SciencesPo
DEPARTMENT OF ECONOMICS

ScPoEconometrics: Advanced

Panel Data

Bluebery Planterose
SciencesPo Paris
2023-03-28

Where Did We Stop Last Time?

- **IV** estimation
- Some important applications
- Some pitfalls

Where Did We Stop Last Time?

- **IV** estimation
- Some important applications
- Some pitfalls

Today

1. Revisit the Ability Bias in an App 😎
2. Introduce Panel Data

Cross-Sectional Data

So far, we dealt with data that looks like this:

County	CrimeRate	ProbofArrest
1	0.0398849	0.289696
3	0.0163921	0.202899
5	0.0093372	0.406593
7	0.0219159	0.431095
9	0.0075178	0.631579

- We have a unit identifier (like `County` here),
- Observables on each unit.
- Usually called a **cross-sectional** dataset
- Provides single snapshot view
- Each row, in other words, is one *observation*.

Panel Data

Now, let's add a `time` index: `Year`.

County	Year	CrimeRate	ProbofArrest
1	81	0.0398849	0.289696
1	82	0.0383449	0.338111
1	83	0.0303048	0.330449
1	84	0.0347259	0.362525
1	85	0.0365730	0.325395
1	86	0.0347524	0.326062
1	87	0.0356036	0.298270
3	81	0.0163921	0.202899
3	82	0.0190651	0.162218

- Next to the unit identifier (`County`) we now have `Year`
- Now a pair (`County`,`Year`) indexes one observation.
- We call this a **panel** or **longitudinal** dataset
- We can track units *over time*.

Crime Rates and Probability of Arrest

- The above data can be loaded with

```
data(crime4,package = "wooldridge")
```

- They are from C. Cornwell and W. Trumball (1994), “Estimating the Economic Model of Crime with Panel Data”.

Crime Rates and Probability of Arrest

- The above data can be loaded with

```
data(crime4,package = "wooldridge")
```

- They are from C. Cornwell and W. Trumball (1994), “Estimating the Economic Model of Crime with Panel Data”.
- One question here: *how big is the deterrent effect of law enforcement*? If you know you are more likely to get arrested, will you be less likely to commit a crime?

Crime Rates and Probability of Arrest

- The above data can be loaded with

```
data(crime4,package = "wooldridge")
```

- They are from C. Cornwell and W. Trumball (1994), "Estimating the Economic Model of Crime with Panel Data".
- One question here: *how big is the deterrent effect of law enforcement*? If you know you are more likely to get arrested, will you be less likely to commit a crime?
- This is tricky: Does high crime *cause* stronger police response, which acts as a deterrent, or is crime low because deterrent is strong to start with?
- This is sometimes called a *simultaneous equation model* situation: police response impacts crime, but crime impacts police response

$$police = \alpha_0 + \alpha_1 crime$$
$$crime = \beta_0 + \beta_1 police$$

Crime Rates and Probability of Arrest

- Most literature prior to that paper estimated simultaneous equations off cross sectional data
- Cornwell and Trumball are worried about **unobserved heterogeneity** between jurisdictions.
- Why? What could possibly go wrong?

- Let's pick out 4 counties from their dataset
- Let's look at the crime rate vs probability of arrest relationship
- First for all of them together as a single cross section
- Then taking advantage of the panel structure (i.e. each county over time).

