EC3320

2016-2017

Michael Spagat

Lecture 11

There is now a large literature that uses regression analysis on cross-country datasets to try to figure out what risk factors are associated with conflict.

Notice the terminology, “associated with”, which refers to correlation, not causation.

The structure of the datasets used is something like the following.

Country - Period	Conflict Onset	Per Capita GDP $	GDP Growth Rate %	More Variables
1	0	3,500	2.1	
2	1	7,000	-1.5	
3	0	350	0.3%	
4	0	6,000	Etc	
5	1	2,300		
6	0	700		
7	1	10,000		
8	0	5,700		
...	...	...	...	

The time periods in the first column can be something like 5-years. So, for example, there can be country periods for DRC 1960 – 64, DRC 1965-69,..., DRC 2000 – 2005, Ethiopia 1960 – 64, ... Or the periods may be just 1-year periods.

The second column in the table above is for what is known as the “onset” of conflict. We code this variable with a “1” if a new conflict starts for that country-time-period and a “0” if a new conflict does not started in that time period.

There is a fairly obvious weakness with this coding scheme – ongoing conflict is coded exactly the same as ongoing peace is coded except in the one time period when the conflict actually starts.

One way to address this weakness is to drop the observations of ongoing conflict, leaving behind just observations of peace and observations of conflict onset.

Standard advice is never to drop data so the above suggestions may sound questionable but here it makes some sense - if you are interested only in conflict onset then observations for which conflict continues are not of great interest.

Sometimes researchers code a different conflict variable known as “incidence” of conflict. This variable is coded “1” for a country-time-period if a conflict is ongoing and “0” if there is no conflict.

“Incidence” is a rather different concept from “onset” and does not come with the tricky coding issues discussed on slide 3 - still incidence is not useful for analysing why conflicts start in the first place.

In this lecture we will only discuss onset but we will touch on incidence in later lectures.

Onset is a 0-1 variable so it is common to use logistic regression (lecture 5) in this literature.

The estimated equation will look something like:

Probability of conflict =

$$\frac{1}{1+e^{-(\text{constant} + \text{a}*\text{GDP per capita} + \text{b}*\text{Growth Rate of GDP} + \text{c}*\text{another variable} + \text{d}*\text{ another variable etc})}}$$

The next slide gives a key table from the Fearon and Laitin paper:

TABLE 1. Logit Analyses of Determinants of Civil War Onset, 1945–99

	Model				
	(1) Civil War	(2) "Ethnic" War	(3) Civil War	(4) Civil War (Plus Empires)	(5) Civil War (COW)
Prior war	−0.954**	−0.849*	−0.916**	−0.688**	−0.551
	(0.314)	(0.388)	(0.312)	(0.264)	(0.374)
Per capita income[a,b]	−0.344***	−0.379***	−0.318***	−0.305***	−0.309***
	(0.072)	(0.100)	(0.071)	(0.063)	(0.079)
log(population)[a,b]	0.263***	0.389***	0.272***	0.267***	0.223**
	(0.073)	(0.110)	(0.074)	(0.069)	(0.079)
log(% mountainous)	0.219**	0.120	0.199*	0.192*	0.418***
	(0.085)	(0.106)	(0.085)	(0.082)	(0.103)
Noncontiguous state	0.443	0.481	0.426	0.798**	−0.171
	(0.274)	(0.398)	(0.272)	(0.241)	(0.328)
Oil exporter	0.858**	0.809*	0.751**	0.548*	1.269***
	(0.279)	(0.352)	(0.278)	(0.262)	(0.297)
New state	1.709***	1.777***	1.658***	1.523***	1.147**
	(0.339)	(0.415)	(0.342)	(0.332)	(0.413)
Instability[a]	0.618**	0.385	0.513*	0.548*	0.584*
	(0.235)	(0.316)	(0.242)	(0.225)	(0.268)
Democracy[a,c]	0.021	0.013			
	(0.017)	(0.022)			
Ethnic fractionalization	0.166	0.146	0.164	0.490	−0.119
	(0.373)	(0.584)	(0.368)	(0.345)	(0.396)
Religious fractionalization	0.285	1.533*	0.326		1.176*
	(0.509)	(0.724)	(0.506)		(0.563)
Anocracy[a]			0.521*		0.597*
			(0.237)		(0.261)
Democracy[a,d]			0.127		0.219
			(0.304)		(0.354)
Constant	−6.731***	−8.450***	−7.019***	−6.801***	−7.503***
	(0.736)	(1.092)	(0.751)	(0.681)	(0.854)
N	6327	5186	6327	6360	5378

