Poli-301:

The Political Science Discipline

TODAY'S AGENDA

1. Practice interpreting
2. Fixed effects

Y = test scores

- Not small vs. small → Small class
- Class doesn't have aide vs. class has aide → Regular + aide class
- Student not white/Asian vs. yes → White or Asian
- Student is boy vs. girl → Girl
- Student does not receive FRL vs. yes → Free/reduced lunch
- Teacher not white/Asian vs. yes → Teacher white or Asian
- Years (actual number) → Years of teacher experience
- Teacher does not have MA vs. yes → Teacher has MA

Table 2: OLS models for four standardized tests

	(1)	(2)	(3)	(4)
VARIABLES	Reading	Math	Listening	Words
Small class	6.47***	8.84***	3.24**	6.99***
	(1.45)	(2.32)	(1.42)	(1.60)
Regular + aide class	1.00	0.42	-0.58	1.27
	(1.26)	(2.14)	(1.32)	(1.42)
White or Asian	7.85***	16.91***	17.98***	7.08***
	(1.61)	(2.40)	(1.70)	(1.91)
Girl	5.39***	6.46***	2.67***	5.03***
	(0.78)	(1.12)	(0.74)	(0.94)
Free/reduced lunch	-14.69***	-20.08***	-15.23***	-15.97***
	(0.91)	(1.33)	(0.90)	(1.07)
Teacher white or Asian	-0.56	-1.01	-3.68	0.46
	(2.66)	(3.80)	(2.59)	(3.07)
Years of teacher experience	0.30**	0.42**	0.25*	0.30**
	(0.12)	(0.20)	(0.15)	(0.14)
Teacher has MA	-0.75	-2.20	0.50	0.24
	(1.25)	(2.08)	(1.24)	(1.46)
School fixed effects	X	X	X	X
Constant	431.69***	475.52***	531.28***	428.97***
	(3.12)	(4.49)	(2.84)	(3.59)

Outcome = temperature (F)

Humidity (%)

Windspeed (knots/sec)

Cloud cover (1 - 8)

Precip = %

Visibility (meters)

	(1)	(2)	(3)	(4)	(5)	(6)
(Intercept)	97.05 ***	33.90 ***	71.80 ***	102.72 ***	74.46 ***	88.24 ***
	(2.63)	(1.65)	(3.70)	(3.67)	(4.41)	(8.50)
humidity	-0.86 ***		-0.56 ***	-0.99 ***	-0.66 ***	-0.76 ***
	(0.05)		(0.05)	(0.08)	(0.07)	(0.08)
monthFebruary		11.85 ***	6.44 ***		6.14 **	5.90 **
		(2.40)	(1.85)		(1.85)	(1.93)
monthMarch		22.75 ***	10.89 ***		10.66 ***	9.62 ***
		(2.33)	(2.05)		(2.04)	(2.20)
monthApril		22.94 ***	11.19 ***		11.07 ***	9.32 ***
		(2.35)	(2.06)		(2.04)	(2.30)
monthMay		35.39 ***	20.70 ***		20.82 ***	19.44 ***
		(2.33)	(2.20)		(2.21)	(2.38)
windSpeed				-0.96	-0.10	-0.04
				(0.53)	(0.44)	(0.43)
cloudCover				0.07 *	0.06 *	0.02
				(0.04)	(0.03)	(0.03)
precipProbability						0.11 **
						(0.04)
visibility						-0.80
						(0.74)
N	151	151	151	151	151	151

Last time

We can close backdoors from X to Y by controlling Z

Use multiple regression to control for Z

Control via multiple regression:
parse out what parts of X and Y *are not explained by Z*

Today (and next couple of weeks)

Different methods beyond simple controls to identify effects from a DAG

Develop *conceptual* understanding of these methods

We won’t be doing “best practices” analysis, but enough to understand what’s going on

Fixed effects

Today we talk about fixed effects

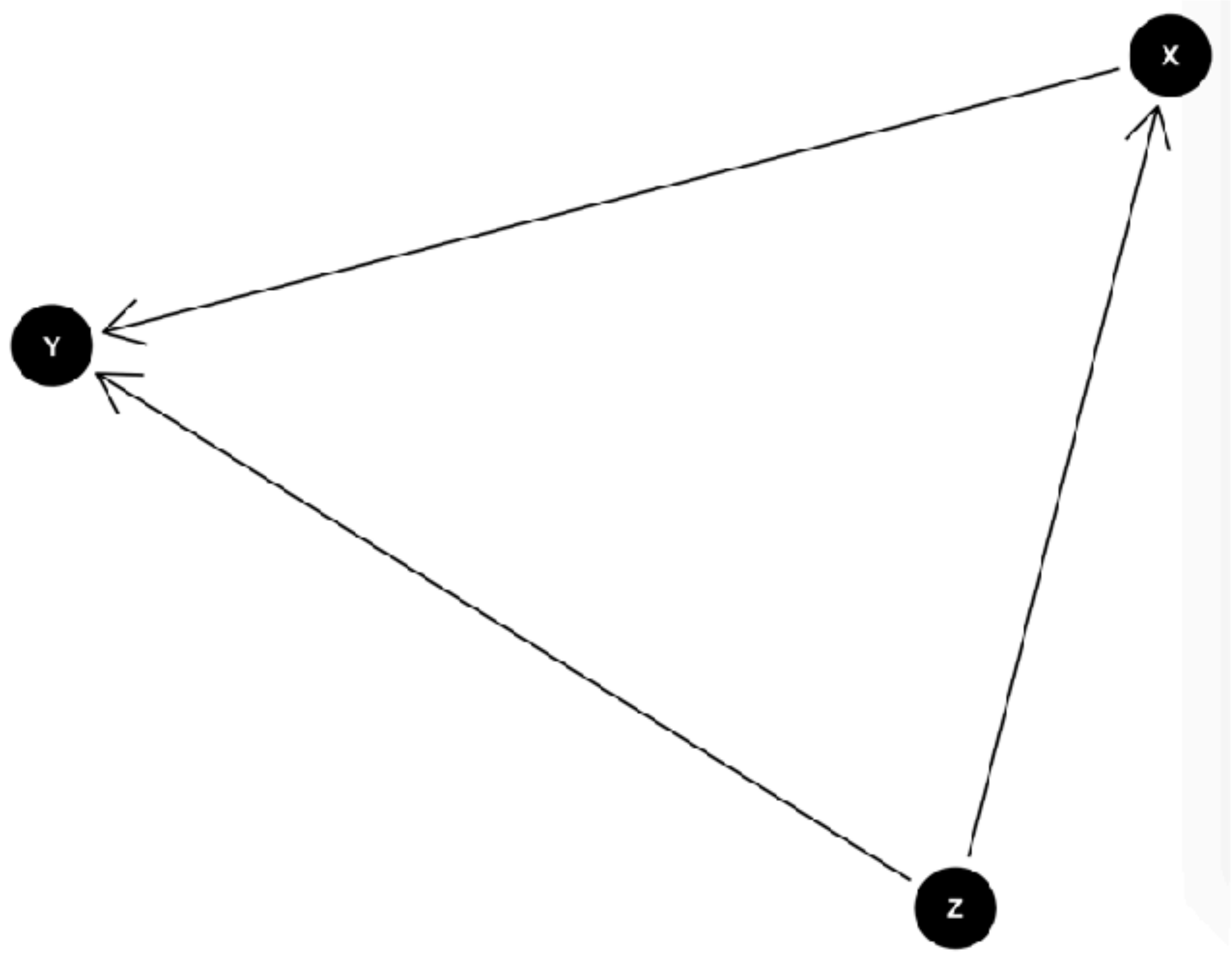
A widely used method that applies to a particular type of DAG

