36-401 Data Exam 2: Invesigation Civil War Outbreaks

Anusha Bhat

2024-02-29

1 Introduction

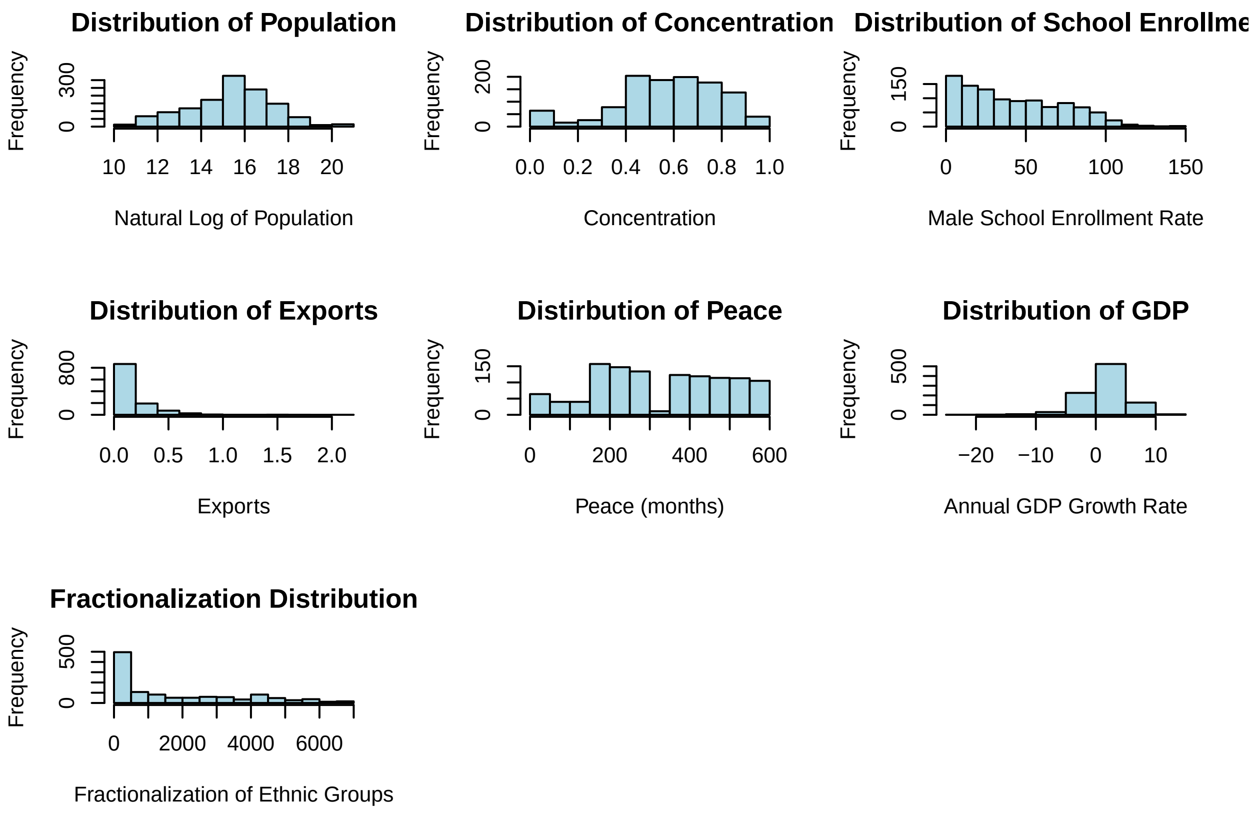
1.1 Identifying the Problem

In this report, we aim to investigate the factors that influence the outbreak of civil wars within nations. Understanding these influences can help us with our understanding of current and past human history, as well as, elucidating potential areas of society that we can target to potentially prevent future outbreaks of civil wars. We will investigate two theories in particular thar hypothesize what vulnerabilities facilitate outbreaks of civil wars within a country. The first theory suggests that for countries that are heavily dependent on commodity exports to sustain their economy, rebels can seize, and sell some of these commodities, allowing for civil wars to become easier to start and maintain in these countries. The second theory suggests that civil wars are more likely to form in countries which have one ethnic group dominating the politics and economy, along with strong ethnic divisions within the society. We will specifically implement a model using quantitative data to evaluate these two claims.