Note: The dependent variable is coded "1" for country years in which a civil war began and "0" in all others. Standard errors are in parentheses. Estimations performed using Stata 7.0. *$p < .05$; **$p < .01$; ***$p < .001$.

[a] Lagged one year.
[b] In 1000's.
[c] Polity IV; varies from −10 to 10.
[d] Dichotomous.

Here are a few things to notice about the Fearon and Laiton table.

They are looking at **Civil War Onset**. That is, “war” rather than “conflict”.

This means that only the start of really big conflicts will cause a coding of “1” for the onset variable.

Specifically, Fearon and Laiton require at least 1,000 people killed in the whole war with at least 100 on each side (state and non-state). What Uppsala defines as “conflict” requires only 25 deaths.

Here are the things that seem to matter according to column 1:

1. Prior War (negative effect).

This variable is potentially confusing.

You might imagine that if there is already a war going then it is impossible for a war to start. If so, then it makes no sense to even have this variable in the first place. But, it *is* actually possible for a new war to start. For example, there can be one rebel group at war with a government (this is the prior war) and then a new rebel group can arise and start a second war against the same government. However, such situations are rare – once one war is active then a second one is unlikely to start - hence the negative estimated coefficient on prior war.

The (fictitious) table on the following slide should clarify the relationship between prior war and war onset. Note that it only allows “prior war” to be coded as a “1” if there was a war going one period before the current period.

	Number of Wars	Onset	Prior War
2000 DRC	0	0	Not applicable
2001 DRC	1	1	0
2002 DRC	1	0	1
2003 DRC	2	1	1
2004 DRC	2	0	1

The onset variable is coded “1” in years when a new war starts and is coded as “0” otherwise.

The “prior war” variable is coded as “1” when there was any war going on in the previous year and is coded as “0” otherwise.

Note, however, that we could easily have other coding rules for “prior war”. For example, we could code this as “1” if there has been a war going on within the last 5 years.

2. Per capita income (negative effect) - higher income is associated with a lower probability of a war starting.

3. Population (positive effect) - more people, more chances for war

4. Mountainous terrain (positive effect) - the usual interpretation is that it is easy for rebels to hide in mountains so the mountains presence increases the risk of war

5. Being an oil exporter (positive effect). However, note that being an oil exporter is close to simply being a Middle Eastern country so it is not totally clear what's being measured here.

6. Being a new state (positive effect) – it may be hard for new states to stabilize and perhaps they slide into war

7. Instability (positive effect) - this is a dummy variable that is one if, during the previous three years, the country had three or more changes to its political regime index according to a project called Polity IV

A surprise is that ethic and religious fractionalization do not seem to matter. (High “Fractionalization” basically means that there are lots of different groups.)

You get broadly similar results in the other columns. For example, restricting to just what are categorized as “ethnic wars” leads to a set of coefficients similar to what we get in column 1.

Here is an important little warning

It is often very easy to tell nice stories about the coefficients in a regression. I have just been doing this on the last few slides.

These stories can help us make sense of what we are finding and even help us to remember the results. We humans relate to stories much better than we relate to numbers.

However, these stories are really just speculations I have created to fit with the estimates. They could easily be wrong. You can probably come up with some other speculations that are just as valid as mine.

This problem is known as "story time", which refers to the moment people switch from describing what their data say to speculating about cause and effect relationships.

To get a sense of the importance of the story time problem suppose that we reanalyse the Fearon and Laitin data and discover that they made a mistake in their table.

