The Path to Never Again

A Logistic Regression Approach to Genocide Forecasting

Charles Costanzo: April 23, 2023

Background

According to Article II of The Convention on the Prevention and Punishment of The Crime of Genocide (1948)[1], genocide is "any of the following acts committed with intent to destroy, in whole or in part, a national, ethnic, racial or religious group, as such:

(a) Killing members of the group;
(b) Causing serious bodily or mental harm to members of the group;
(c) Deliberately inflicting on the group conditions of life calculated to bring about its physical destruction in whole or in part;
(d) Imposing measures intended to prevent births within the group;
(e) Forcibly transferring children of the group to another group."

After the Holocaust in the 1940's, the international community created this convention to prevent genocide from happening ever again. Despite these efforts, there have been many genocides since the Holocaust. Many are ongoing. Clearly, the international community has failed to eradicate genocide, despite its repeated promises of "Never Again". In an attempt to begin to answer the extremely complex question of "why," one tool that researchers use is statistical modeling. This strategy allows for exploration of risk factors that influence the onset of genocide. More specifically, regression analysis has been used to test if there are statistically significant associations between certain risk factors and onset of genocide, as well as to determine the direction of any such relationship. Some of the core risk factors of genocide include prior atrocity, political upheaval, and threats to the perpetrating regime.[2] This project will focus on the impact that political upheaval in particular has on onsets of mass killing episodes. However, other variables will be included to take into account population, economic upheaval, and the social situation. This allows us to determine what impact (if any) they have on the outcome. In this project, variables that measure these factors will be used to build a logistic regression model that estimates the probability that a mass killing episode will start.

[1] https://ihl-databases.icrc.org/en/ihl-treaties/genocide-conv-1948/article-2

[2] Hollie Nyseth Brehm (2017) Re-examining risk factors of genocide, Journal of Genocide Research, 19:1, 61-87, DOI: 10.1080/14623528.2016.1213485

Research Question

What are factors/elements of a country and its society that affect the likelihood of genocide/mass killings more generally? In particular, what political factors are important? How can we assess the risk of genocide based on these factors?

Variables of Interest

anymk.ever	1 if any mass killing ever; 0 if not	Prior Atrocity
popsize.ln.combined	population size (log)	Population
coup.try.5yr	1 if coup attempt (successful or not) in the last five years; 0 if not	Threat to Regime/Political Upheaval
pol_killing_approved	1 if political killings are practiced systematically and are typically incited and approved by top leaders of government; 0 if not	Threat to Regime/Political Upheaval
freemove_men4	1 if virtually all men enjoy full freedom of movement; 0 if not	Threat to Regime/Political Upheaval
imr.sqrt	Infant mortality per 1000 live births, square root	Economic Upheaval/Public Health
social_inequality	1 if members of some social groups enjoy much fewer civil liberties than the general population; 0 if not	Social Dynamic
candidaterestriction	1 if there are no restrictions; 0 if candidates are restricted	Threat to Regime/Political Upheaval

Note: Infant Mortality is highly (negatively) correlated with GDP.[3]

Hypotheses

H1: Higher population is associated with higher odds an onset of a mass killing episode.

H2: A coup attempt (unsuccessful or not) within the last five years is associated with higher odds of an onset of a mass killing episode.

H3: Political killings are associated with higher odds an onset of a mass killing episode. H4: Higher infant mortality rates are associated with higher odds an onset of a mass killing episode.

H5: Social inequality is associated with higher odds of an onset of a mass killing episode.

H6: Restrictions on political candidates are associated with higher odds of an onset of a mass killing episode.