1.2 Elementary Data Analysis

Our data file was provided by Professor Cosma Shalizi who gathered this data from a study conducted by another professor regarding the causes of civil wars. The variables in the data set are given country name, year the data was collected (over a five year period), an indicator of whether a civil war began during that period, or was ongoing, or if there was peace, exports reflecting a country's dependency on commodity exports, male secondary school enrollment rate, growth rate for the gross domestic product (GDP), an index of geographic concentration of the country's population, how many months it has been since the country's last war or the end of world war 2 based on which one was more recent, natural logarithm of the population, an index of the country's divide along ethnic lines, and an index of ethnic dominance. Essentially, one row highlights for a particular country, whether there was outbreak of war, an ongoing war, or peace in that time period, as well as other social and economic metrics.

We note that there are 600 rows containing NAs in our data set, out of a total of 1,288 rows. In particular, NA can indicate that a civil war was ongoing for the third variable mentioned or it can indicate there is missing data for the other metrics. We also note that there are some percentages over 100% for male secondary school enrollment rate, though, there is no answer as to why this is. The variables of importance to us are indication of civil war outbreak, exports, male secondary school enrollment rate, GDP, population, fractionalizaton, peace, concentration and ethnic dominance. In figure 1, we display the distributions of some of these variables. We note that some of the distributions are skewed and non-normal, indicating that we may need to perform a possible log transformation for these variables in our model.



Additionally, we see that approximately 9.39% of the data had an ongoing civil war during the time period observed, approximately 6.06% of the data had no civil war outbreak during the period, and approximately 84.55% of the data had a civil war outbreak within the five year period.

2. Model

2.1 Model Selection

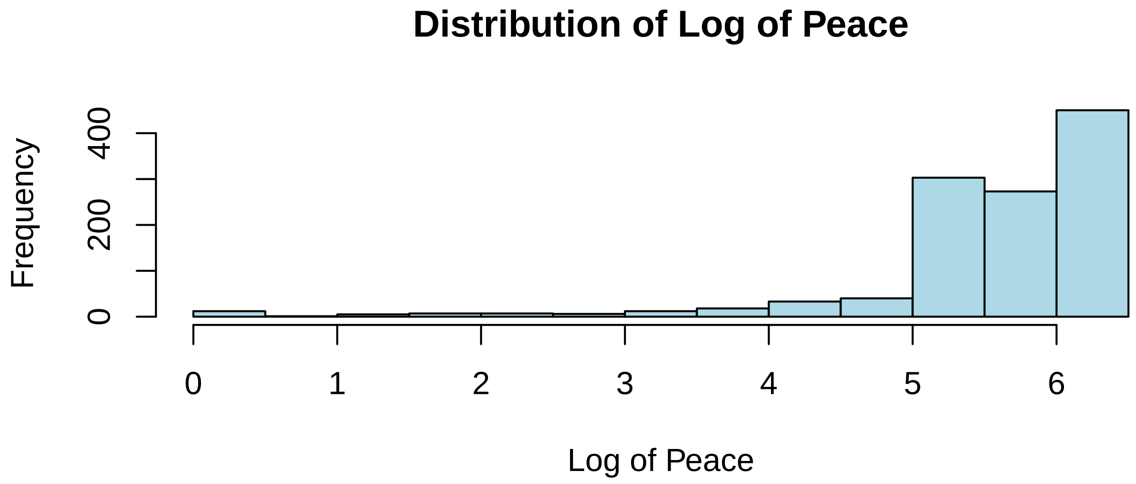
Since we are not particularly investigating the linear associations between these variables and civil war outbreak, we decided to forgo using linear regression as our model. Due to the use of several predictor variables, a generalized additive model will be useful to us. In the model, the response variable “start” of civil war is categorical, so we set the family as binomial in our model. Due to the binomial family, our model will thus predict the log offs of success as well as the probability of success. Additionally, since the model is additive, we can observe the different contributions of each predictor variable to the response through the partial response functions. For the model, we can write the formula as

$$g(\mu(x)) = \alpha + \sum_{j=1}^{\alpha} f_j(x_j)$$

where each f_j is a smooth function of the predictor x_j. We note that these smooth functions may possibly be linear, but not necessarily.

