36-402 Exam 1

Eu Jing Chua, eujingc

Introduction

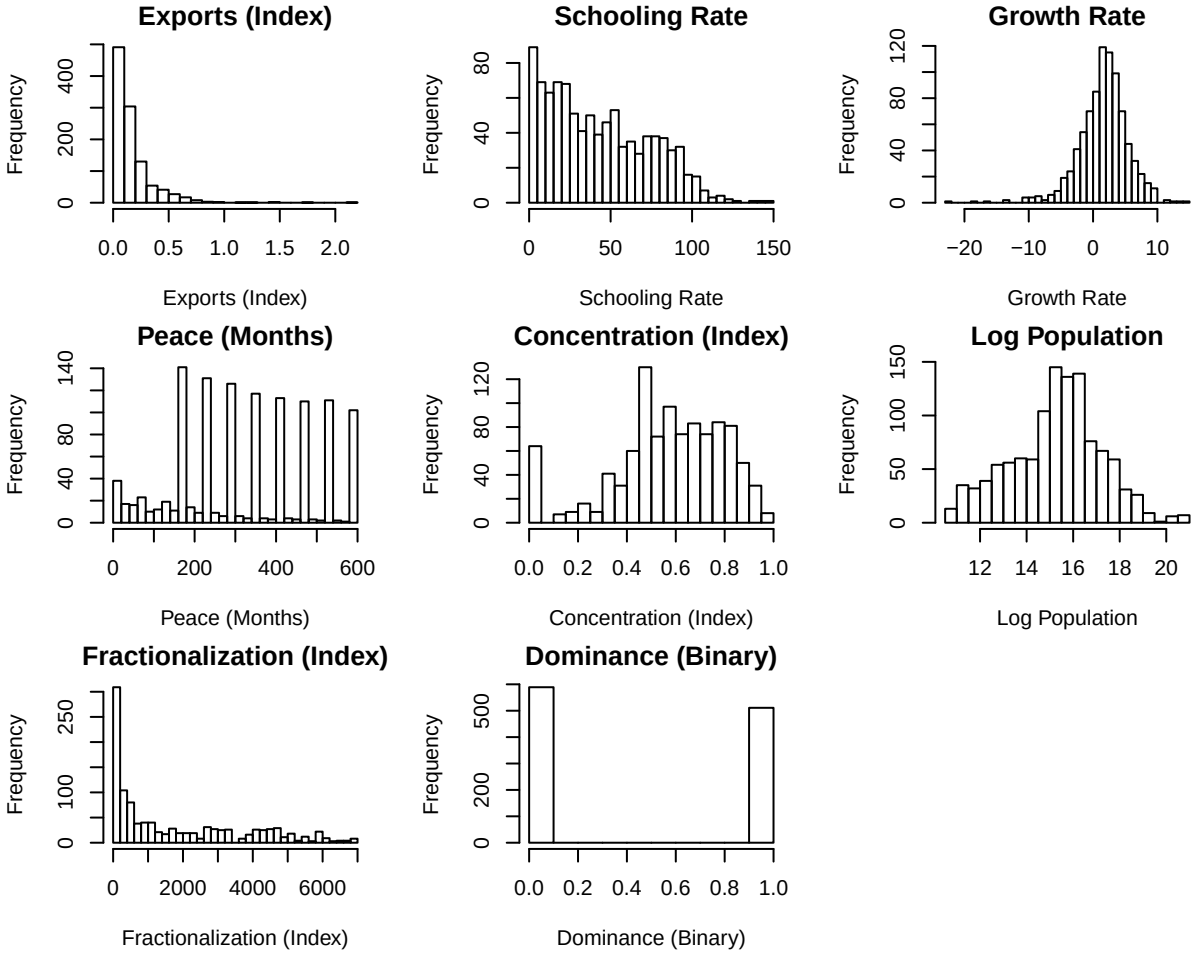
Civil wars, in contrast to wars between states, have been happening more frequently ever since the end of World War II. If we can predict the outbreak of civil wars, we can better allocate resources to helping those that will be affected and potentially save lives and protect important infrastructure. In particular, the results of our modelling will be used to discuss two leading-theories for likelihood of civil wars:

1. Civil wars are easier to start and maintain in countries whose economies are heavily dependent on commodity exports, where rebels can seize, and sell, some part of the commodity production.
2. Civil wars tend to start in countries where there are strong ethnic divisions, and one ethnic group dominates the government and economy.