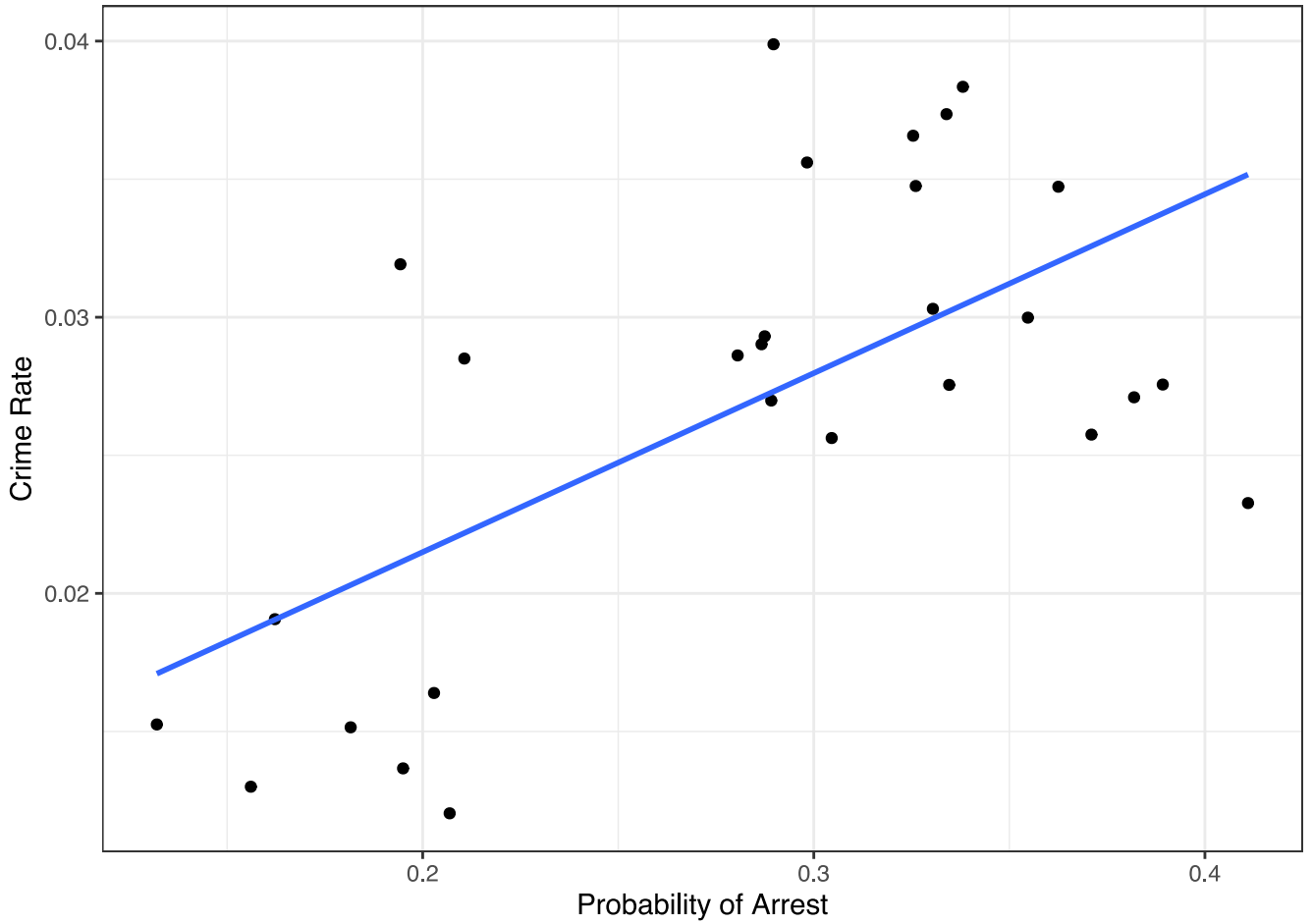
Crime vs Arrest in Cross Section

1. Subset data to 4 counties

2. plot probability of arrest vs crime rate.

```
css = crime4 %>%
  filter(county %in%
           c(1,3,145, 23))

ggplot(css,
       aes(x = prbarr,
           y = crmrte)) +
  geom_point() +
  geom_smooth(method="lm",
              se=FALSE) +
  theme_bw() +
  labs(x = 'Probability of Arrest
       y = 'Crime Rate')
```



Crime vs Arrest in Cross Section: Positive Relationship!

- We see an upward sloping line!
- Higher probability of arrest is associated to higher crime rates.
- How strong is the effect?

Crime vs Arrest in Cross Section: Positive Relationship!

- We see an upward sloping line!
- Higher probability of arrest is associated to higher crime rates.
- How strong is the effect?

```
xsection = lm(crmrte ~ prbarr, css)
coef(xsection)[2]  # gets slope coef
```

```
##     prbarr
## 0.06480104
```

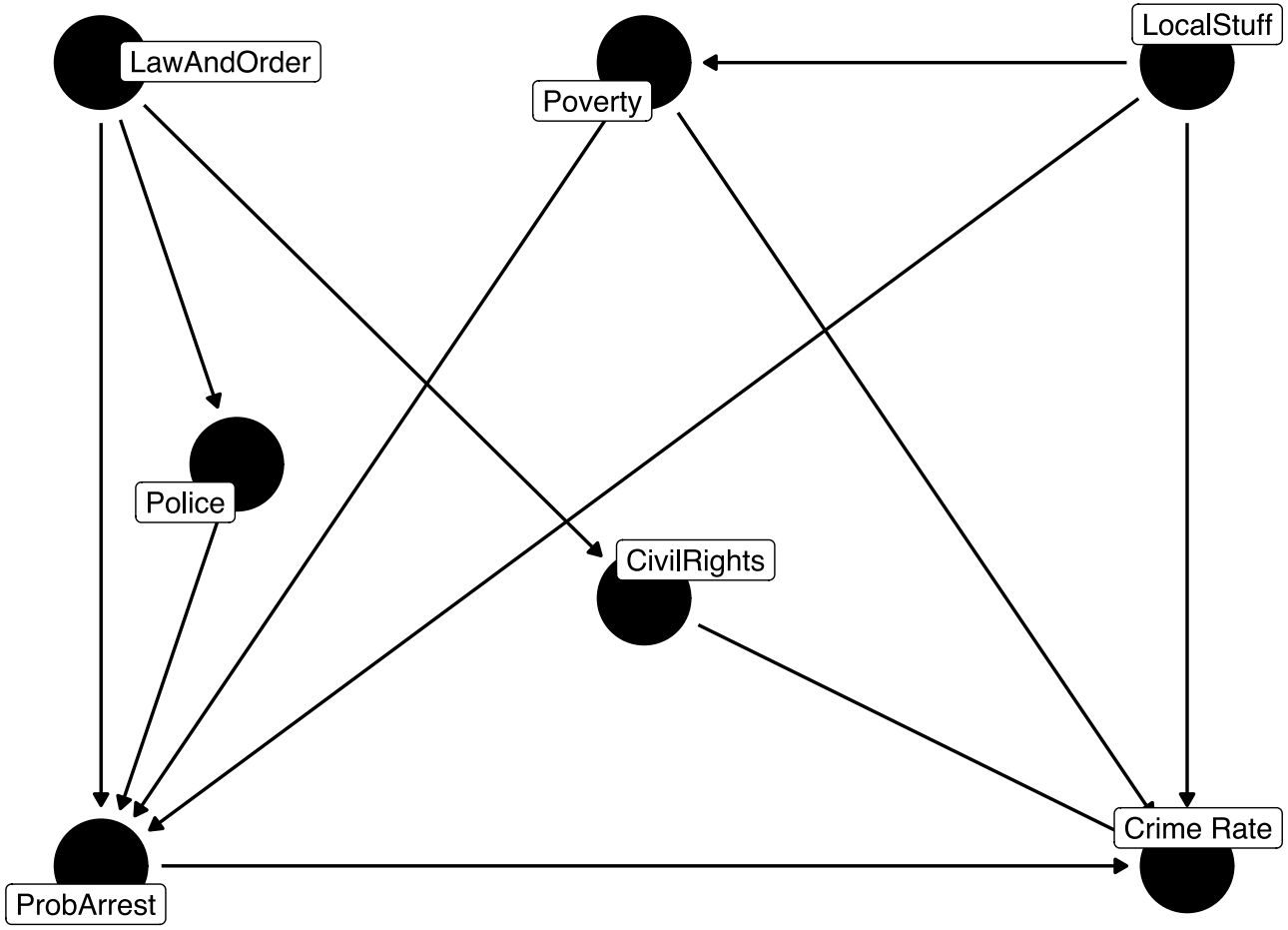
- Increasing probability of arrest by 1 unit (i.e. 100 percentage point), increases the crime rate by 0.064801. So, if we double the probability of arrest, crime would increase by 0.064 crimes per person.
- Increase of 10 percentage points in the probability of arrest (e.g. `prbarr` goes from 0.2 to 0.3) ...
- ... is associated with an increase in crime rate from 0.021 to 0.028, or a 33.33 percent increase in the crime rate.