When determining what predictor variables to use, we began with a model including all variables and eliminated one variable at a time. If the cross-validated mean squared error (MSE) decreased after removing the variable, then that variable was omitted from our model. We note that all cross-validated mean squared errors reported in this analysis is produced using 5 folds. Below is the construction of the model. We note that code used for this report was sourced from the sample student paper provided by Dr. Shalizi and from the 36-402 course notes.

This final combination of the predictor variables selected to construct our model shown above, resulted in the lowest cross validated MSE of approximately 0.0588. Furthermore, the cross validated MSE without the log transformation of exports was approximately 0.0602. Thus, we decided to include the log transformation as it improved the model's predictive accuracy. We note that peace has a non-normal distribution which could cause bias in our model's prediction. However, a log transformation resulted in a right skewed distribution (shown in figure 2), so we omitted this transformation from our model.



2.2 Model Analysis

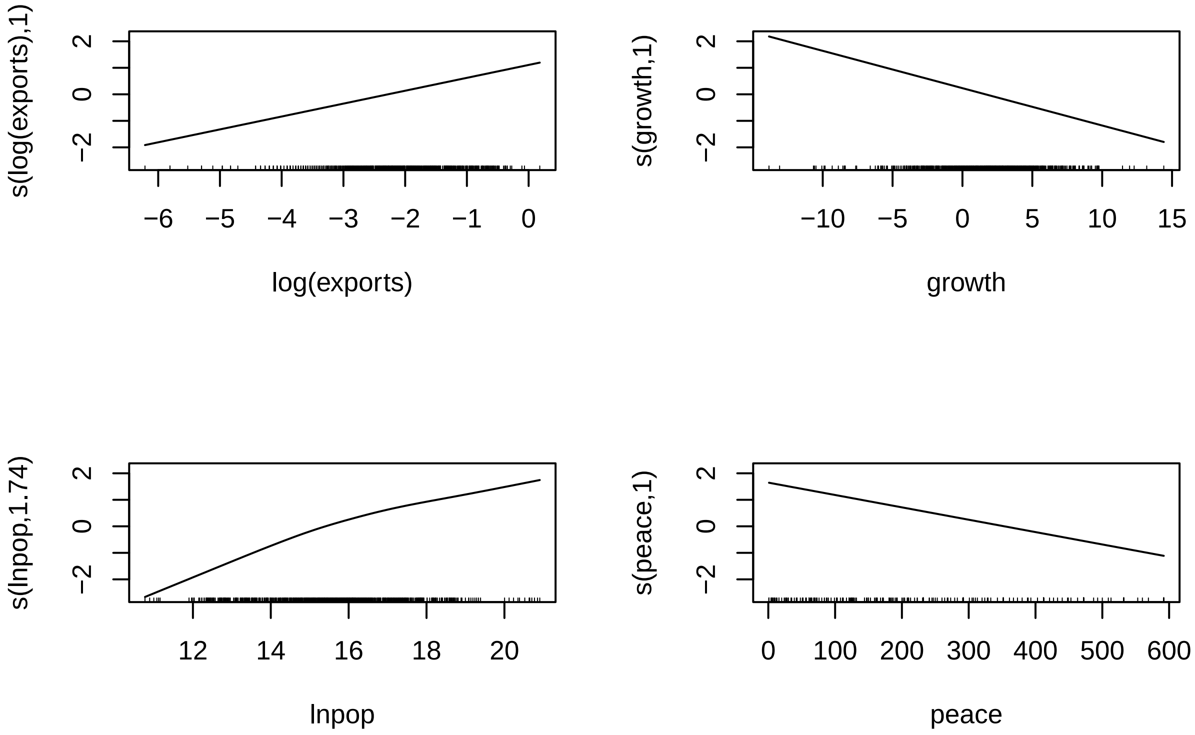
In our final model, we have one parametric model, dominance, since it is binary, and four nonparametric variables, peace, exports, male school enrollment rate, and GDP growth. A confidence interval, along with a point estimate, for dominance is included in table 1.

	lo	est	hi
(Intercept)	-3.60	-3.37	75.52
dominance	-0.61	0.24	0.79

The confidence intervals were constructed using bootstrapping for resampling cases. We chose to use bootstrapping with resampling cases since our response variable is categorical and we do not want to rely on the correctness of the model in our simulation.