We attempt to model the outbreak of civil war from a dataset with a binary response with 1 for outbreak and 0 for no outbreak, to be predicted from the following possible predictors:

1. Exports: Index of country's dependence on commodity exports, with higher values indicating higher dependence
2. Schooling: Secondary school enrollment rate for males
3. Growth: Growth rate of GDP
4. Concentration: Index of population's geographical concetration, where higher values indicate higher concentration
5. Natural log of population size
6. Peace: Number of months of peace since last war (or end of WWII, taking the most recent)
7. Fractionalization: Index of fractionalization of population, where higher values indicate more ethnic and religious division
8. Dominance: Precense of ethnic dominance (binary).

EDA



Models

Baseline Model

In the baseline, we use a general linear model (logistic regression) to model the relationship between all the predictors, $\mathbf{X}$, and the binary outcome Y, whether a civil war is predicted to start or not.
Let $p_i = P(Y = 1 \mid \mathbf{X} = \mathbf{x}_i)$, then

$$\log \frac{p_i}{1 - p_i} = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \beta_4 x_{i,4} + \beta_5 x_{i,5} + \beta_6 x_{i,6} + \beta_7 x_{i,7} + \beta_8 x_{i,8} + \epsilon_i$$

where $x_{i,1}$ represents the dependence on exports, $x_{i,2}$ the schooling rate, $x_{i,3}$ the GDP growth rate, $x_{i,4}$ the duration of peace, $x_{i,5}$ the concentration index, $x_{i,6}$ the log population, $x_{i,7}$ the fractionalization index, $x_{i,8}$ the presence of ethnic dominance, and ϵ_i is a noise term.

\begin{table}[!h]

\caption{Baseline Model Estimated Coefficients with 95% C.I}

	Coefficient	Lower	Upper
(Intercept)	-9.1753	-12.3012	-4.2993
exports	3.2801	1.2916	4.7716
schooling	-0.0262	-0.0407	-0.0054
growth	-0.1360	-0.2202	-0.0614
peace	-0.0041	-0.0058	-0.0020
concentration	-1.7644	-3.5679	-0.0644
lnpop	0.5706	0.3176	0.7444
fractionalization	-0.0001	-0.0003	0.0001
factor(dominance)1	0.6021	-0.2407	1.2013