All methods we discuss now are like this: apply to particular DAGs

That means our choice of method will depend on the DAG

The problem

You want to identify effect of X on Y and are worried about confounding



But you don't have data on Z, or can't measure it, or maybe you're not sure what Z is

The particular DAG

If your data is observing some *unit* multiple times, you might not need to know Z**

Unit: person, country, company, state, county
Multiple times: across years, within states, etc.

You can just control for the *unit*

This will control for everything unique to that *unit*, whether we can measure it or not

Panel data

What's the effect of GDP on LifeExp?

Table 1

country	continent	year	lifeExp	pop	gdpPercap
China	Asia	1952	44	556263527	400
China	Asia	1957	50.5	637408000	576
Vietnam	Asia	1952	40.4	26246839	605
Vietnam	Asia	1957	42.9	28998543	676
Zambia	Africa	1952	42	2672000	1.15E+03
Zambia	Africa	1957	44.1	3016000	1.31E+03

Example

```r
# regression (plain)
m1 = lm(lifeExp ~ log(gdpPercap), data = gapminder)
moderndive::get_regression_table(m1)
```

	term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
1	intercept	-9.10	1.23	-7.41	0	-11.5	-6.69
2	log(gdpPercap)	8.40	0.149	56.5	0	8.11	8.70

```r
# regression (fixed effects)
m2 = lm(lifeExp ~ log(gdpPercap) + country, data = gapminder)
moderndive::get_regression_table(m2)
```

Table 1-1

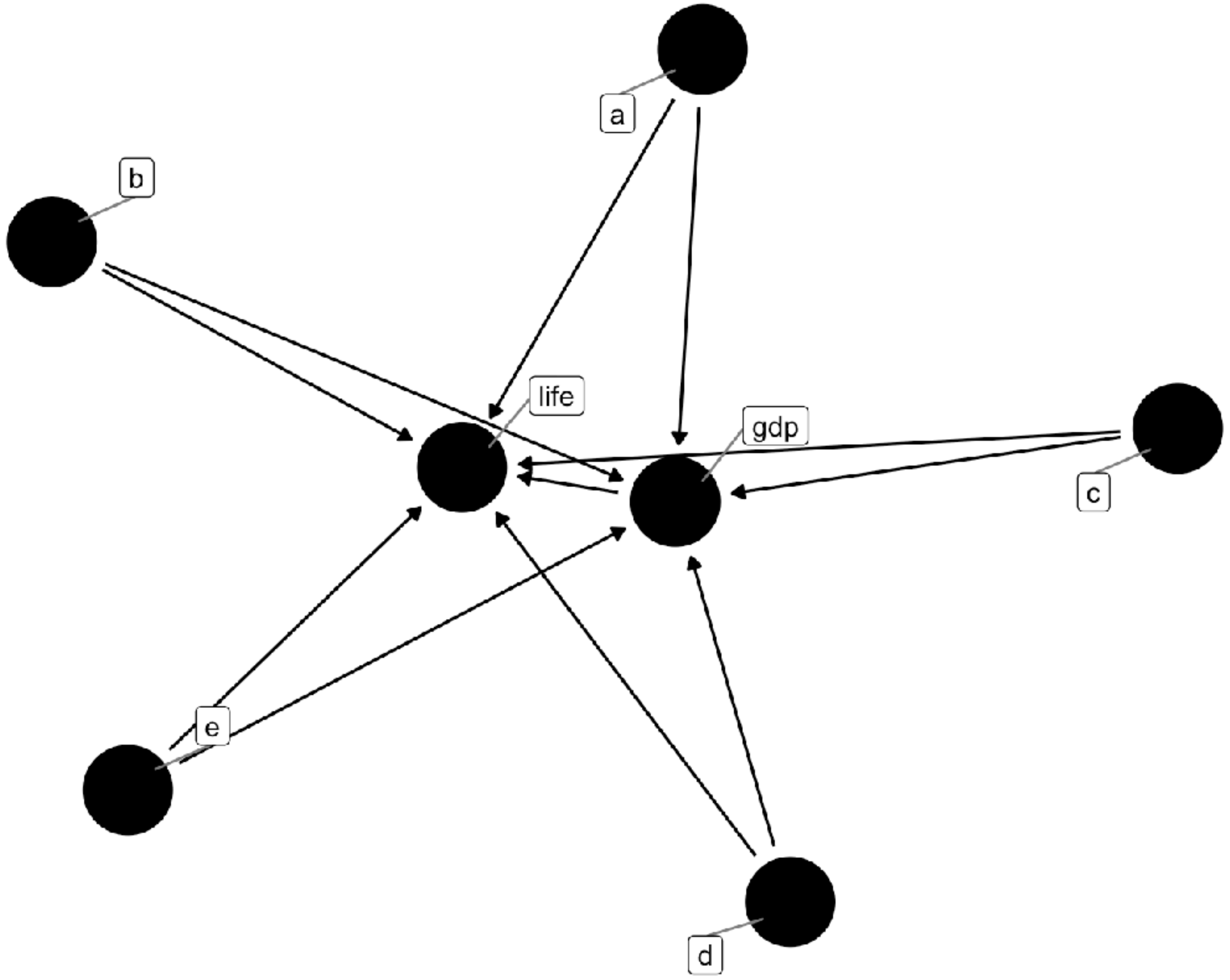
	term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
1	intercept	-27.8	2.50	-11.1	0	-32.7	-22.9
2	log(gdpPercap)	9.77	0.297	32.9	0	9.19	10.4
3	countryAlbania	17.8	2.19	8.10	0	13.5	22.1
4	countryAlgeria	5.24	2.21	2.37	0.018	0.897	9.59
5	countryAngola	-13.9	2.20	-6.32	0	-18.2	-9.59
6	countryArgentina	8.13	2.27	3.58	0	3.67	12.6
7	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