The coefficient for dominance is positive, indicating that the odds of a civil war breaking out is approximately 1.27 more likely if the country has one ethnic group dominating the government and economy compared to a country that does not have a dominating ethnic group. We note that the 95% confidence interval for this point estimate covers 0.

The partial response functions are displayed in figure 3, along with their confidence intervals in figure 4 determined from the same bootstrapping method mentioned previously.

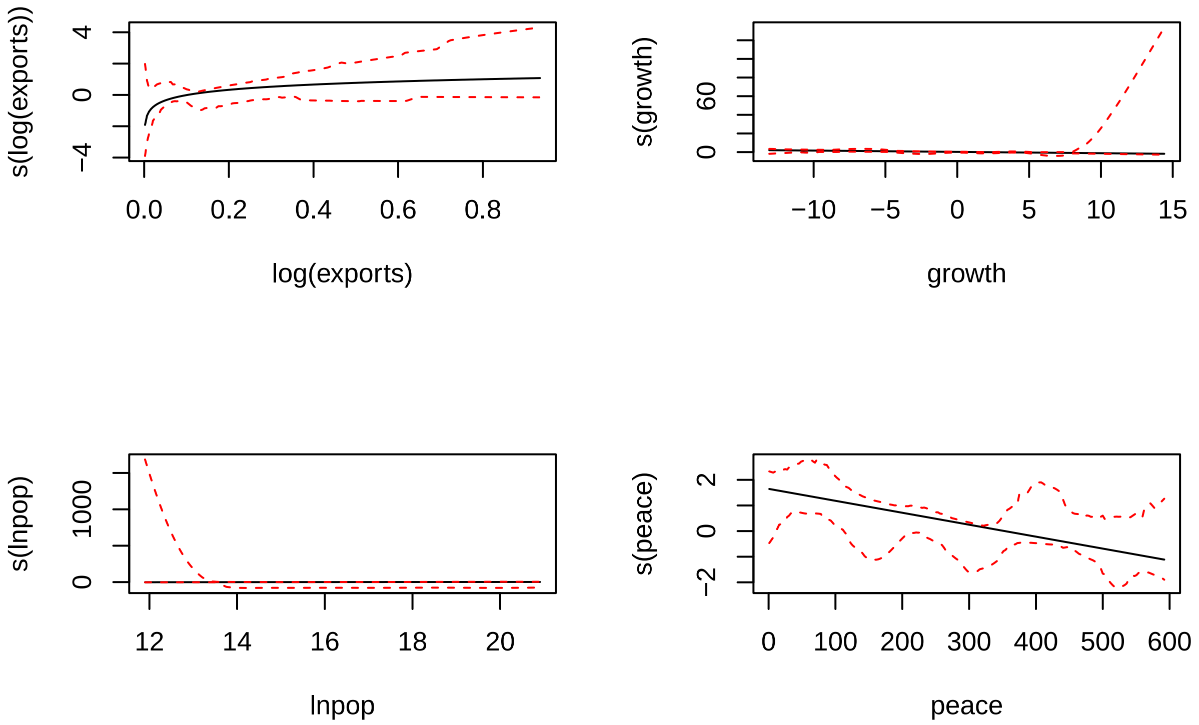


The partial response to the log of exports increase as the log of exports increases. The response function increases from a low level of odds to a high level of odds as the log of exports increases. Since the log of exports increases as exports increases, we observe that civil wars are more likely to start when there are more commodity exports.

The partial response to the growth of GDP decreases as the growth rate decreases. The response function decreases from a high level of odds to a low level of adds as the growth rate increases. This suggests that civil wars might be less likely to start if the growth rate of the GDP is lower. This is an interesting result in contrast to the exports, as one could suspect that an increase in exports is in tandem with an in crease in growth. Further analysis into this topic could yield interesting results.

The partial response to the log of the population increases from low odds to high odds as the log of the population increases. This suggests that civil wars are more likely to start if there are more people in the country.

The partial response to pease decreases, beginning at high odds and decreases to low odds as peace increases. This suggests that civil wars are less likely to start as the duration of time since the country's last war or World War 2 (depending on which event was more recent) occurred increases. This makes sense as people may be more hesitant to disrupt the peace the longer there is no war, and/or the society may be working towards preventing civil wars to sustain peace. This could also be an interesting area for further research.



