Data challenge 3

Pablo Argote

April 17, 2018

Research questions

The dataset $AllViolenceData171220.csv$ 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 amount of people detained in an event, since it indicates the extent in which the police was able to put criminals in jail without harming them, which may be the ultimate goal of law enforcement. Likewise, there may be several conditional relations that predict the number of people detained, such as the presence of certain branches of the army and some characteristics of the event, such as the number of people dead. In this sense, my two research questions are the following:

1. What is the effect of army participation conditional on the number of people dead on the event on the number of people detained?
2. What is the effect of both army and federal police participation, conditional on the number of people dead in the event, on the number of people detained?

For the first hypothesis, I will estimate the following OLS regression model:

$$Detained = \alpha + \beta_1(army)_i + \beta_2(dead)_i + \beta_3(dead * army)_i + \beta_4(controls)_i + \epsilon$$

Where *Detained* -the dependent variable- is a a count variable equal to the number of people detained per event, *army* is a indicator variable equals to 1 if the army participated in the event, *dead* is the number of people dead in the event, $type * army$ is the interaction between the two and $controls$ are several covariates defined as controls variables.[1] The marginal effect of interest is $\beta_1 + \beta_3$, which indicates the effect of army participation conditional on number of people dead. More specifically, $\beta_1 + \beta_3$ shows the effect of army in at different level of people dead.

[1] See R code for a list of the control variables.

For the second hypothesis, I will estimate the following OLS regression equation:

$$Detained = \alpha + \beta_1(army)_i + \beta_2(dead)_i + \beta_3(fp)_i + \beta_4(army * dead)_i +$$
$$\beta_5(army * fp)_i + \beta_6(dead * army * fp)_i + \beta_7(fp * dead)_i + \beta_8(controls)_i + \epsilon$$

In this case, the marginal effect of interest is $\beta_1 + \beta_4 + \beta_5 + \beta_6$, because it shows the marginal effect of army participation, conditional on federal police participation, at different levels of people dead. It is worth noting that β_6 represents the triple interaction.

The main assumptions of both models are the standard OLS assumptions: that the relationship between the outcome and the predictors is linear in the coefficients, no autocorrelation, homocesdasticity, zero mean of the error term and no omitted variable bias. Certainly, the latter assumptions is the less realistic, since there may be several variables that correlates both to the number people detained and some of the regressors.

Results

Figure 1 shows a visualization of the first regression equation, showing the predicted values of army participation at different levels of total people dead.[2] In one sentence, the higher the number of people dead, the more negative is the effect of army participation on the number of people detained.

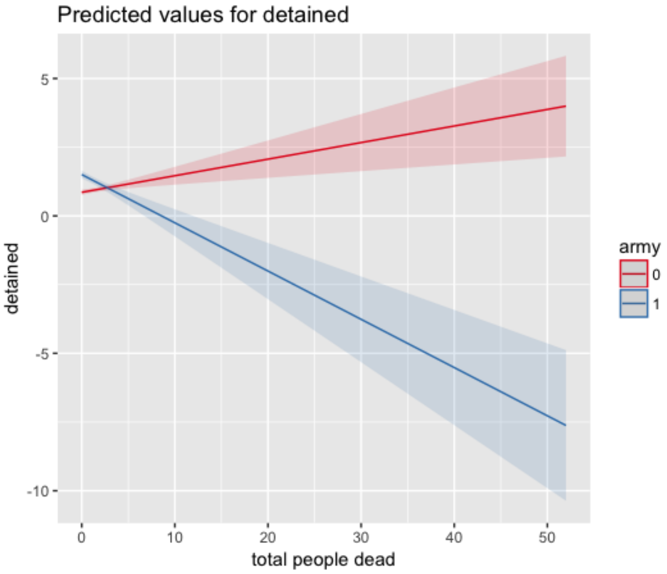
For instance, we observe that when there is no dead people, army participation appears to have a positive effect on people detained. Indeed, with 0 dead people, there is more detained people when the army participated compared when it did not. However, when there is, say, 30 dead people, the marginal effect of army participation is clearly negative. This suggest that the army may help detained people only in less violent events; in turn, in more violent events, army participation might reflect more people wounded or dead instead of detained.

Of course, the main limitation of this analysis is the risk of reverse causality, since we do not know whether people were not detained because the army participated in violent events, or whether the army participated precisely in events where is likely to observe more dead people and just a few detained.

Figure 2 shows the predicted values of federal police participation at different number of total people dead, at both levels of army involvement. The left panel describes the effect of federal police involvement when the army was not involved, while the right panel shows the marginal effect of federal police involvement when the army was involved.