Why do this?

There might be a ton of backdoors between GDP and life-expectancy

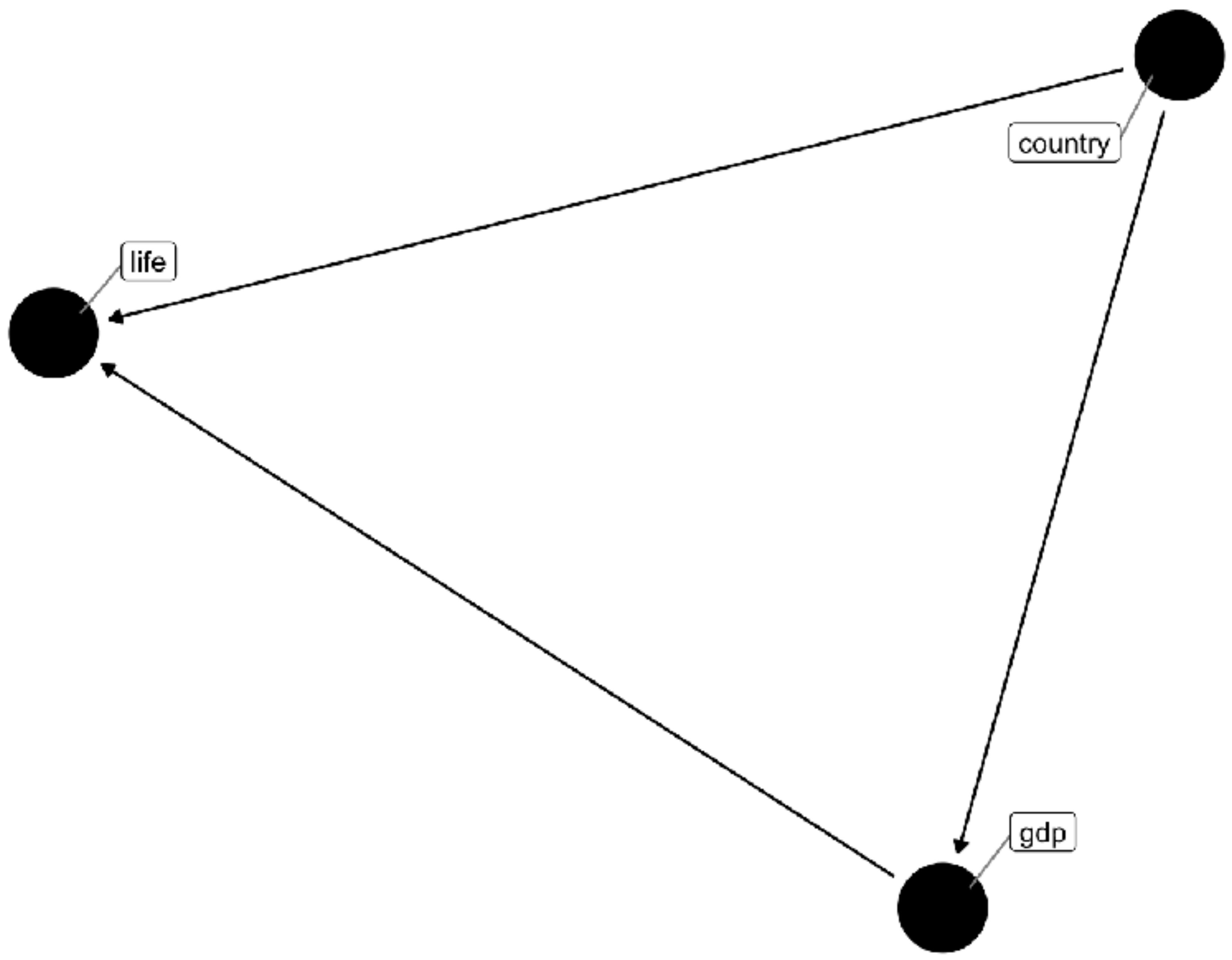
War, natural resources, disease, the kinds of institutions the countries have

No data on all of these



The DAG

But if these are *constant within country* then we can condense all these backdoors into one:



E.g.: some country characteristic(s) might be a backdoor from GDP —> life

“Constant within unit”

A variable that only varies ACROSS units, not WITHIN them

country	continent	year	lifeExp	pop	gdpPercap	ww2
Afghanistan	Asia	1952	28.8	8425333	779	no
Afghanistan	Asia	1957	30.3	9240934	821	no
Afghanistan	Asia	1962	32	10267083	853	no
Afghanistan	Asia	1967	34	11537966	836	no
Afghanistan	Asia	1972	36.1	13079460	740	no
Afghanistan	Asia	1977	38.4	14880372	786	no
Afghanistan	Asia	1982	39.9	12881816	978	no
Afghanistan	Asia	1987	40.8	13867957	852	no
Afghanistan	Asia	1992	41.7	16317921	649	no
Afghanistan	Asia	1997	41.8	22227415	635	no
	Constant		**Varies**	**Varies**	**Varies**	**Constant**

The data

In our data, we have variation in GDP (and LifeExp):

ACROSS countries (e.g., China richer than Cambodia)

WITHIN countries (e.g., Canada higher LifeExp in 2005 than 1975)

By controlling for country, we are subtracting out all of the differences *between* countries in GDP and LifeExp

~~ACROSS countries (e.g., China richer than Cambodia)~~

WITHIN countries (e.g., Canada higher LifeExp in 2005 than 1975)

The payoff

This means we are effectively only looking at differences *within* countries

WITHIN countries (e.g., Canada higher LifeExp in 2005 than 1975)

Essentially comparing countries to themselves at different time periods

As GDP rises within countries, does LifeExp rise/fall?