The confidence intervals for the partial response functions for log of the population and growth seem to be similar but opposite–the interval starts out wide then narrows around the function for population, whereas, the interval starts narrow then widens for growth. We observe wide bands for the log of exports and peace, although the bans widen and narrow and are more wiggly for peace. We also observe that all of the confidence bands for each of the partial response functions include 0 in the entirety of the intervals. Despite this, we can not conclude whether or not the predictors should be removed from the model without further analysis. Additionally, we can not conclude whether any particular predictor is significant since all of the intervals overlap zero (whereas we could determine significance if the interval did not include 0).

2.3 Model Diagnostics

In this section we will check the goodness of fit of our model to determine validity of the analysis from section 2.2.

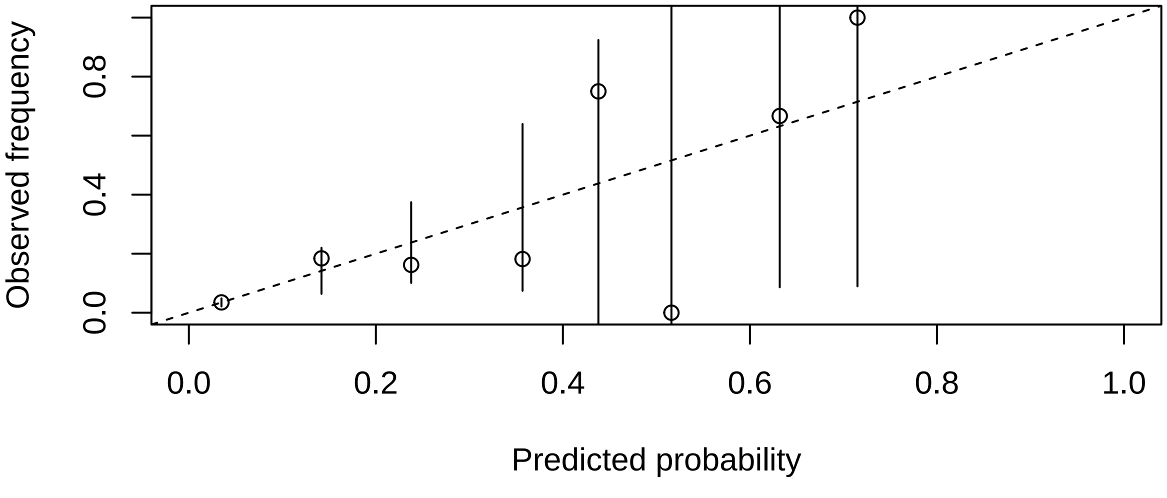
2.3.1 Prediction Classification

First, we will check whether the model can predict the outcome accurately. The in-sample error of our model is approximately 0.0623, while, the in-sample error of a basic model which predicts the majority class everytime has an in-sample error of approximately 0.33. Therefore, our model predicts better than the basic model for in-sample predictions. Our model's error rate improves under five-fold cross-validation, with an MSE of 0.0588. This indicates that the model is wrong 5.88% of the time under cross-validation, showing that we have predictive power in our model.

2.3.2 Calibration

Next, we will check the calibration of the probabilities.

