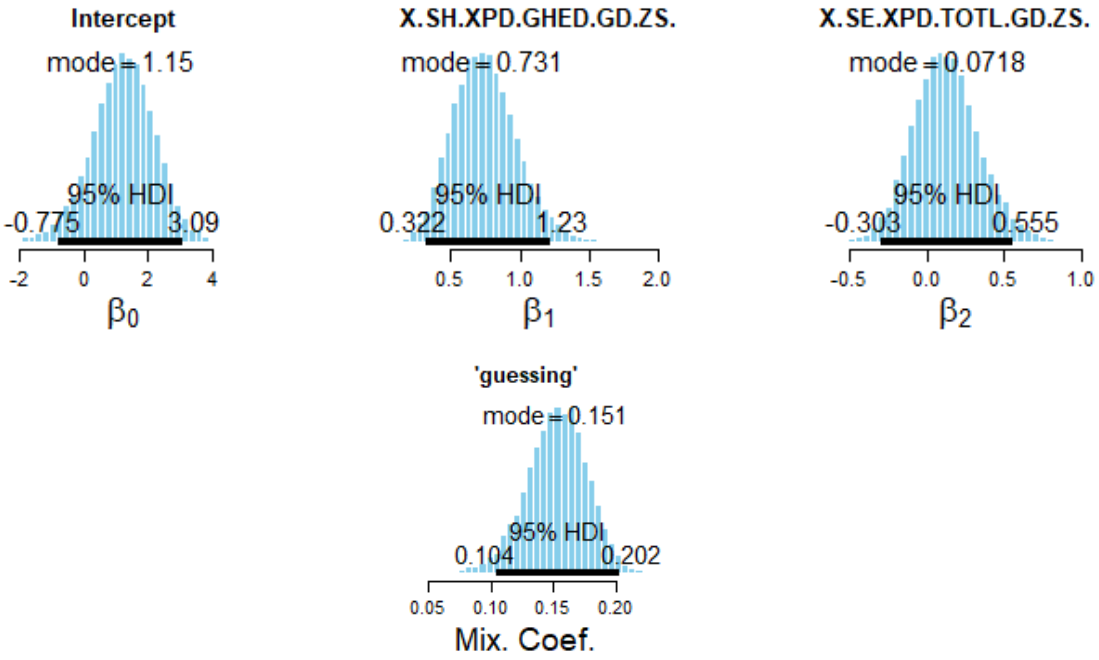
For our analysis, we opted for the GLM extension, robust logistic regression (to be robust to outliers). Why not use the simple Bayesian logistic form? As we found in our Hierarchical analysis, our data is not well balanced in that there are many Terror_Groups with a low "N", nearly always equaled by "z". Accordingly, we will describe the data as being a mixture of two different sources: a logistic function of the predictor(s) and another to account for randomness or "guessing" which is expressed as alpha below. With the two sources combined, the predicted probability that y = 1.

The Mu for the robust logistic model takes the following form:

$$\mu = \alpha \cdot \frac{1}{2} + (1 - \alpha) \cdot \text{logistic}\left(\beta_0 + \sum_j \beta_j x_j\right)$$

Model 1: Government Expenditure in Terror Group Native Country
Region: Middle East & North Africa

(Y): Success = 1, Not Success = 0
(X_1): X.SH.XPD.GHED.GD.ZS - Domestic general government health expenditure (% of GDP)
(X_2): X. SE.XPD.TOTL.GD.ZS - Government expenditure on education, total (% of GDP)