FE also called the “within” estimator

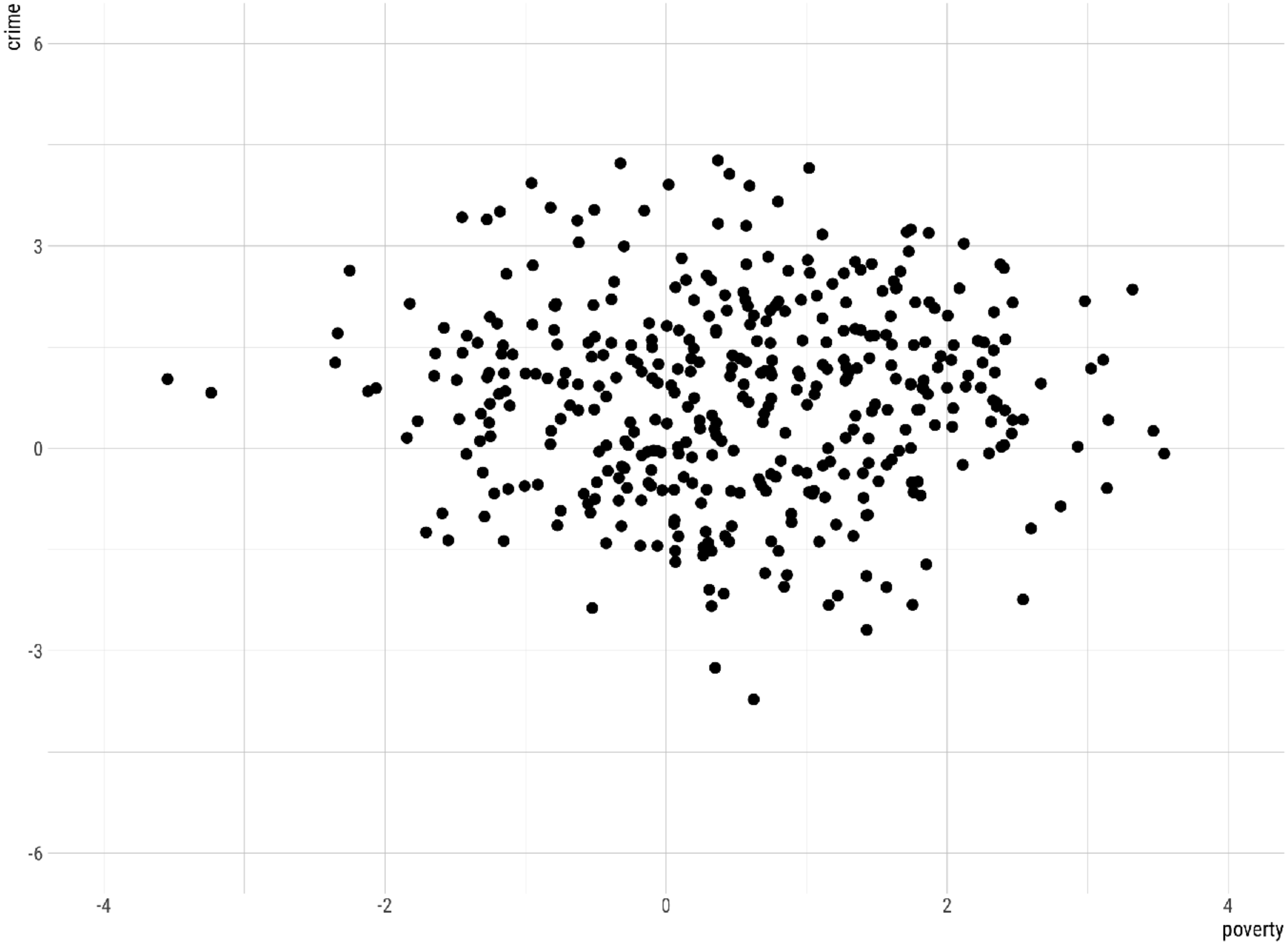
(made up) state poverty & crime data

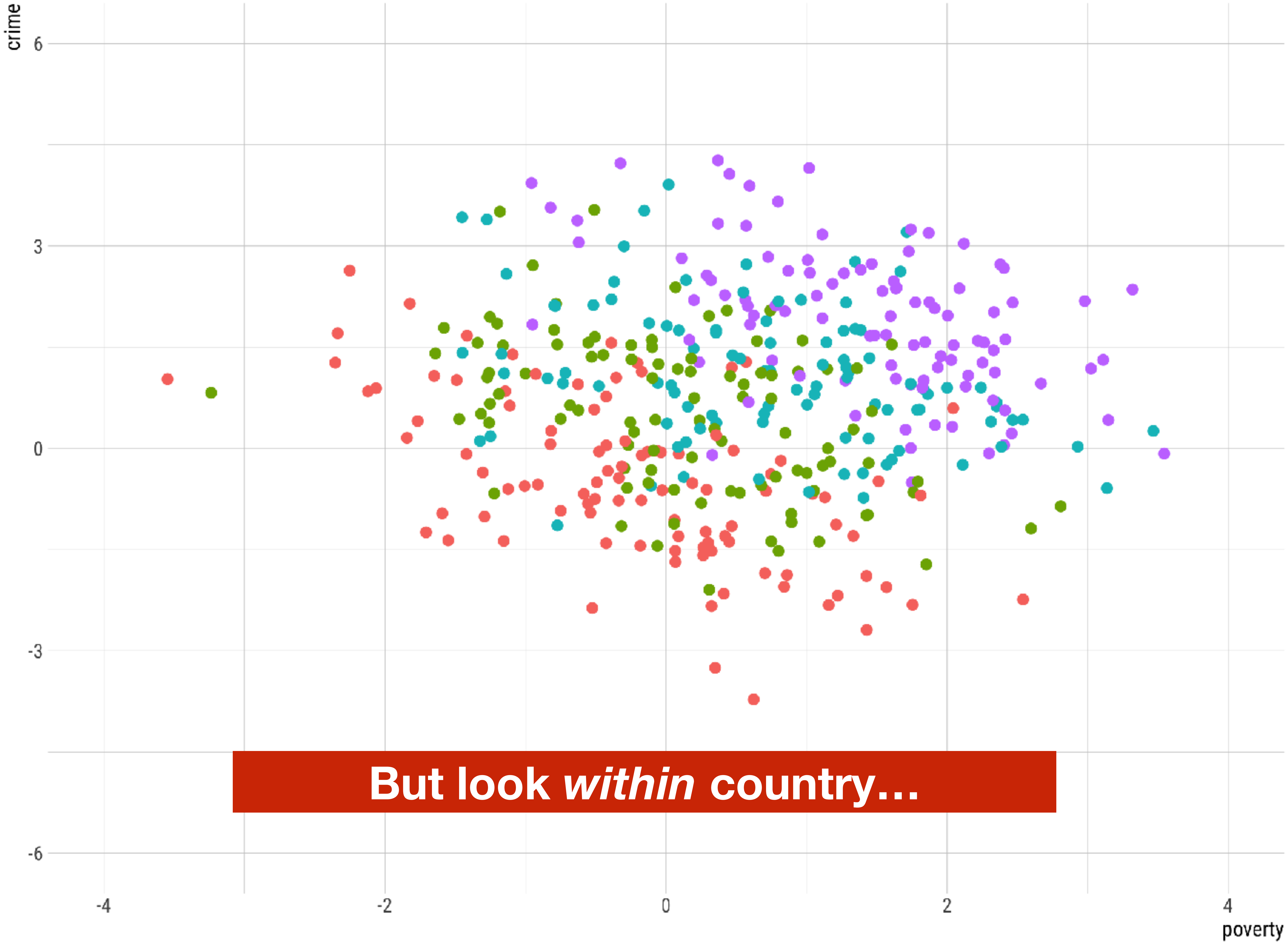
state	poverty	crime	year
1	1.16	-2.32	1992
1	-1.77	0.404	1993
2	1.33	0.283	1992