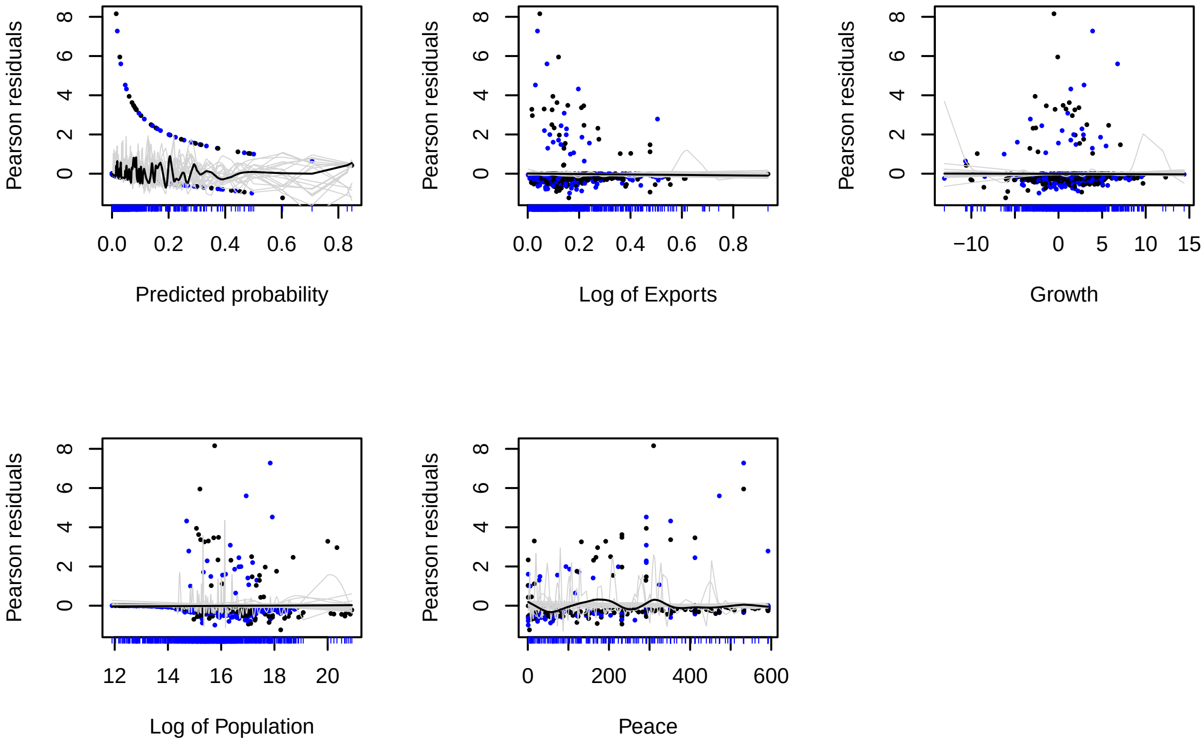
In figure 5, we display the calibration plot for the logistic regression done by our model, using probability bins of width 10 percentage points. We can see that the vertical lines, which indicate the 95% confidence intervals for the observed frequency of each predicted probability,