Ok, but what does that *mean*?

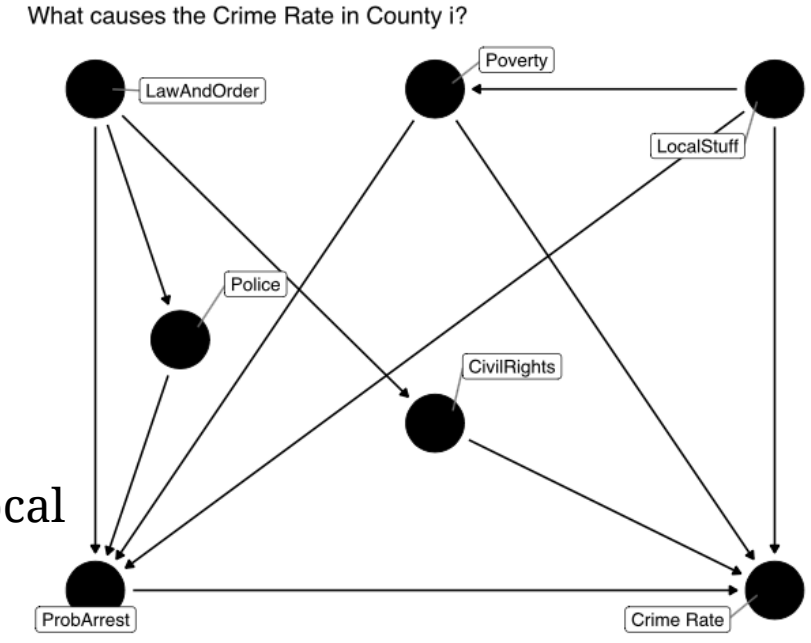
- Literally: counties with a higher probability of being arrested also have a higher crime rate.
- So, does it mean that as there is more crime in certain areas, the police become more efficient at arresting criminals, and so the probability of getting arrested on any committed crime goes up?
- What does police efficiency depend on?
- Does the poverty level in a county matter for this?
- The local laws?
- 🤯 wow, there seem to be too many things left out of this simple picture.

Crime in a DAG

What causes the Crime Rate in County i?



Crime in a DAG



Fixed Characteristics: vary by county

- `LocalStuff` are things that describe the County, like geography, and other persistent features.
- `LawAndOrder`: commitment to *law and order politics* of local politicians
- `CivilRights`: how many civil rights you have

Time-varying Characteristics: vary by county and by year

- `Police` budget: an elected politician has some discretion over police spending
- `Poverty` level varies with the national/global state of the economy.

Within and Between Variation

You will often hear the terms *within* and *between* variation in panel data contexts.