2	-0.0619	-1.45	1993
3	1.28	1.2	1992
3	0.126	-0.427	1993
4	1.57	1.68	1992
4	1.62	2.48	1993

True effect of poverty on crime = -0.5

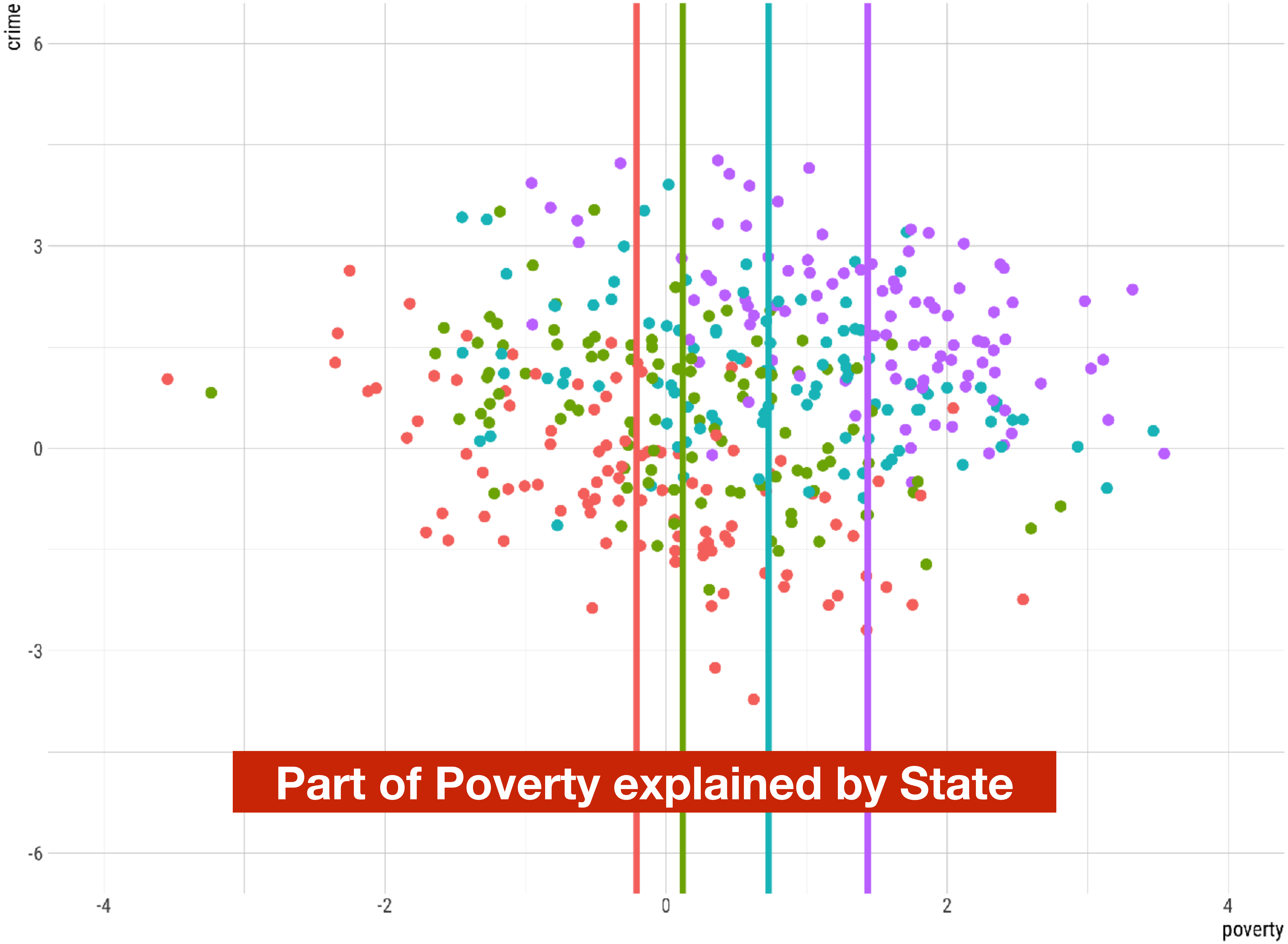
Poverty and Crime

What might be confounding this relationship?

