Effect Heterogeneity and Variable Selection

? sketch the contours of mechanisms through which the broad relationship in their study - between pre-colonial conflict and modern day conflict - is realized. They argue that historical conflict affects modern outcomes through its consequences on subsequent economic, political and social outcomes. However, it remains to assess which of these mechanisms matter when and the authors demonstrate only that this pre-colonial conflict correlates with some indicators of trust and certain measures of local economic development.

There are several feasible ways to parse the mechanisms through which pre-colonial conflict matters for modern development outcomes, both theoretical and empirical. In this section, we discuss an automated variable selection method using the lasso algorithm (**?**). The basic idea behind such variable selection approaches is simple: given a set of candidate independent variables, what combination of them best explains the observed variance in the outcome variable?

The intuition behind lasso (the least absolute shrinkage and selection operator) is that we run a series of models and penalize the absolute size of the regression coefficients. In so doing, we constrain the sum of regression coefficients to be below a certain threshold (which we specify using cross-validation, as explained below), and if this threshold is exceeded then we penalize particular parameters by setting them to zero. The basic process is to sequentially add parameters to the model in the order of their bivariate correlation with the outcome variable, and gradually increase the coefficients associated with these variables until they are no longer the most strongly correlated with the outcome. Then, introduce additional explanatory variables until no other variables can be introduced into the model which are better correlated with the outcome subject to the coefficient size constraint.

In all models, we use cross-validation to select the optimal number of coefficients to keep in the model. Cross-validation randomly selects a subset of the overall dataset as a training set, then fits a model and assesses its performance on the rest of the dataset (the validation data). If the model predicts out-of-sample observations with a relatively low mean-squared error then it is selected. This enables us to select the optimal lasso-generated set of independent variables.

We argue that using the lasso is useful in the context of **?** for two distinct applications:

1. When the order of candidate independent variables is a similar order to the number of observations in a dataset, and including all of them would badly overfit the model.

2. When we believe that the way key explanatory variables interact with others is important for the underlying mechanism, but it is theoretically hard to predict which deserve focus.

These two applications allow us first to select variables in models that are otherwise overdetermined, and secondly to detect effect heterogeneity that the authors do not explore.

Cross-country analysis

Regarding the first application, we apply the lasso to the cross-country regressions the authors describe in Table 2 of their paper, which show that the number of years a given African country has been in civil war since independence is highly correlated with the number of wars in that territory between 1400-1700.

However, in their baseline specification they fit 9 independent variables to their 49 observations, and in their secondary specification they fit 29 independent variables to the same sample size. This represents a clear case of overadjustment (**?**) and the resultant models are highly overfitted to the data, with many regressors sharing high covariance. So, we use the lasso to define the subset of these many variables which best explain variation in the dependent variable, then re-run the regression to assess how the coefficients of interest are affected.

Column 1 in Table **WHAT** replicates the baseline cross-country specification of the authors, and column 2 uses the same set of regressors, applies the lasso, then re-runs the regression with only the selected variables. Column 3 replicates their saturated specification, and column 4 applies the lasso to the larger set of independent variables and run-runs the regression accordingly. In column 3, most of the regressors are excluded for display purposes.

The results of the lasso process suggest that, even in spite of the overdetermined model, their key explanatory variable is explaining parts of variation in the outcome that the other variables - many of which are post-treatment - do not. This derives from the fact that the pre-colonial conflict variable is selected in both columns 2 and 4, and the coefficient size and significance is stable across specifications.