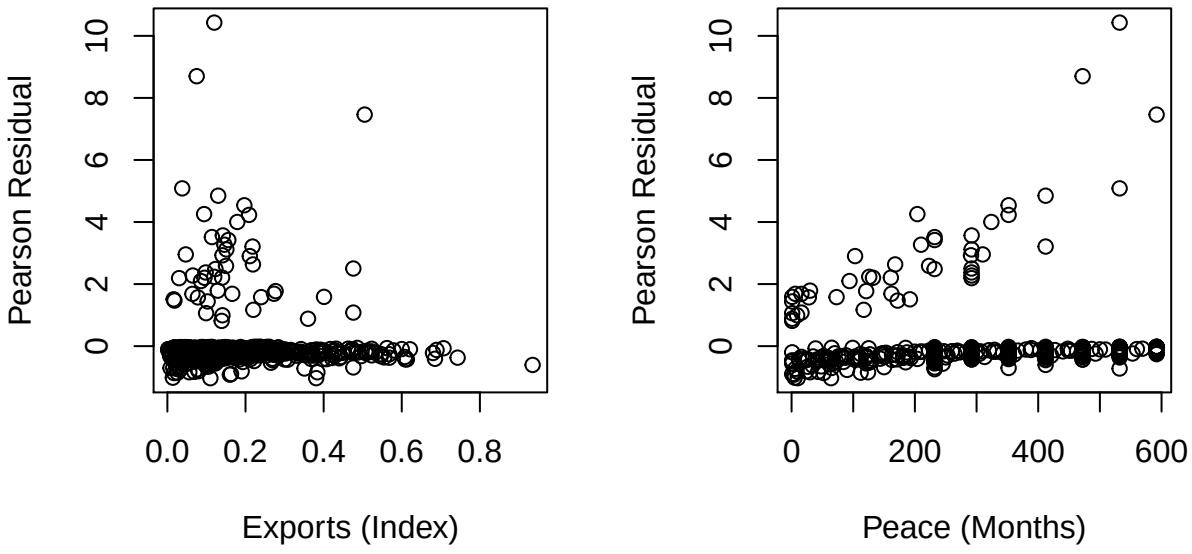
\end{table}

The estimated coefficients of the baseline model as shown above, using case-resampling and bootstrapping to generate 95% confidence intervals for the point estimates to reduce assumptions made about the data. On the whole, the model is dominated by the intercept and exports which have the biggest coefficients, with the other coefficients being relatively small. Taking the confidence intervals into account, we can see that some predictors such as concentration, fractionalization and dominance are not significant in the presence of all the other predictors.

Table 1: Baseline Model Properties

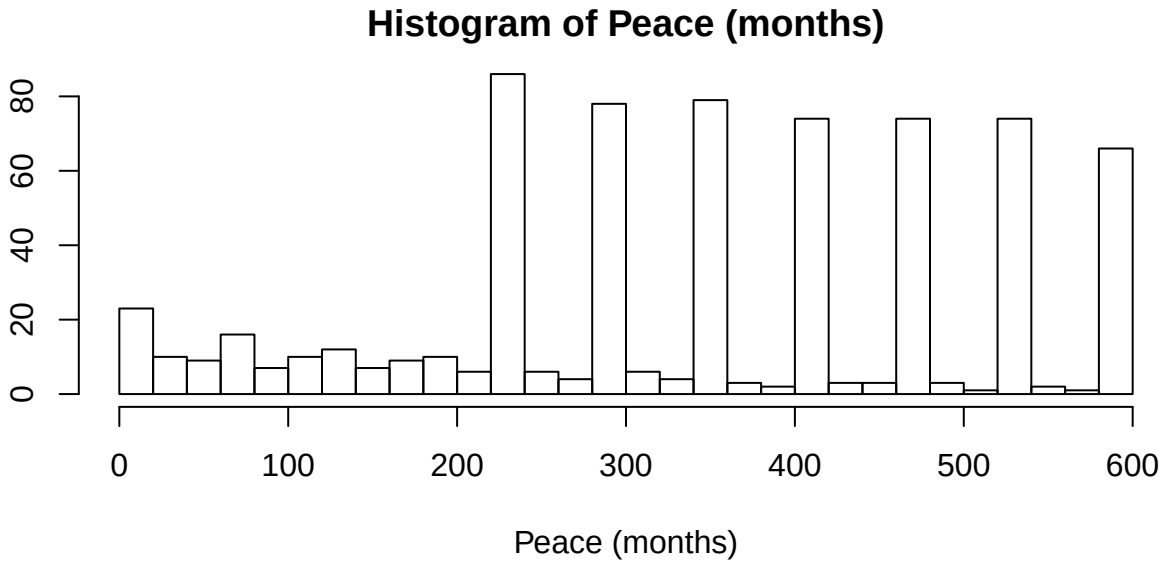
	x
CV MSE	0.0581
Brier Score	0.0543
Classification Err.	0.0669
Deviance	266.3587

Several metrics of the baseline model are also calculated above. It should be noted that the predicting the majority class in the dataset, where a civil war did not start, has a naive classification error of 0.0669, which is is closely reflected in the classification error of the baseline model. This implies that the baseline model is hardly an improvement over the naive predictor, showing very little predictive power.