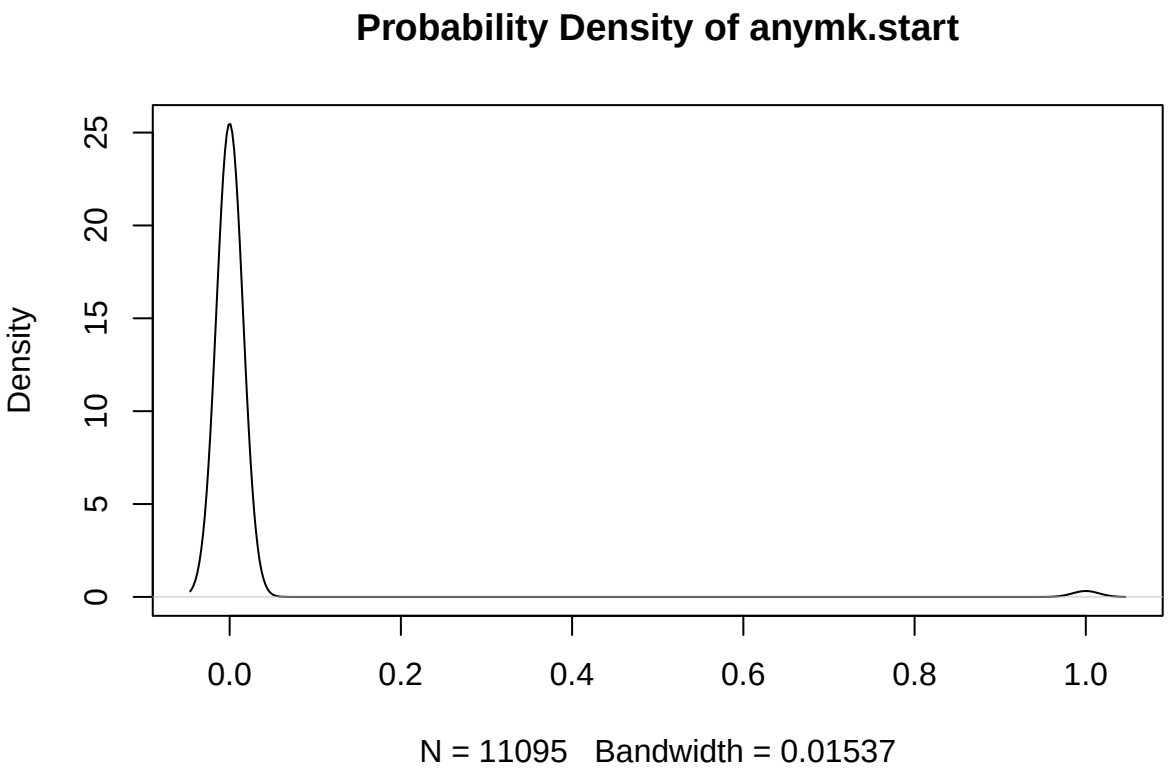
H7: Previous mass killings are associated with higher odds of an onset of a mass killing episode.

H8: Freedom of movement for men is associated with lower odds of an onset of a mass killing episode.

[3] O'Hare B, Makuta I, Chiwaula L, Bar-Zeev N. Income and child mortality in developing countries: a systematic review and meta-analysis. J R Soc Med. 2013 Oct;106(10):408-14. doi: 10.1177/0141076813489680.

Data

Assessment of Outcome Variable `anymk.start`



When we are addressing a research question involving genocide, ideally our outcome variable would be binary with 1 if a genocide occurred in a specific country/region in a specific year, and 0 if not. However, due to the relative rarity of genocide, restricting our analysis to only events that meet the strict UN definition of genocide would result in a very small sample size. Thus, for our (less than ideal) empirical measure of the outcome, we use the binary variable mass killings. According to the Early Warning project, "...a mass killing episode [has] occurred when the deliberate actions of armed groups, including but not limited to state security forces, rebel armies, and other militias, result in the deaths of at least 1,000 noncombatant civilians targeted as part of a specific group over a period of one year or less."[4] For an event to be considered a genocide, it must meet the more restrictive legal criteria (relying more on the perpetrators' intentions) defined at the beginning of this project. Most genocides are mass killings, but not all mass killings are genocides. However, it is reasonable to assume that the risk factors for genocide may be similar to those for mass killings since, at their core, they are both acts of violence against groups of people.