It was a simple transcription error – the correct coefficient for prior war is actually +0.954, not

-0.954 as written. (I repeat that this is not true – I'm just considering this possibility for the sake of illustrating the story time idea.)

I can easily make up a story that could rationalize why it makes total sense for the coefficient to be positive rather than negative. "Violence begets further violence. When there's a war on lots of people are hurt – they build up grievances and start new wars, even in the middle of existing wars."

This story is wrong but plausible.

Now if I suddenly discover that Fearon and Laitin's coefficients were correct after all I could just switch right back to my original story.

My point is:

1. You should become aware of the strong psychological hold that stories exert over you.

2. Realize that many of the stories you hear about data can easily be wrong and are often conjured up with little thought.

3. Consequently, you should pay attention to these stories but be sceptical of them at the same time.

This war onset table from the Collier and Hoeffler paper is similar to the Fearon-Laitin table:

576 P. COLLIER AND A. HOEFFLER

Table 4 Grievance model

	1	2	3
Ethnic fractionalization	0.010 (0.006)*	0.011 (0.007)*	0.012 (0.008)
Religious fractionalization	−0.003 (0.007)	−0.006 (0.008)	−0.004 (0.009)
Polarization $\alpha = 1.6$	−3.067 (7.021)	−4.682 (8.267)	−6.536 (8.579)
Ethnic dominance (45–90%)	0.414 (0.496)	0.575 (0.586)	1.084 (0.629)*
Democracy	−0.109 (0.044)***	−0.083 (0.051)*	−0.121 (0.053)**
Peace duration	−0.004 (0.001)***	−0.003 (0.001)***	−0.004 (0.001)***
Mountainous terrain	0.011 (0.007)	0.007 (0.009)	−0.0001 (0.009)
Geographic dispersion	−0.509 (0.856)	−0.763 (1.053)	−1.293 (0.102)
Ln population	0.221 (0.096)**	0.246 (0.119)**	0.300 (1.133)**
Income inequality		0.015 (0.018)	
Land inequality			0.461 (1.305)
N	850	604	603
No of wars	59	41	38
Pseudo R^2	0.13	0.11	0.17
Log likelihood	−185.57	−133.46	−117.12

Notes: All regressions include a constant. Standard errors in parentheses. ***, **, * indicate significance at the 1, 5 and 10% level, respectively.
Column 1: the two measures of fractionalization and ethnic dominance are not jointly significant.

Notice that Collier and Hoeffler's variable list is rather different from Fearon and Laitin's. Look at the variables that are in both.

1. Population size is still positive and significant.

2. Religious and ethnic fractionalization still do not matter (maybe ethnic fractionalization does matter slightly)

3. Now democracy comes out with a highly significant and negative sign rather than coming out positive but insignificant as it does for Fearon and Laitin.

4. Mountainous terrain is now insignificant rather than positive and significant as it was for Fearon and Laiton.

Let’s look at a second Collier and Hoeffler table:

Table 3 Opportunity model

	1	2	3	4	5
Primary commodity exports/GDP	18.149 (6.006)***	18.900 (5.948)***	16.476 (5.207)***	17.567 (6.744)***	17.404 (6.750)***
(Primary commodity exports/GDP)^2	−27.445 (11.996)***	−29.123 (11.905)***	−23.017 (9.972)**	−28.815 (15.351)*	−28.456 (15.366)*
Post-coldwar	−0.326 (0.469)	−0.207 (0.450)	−0.454 (0.416)		
Male secondary schooling	−0.025 (0.010)**	−0.024 (0.010)**			
Ln GDP *per capita*			−0.837 (0.253)***	−1.237 (0.283)***	−1.243 (0.284)***
GDP growth	−0.117 (0.044)***	−0.118 (0.044)***	−0.105 (0.042)***		
Peace duration	−0.003 (0.002) $p = 0.128$	−0.004*** (0.001)	−0.004 (0.001)***	−0.002 (0.001)	−0.002 (0.001)
Previous war	0.464 (0.547) $p = 0.396$				
Mountainous terrain	0.013 (0.009) $p = 0.164$	0.014 (0.009)	0.008 (0.008)		
Geographic dispersion	−2.211 (1.038)**	−2.129 (1.032)**	−0.865 (0.948)		
Social fractionalization	−0.0002 (0.0001) $p = 0.109$	−0.0002 (0.0001) $p = 0.122$	−0.0002 (0.0001)**		
Ln population	0.669 (0.163)***	0.686 (0.162)***	0.493 (0.129)***	0.295 (0.141)**	0.296 (0.141)**
Diaspora/peace				700.931 (363.29)**	
Diaspora corrected/peace					741.155 (387.636)*
(Diaspora-diaspora corrected)/peace					823.941 (556.024)
N	688	688	750	595	595
No of wars	46	46	52	29	29
Pseudo R^2	0.24	0.24	0.22	0.25	0.25
Log likelihood	−128.49	−128.85	−146.86	−93.27	−93.23