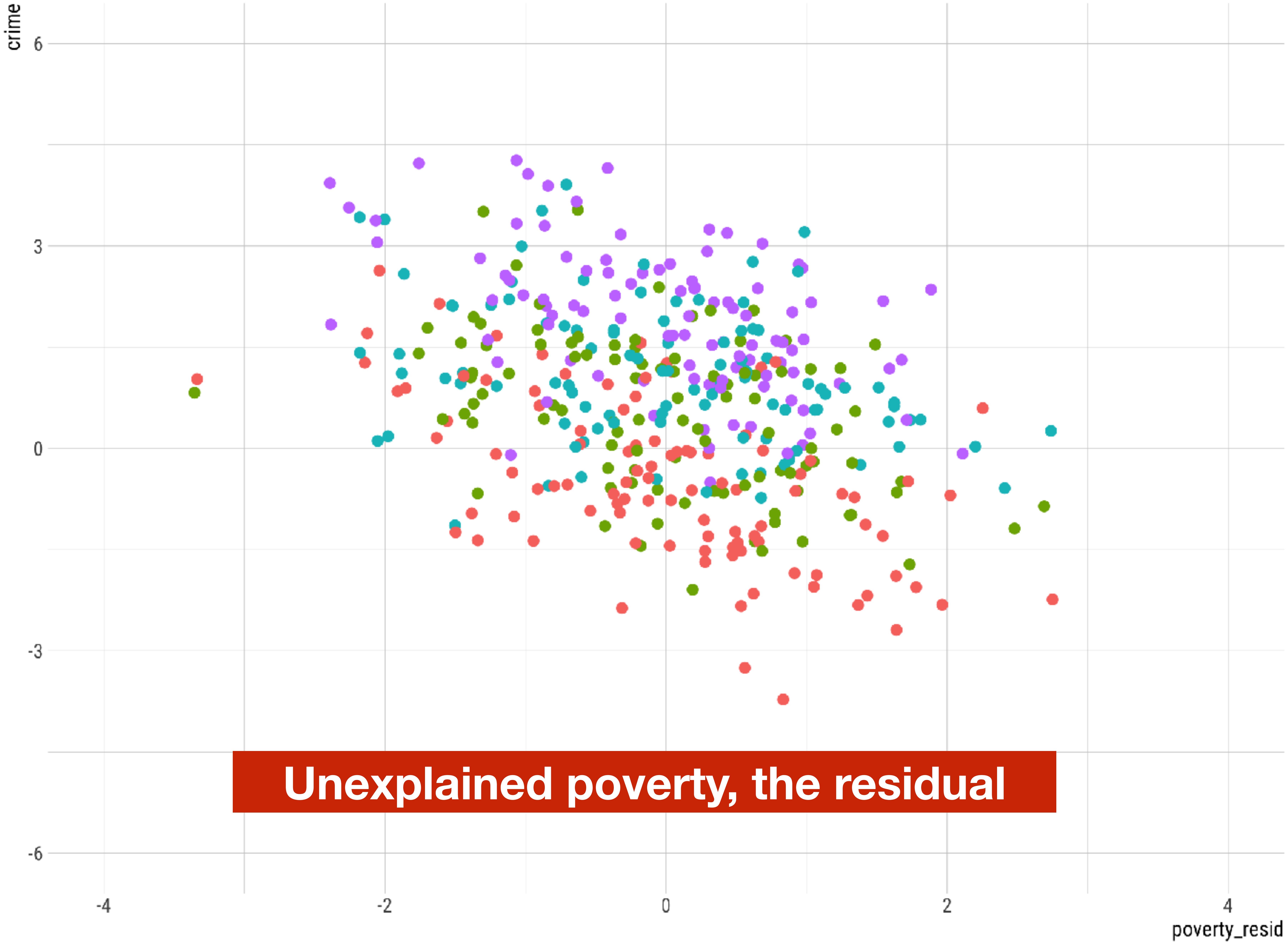

crime
6
3
0
-3
-6
-4
-2
0
2
4
poverty



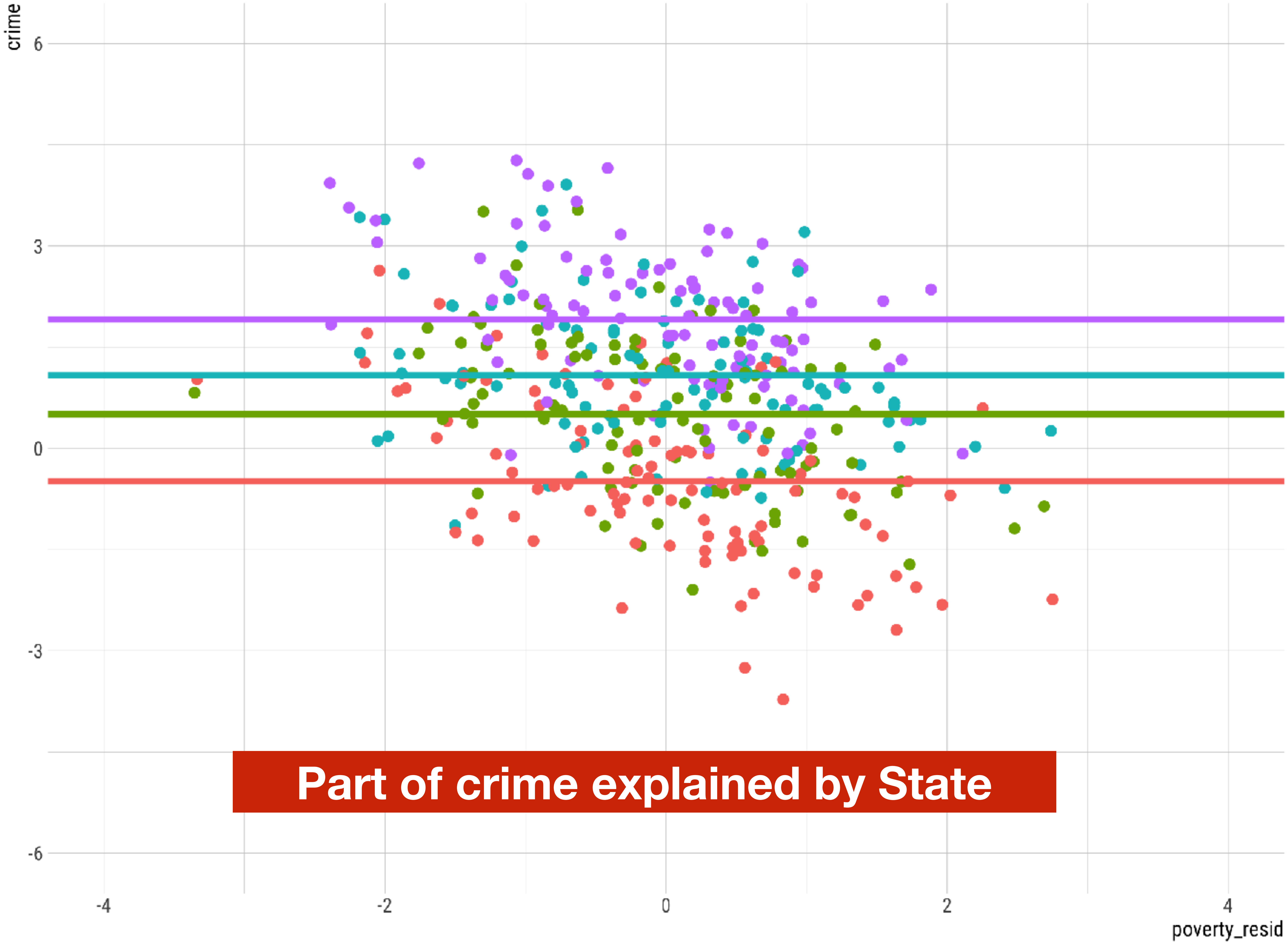
But look *within* country…



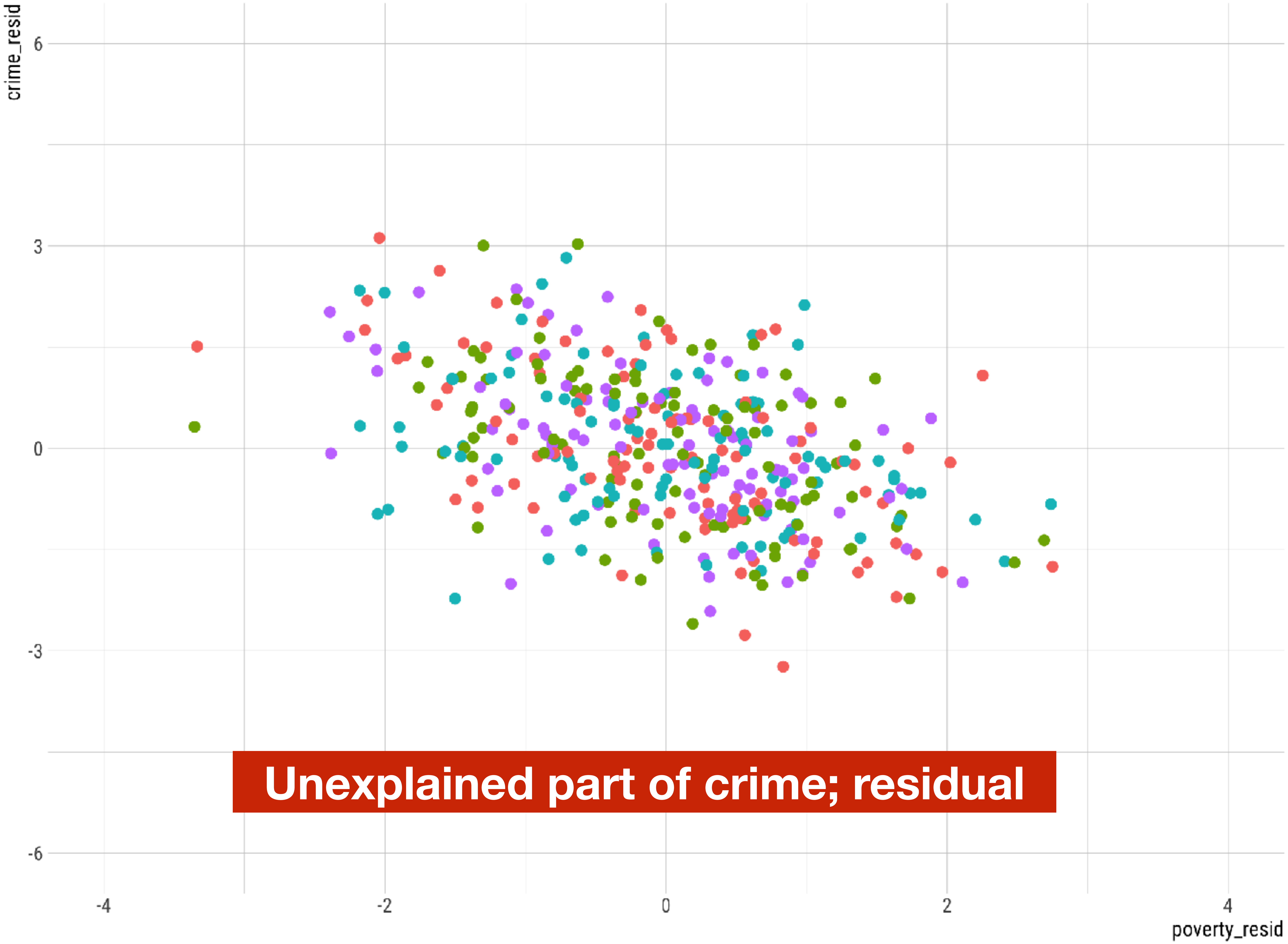
Part of Poverty explained by State



Unexplained poverty, the residual



Part of crime explained by State



Unexplained part of crime; residual


crime_resid
6
3
0
-3
-6
-4
-2
0
2
4
poverty_resid
The answer we were looking for!