Table 1: Cross-country lasso

	Dependent variable:			
	Civil war incidence			
	(1)	(2)	(3)	(4)
WarPrevalence14001700	0.12^{**}	0.13^{***}	0.13^{**}	0.14^{***}
	(0.05)	(0.05)	(0.06)	(0.04)
f_french	-2.91	-2.60	-4.12	-1.19
	(2.40)	(1.90)	(3.32)	(1.30)
f_pothco	5.65	5.69^{*}	9.96^{**}	6.90^{**}
	(3.62)	(3.35)	(4.68)	(3.37)
f_belg	1.04		-0.58	
	(2.99)		(3.99)	
f_italy	-3.45		-9.64^{*}	
	(3.33)		(5.54)	
f_germ	-3.34	-3.26	-4.44	
	(2.96)	(2.60)	(3.55)	
region_nNUNN	-2.52		-15.74^{**}	
	(3.29)		(5.55)	
region_sNUNN	-3.87		-3.08	
	(2.87)		(5.09)	
region_wNUNN	-5.07^{**}	-1.54	-5.25	-3.00^{**}
	(2.07)	(1.55)	(3.79)	(1.20)
region_eNUNN	0.32	3.92	13.22^{*}	2.35
	(2.69)	(3.26)	(6.52)	(2.72)
region_cNUNN		3.65^{*}		
		(2.08)		
Lasso	No	Yes	No	Yes
Specification	Baseline	Baseline	Secondary	Secondary
Variables omitted	No	No	Yes	No
Note:				*p<0.1; **p<0.05; ***p<0.01

The procedure suggests that the only other variables that meaningfully explain variation are whether the country is in the west or east of Africa, and whether it was a French or Portuguese colony. All the other controls are shown to contribute little to the explanatory power of the model.

Disaggregated analysis: Social attitudes

Regarding the second application of the lasso, in their more disaggregated analysis the authors use samples of a much larger size to demonstrate that historical conflict affects modern social attitudes but do not explore how the effect is moderated by any other factors. To try and get at the mechanisms through which pre-colonial conflict matters, we argue that looking at interaction effects between the key explanatory variable and other explanatory variables is an appropriate strategy.

For example, we may believe that pre-colonial conflict would affect the development of political institutions in a given country, which then affect the probability of civil wars (CITE SOMEONE). So, pre-colonial conflict may be expected to especially impact social outcomes in areas which subsequently became autocratic, or which were later exploited by the slave trade. However, given the number of possible interaction effects in their models, it is hard to choose *ex ante* which ones should be assessed.

The lasso approach is highly applicable here: we generate models based on their most saturated specifications of their key results, and interact every independent variable with the explanatory variable of interest, pre-colonial conflict. Then, using the lasso, we select the subset of interactions that together best explain variation in the outcome variable. While we cannot apply such an approach to the cross-country regressions due to power constraints, the Afrobarometer survey response, with $n > 17,000$, is more amenable to such approach.

The authors use this data to show that respondents who live in locations which suffered more pre-colonial civil wars display significantly lower levels of inter-group trust, significantly higher levels of ethnic identification, and significantly lower levels of national identification. In these regressions, they use 76 independent variables (including pre-colonial conflict) to control for omitted variables.

From these 76 variables, we exclude all occupation and country-level fixed effects and interact

the remaining 18 independent variables with the variable for pre-colonial conflict. Then, running the lasso algorithm on the entire model, we assess which interaction terms are selected by the lasso. The reason for excluding the fixed effects is that those interactions are hard to meaningfully interpret - we have no reason to believe that the effects of pre-colonial conflict should interact with the particular profession of the survey respondent, and looking at the interaction with country fixed effects just indicates which specific countries have had both pre-colonial conflicts and lower trust levels today.

Table **WHAT** presents the results, where 'Positive' means that the interaction of the variable listed (when interacted with pre-colonial conflict) is selected by the lasso and has a coefficient greater than zero, 'Negative' means the coefficient is less than zero, and an empty cell means the interaction is not selected by the lasso.