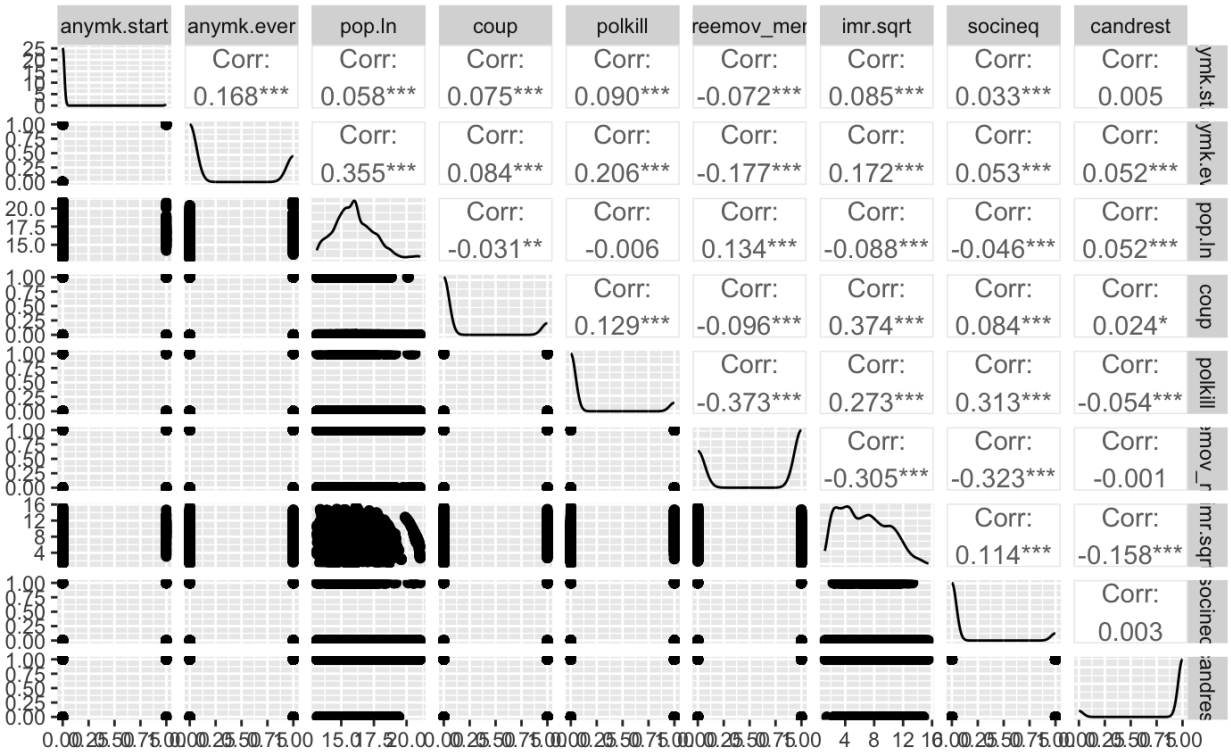
In order to test our hypotheses, we used data[5] from the Early Warning Project's 2022 Statistical Risk Assessment.[6] This dataset includes state-level data by year from 1945 - 2021.

[4] https://earlywarningproject.ushmm.org/definitions

[5] https://github.com/EarlyWarningProject/2022-Statistical-Risk-Assessment/blob/d3345713f4bc1d135834c8a8a6c1e8c1dce988f6/Make%20data/output/prepared2021predictors-2022-09-30.csv

[6] https://earlywarningproject.ushmm.org/reports/countries-at-risk-for-mass-killing-2022-23-early-warning-project-statistical-risk-assessment-results#:~:text=The%20Early%20Warning%20Project's%20Statistical,preventive%20actions%20may%20be%20needed.