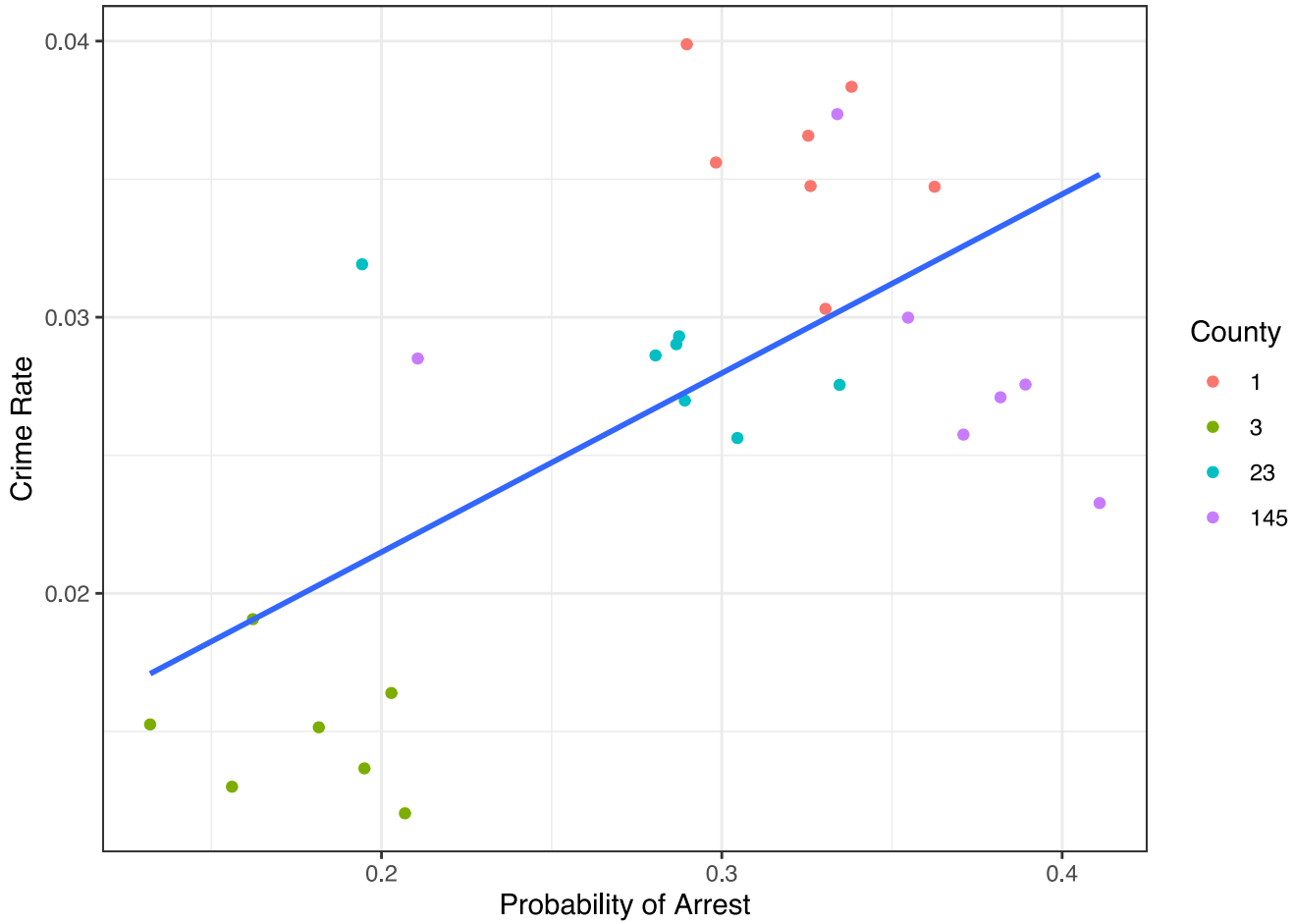
Within Variation

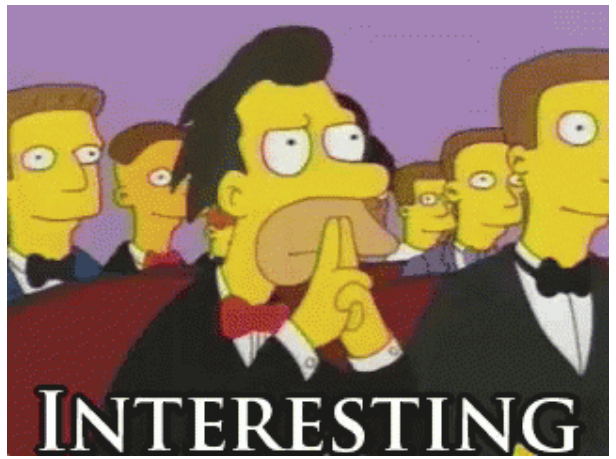
- things that change *within each group* over time:
- here we said police budgets
- and poverty levels would change within each group and over time.

Between Variation

- Things that are **fixed** for each group over time:
- `LocalStuff`
- `LawAndOrder` and
- `CivilRights`
- differ only across or **between** groups

Within and Between Variation: Give us a Plot.



