Data Science for Economists

Fixed Effects - control for what you can't see

Kyle Coombs

Bates College | ECON/DCS 368

Table of contents

Prologue

Goals today

1. Review omitted variable bias
2. Introduce fixed effects
3. Go over an example
4. Play with fixed effects in R

Attribution

- These slides are adapted from work by Nick Huntington-Klein and Ed Rubin

- They're both superb econometric instructors and I highly recommend their work

Omitted Variable Bias

Causal Inference Review

- Last week, we discussed **omitted variable bias**.

- We worked through using control variables to isolate the relation between our treatment and our outcome.

- What is endogeneity?

Causal Inference Review

- Last week, we discussed **omitted variable bias**.

- We worked through using control variables to isolate the relation between our treatment and our outcome.

- What is endogeneity?

- $cov(X, \varepsilon) = 0$

- No relationship between the error term and the independent variable

Forms of endogeneity

1. Reverse causality: $X \to Y$ but also $Y \to X$

2. Selection bias: $X \to Y$ but also $Z \to X$ and $Z \to Y$

3. Omitted variable bias: $X \to Y$ but also $Z \to Y$

4. Measurement error: $X \to Y$ but also $X \to \hat{X}$ (this one is a little different, but it's still a form of endogeneity)

Conditional Independence Assumption

- Last week we also talked about conditional independence assumption

Conditional Independence Assumption

- Last week we also talked about conditional independence assumption

- After controlling for all the variables that are correlated with both the treatment and the outcome, the treatment is independent of the error term

- This is a way to minimize selection bias/omitted variable bias

- But what if we miss a variable in our model?

- What if we know a variable, but we cannot measure it?

- Any guesses?

You cannot measure what you cannot see

- What if our data vary by time and individual and our regression model is:

$$Earnings = \beta_0 + \beta_1 Edu + \beta_2 Ability + \beta_3 Experience + \cdots +$$
$$\beta_{k-2} Parent Attentive + \beta_{k-1} Race + \beta_k Gender + u$$

- ParentAttentive captures how attentive a child's parents are, which definitely matters to upbringing

- We probably do not have data on parental attention or ability, but we know they matter

- So the regression we run is

$$Earnings = \beta_0 + \beta_1 Edu + \beta_2 Race + \beta_3 Experience \cdots +$$
$$+ \beta_k Gender + (\beta_{k-1} Ability + \beta_k Parent Attentive + u)$$

- But after childhood, parental attention is fixed and ability is slow to change

- How can we control for them?

Causal Inference Unit

- Much of causal inference is about *finding ways to control for stuff that we can't measure*

Causal Inference Unit

- Much of causal inference is about *finding ways to control for stuff that we can't measure*

John Cena cannot hide from fixed effects.

Causal Inference Unit

- Much of causal inference is about *finding ways to control for stuff that we can't measure*

John Cena cannot hide from fixed effects.

- Seems impossible! But it is possible, at least in some circumstances

- Today, we will talk about *within* and *between* variation, controlling for all *between variation* using *fixed effects*

Fixed Effects

Fixed effects

- Fixed effects are a way to control for *some* **endogeneity**

- Within a **group of observations**, or **dimension**, we remove the *within-group* averages

- A group could be a person, a company, a state, a country, a time period, etc. -- you just need multiple observations within the group

- Any leftover variation in the data is not related to differences between groups

- This is all based on the Frisch-Waugh-Lovell theorem, which we won't cover formally

- FEs are powerful for causal inference and simplifying big, multi-dimensional data

Panel Data

- We are working in the domain of *panel data*

- Panel data is when you observe the same individual over multiple periods

- "Individual" could be a person, or a company, or a state, or a country, etc. There are N individuals in the panel data

- "Time period" could be a year, a month, a day, etc.. There are T time periods in the data

- For now we'll assume we observe each individual the same number of times, i.e. a *balanced* panel (so we have $N \times T$ observations)

- This works with unbalanced panels too, but it's more complicated

Panel Data

- Here's what (a few rows from) a panel data set looks like - a variable for individual (county), a variable for time (year), and then the data

```r
data(crime4)
crime4 %>%
  select(county, year, crmrte, prbarr) %>%
  rename(County = county,
         Year = year,
         CrimeRate = crmrte,
         ProbofArrest = prbarr) %>%
  slice(1:9) %>%
  knitr::kable(note = '...') %>%
  kableExtra::add_footnote('9 rows out of 630. "Prob. of Arrest" is estimated probability of being
```

Panel Data

- Here's what (a few rows from) a panel data set looks like - a variable for individual (county), a variable for time (year), and then the data

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

Note: 9 rows out of 630. "Prob. of Arrest" is estimated probability of being arrested when you commit a crime

Crime and Arrests

- Let's ask how increased probability of arrest affects crime

- Certainly we'd expect there to be correlation between the two!