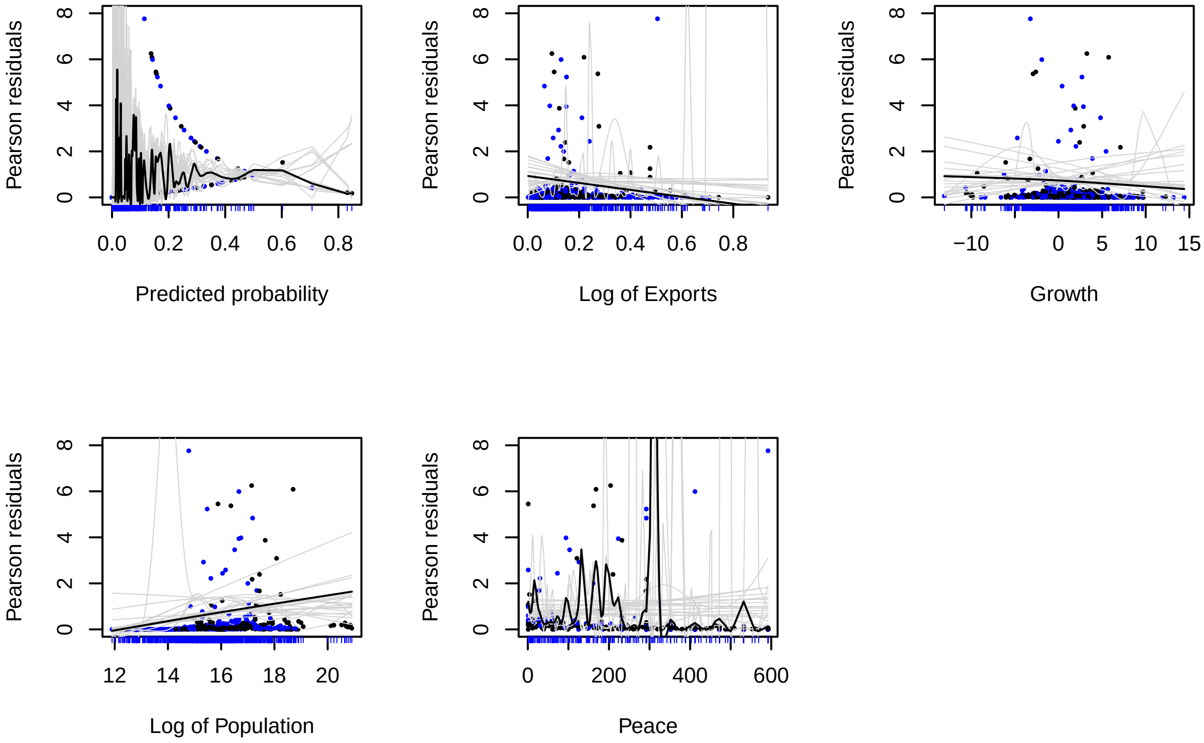


As the predicted probabilities increase, there is a fluctuation between an increase and decrease in observed frequency. The observed frequencies of the predicted probabilities increase until 0.2, then decrease and stay the same from 0.2-0.4, increase at 0.4, decrease at 0.5, then again increase from 0.6 onwards. Moreover, the 95% confidence intervals of the points are quite large, indicating that our model may not be as well calibrated as we would hope.

2.3.3 Residuals

Next, we will check the residuals of our model against each predictor using pearson residuals. Since our outcomes are categories, we need to standardize the residuals to pearson residuals. We standardize by using the equations $\frac{y_i - p_i}{(p_i(1-p_i))^{1/2}}$. Figure 5 shows plots of the regular pearson residuals, whereas, figure 6 shows plots of the squared pearson residuals. For figure 5, we expect the residuals to fluctuate around y = 0 with no pattern to the fluctuation. In gray are the simulated new responses based on our fitted model. For figure 6, we expect the residuals to fluctuate around y = 1. In gray are the simulated new responses based on our fitted model. In both figures 5 and 6, the predicted response lines do not warrant further attention. The squared residuals do seem of concern as they do not evenly fluctuate about y = 1. Furthermore, the pearson residuals (not squared) observe some heteroskedascity which could potentially influence our model. However, they are not distributed in a strong pattern, so it is fine for our purposes.





3 Conclusions

The positive point estimate of dominance suggests that as dominance of one ethnic increases, civil war is more likely to start. We also observe in the partial response functions that an increase in population and an increase in commodity exports is associated with higher odds of a civil war starting, whereas, a higher GDP growth rate and a longer duration of peace are associated with lower odds of a civil war outbreak. However, we do note that all of the confidence intervals for the point estimate of dominance and for the four partial response functions overlapped 0. Due to this we can not readily determine significance of variables and whether or not certain predictors should be removed. From our analysis, it appears that a country with low ethnic dominance, a high GDP growth rate, a long duration of peace, and low commodity exports would have low odds of a civil war outbreaking in a five-year time period.

Our results support the two theories that countries with more ethnic dominance are more prone to civil war outbreaks as well as countries who have a higher economic dependence on commodty exports. Specifically, we found that dominance, log population, (log) exports, peace, and growth, can be used to predict the start of civil wars. Compared to other variables, these predictors improved the prediction accuracy of our model, whereas, the excluded predictors in our dataset served to worsen the model.

We must take these results and conclusions with a grain of salt. Through model checking, we determined that our model may not be well calibrated. Furthermore, our residuals observed some heteroskedascity that may influence the model, although it is not to an extreme amount. Since the residuals and simulated responses in the residuals are doable for now, we proceeded with this model for our report. Additionally, through observing classification errors for in-sample and out-of sample, we found that our model performs better than a baseline model that only predicts the most common class. Our model makes an out-of-sample error less than 6% of the time, highlighting its predictive power. However, future work can be done to improve the model to further validate these conclusions.

Future analysis could take into consideration interactions which our model did not account for as well as all possible combinations of predictor variables. We only tested a few predictor variables and their impact on cross-validated MSE due to time constrains, however, testing all possible combinations (or at the very least, more than what this report did) could lead to a better predictive model that can more accurately answer these questions. Further data collection could also look into other aspects of society that influence politics, economics, and strike (such as percentage of people who are poor) to identify more metrics that may help us elucidate the accuracy of the two theories posed.