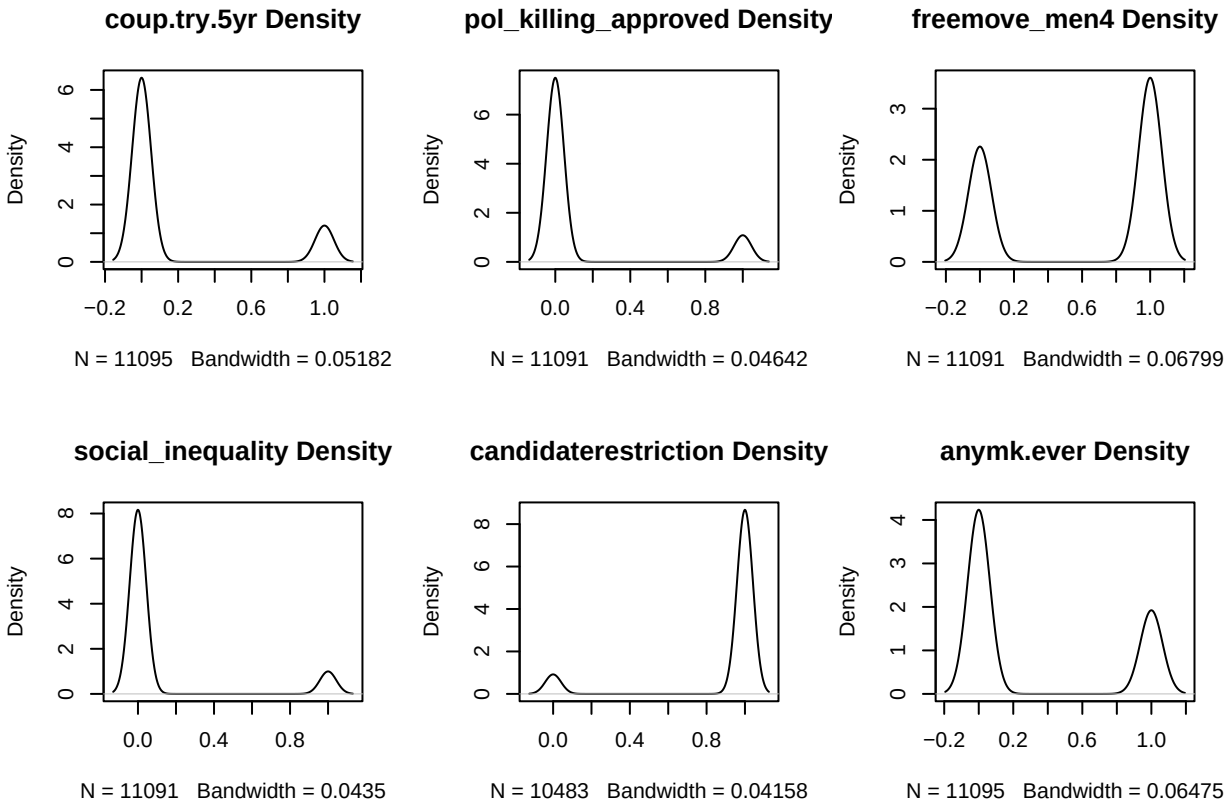
Data Exploration



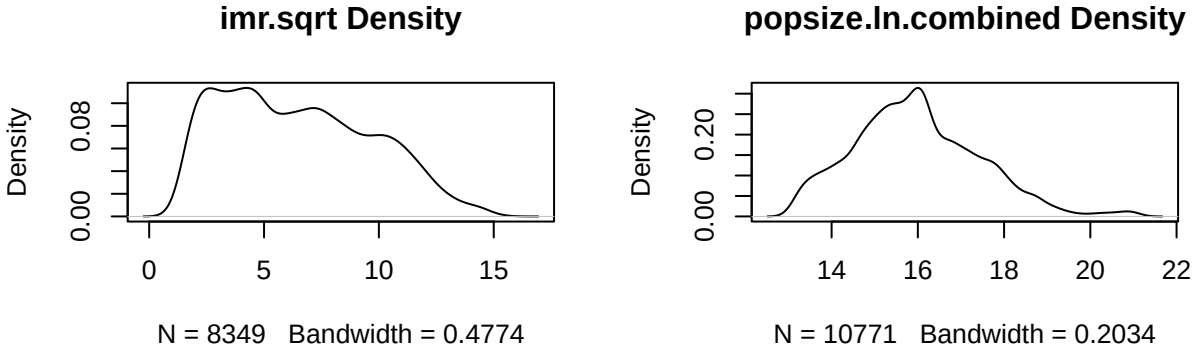
The `ggpairs` function from the `GGally` package can be used to visualize relationships between the variables of interest. As shown in the plot, there appear to be statistically significant correlations between the start of mass killings and all of the predictor variables (except for `candidaterestriction`). It is difficult to visualize most of the predictor variables because they are binary.