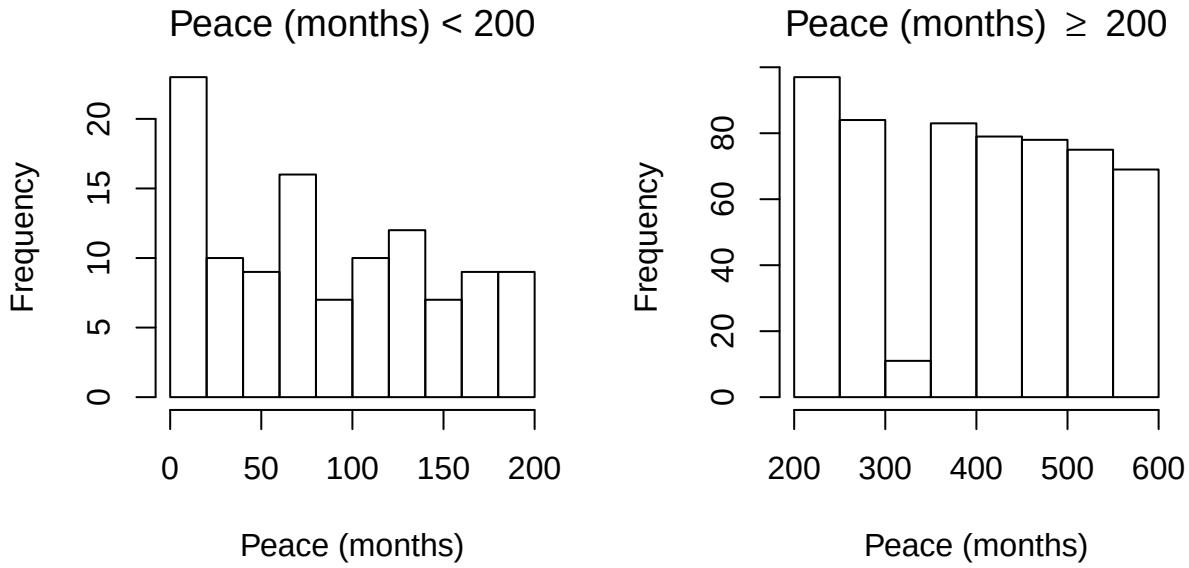


Residual analysis of the baseline model show that residuals on the whole are hardly centered around 0, with obvious clusters or outliers, and patterns as seem from the residual plots against exports and peace. The residuals seem to indicate that there are two clusters of data, with the larger cluster having residuals roughly centered around 0 without any pattern, and the smaller cluster disrupting the baseline model (examples shown above).

Splitting the Data



Building on the failures of the baseline model, we investigate further as to what could be causing the model to fail. The patterns in the residuals from the baseline model opened up the possibilities of there being two clusters in the data, which we may not be able to model both with just one model. By looking at the initial data exploration, we take a closer look at the distribution of peace duration across the data and postulate that it is not as smooth as the rest, possibly coming from two overlapping smooth distributions instead.



We explore the possibility of there being two separate distributions for peace, and attempt to simply split

them at a threshold of 200 months. By splitting the data, we can evaluate separate models on each split to see if indeed they are better modelled separately. We refer to the split with lesser than 200 months of peace as the lower data set, and the complement the greater data set.

Baseline Model on Each Split

Estimates

The original general linear model is fitted back on each of the splits to produce two separate models, the lesser model from the lesser data and the greater model from the greater data. Once again, we use case-resampling and bootstrapping to generate 95% confidence intervals for the point-estimates of the coefficients.

\begin{table}[H]

\caption{Split Models Estimated Coefficients with 95% C.I}

	Lesser Model			Greater Model		
	Coefficient	Lower	Upper	Coefficient	Lower	Upper
(Intercept)	-11.4433	-16.7978	-8.9574	-11.3207	-16.4825	-9.3684
exports	11.3890	17.6642	21.5435	2.2732	-0.8696	2.7238
schooling	-0.0527	-0.0917	-0.0594	-0.0237	-0.0334	0.0035
growth	-0.2306	-0.3888	-0.2316	-0.0999	-0.1491	0.0226
peace	0.0042	0.0106	0.0147	-0.0047	-0.0076	-0.0030
concentration	-5.0146	-9.7076	-6.2460	-1.5421	-2.8933	0.4000
lnpop	0.7641	0.6981	1.1479	0.7085	0.6093	1.0449
fractionalization	-0.0003	-0.0007	-0.0004	0.0000	-0.0001	0.0002
factor(dominance)1	1.3631	1.4325	2.8491	0.4004	-0.3997	0.8054

\end{table}

At a glance, the two models have quite different coefficients. The lesser model has relatively larger exports, concentration, and dominance coefficients as compared to the greater model, which just has smaller coefficients on the whole. Both models have very small fractionalization coefficients, which is even non-significant in the greater model as seen from its 95% confidence interval that contains 0. Several other coefficients in the greater model are non-significant, overall being a model that is dominated by its intercept.