What does FE do?

Table 1

state	poverty	crime	unobserved
1	1.16	-2.32	0.186
1	-1.77	0.404	0.186
2	1.33	0.283	0.503
2	-0.0619	-1.45	0.503
3	1.28	1.2	0.946
3	0.126	-0.427	0.946
4	1.57	1.68	-0.89
4	1.62	2.48	-0.89

What does FE do?

Table 1

state	poverty	crime	unobserved	poverty_mean	crime_mean	unobserved_m
1	1.16	-2.32	0.186	-0.211	-0.486	0.186
1	-1.77	0.404	0.186	-0.211	-0.486	0.186
2	1.33	0.283	0.503	0.118	0.505	0.503
2	-0.0619	-1.45	0.503	0.118	0.505	0.503
3	1.28	1.2	0.946	0.728	1.08	0.946
3	0.126	-0.427	0.946	0.728	1.08	0.946
4	1.57	1.68	-0.89	1.43	1.91	-0.89
4	1.62	2.48	-0.89	1.43	1.91	-0.89

What does FE do?

Table 1

state	poverty	crime	unobserved	poverty_resi	crime_residu	unobserved_r
1	1.16	-2.32	0.186	1.37	-1.84	0
1	-1.77	0.404	0.186	-1.56	0.89	0
2	1.33	0.283	0.503	1.22	-0.222	0
2	-0.0619	-1.45	0.503	-0.18	-1.95	0
3	1.28	1.2	0.946	0.553	0.117	0
3	0.126	-0.427	0.946	-0.602	-1.51	0
4	1.57	1.68	-0.89	0.132	-0.228	0
4	1.62	2.48	-0.89	0.185	0.567	0

In regression

```
lm(crime ~ poverty, data = new_df)
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.757165   0.077658   9.750   <2e-16 ***
poverty     -0.007792   0.058918  -0.132    0.895
---
```

```
lm(crime ~ poverty + state, data = new_df)
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.59270    0.10083  -5.878 8.82e-09 ***
poverty     -0.50599    0.04835 -10.464  < 2e-16 ***
state2       1.15690    0.14275   8.104 6.70e-15 ***
state3       2.04552    0.14895  13.733  < 2e-16 ***
state4       3.22820    0.16264  19.849  < 2e-16 ***
---
```

```
lm(crime ~ poverty + state + unobserved, data = new_df)
```

```
Coefficients: (1 not defined because of singularities)
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.59270    0.10083  -5.878 8.82e-09 ***
poverty     -0.50599    0.04835 -10.464  < 2e-16 ***
state2       1.15690    0.14275   8.104 6.70e-15 ***
state3       2.04552    0.14895  13.733  < 2e-16 ***
state4       3.22820    0.16264  19.849  < 2e-16 ***
unobserved        NA         NA      NA       NA
```

Caveat

Fixed effects account for all differences ACROSS units

So FE can account for any backdoor that is constant within units

But if a backdoor varies within units (e.g., changes over time), FE won't help!

"Constant within unit"

A variable that only varies ACROSS units, not WITHIN them

country	continent	year	lifeExp	pop	gdpPercap	ww2
Afghanistan	Asia	1952	28.8	8425333	779	no
Afghanistan	Asia	1957	30.3	9240934	821	no
Afghanistan	Asia	1962	32	10267083	853	no
Afghanistan	Asia	1967	34	11537966	836	no
Afghanistan	Asia	1972	36.1	13079460	740	no
Afghanistan	Asia	1977	38.4	14880372	786	no
Afghanistan	Asia	1982	39.9	12881816	978	no
Afghanistan	Asia	1987	40.8	13867957	852	no
Afghanistan	Asia	1992	41.7	16317921	649	no
Afghanistan	Asia	1997	41.8	22227415	635	no
	Constant		Varies	Varies	Varies	Constant

Example: population

FE will account for differences ACROSS countries in population

But it can't account for changes in population WITHIN countries

E.g., if population growth is *causing* both GDP and Life-expectancy, then FE results will still be confounded

This doesn't mean the FE "don't work"
we just need to also control for population

What other things might change over time?

Practice

Is it worth having teachers get Master's degrees? Does this make them better at teaching?

Want to know the effect of Master's degree on student scores

Each student took multiple tests

Draw a DAG and think about what variables might be backdoors

Think about: what backdoors would FE block?

Practice

School grades study

name	sex	race	income	grade	reading(min)	bullied	age	weather	masters	teacher(t-1)
Jaunice	Male	White	4.04	88.8	40	no	12	79.2	no	Horrace
Jaunice	Male	White	4.04	88.6	50	no	12	76.7	yes	Burdette
Jaunice	Male	White	4.04	93.5	29	no	14	79.3	no	Shakira
Joleigha	Female	White	6.44	90.8	51	no	10	74.1	yes	Havala
Joleigha	Female	White	6.44	98.3	38	no	14	78.2	no	Anistynn
Joleigha	Female	White	6.44	95.4	49	no	16	79.7	no	Novela
Orlin	Male	White	4.81	85.7	30	no	10	82.5	no	Twana
Orlin	Male	White	4.81	90.5	38	no	12	81	yes	Branae
Orlin	Male	White	4.81	90.4	41	no	14	81.1	no	Krithik
Shadow	Male	White	6.25	92.7	48	no	15	80	no	Danait
Shadow	Male	White	6.25	91.3	54	no	15	82.6	yes	Kately
Shadow	Male	White	6.25	87.1	36	no	16	91.3	no	Shainah
Threasa	Male	Hispanic	4.26	86.3	42	no	15	78.6	no	Selmer
Threasa	Male	Hispanic	4.26	87.1	33	yes	16	71.5	yes	Jasmely
Threasa	Male	Hispanic	4.26	84.2	58	no	16	78.4	yes	Jaxi