The binary predictor variables `coup.try.5yr`, `pol_killing_approved`, `freemove_men4`, `social_inequality`, `candidaterestriction`, and `anymk.ever` appear to be binomially distributed. However, the continuous predictor variables `popsize.ln.combined` and `imr.sqrt` may be (approximately) Normally distributed since the number of observations n is large. However, when the probability densities are plotted (below), it appears that the distribution for `imr.sqrt` is right-skewed and multimodal. The distribution for `popsize.ln.combined` also appears right-skewed and unimodal.

Probability Density of Binary Variables



Probability Density of Continuous Variables

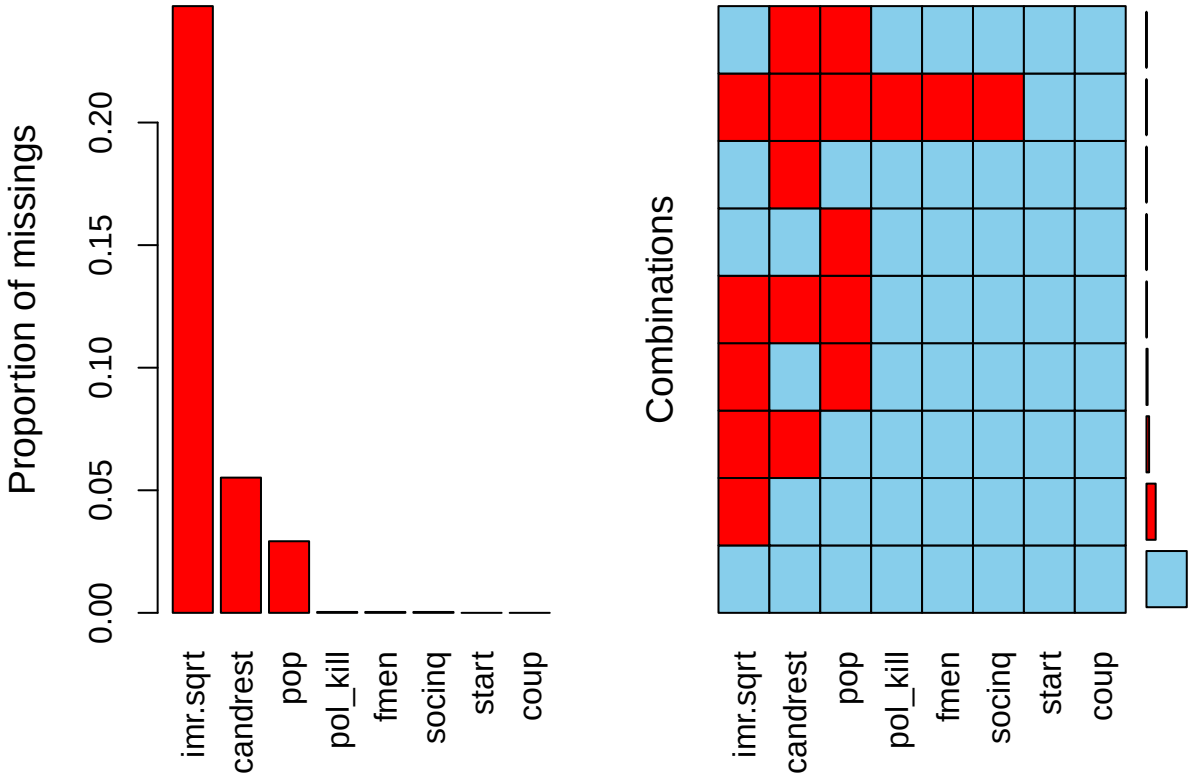


Both `imr.sqrt` and `popsize.ln.combined` have already been transformed (using a `sqrt()` and `log()` transformation, respectively). Reversing the transformations results in a density that appears more right-skewed, so the transformations should be kept as is. The remaining variables are binary, so performing a transformation on any of them would not necessarily be helpful.