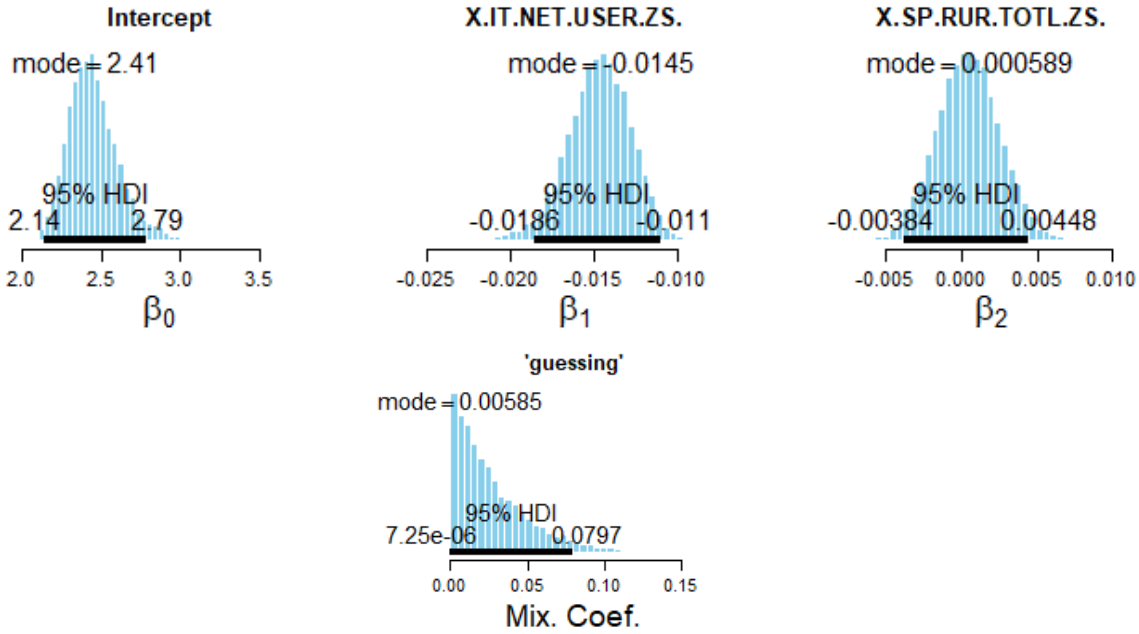
Consider a Terror_Group from a country with Government Expenditures (as % GDP) on Health (**X_1 = 2.5**) and Education (**X_2 = 2.5**). The predicted probability of a successful attack is:

$$\text{Guess} * (0.5) + (1 - \text{Guess}) * \text{logistic}(\beta_0 + \beta_1 x_1 + \beta_2 x_2)$$
$$= 0.15 * (0.5) + (1 - 0.15) * \text{logistic}(1.15 + 0.7310 * 2.5 + 0.0718 * 2.5) = 0.8898$$

That probability is a log odds of 0.15 * (0.5) + (1 – 0.15) * log(0.89015/(1 – 0.89015)) = 2.756. If Government Expenditure on Education increased by 1% (3.5), then the predicted probability of a successful attack is 0.8922, which has a log odds of 2.7558. Thus, when X_2 was increased by 1 unit, the probability went up 0.002 (from 0.8898 to 0.8922), and the log odds increased 0.0610 (from 2.7558 to 2.8168) or the value of β_2 after accounting for the Guess coefficient. Given the positive β_1 value, we can infer that Government Expenditure on Health has the same directional impact on probability (per 1 unit increase in X_1) yet to a greater degree (from 0.8898 to 0.9075). These are interesting results because as a "terror sponsoring" nation spends more on its populous, their probability to complete a successful attack increases.

Model 2: Population Characteristics in Terror Group Native Country
Region: Middle East & North Africa

(Y): Success = 1, Not Success = 0
(X_1): X. IT.NET.USER.ZS - Individuals using the Internet (% of population)
(X_2): X. SP.RUR.TOTL.ZS - Rural population (% of total population)



Consider a Terror_Group from this Region with Internet Users (% of population) (**X_1 = 50**) and Rural Population (% of population) (**X_2 = 50**). The predicted probability of a successful attack is:

$$\text{Guess} * (0.5) + (1 - \text{Guess}) * \text{logistic}(\beta_0 + \beta_1 x_1 + \beta_2 x_2)$$
$$= 0.00585 * (0.5) + (1 - 0.00585) * \text{logistic}(2.41 + -0.0145 * 50 + 0.0006 * 50) = 0.8764$$

Here we find a negative relationship for Internet Users; for a 10% increase (**X_1 = 60**), the predicted probability of a successful attack decreases by 0.0162 to 0.8602. This is an interesting result as one could reason that with more access to the internet, attacks would be better coordinated and successful with greater probability. We propose an alternative hypothesis in that as internet usage increases, talented individuals otherwise susceptible to radical ideologies, are made moderate via digital outlets, leaving the gap in recruiting for terror organizations to be filled with less competent members to commit attacks, all other variables equal.

Concluding Remarks

Our goal through this project was to apply a Hierarchical model and Bayesian logistic regression using JAGS to evaluate the impact of geographic location as well as economic & demographic indicators on the probability of success of a terrorist attack. Our Exploratory Data Analysis revealed a rather alarming trend of exponential increases in the number of attacks beginning at the turn of the century as well as a concentration in the Middle East & North Africa and South Asia Regions. Though a Hierarchical model build, we found that a given Terror_Group's "success rate" is pulled toward a categorical mode per the Region in which the attack took place. We opted to try a robust logistic linear regression model to evaluate the combined influence of predictors (local to the nationality of a terror perpetrators) on the probability of success of an attack. We acknowledge additional methods such as variable selection would nicely augment this analysis as well as including more combinations of explanatory variables.

***APPENDIX Included separately for MCMC Diagnostics from Analysis #1 and #2.**