Table 2: Interaction effects selected by lasso (Afrobarometer)

Interacted variable	Inter-group trust	Ethnic identity	National identity
Age		Positive	Negative
Age squared	Negative		Negative
Male		Negative	Positive
Urban			Negative
Civil war incidence			
Log slave exports	Negative	Positive	Negative
Log pop density (1400)			Negative
GDP per capita			Negative
Latitude			Negative
Longitude	Negative		
Rainfall			Negative
Humidity			Negative
Temperature			
Log coastline area	Positive		
Islam		Positive	
Log gold per capita			
Log oil per capita	Positive		Positive
Log diamonds per capita			
Ruggedness			

Even though this is a fully automated approach to variable selection, the algorithm produces a number of interesting results. First, the only interaction term selected across the three dependent variables measures the exposure of that region to the slave trade. So, places with long histories of conflict and high amounts of slaves exported demonstrate lower levels of inter-group trust, higher

ethnic identification, and lower national identification. This aligns strongly with eviden ce on the persistent impact of the slave trade on modern attitudes, as argued by **?**.

Second, the interaction on oil reserves is selected in two of the models. This again makes sense: if we believe that pre-colonial conflict weakened institutional structures, then the presence of oil in that region ought to act as an exacerbator for conflict once the resource becomes valuable. A well-established literature on how institutions facilitate the resource curse in developing countries, which may impact social attitudes, leads credence to this heterogeneity (**?**).

Third, that the interaction with civil war incidence is not selected in any of the models. This suggests, perhaps surprisingly, that there is no meaningful interactive effect between pre-colonial conflict and post-colonial conflict on modern social attitudes.

Other implications are less clear - that male respondents in areas with more history of conflict display stronger senses of ethnic identity and weaker senses of national identity may suggest something about the implications of conflict on gender norms, but it is hard to say much more.

Disaggregated analysis: Grid cell results

In their paper, the authors provide more granular evidence that the existence of pre-colonial conflict in a given 125 x 125 km grid cell is associated with more modern conflict in that grid cell, as well as significantly lower levels of night light intensity. As outlined in **?**, the use of night light density is argued to strongly proxy for subnational income indicators in developing countries - so, grid cells with more pre-colonial conflict have lower income today.

Here, we apply the same approach as to the survey data: we take the full set of 119 regressors from their from their most saturated models, then drop all the fixed effects for countries and economic activity in that grid cell. Then, with the remaining independent variables, we interact all of them with the central independent variable, apply the lasso, and assess the interaction terms that are selected as explaining important heterogeneity.

While the selected interaction terms are less interpretable than those from the social attitudes data, it still provides some suggestive evidence on the mechanisms at work. For interactive effects on the presence of modern conflict, areas with pre-colonial conflict far from capital cities have more

Table 3: Interaction effects selected by lasso (grid cell)

Interacted variable	Conflict	Light density
Distance to coast		
Average elevation		Positive
Ruggedness		
Average temperature	Positive	
Average precipitation		
Area		
Log pop density (1990)		Negative
City in cell in 1400		
More than one ethnicity in cell		Negative
Log slave exports		Negative
Capital in cell		Positive
Distance to capital 0-10 percentile		
Distance to capital 10-25 percentile		Negative
Distance to capital 25-50 percentile		
Distance to capital 50-75 percentile	Positive	Positive
Distance to capital 75-90 percentile		
Jurisdictional hierarchy		
Log night lights (1992)		
Mineral share		

modern conflict, as well as those areas with higher average temperatures - which perhaps suggests a historical component to the large literature demonstrating a link between climate and violence (**?**). Regarding local economic activity, more interactive effects appear important: again, the interaction with the slave trade is selected by the algorithm, as is the interaction with the extent of ethnic division in a given grid cell. Other interactions are harder to interpret.

Summary

The results of the variable selection analysis demonstrate two points. First, that in spite of the specification issues of their coarse cross-country regressions, the incidence of pre-colonial conflict appears to explain variation in modern conflict distinct from their set of plausibly overfitted set of independent variables. And second, that the process detects intuitive interactive effects in their data that the authors do not explore in their analysis - that the effect is especially magnified by subsequent exposure to the slave trade and the existence of oil, which supports the academic literature pointing to the central importance of historical institutional development in determining modern outcomes.