Missingness

Number of `NA`'s in each variable:

anymk.start	0
popsize.ln.combined	324
coup.try.5yr	0
pol_killing_approved	4
freemove_men4	4
imr.sqrt	2746
social_inequality	4
candidaterestriction	612




```
##
##  Variables sorted by number of missings:
##  Variable        Count
##  imr.sqrt 0.2474988734
##  candrest 0.0551599820
##       pop 0.0292023434
##  pol_kill 0.0003605228
##      fmen 0.0003605228
##    socinq 0.0003605228
##     start 0.0000000000
##      coup 0.0000000000
```

There appears to be missingness in three variables: `popsize.ln.combined`,`imr.sqrt`, and `candidaterestriction`. Due to the sheer volume of missingness in `imr.sqrt` in particular, it is not reasonable to assume that this missingness is completely at random (MCAR).

Missing Data Imputation

```
set.seed(123456)
imputed_dt <- mice(earlywarning_subset, m = 5)
# summary(imputed_dt)

# You can select one of the 5 data sets or leave them as a list
final_impute <- complete(imputed_dt,5)

# Check for na (should be 0)
sum(is.na(final_impute))
```

Analysis

Modeling Strategy

The outcome of interest Y is a binary variable that can take on values of 0 or 1. Thus, it is reasonable to choose a model whose predicted outcome can only take on values between 0 and 1. An appropriate modeling strategy is to fit a logistic regression model to assess whether there is a statistically significant association between the predictor variables and the outcome of interest, as well as the direction of these potentially significant associations. The use of a regular linear regression model would be inappropriate because this model does not satisfy the constraint that the outcome $Y \in [0, 1]$.

Model Summary[7]

Table 1:

	Dependent variable:
	as.factor(anymk.start)
popsize.ln.combined	0.183***
	(0.067)
as.factor(coup.try.5yr)1	0.671***
	(0.193)
as.factor(pol_killing_approved)1	0.348*
	(0.200)
as.factor(freemove_men4)1	−0.495**
	(0.212)
imr.sqrt	0.173***
	(0.036)
as.factor(social_inequality)1	0.122
	(0.231)
as.factor(candidaterestriction)1	0.213
	(0.335)
as.factor(anymk.ever)1	18.801
	(530.148)
Constant	−26.810
	(530.150)
Observations	11,095
Log Likelihood	−532.759
Akaike Inf. Crit.	1,083.517
Note:	*p<0.1; **p<0.05; ***p<0.01

[7]Hlavac, Marek (2022). stargazer: Well-Formatted Regression and Summary Statistics Tables.

Summary of Results and Causality

Model Interpretation

Odds Ratio for Statistically Significant Predictor Variables ($\alpha = 0.05$):

