Miranda Gibbons

Project Four – Modeling Terror using the GTD

The Global Terrorism Database, maintained by the National Consortium for the Study of Terrorism and Responses to Terrorism (START) at the University of Maryland, records and catalogs terrorist attacks as culled from media reporting and previously maintained databases. Their data currently spans 1970 to 2015, with 138 variables to describe each incident, and the entire dataset comprises over 150,000 incidents.

The first stage to my analysis was to visualize the data, both literally and by looking over summary statistics for the variables. I often complete this primary stage of EDA before beginning to clean the data, in order to understand what story the data is telling – the patterns that emerge from this cursory analysis give rise to more qualitative theories about why the data might look the way it does, and help to inform exactly *how* I want to clean my data.

For many of the records there were missing values in numerous columns – unlike in previous projects, I did not drop any null values nor impute any numbers. The nature of the data collection for this database, and the dependence of some variables on others within a given row, suggested that these values were missing not because of unclean data but because those attacks did not generate information that would be recorded in all variables (or the nature of the data collection from media sources meant that no information could be found).

The variables I found to be of the greatest importance, with regard to grouping similar events together (or rather differentiating between groups of events), were date, region, attack type, weapon type, and country. For the most part, I summed attacks over year periods rather than utilizing the day-specific information provided – such detail introduced a lot of noise.

In order to compare two populations, and estimate the degree to which those populations differed, I utilized the pymc3 library in Python and Bayesian inference. From Bayes' Rule, we know that the posterior distribution of some parameter (or set of parameters) is equivalent to the likelihood function of our data multiplied by our prior distribution and scaled by (divided by) the marginal likelihood of the data:

$$P(\theta|y) = \frac{P(y|\theta)P(\theta)}{P(y)}$$

I was interested in looking at the difference in level of violence when comparing two countries within a given region for a given year – I specifically chose to only look at values within a given year due to the dependent nature of our attack data on time. I chose 1990 when considering Sub-Saharan Africa for two (not necessarily independent) reasons – the year 1990 saw a larger quantity of attacks within the region as compared to the surrounding years, and 1990 marked the official end (or beginning of the end) of Apartheid in South Africa. Nelson Mandela was also freed from prison this year. I decided to compare South Africa to Uganda – the two countries are close but not bordering each other, each experienced political change during the late 80s and early 90s, yet Uganda saw far fewer terrorist attacks within 1990 than South Africa. I was interested in choosing two populations of disparate sizes, to see how that size difference might effect my estimation. For my prior distribution I chose number of people wounded in terrorist attacks in 1990 for all countries in Sub-Saharan Africa *except* South Africa and Uganda – the hypothetical scenario might be not having access to such data until some large-scale political change occurred within those countries. The posterior distributions generated by running the probabilistic model demonstrated that the two populations did not have significantly different means, whereas the spread of the data was notably different. Overall, the two populations followed similar distributions. Number of individuals wounded per attack followed a normal distribution for both countries, whereas the standard deviation (measure of spread) for each country was near normal, but skewed to the right.

One notable weakness of the dataset is the lack of data for the year 1993 – as records were being transferred over the years, all but 15% of this information was lost. In order to not misrepresent data for that year, all data for 1993 was removed. An appendix to the START GTD codebook notes country-level statistics for the year, with a total number of incidents at 4954.

In order to impute values for 1993, specifically the number of bombings, I explored relationships and dependencies between variables in the dataset at large and for the years 1991-1995. A heat map of just those records where the attack type was a bombing indicated that year, region, and specific latitude and longitude were the strongest indicators of whether an attack was a bombing. However, given that over the entire dataset 48% of attacks were bombings, these variables may just be good indicators of whether any attack will occur. The proportion of attacks that were bombings varied over the years, whereas the change in proportion of attacks remained fairly stationary for the time period of interest. These facts indicated that perhaps a time series model such as an autoregression might accurately predict the number of bombings in 1993.

I performed a Dickey-Fuller test for stationarity on number of bombings per year, per month, and per day, as well as first order differences for each of those. I fit my model specifically on the first order difference of the log of total bombings per month for 1970 to 1992 (I chose this transformation due to its stationarity). I then predicted values for 1990 to 1995 in order to compare my actual and predicted values in a meaningful way. Unsurprisingly, having fit my autoregression model on the years prior to 1993, the predictions for 1994 and 1995 do not match the actual values well – this indicates overfitting, perhaps, or just the weakness of the model. The sum of bombings for the year 1993, via this method, came out to 1495.