Notes: All regressions include a constant. Standard errors in parentheses. ***, **, * indicate significance at the 1, 5, and 10% level, respectively.

Again some common variables behave differently than they do in the other regressions.

Collier and Hoeffler find a strong effect of the ratio of primary commodity exports to GDP but Fearon and Laitin disagree (actually the table above does not show Fearon & Laitin's treatment of this variable but see page 87 of their paper if you are curious).

Fearon & Laitin disagree with Collier and Hoeffler over mountainous terrain and previous war although they agree on population and GDP per capita.

So what is going on? Why are we having these disagreements?

For at least two reasons it is, in fact, pretty hard to get to the bottom of these disagreements.

1. The two papers use slightly different definitions of war and different time periods (Collier and Hoeffler use five-year periods and Fearon & Laitin use 1-year periods.)

2. More than one variable changes as you move from specification to specification so it is impossible to isolate the impact of each particular change.

Subsequent to these papers a large literature grew up with people running all sorts of logistic regressions like the ones above on all sorts of collections of variables.

The results frequently contradict one another.

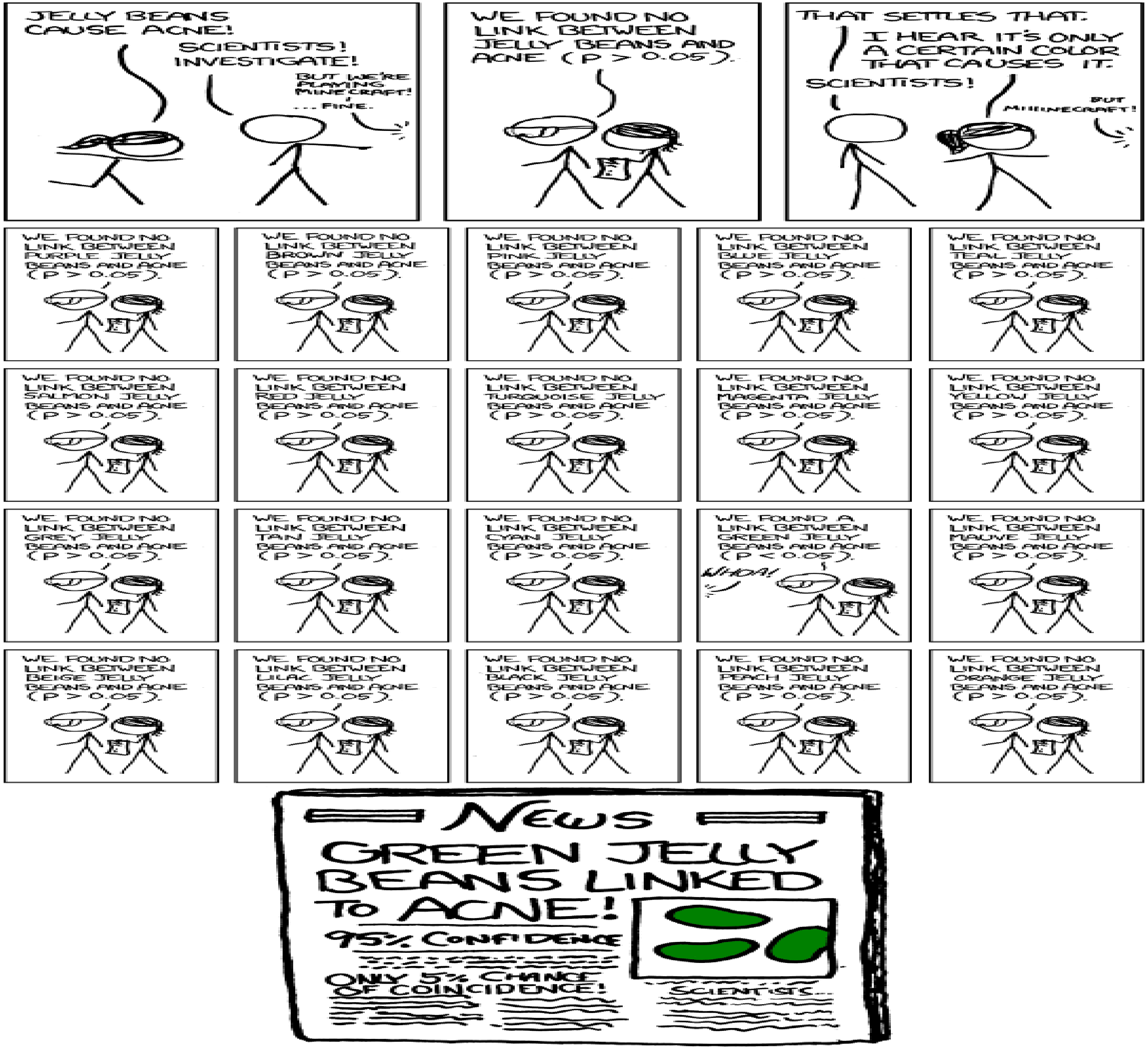
A general problem with this sort of work is that researchers might start with a bunch of variables, try lots of combinations and then only report regressions that have several statistically significant variables.

These reported results may just have come out that way by luck or chance.

Proceeding like this would be analogous to a drug company testing their drug in 20 separate clinical trials, getting a positive result in 1 out of the 20 and then reporting only that one positive result.

This practice of trying lots of things until you get a significant p value and reporting only that is sometimes known as “p hacking”

The cartoon on slide 22 illustrates the problem.


JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!
BUT WE'RE PLAYING MINECRAFT!
... FINE.
WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE (P > 0.05).
THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.
SCIENTISTS!
BUT MIIIINECRAFT!
WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!
WE FOUND NO LINK BETWEEN MAUVE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).
NEWS
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...

The paper by Hegre and Sambanis addresses this issue by looking at 88 variables that have appeared in the literature. Hegre & Sambanis systematically range across regressions containing all sorts of combinations of these variables and attempt to figure out which ones are consistently correlated with the onset of war. Basically they wind up with three variables:

1. GDP per capita is negatively associated with civil war onset

2. Having had a previous war is positively associated with civil war onset – the more recent the war the stronger the association. This result may appear to contradict the Fearon and Laitin results on prior war but actually does not. Hegre & Sambanis code ongoing war as a missing observation rather than as a "0" as Fearon & Laitin do.

    Fearon and Laitin find that when there is a civil war going on in the previous time period then it is unlikely that second war will start in the current time period.

    Hegre & Sambanis find that when there is no current civil war that having had a recent civil war increases the chances of a civil war breaking out in the current period.

3. Country size (population and territory) is positively associated with civil war onset.

Some other variables, such as the growth rate of GDP per capita, may have real relationships with civil war onset but such relationships do not shine through as robustly and consistently as the three variables above.

Finally, we have to bear in mind that correlation does not imply causation. For example, just because low GDP is correlated with conflict does not imply that low GDP causes conflict. The causality could just as easily run the other way. For example, GDP could be low in part because potential investors anticipate that there might be a war in the future so they withhold their investment money.

Next week we will look more closely at the causation question.