Properties

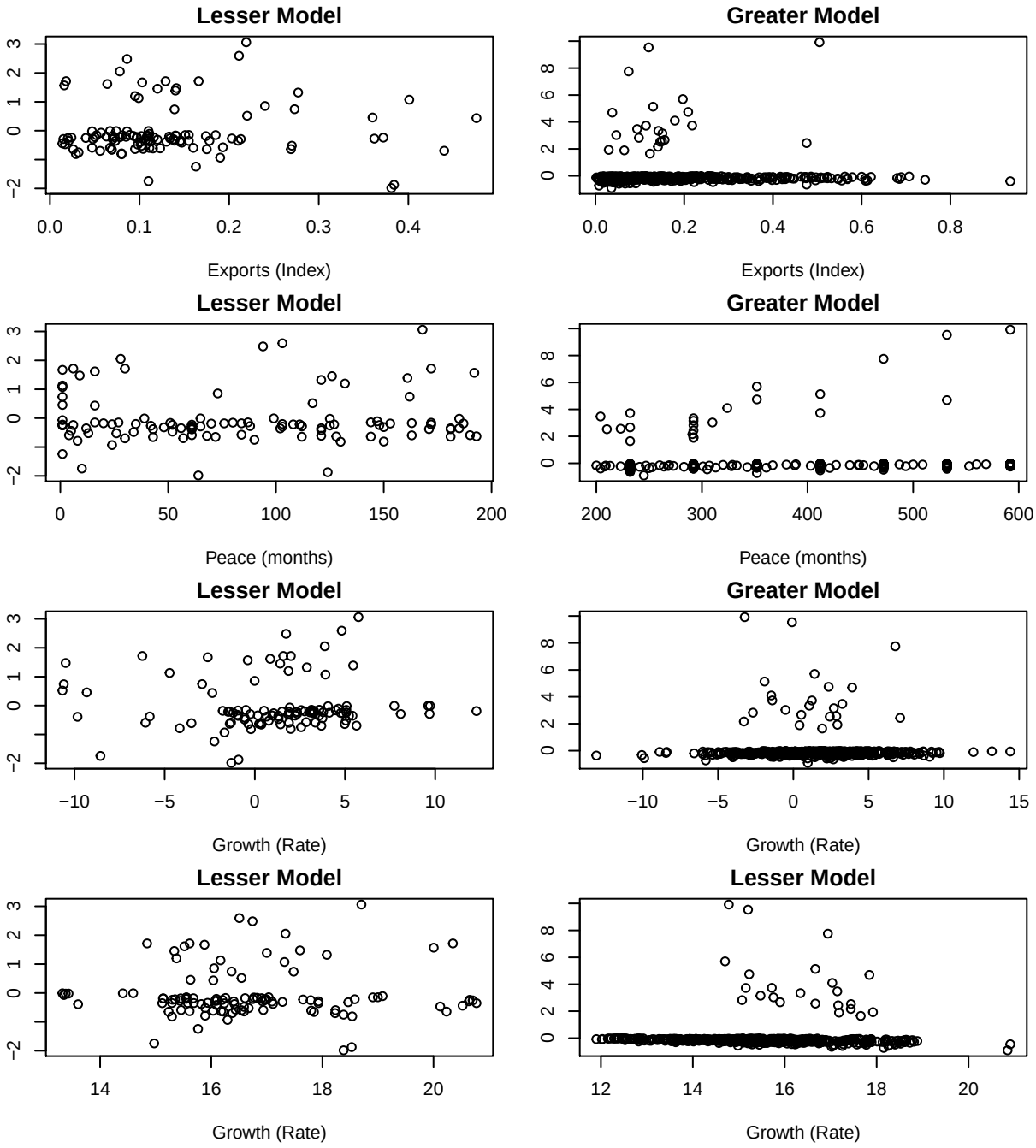
In this case, we know have two separate naive classification errors. for each split, as each split has a different proportion of positive responses. The lesser model has an in-sample classification error that is smaller than that of the naive classification error, but the greater model's in-sample classification error is still similar to its naive classification error. Although the greater model has a lower 10-fold cross-validation MSE as compared to the lesser model, it does not say much if it performs similarly to the naive classifier that predicts the majority class every time. The great difference in naive classification error shows that the class imbalance in the greater model is much more significant than the one in the lesser model. This could be a reason why the lesser model is able to have improved classification, relative to the greater model.