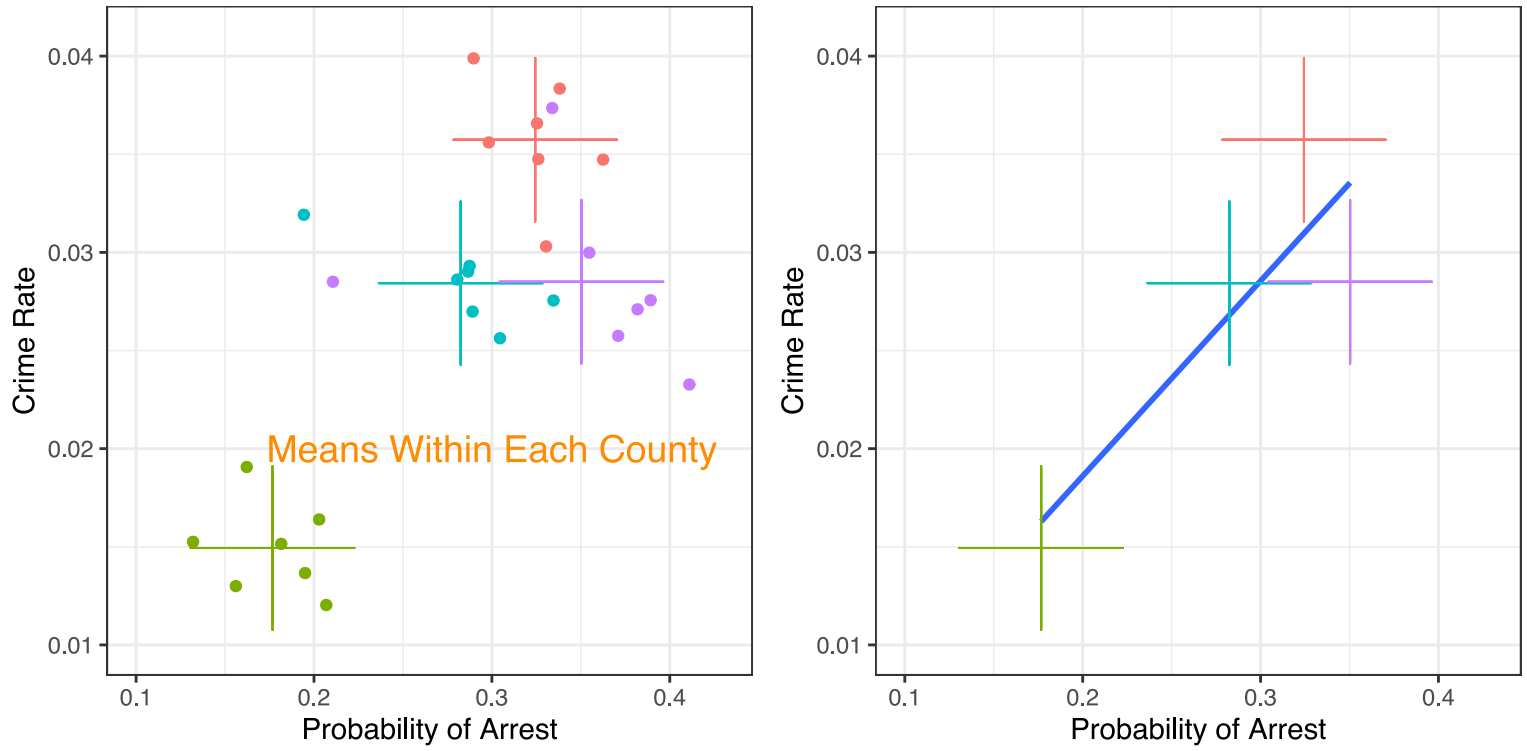


Pooled OLS recovers *between* variation

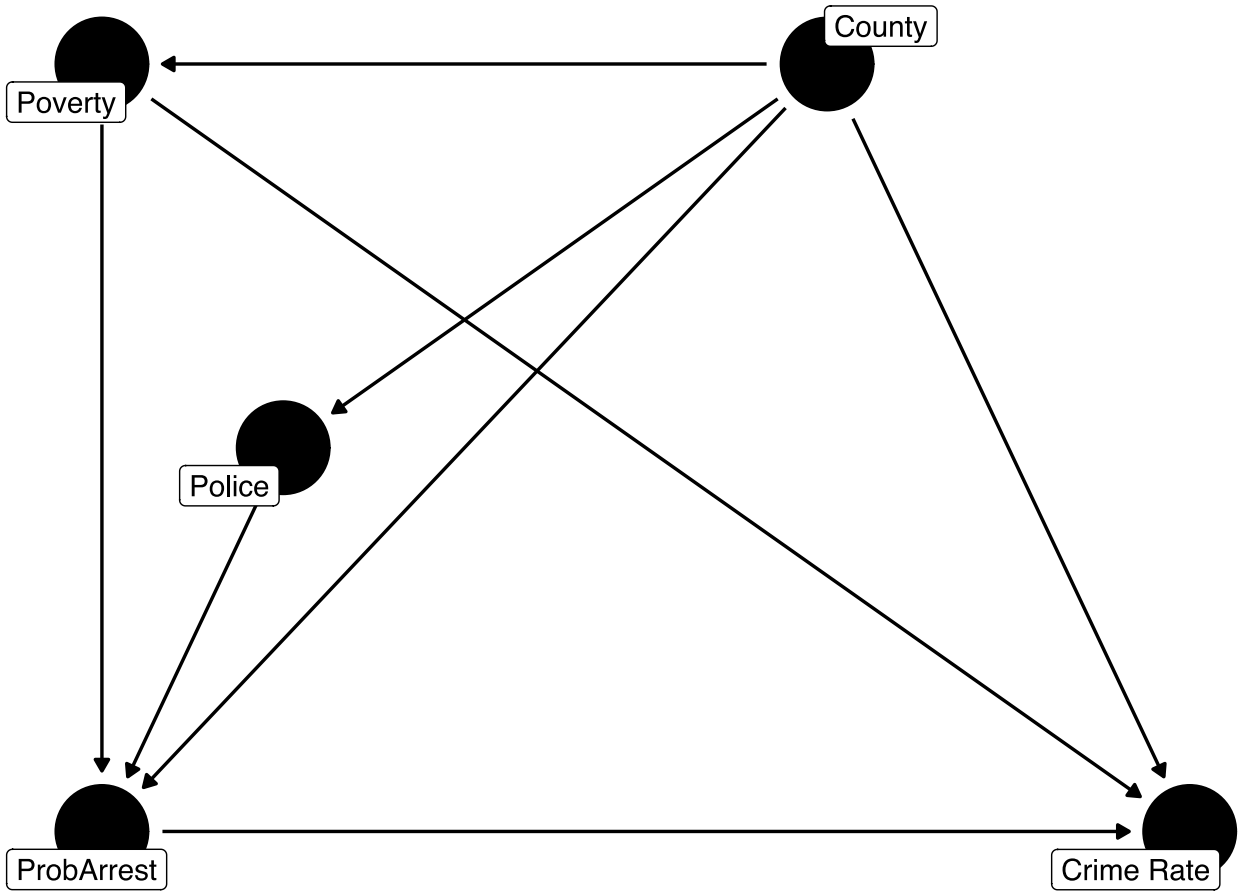
- Let's add the mean of `prbarr` and `crmrte` for each of those counties to the scatter plot!
- And then a regression through those 4 points!

Pooled OLS recovers *between* variation

- Let's add the mean of `prbarr` and `crmrte` for each of those counties to the scatter plot!
- And then a regression through those 4 points!



Accounting for Grouped Data: Introducing the *Fixed Effect*



- Collect all group-specific time-invariant features in the factor `County`.
- Takes care of all factors which do *not* vary over time within each unit.
- We can **net out** the group effect!
- We call `County` a **fixed effect**.

Fixed Effects Estimation in R

OVB, IV and Panel Data

We've seen *unobserved variable bias* (OVB). For example, if the true model read:

$$y_i = \beta_0 + \beta_1 x_i + c_i + u_i$$

if c_i unobservable and $Cov(x_i, c_i) \neq 0 \Rightarrow E[u_i + c_i | x_i] \neq 0,$ with $u_i + c_i$ total unobserved component.

OVB, IV and Panel Data

We've seen *unobserved variable bias* (OVB). For example, if the true model read:

$$y_i = \beta_0 + \beta_1 x_i + c_i + u_i$$

if c_i unobservable and $Cov(x_i, c_i) \neq 0 \Rightarrow E[u_i + c_i | x_i] \neq 0,$ with $u_i + c_i$ total unobserved component.

Cross-Sectional Solution

- where $c = A_i$ and $x = s$ was schooling.
- *ability bias*.
- Find IV correlated with schooling but not ability

OVB, IV and Panel Data

We've seen *unobserved variable bias* (OVB). For example, if the true model read:

$$y_i = \beta_0 + \beta_1 x_i + c_i + u_i$$

if c_i unobservable and $Cov(x_i, c_i) \neq 0 \Rightarrow E[u_i + c_i | x_i] \neq 0,$ with $u_i + c_i$ total unobserved component.