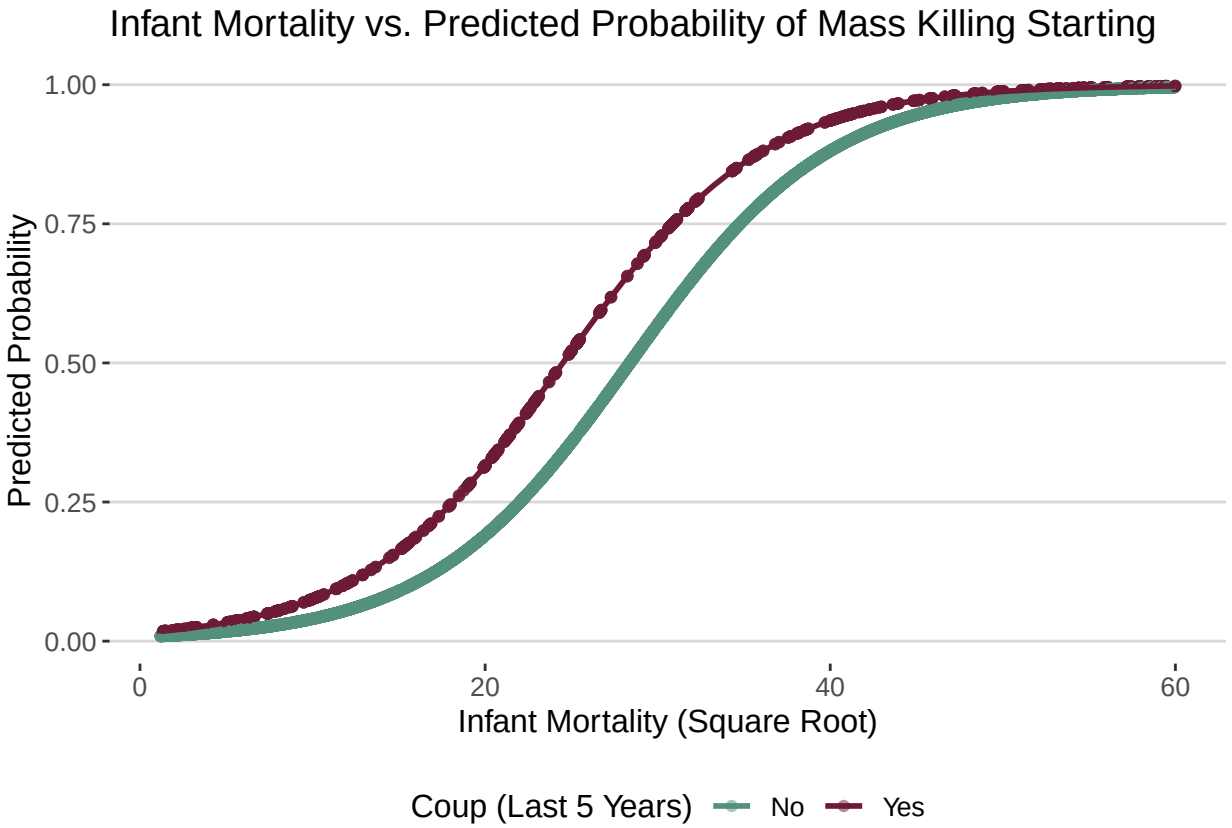
> Holding all other predictors constant, a **one-unit increase in log population** corresponds to a multiplicative change of **1.201** in the **odds of a mass killing event starting** (a 20.1% increase).
>
> Holding all other predictors constant, the **odds of a mass killing event starting** if there has been a **coup attempt** in the current year or the prior four is **1.957 times higher** relative to if there has not been a coup attempt (95.7% higher).
>
> Holding all other predictors constant, the **odds of a mass killing event starting** if virtually all **men enjoy full freedom of movement** is **0.61 times lower** relative to if men do not enjoy full freedom of movement (39% lower).
>
> Holding all other predictors constant, a **one-unit increase in infant mortality (square root)** corresponds to a multiplicative change of **1.188** in the **odds of a mass killing event starting** (a 18.8% increase).

Odds Ratio for Non-Statistically Significant Predictor Variables ($\alpha = 0.05$):

> Holding all other predictors constant, the **odds of a mass killing event starting** if **political killings** are practiced systematically and they are typically incited and approved by top leaders of government is **1.416 times higher** relative to if there are not political killings (41.6% higher).
>
> Holding all other predictors constant, the **odds of a mass killing event starting** if members of some **social groups enjoy much fewer civil liberties** than the general population is **1.129 times higher** relative to if this is not the case (12.9% higher).
>
> Holding all other predictors constant, the **odds of a mass killing event starting** if **candidates are restricted** is **1.237 times higher** relative to if candidates are not restricted (23.7% higher).
>
> Holding all other predictors constant, the **odds of a mass killing event starting** if there has been **any mass killing episode ever** is 1.4629154×10^8 **times higher** relative to if there has not been any mass killing episode ever (1.4629154×10^{10}% higher).