- Why can't we just estimate this regression?

$$\text{Crime Rate} = \beta_0 + \beta_1 \text{Prob. of Arrest} + \epsilon$$

Crime and Arrests

- Let's ask how increased probability of arrest affects crime

- Certainly we'd expect there to be correlation between the two!

- Why can't we just estimate this regression?

$$\text{Crime Rate} = \beta_0 + \beta_1 \text{Prob. of Arrest} + \epsilon$$

1. Reverse causality -- more crime leads to more arrests

2. Selection bias -- counties with more crime might institute

3. Omitted variable bias -- counties with higher property values may have higher crime rates, but also more tax revenue to spend on police

Do you notice that these problems have overlap? That's cause they're all forms of **endogeneity!**

Between and Within

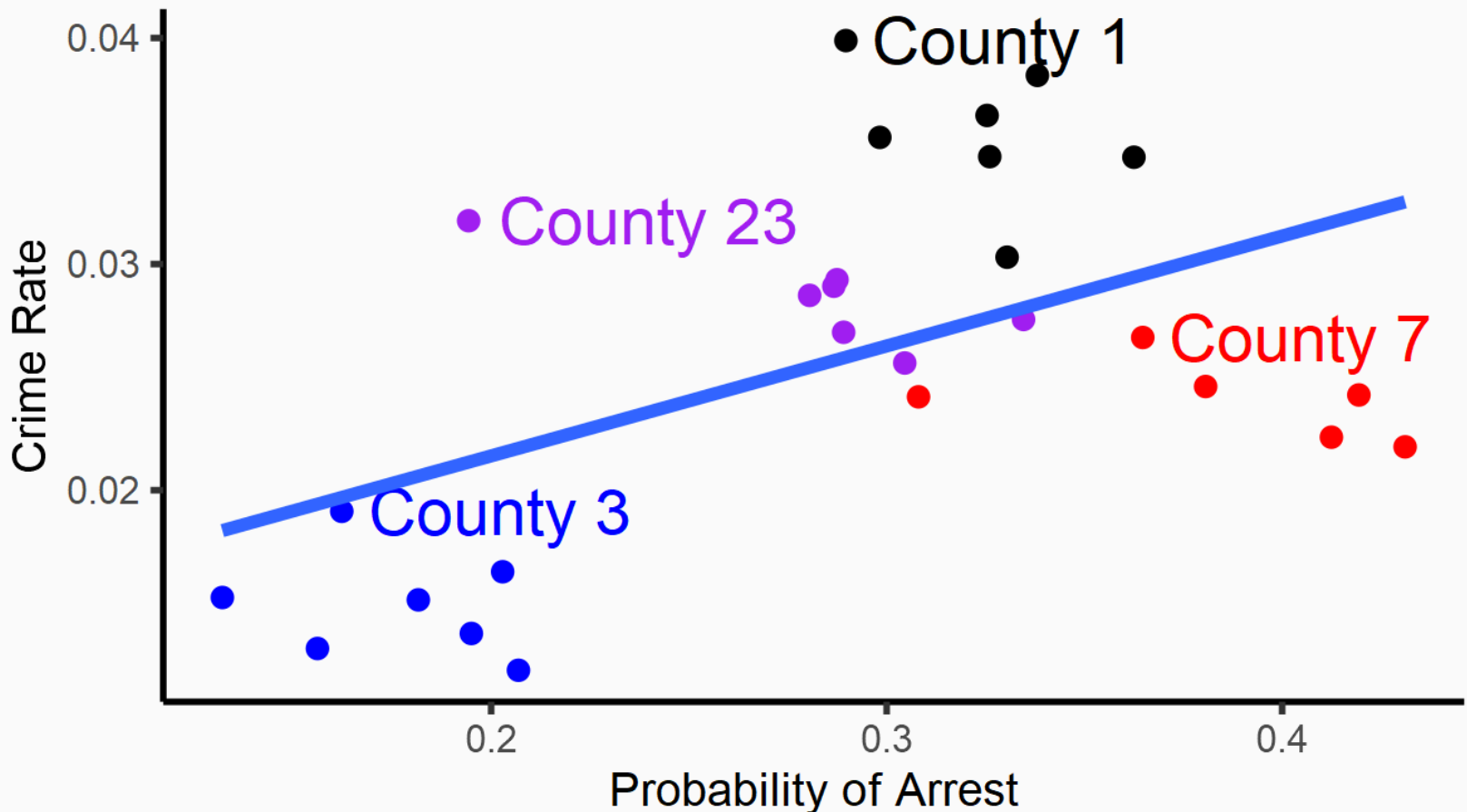- Let's pick a few counties and graph this out
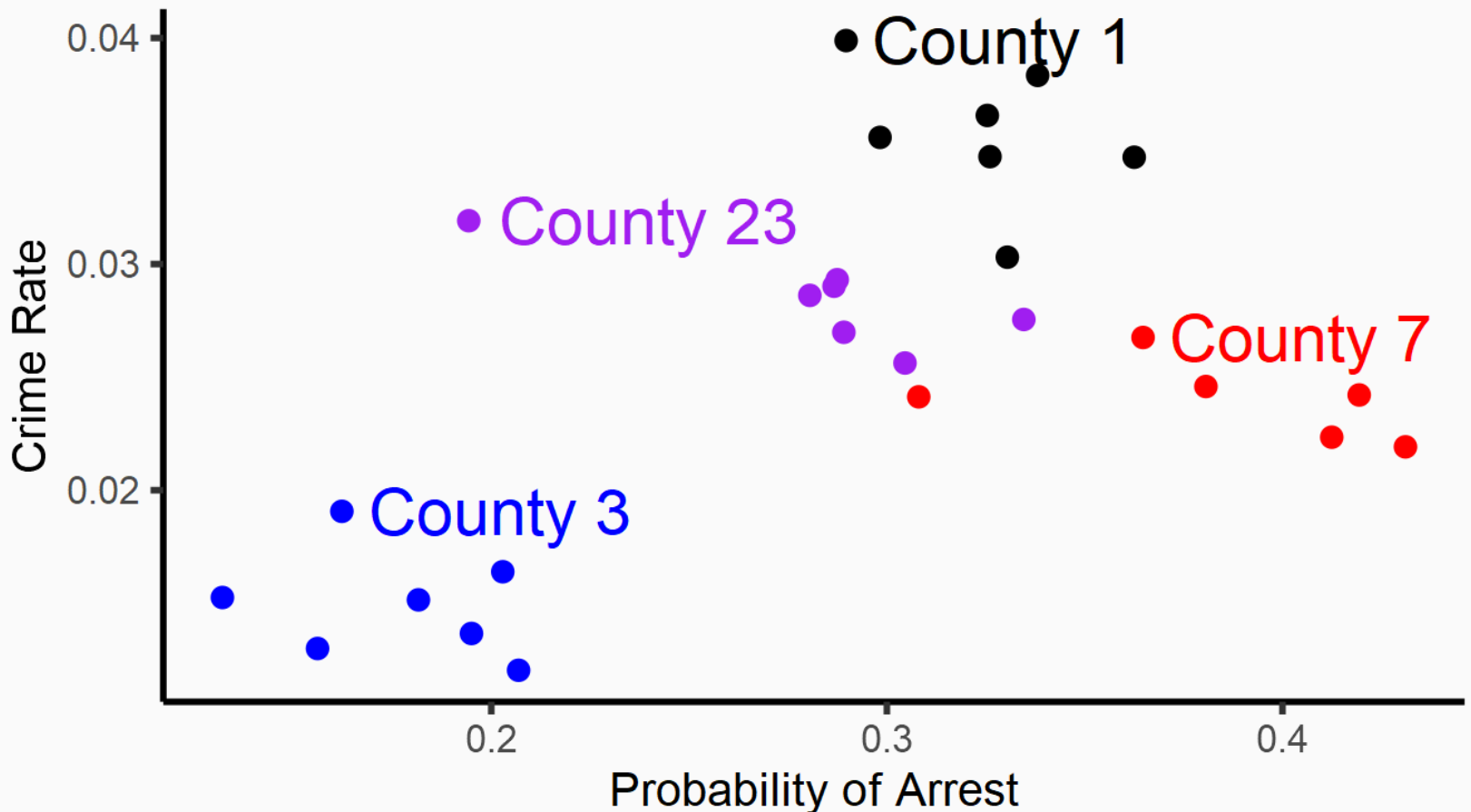
One outlier eliminated in County 7.

