QRM Homework 2

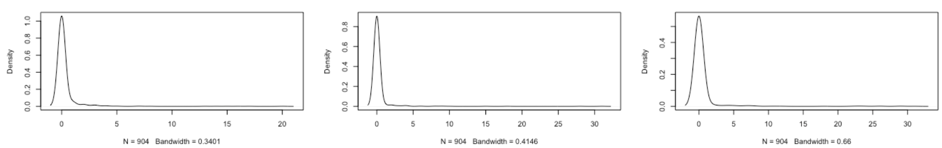
GLM on Terrorists Attacks Data

Işık Topçu

March 28, 2022

1 Method

After inspecting the dataset, we can see that the outcome variables are counts which leaves us with Poisson or Negative Binomial. The density plots for the outcome variables $attack_p22, nkill_{np}22, nwound_{np}22$ illustrate a zero-inflated, positively skewed spikey data. Using Poisson regression assume that the mean and variance are equal ($\mu = \sigma^2$). To test this, we found the mean and variance of the outcomes, and realized that these values are far from similar. This left us with the negative binomial. We also checked for zero inflation in the data with the `100*sum(dataNewC[,var] == 0)/nrow(dataNewC)` command which showed high zero inflation for all three.



Plot 1: *Density plots of the outcomes: $attack_p22, nkill_{np}22, nwound_{np}22$ respectively.*

2 Models

2.1 Choice of controls

Theoretically, *the overall population of the district*, *urbanization rate* and whether the district is a *border* to Iraq, Iran or Syria as well as vote *margin* of the last election made sense to be used as controls. We assumed that more populous,urbanized and close to border districts are more prone to submit to terror attacks because of the nature and goal of terror attacks against the civilians. We left out lagged outcome variables as explanatory variables since they tend to bias one's estimate downwards, so is considered over-controlling. We also left out coordinates because of the fact that they are not manipulable, and as such cannot be causally prior.

2.2 Negative Binomial

We applied negative binomial regression on the outcomes. On our choice of outcome, $attack_p22$, *margin* and *border* were significant.

We also tried Zero-Inflated Negative Binomial, but we scaled the variables because because `R` wasn't able to compute the Hessian otherwise, the SE were NA. *margin* was also significant for this model. This model had also the dispersion statistic closest to 1.[1]

2.3 Model Fit

We have performed likelihood-ratio test on the variable with the highest p-value, to see if we can run a model without it. The test showed we shouldn't exclude the urbanization rate from the model.[2]

[1] 0.927

[2] Pr(>Chisq): 2.2e-16***

	Model 1
(Intercept)	0.18
	(0.62)
margin	-3.13^{***}
	(0.57)
population	0.00^{*}
	(0.00)
urbanization_rate	0.01
	(0.01)
as.numeric(border)	3.16^{***}
	(0.76)
AIC	812.80
BIC	841.64
Log Likelihood	−400.40
Deviance	244.25
Num. obs.	904

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 1: Negative Binomial Regression for $attack_p22$

	Model 2
(Intercept)	0.95
	(0.94)
margin	-4.60^{***}
	(0.88)
population	0.00
	(0.00)
urbanization_rate	0.01
	(0.01)
as.numeric(border)	3.52^{**}
	(1.23)
AIC	579.51
BIC	608.35
Log Likelihood	−283.75
Deviance	135.19
Num. obs.	904

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 2: Negative Binomial Regression for $nkill_{np}22$

	Model 3
(Intercept)	1.07
	(0.98)
margin	-3.83^{***}
	(0.91)
population	0.00
	(0.00)
urbanization_rate	0.01
	(0.01)
as.numeric(border)	3.36^{*}
	(1.32)
AIC	770.41
BIC	799.25
Log Likelihood	-379.21
Deviance	156.20
Num. obs.	904

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 3: Negative Binomial Regression for $nwound_{np}22$

	Model 1
Count model: (Intercept)	-6.00^{**}
	(2.23)
Count model: log10(1 + margin)	4.40
	(2.29)
Count model: log10(1 + population)	0.86^{*}
	(0.40)
Count model: log10(1 + urbanization_rate)	0.62
	(0.90)
Count model: log10(1 + as.numeric(border))	3.84^{*}
	(1.82)
Count model: Log(theta)	-0.98^{**}
	(0.38)
Zero model: (Intercept)	-11.54
	(7.17)
Zero model: log10(1 + margin)	38.17^{**}
	(14.52)
Zero model: log10(1 + population)	0.60
	(0.72)
Zero model: log10(1 + urbanization_rate)	0.12
	(1.32)
Zero model: log10(1 + as.numeric(border))	-34.47
	(106.94)
AIC	759.52
Log Likelihood	-368.76
Num. obs.	904

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 4: Zero-inflated Negative Binomial with log(1+x) scale.

Table 5:

	Dependent variable:	
	attack_p22	
	(1)	(2)
margin	-3.135^{***}	
	(0.568)	
population	0.00001^{**}	0.00001^{***}
	(0.00000)	(0.00000)
urbanization_rate	0.006	0.004
	(0.008)	(0.009)
as.numeric(border)	3.156^{***}	2.799^{***}
	(0.761)	(0.844)
Constant	0.182	-2.265^{***}
	(0.615)	(0.395)
Observations	904	904
Log Likelihood	-401.399	-413.603
θ	0.097^{***} (0.017)	0.079^{***} (0.013)
Akaike Inf. Crit.	812.797	835.207
Note:		*p<0.1; **p<0.05; ***p<0.01

AIC and BIC statistics favor the null model.