This model cannot be used to make causal claims. The dataset used is made up of observational data, and missing datapoints that are not missing completely at random. However, this model does shed light on variables that may be associated with onsets of a mass killing episode.

Predicted Probability Plot



The predicted probability of a mass killing starting increases as infant mortality (square root) increases. For countries that have experienced a coup or a coup attempt within the last five years, the predicted probability of a mass killing start is slightly higher.

[8]

According to the Early Warning Project, the Democratic Republic of the Congo is number 18 on the list of countries most at risk for a new mass killing. Here, we will compare the predicted probability of a new mass killing episode from our model and from the Early Warning Project.

Model Predicted Probability of a New Mass Killing Episode:

United States: $9.3 \times 10^{-9}\%$

Democratic Republic of the Congo: 2.8%

Our model predicts a smaller probability (2.8%) than the Early Warning Project (4.3%), which may be due to the fact that our model includes a much smaller subset of the variables includes in the Early Warning Project's model. However, it still captured the general trend, with the United States (a country not at risk for mass killing) having a very small predicted probability, especially compared to the probability for the Democratic Republic of the Congo. While 2.8% may seem like a small probability, it is much higher than low risk countries like the United States who have a near-zero probability of a new mass killing episode.

[8] https://www.ushmm.org/genocide-prevention/blog/countries-at-risk-for-mass-killing-2022-23

Conclusion

Hypotheses

H1: Higher population is associated with higher odds an onset of a mass killing episode.

H2: A coup attempt (unsuccessful or not) within the last five years is associated with higher odds an onset of a mass killing episode.

H3: Political killings are associated with higher odds an onset of a mass killing episode.

H4: Higher infant mortality rates are associated with higher odds an onset of a mass killing episode

H5: Social inequality is associated with higher odds of an onset of a mass killing episode.

H6: Restrictions on political candidates are associated with higher odds of an onset of a mass killing episode.

H7: Previous mass killings are associated with higher odds of an onset of a mass killing episode.

H8: Freedom of movement for men is associated with lower odds of an onset of a mass killing episode.

Overall, the direction of the associations between our predictor variables and the response estimated by our model line up with those of our hypotheses. However, not all of our predictor variables were statistically significant. Out of the eight, only four were significant at a significance level of $\alpha = 0.05$: log-population, coup (successful or not), freedom of movement for men, and infant mortality. Political killings, social inequality, candidate restrictions, and previous mass killings were not significant. The coefficient for previous mass killing episode(s) was very large and positive (previous mass killing is associated with a higher risk for onset of a mass killing episode), so the lack of statistical significance for this variable was surprising. These findings provide support for the argument that both political and economic upheaval influence the odds of an onset of a mass killing episode.

While mass killings episodes and genocide are not random, this model (and other models) are not based on experimental data. The world is not a laboratory, and there are many factors that can impact when, where, and how these events occur. However, this model does provide insight into risk factors that impact the likelihood of a mass killing episode or genocide. Informed by studies of historical instances of mass killing and genocide, regression models can forecast when and where these events are more or less likely to occur. While our inability to make causal arguments is a limitation, these models are one of the best tools the world has. They allow us to better anticipate genocides and mass killings, at least in theory making the international community better equipped to prevent them from occurring.

Furthermore, our analysis is limited in that our data format and modeling approach is at the state-and-year-level. While studying the impact that state-level dynamics have on mass killing episodes and genocide is extremely important, there is a need for more research on risk factors at the subnational level. Genocide is not always committed by state actors (e.g. ISIL in Iraq/Syria); conducting research using a unit of analysis of individual conflicts and mass killing episodes, as well as on risk factors within regions and communities, is needed to more fully understand genocide and mass killing episodes.[9]

[9]Nyseth Brehm, Hollie (2019) "Moving Beyond the State: An Imperative for Genocide Prediction," Genocide Studies and Prevention: An International Journal: Vol. 13: Iss. 3: 64-78. DOI: https://doi.org/10.5038/1911-9933.13.3.1673