Table 2: Split Models Properties

	Lesser Model	Greater Model
CV MSE	0.1500	0.0385
Brier Score	0.1268	0.0363
Classification Err.	0.1875	0.0399
Naive Classification Err.	0.2054	0.0399
Deviance	85.7102	165.4125

Residuals

Analyzing the residuals for each of the models shows that in general, the lesser model exhibit better-behaving residuals compared to the Greater model, in the context of a general linear model. The residuals of the lesser model are morecentered around 0 with random spread (examples in left column below), while the residuals of the greater model still display the same problems we had in the original whole data set (examples in right column below). Namely, they are still hardly centered around 0, with a noticeable but smaller separate cluster with consistently higher residuals from the majority of the residuals around 0.



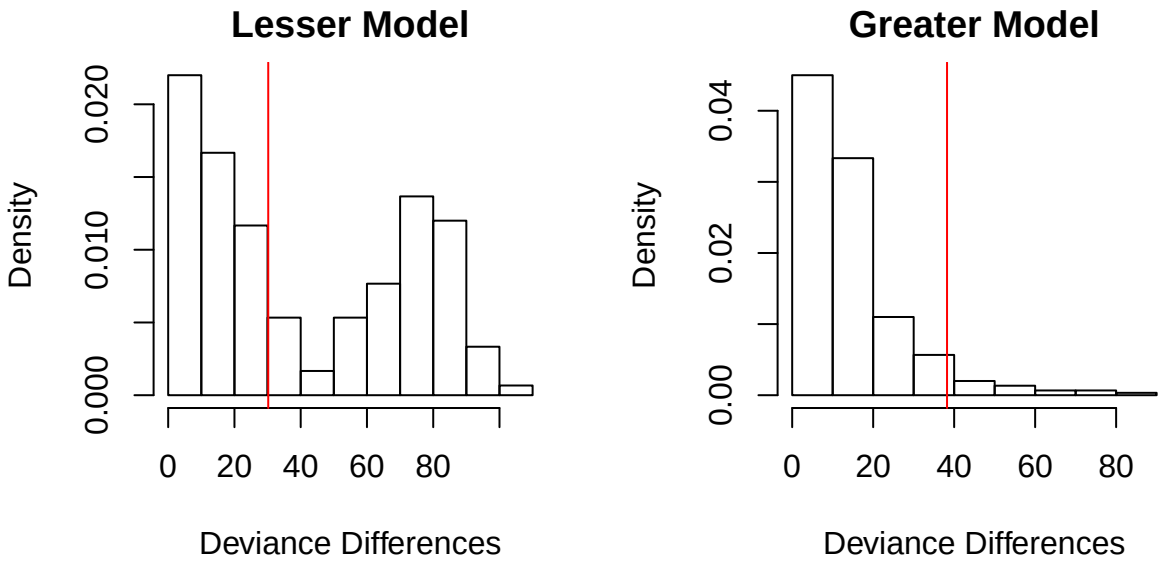
Model Checking

Goodness-of-fit can be checked by testing for the difference in deviance of our general linear models against general additive models that do not require the log-odds of civil war starting to follow a linear function, but instead just additive in smooth functions. We test both of the general linear models against a general additive model trained on the respective split of data, using spline smoothing each of the same continuous predictors, while keeping the binary predictors the same (dominance). The hypothesis of each test is H_0: The log odds follows a linear function of the predictors versus H_a: The log odds does not follow a linear function of the predictors. The test statistic of this test would be the difference in deviance between the general linear and general additive models. We use simulate generate data under the null hypothesis and fit

both models to the data to find the distribution of the difference in deviance under the null. We can then get a p-value for our observed differences in deviances from the simulated distributions.