[2] All the other predictors are set at their means or proportions

Figure 1: Marginal effects of army participation at different values of people dead



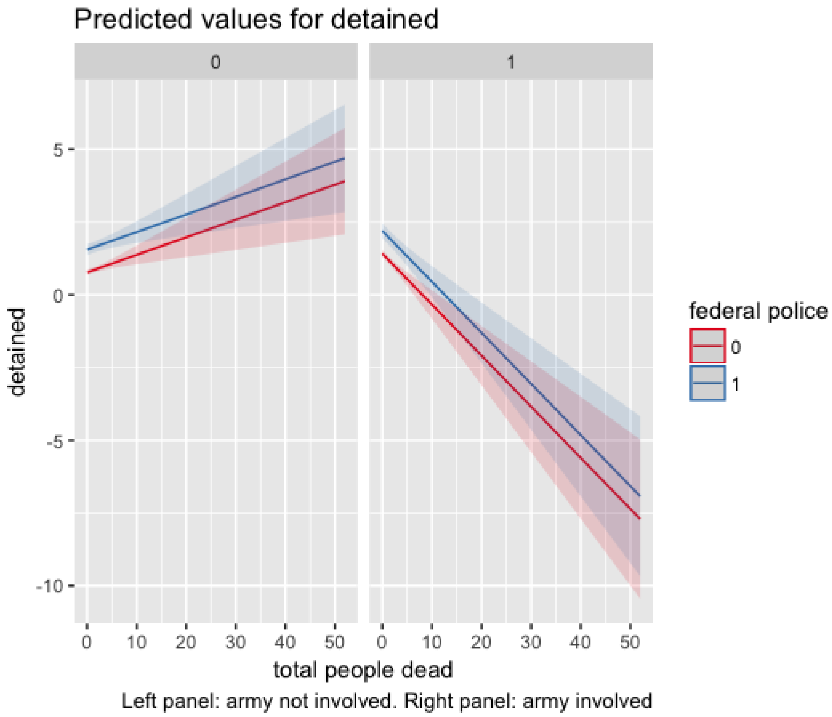
In one sentence, we observe that at every value of people dead and army involvement, the involvement of the federal police has a positive effect on the number of people detained.

In the left panel, we see that when the army is not involved, more deaths imply more detained, at both values of federal police involvement. Moreover, we observe a persistent gap concerning federal police involvement: when the police is involved, more people is detained. In the right panel, we see that when the army is involved, number of deaths negatively correlates to number of people detained, although there is still a positive gap of police involvement.

As before, the main limitation of this analysis is that we do not know the direction of the correlation and that the standard errors are too big at higher number of dead people, so the effect of federal police may not be statistically significant, as shown in the graph.

Overall, the two models show that the effect of army involvement is conditional on the number of deaths, although the positive effect of federal police involvement holds regardless of army involvement and the number of deaths. This makes complete intuitive sense: police involvement always produce more people detained, but when the army is involved, it only correlates with more people detained in non-violent events.

Figure 2: Marginal effects of army participation at different values of people dead



```
# Data challenge 3

rm(list=ls())
setwd("/Users/pabloargote/Dropbox/Columbia/02. Semester/Data_Science/Data challenge 3")
violence <- read.csv("AllViolenceData_171220.csv")
library(data.table)
library("margins")
options("scipen" = 5)
library(popbio)
library(bestglm)
library(leaps)
library(caret)
library(lattice)
library(ggplot2)
library(e1071)
library(interplot)
library(sjPlot)
library(sjmisc)

set.seed(1183)

# a) Estimating OLS 1

violence$source1 <- factor(violence$source)

summary(lm(detained ~ army*total_people_dead + small_arms_seized +
federal_police + cartridge_sezied + clips_seized
           + vehicles_seized + afi + long_guns_seized + ministerial_police +
            municipal_police + navy + state_police
           + perfect_lethality,  data = violence))

ols1 <- (lm(detained ~ army*total_people_dead + small_arms_seized +
 federal_police + cartridge_sezied + clips_seized
           + vehicles_seized + afi + long_guns_seized +
            ministerial_police + municipal_police + navy + state_police
           + perfect_lethality,  data = violence))

plot_model(ols1, type = "pred", terms = c("total_people_dead", "army"))



# b) Estimating OLS 2

summary(lm(detained ~ army*source1*federal_police + small_arms_seized  +
 cartridge_sezied + clips_seized
           + vehicles_seized + afi + long_guns_seized +
           ministerial_police + municipal_police + navy + state_police
           + perfect_lethality,  data = violence))

ols2 <- (lm(detained ~ army*total_people_dead +
small_arms_seized + federal_police + cartridge_sezied + clips_seized
```

```
            + vehicles_seized + afi + long_guns_seized +
             ministerial_police + municipal_police + navy + state_police
            + perfect_lethality,  data = violence))

p <- plot_model(ols2, type = "pred", terms = c("total_people_dead", "federal_police", "army"))
p + labs(caption = "Left panel: army not involved. Right panel: army involved")
```