Data challenge 2

Pablo Argote

March 20, 2018

Research questions

The dataset *AllViolenceData*171220.*csv* allows to answer relevant questions about the characteristics of violent events and the type of victims of such events. A particularly interesting variable is the violence against civilians, because it indicates the extent in which civilians are affected by this climate of violence. Moreover, policy-makers should be interested in minimizing civilians victims, regardless of the strategy for dealing with crime and armed violence. Thus, my two research questions are the following:

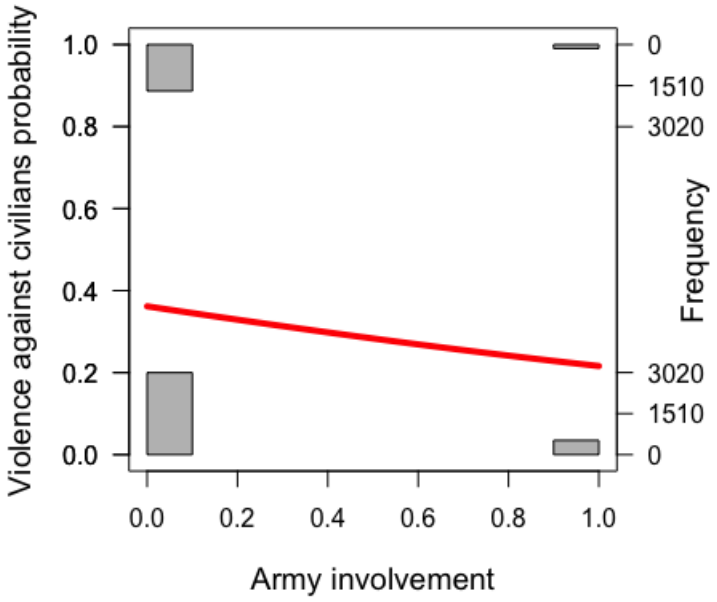
1. What is the correlation between whether the army participated in the event and violence against civilians? This is my inferential question.
2. Are the set of variables that characterize the event predictive of violence against civilians? This is my predictive question.

Measures

To measure the dependent variable, I created an indicator variable equal to 1 if a civilian was either wounded or killed in an event and 0 otherwise (See code here). It was necessary to perform this transformation because the raw variables contained separate measures of whether a person was killed or wounded. Given that in most of the cases no civilian was hurt, it was necessary to combine both variables and create a simple indicator of violence that maximizes variation.

Even including either wounded and killed as a criteria for being a victim of violence, there are not many events where civilians where victims of violence. Indeed, in only 12.5% of events a civilian was victim of violence under the definition described above.

Figure 1: Change in probability of violence against civilians when the army is involved



Inferential analysis

To answer my inferential question, I will estimate a logistic regression model, with violence against civilians as the dichotomous dependent variable, and the following predictors: detained, long guns seized, small arms seized, cartridge seized, clips seized, vehicles seized, afi, army, federal police, ministerial police, municipal police, navy, state police and perfect lethality (See dictionary here for a description of the variables). I decided to use as predictors variables that do not directly incorporate information about civilian casualties. For example, I am not using the variable *totalpeopledead* as predictor because by construction, it can use information of my dependent variable. Given that my substantive interest focuses on the effect of the army participation, I will pay particular attention of the effect of the variable *army*.

After estimating the model, we observe that indeed, the effect of army participation is significant at the 0.001 level. Indeed, when computing the marginal effect of army participation, we see that when the army participates, the probability of having violence against civilians decreased in 0.11. In other words, when the army participated, civilians were 11 percentage points less likely of being victims of violence. (See code here). In figure 1, we can visually see this change in probability when the army participate in these events.

This correlation suggest that army involvement tend to protect people for violence, since civilians are much less likely of being harmed when the army is involved. However, we cannot be sure if civilians are less likely of being harmed because the army protected them, or because the army is less likely to be located in places where civilians tend to receive violence.

This analysis has two main limitations. First, the dependent variable is dichotomous, so I am not distinguishing between the amount of people harmed. There could be some differences between events with one wounded and events with more than one civilian killed. Second, I do not have evidence to provide causal claims, so there might several endogeneity issues with this model. For example, we are not sure whether civilians are less harmed because the army was involved, or given that the army was involved, is unlikely to observe civilians participants.

Predictive analysis

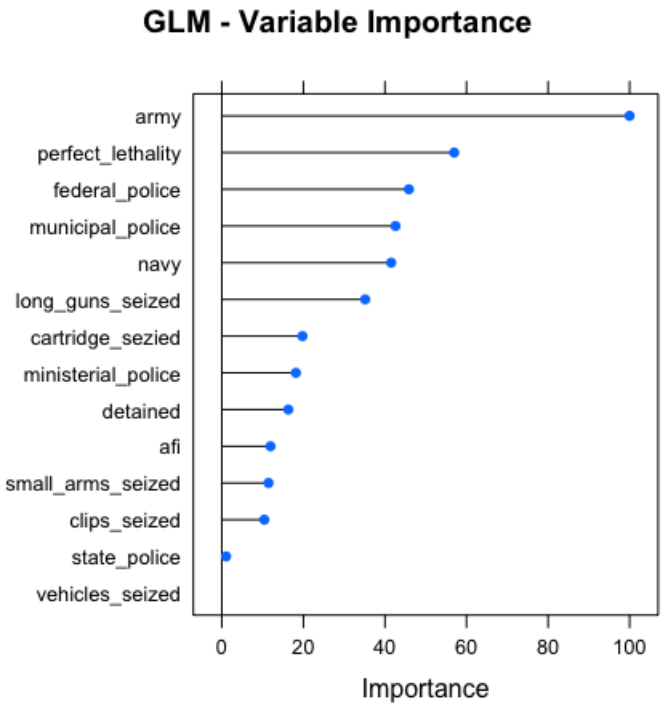
For my predictive analysis, I will perform a 10-fold cross validation procedure, in order to test the accuracy of my model [1]. First, I will cross-validate the model that I describe above (model 1) and compute the accuracy. Then, I will compare this measure of accuracy with three alternative models: a) model without the non-significant predictors from the previous regression (model 2), b) same than model 2 plus interactions between the army variable and the other predictors (model 3) and c) model with all the predictors from model 1 plus interactions from model 2 (See code here).

In all the models, the accuracy measure is around 0.875, which means that I am not gaining any improvement with the interaction terms, or by removing the non-significant variables. Moreover, this number suggest that this set of variables are not predictive of the outcome, since the accuracy is almost identical to the baseline accuracy of $100 - 12.5 = 87.5$. Thus, we can conclude that this set of variables are not predictive of violence against civilians. Probably, it would be necessary to collect more detail measures about the characteristics of the locality where a civilian was harmed, individual characteristics of civilians and characteristics of the criminal gang involved in the event. For example, perhaps a particular type of band uses violence against civilians as a strategy, so it would be helpful to know the features of such band.

Even if this set of variables that characterized the event are not predictive of my outcome of interest, we can still show how important were each of these variables. Indeed, figure 2 shows the relevance of each variable of the model. Clearly, the involvement of the army is the most relevant predictor of violence against civilians.

[1] We have not cover specific aspects of predictive methods in classes, so I am not very familiar with this method. Thus, I will probably use inaccurate terminology

Figure 2: Change in probability of violence against civilians when the army is involved



However, as explained above, the lack of predictive capacity of this model suggest that this effect might disappear if we include more predictive measures of the outcome. Therefore, an interesting future avenue of research could be to study what determines violence against civilians, and collect variables that could capture such variables.