Between and Within

- If we look at the overall variation, just pretending this is all together, we get this

One outlier eliminated in County 7.

Between and Within

- BETWEEN variation is what we get if we look at the relationship between the *means of each county*

One outlier eliminated in County 7.
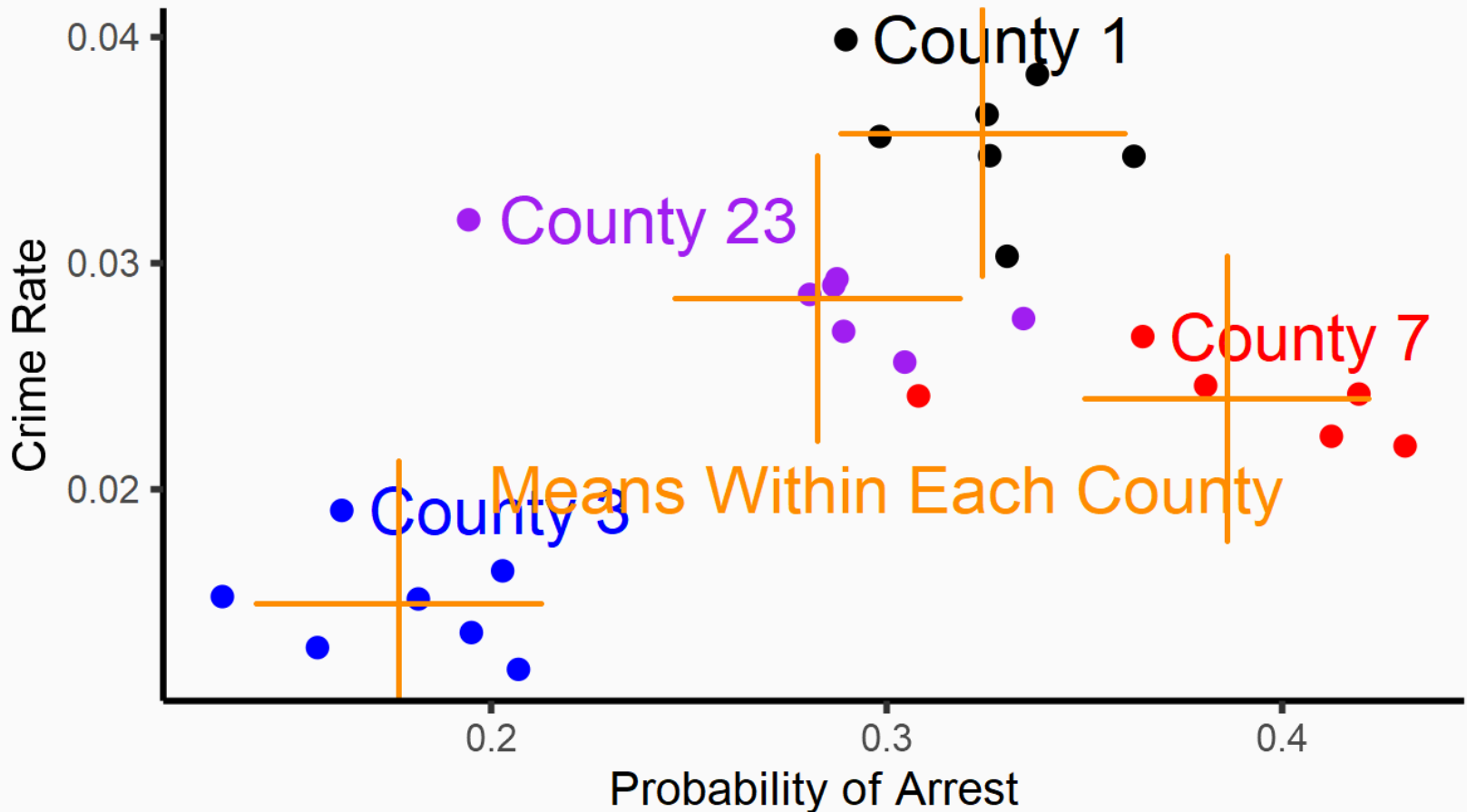
Between and Within

- And I mean it! Only look at those means! Individual year-to-year variation within county doesn't matter.