Cross-Sectional Solution

- where $c = A_i$ and $x = s$ was schooling.
- *ability bias*.
- Find IV correlated with schooling but not ability

Panel Data

$$y_{it} = \beta_1 x_{it} + c_i + u_{it}, \quad t = 1, 2, \ldots T$$

- c_i: *individual fixed effect, unobserved effect* or *unobserved heterogeneity*.
- c_i: is fixed over time (ability A_i for example), but can be correlated with x_{it}!

Dummy Variable Regression

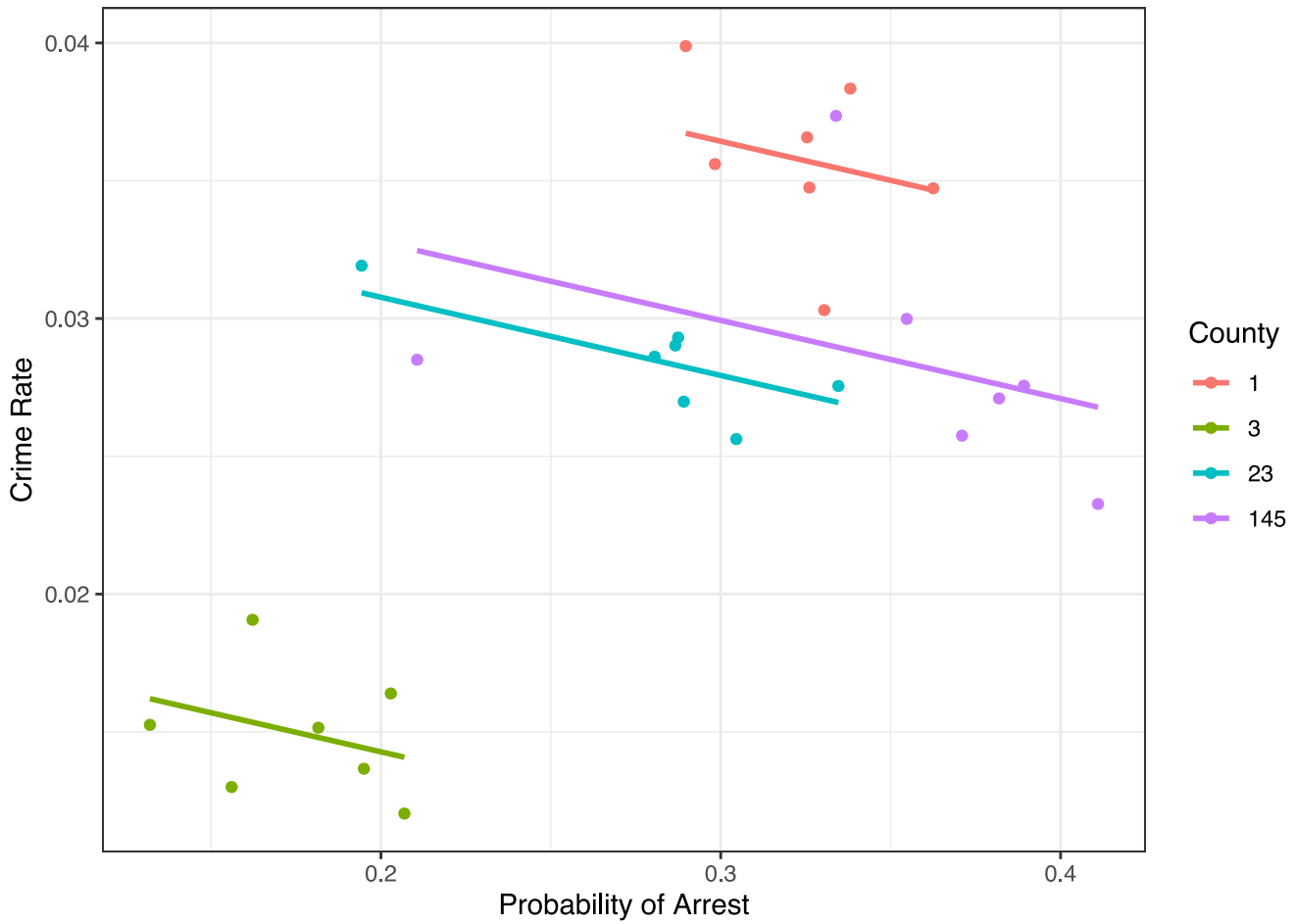
- Simplest approach: include a dummy variable for each group i.
- This is literally *controlling for county* i
- Each i has basically their own intercept c_i
- In `R` you achieve this like so:

$$y_{it} = \beta_1 x_{it} + c_i + u_{it}, \quad t = 1, 2, \dots T$$

```
mod = list()
mod$dummy <- lm(crmrte ~ prbarr + factor(county), css
broom::tidy(mod$dummy)
```

```
## # A tibble: 5 × 5
##   term              estimate std.error statistic  p.value
##   <chr>                <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)         0.0449   0.00456      9.87 9.85e-10
## 2 prbarr             -0.0284   0.0136      -2.08 4.86e- 2
## 3 factor(county)3    -0.0250   0.00254     -9.82 1.07e- 9
## 4 factor(county)23   -0.00850  0.00166     -5.13 3.41e- 5
## 5 factor(county)145  -0.00650  0.00160     -4.07 4.70e- 4
```

Dummy Variable Regression



- *Within* each county, now is a **negative** relationship!!
- Different intercepts (county 1 is the reference group),
- Unique slope coefficient β. (you observe that the lines are parallel).
- We are shifting lines down from the reference group 1.