Table 3: P-value of each test

	x
Lesser	0.4967
Greater	0.0567



Both tests have p-values that would not be rejected under 0.05 significance, so in both tests we fail to reject the null hypothesis that the log odds follow a linear function of the predictors. However, we can see that the deviance difference of the lesser model is well within its simulated null distribution, in comparison to the greater model that has a deviance difference relatively further from its simulated null distribution. On further inspection, we calculate the generel additive model's in-sample classification error to be 0.0347, which is quite close to the naive classification error of the greater data. This implies that the additive model did not gain much classification power from the greater data, and hence would have a higher deviance. Thus, due to the inability of both the linear and additive model to classify the greater data well, both have higher deviances that would seem similar in the hypothesis testing.

Results

From analysis of the two splits of data and the general linear models we have fit to each, we can conclude that for the split with peace duration less than 200 months, a logistic regression on all the predictors gives a good fit of the data. It has a low deviance, is an improvement on the naive classification error, has well-behaved residuals that are roughly centered around 0 that are quite uncorrelated with predictors, and satisfies our model-checking against a general additive model. The same cannot be said about the general linear model for the greater data set. Its residuals are not centered around 0, and display higher correlation with the predictors. Also, it demonstrably does not improve on the naive prediction error of the greater data set. Since the residuals do not satisfy the assumptions of logistic regression, the greater model cannot be reliably used for inference. Thus we will conduct inference with the lesser model, for peace durations less than 200 months.

We evaluate the two theories for patterns of civil wars starting in countries in the context of our lesser model, so we restrict ourselves to countries that experience less than 200 months of peace.

\begin{table}[H]

\caption{Lesser Models Estimated Coefficients with 95% C.I}

	Coefficient	Lower	Upper
(Intercept)	-11.4433	-16.7978	-8.9574
exports	11.3890	17.6642	21.5435
schooling	-0.0527	-0.0917	-0.0594
growth	-0.2306	-0.3888	-0.2316
peace	0.0042	0.0106	0.0147
concentration	-5.0146	-9.7076	-6.2460
lnpop	0.7641	0.6981	1.1479
fractionalization	-0.0003	-0.0007	-0.0004
factor(dominance)1	1.3631	1.4325	2.8491

\end{table}

The predictors of interest are index of dependence on exports, index of fractionalization and the presence of ethnic domination. The first theory postulates that civil wars are more likely to start in countries with higher dependence on exports. This is supported by the estimate of the coefficient of index of dependence on exports, which is positive, considering its uncertainty. When the index of dependence on exports increases while keeping everything else constant, the model predicts that log-odds of a civil war starting increases.

The second theory postulates that civil wars are more likely to start in countries with higher fractionalization and the presence of domination. However, this is not supported by the respective estimates of the coefficients. The coefficient of fractionalization is negative, considering its uncertainty, which implies that with higher levels of fractionalization while keeping everything else constant decreases the log-odds of civil war starting, which is contrary to the theory. It should be noted however, that the model predicts that the presence of domination does increase the log-odds of civil war starting.

Conclusion

From the results of the baseline model, we find that a general linear model is unable to effectively predict the outburst of civil war from all of the predictors. Further analysis revealed that there are potentially two clusters of data that a single general linear model might be unable to account for. The simple method of splitting the data on a single predictor (peace) and fitting separate models to each split proved to be fruitful, allowing us to fit a general linear model to the split with lower peace durations to carry out inference and evaluate the two theories. In the context of countries with less than 200 months of peace, we found that our model supports the exports dependence theory of civil wars starting, but not the fractionalization and domination theory.

However, no effective model was found for the other split with greater peace durations, which suffered from the same ill-fitting problems as the original baseline model. In order to generalize the results further, more work can be done to investigate if there are indeed clusters in the data that cannot be well modelled with a single model, like a general linear model, and if so how we can better split the clusters to fit seperate models for all the clusters to evaluate the two theories across all of the data instead.