First Differencing Solution

If we only had $T = 2$ periods, we could just difference both periods, basically leaving us with

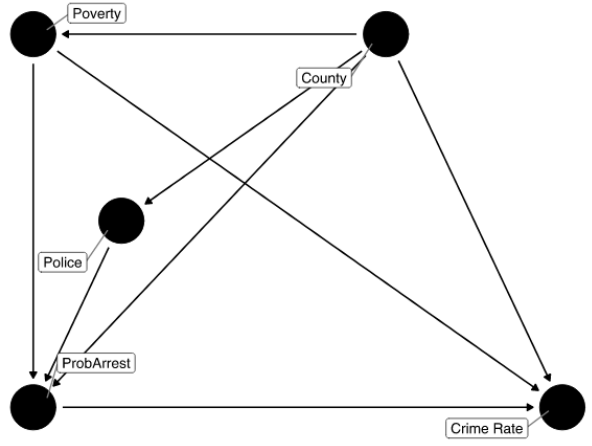
$$
\begin{aligned}
y_{i1} &= \beta_1 x_{i1} + c_i + u_{i1} \\
y_{i2} &= \beta_1 x_{i2} + c_i + u_{i2} \\
&\Rightarrow \\
y_{i1} - y_{i2} &= \beta_1 (x_{i1} - x_{i2}) + c_i - c_i + u_{i1} - u_{i2} \\
\Delta y_i &= \beta_1 \Delta x_i + \Delta u_i
\end{aligned}
$$

where Δ means *difference over time of* and to recover the parameter of interest β_1 we would run

```
lm(deltay ~ deltax, diff_data)
```

The Within Transformation

- With $T > 2$ we need a different approach
- One important concept is called the *within* transformation
- So, *controlling for group identity and only looking at time variation*
- Remember DAG!



- Let $\bar{x}_i$ the average *over time* of i's x values:

$$\bar{x}_i = \frac{1}{T} \sum_{t=1}^{T} x_{it}$$

1. for all variables compute their time-mean for each unit i: $\bar{x}_i, \bar{y}_i$ etc
2. for each observation, substract that time mean from the actual value and define $(x_{it} - \bar{x}_i), (y_{it} - \bar{y}_i)$
3. Finally, regress $(x_{it} - \bar{x}_i)$ on $(y_{it} - \bar{y}_i)$

The Within Transformation in R: Manual Solution

This *works* for our problem with fixed effect c_i because c_i is not time varying by assumption! hence it drops out:

$$y_{it} - \bar{y}_i = \beta_1(x_{it} - \bar{x}_i) + c_i - c_i + u_{it} - \bar{u}_i$$

It's easy to do yourself! First let's compute the demeaned values:

```
cdata <- css %>%
  group_by(county) %>%
  mutate(mean_crime = mean(crmrte),
         mean_prob = mean(prbarr)) %>%
  mutate(demeaned_crime = crmrte - mean_crime,
         demeaned_prob = prbarr - mean_prob)
```

Then, run both models with simple OLS:

```
mod$xsect <- lm(crmrte ~ prbarr, data = cdata)
mod$demeaned <- lm(demeaned_crime ~ demeaned_prob, data = cdata)
```

The Within Transformation in R: Manual Solution

We get this table:

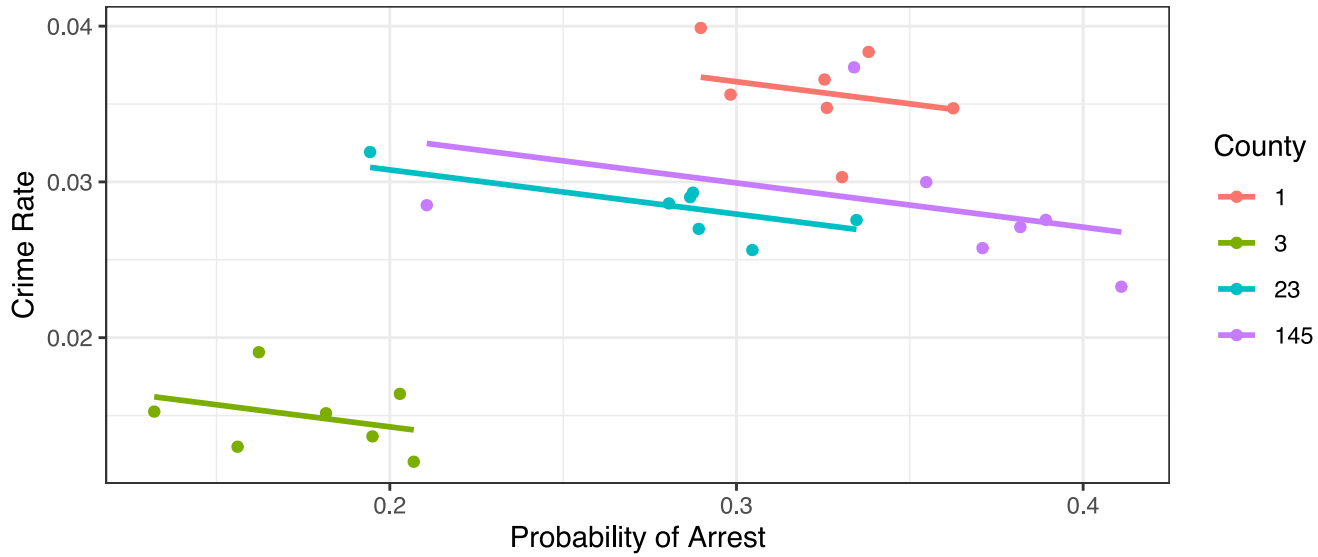
	xsect	dummy	demeaned
(Intercept)	0.009	0.045	0.000
	(0.005)	(0.005)	(0.001)
prbarr	0.065	-0.028	
	(0.016)	(0.014)	
demeaned_prob			-0.028
			(0.013)
R2	0.390	0.893	0.159

- Estimate for `prbarr` is positive in the cross-section
- Taking care of the unobservered heterogeneity c_i...
- ...either by including an intercept for each i or by time-demeaning the data
- we obtain: -0.028 .

Interpreting the Within Estimates

- How to interpret those negative slopes?
- We look at a single unit i and ask:

> if the arrest probability in i increases by 10 percentage points (i.e. from 0.2 to 0.3) from year t to $t+1$, we expect crimes per person to fall from 0.039 to 0.036, or by -7.69 percent (in the reference county number 1).



Fixed Effects Estimation in R: use a Package!

- In real life you will hardly ever perform the within-transformation by yourself
- and use a package instead!
- There are several options (`fixest` is fastest). In our context:

```
mod$FE = fixest::feols(crmrte ~ prbarr | county, cdata)
```

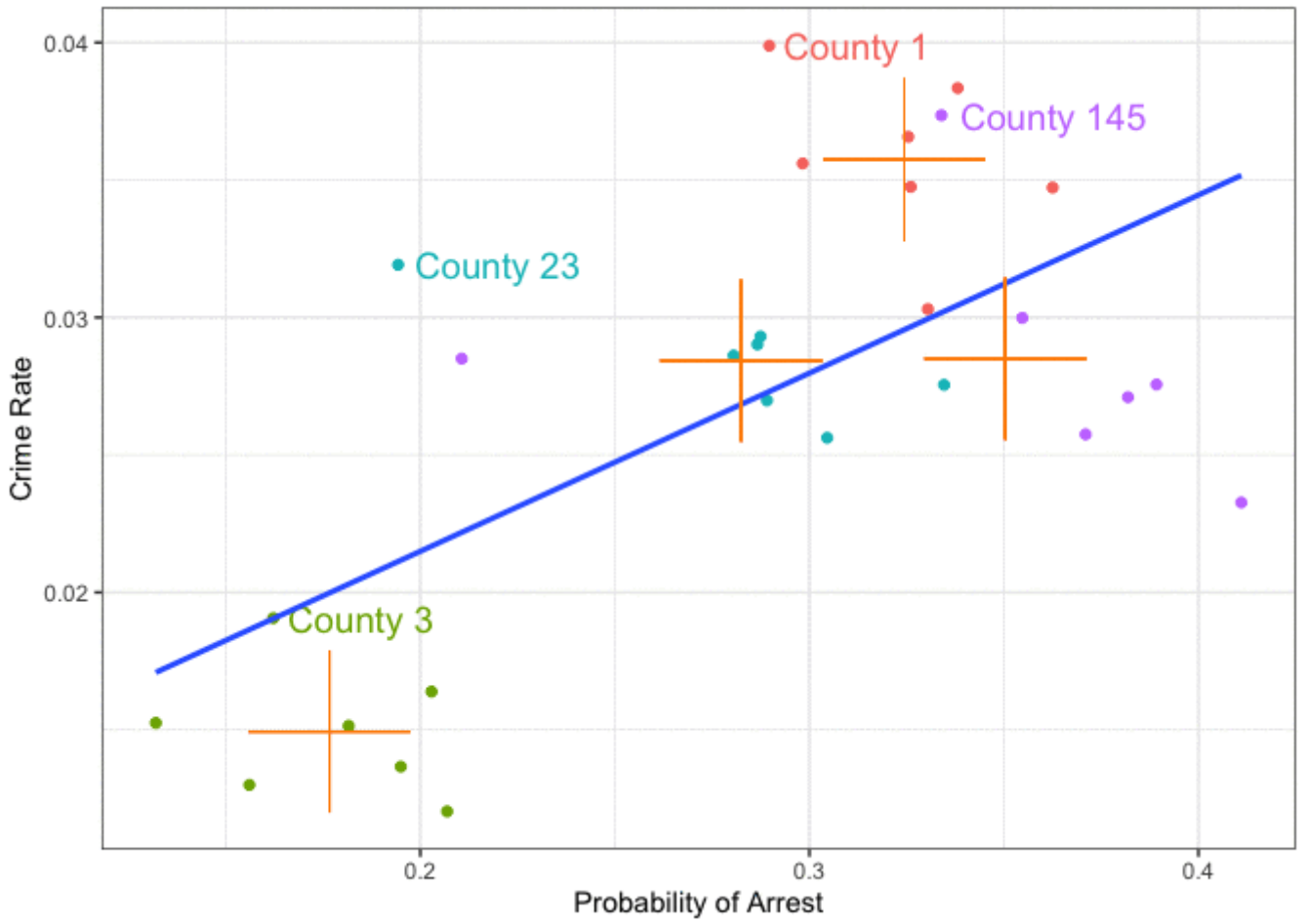
- Notice the similar setup to the `estimatr::iv_robust` *two-part formula*. Here the fixed effects come after the `|`.
- Also, we can have *more than one fixed effect*! For a cool example with *three* fixed effects see the package vignette

Fixed Effects Estimation in R: use `fixest` 🙂

	xsect	dummy	demeaned	FE
(Intercept)	0.009	0.045	0.000	
	(0.005)	(0.005)	(0.001)	
prbarr	0.065	-0.028		-0.028
	(0.016)	(0.014)		(0.005)
demeaned_prob			-0.028	
			(0.013)	
R2	0.390	0.893	0.159	0.893

- Same estimates! 😅
- Notice the standard errors: *robust*?!
- `fixest` computes **cluster-robust** se's.
- We suspect there is strong correlation in residuals *within* each county (over time).

Within Transformation Animated



- The within transformation **centers** the data!
- By time-demeaning y and x, we *project out* the fixed factors related to *county*
- Only *within* county variation is left.
- Made by Nick C Huntington-Klein. 🙏

SciencesPo
DEPARTMENT OF ECONOMICS

END

	bluebery.planterose@sciencespo.fr
	Original Slides from Florian Oswald
	Book
	@ScPoEcon
	@ScPoEcon