One outlier eliminated in County 7.
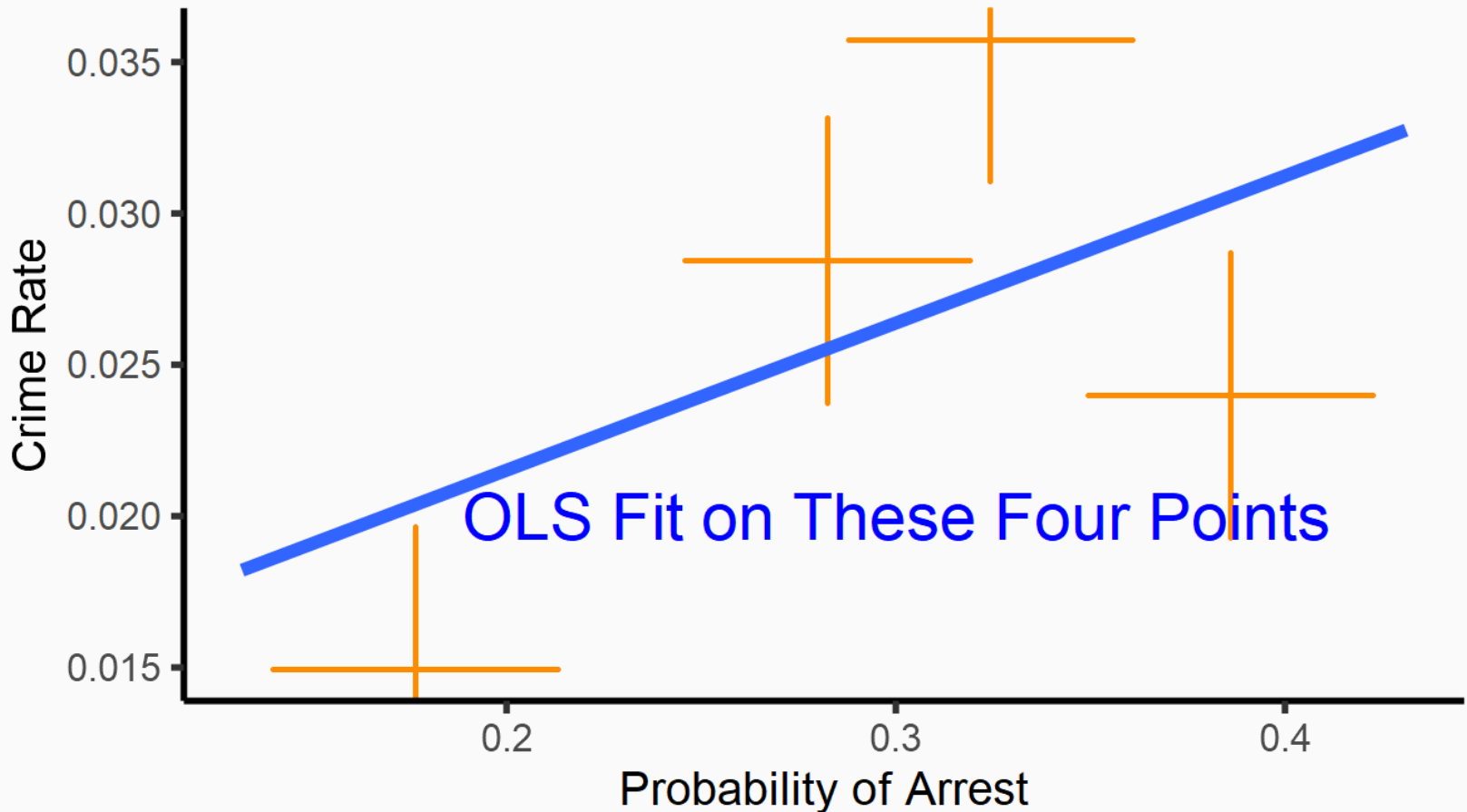
Between and Within

- Within variation treats the orange crosses as individualized axes for the variation *within* county from year-to-year only!
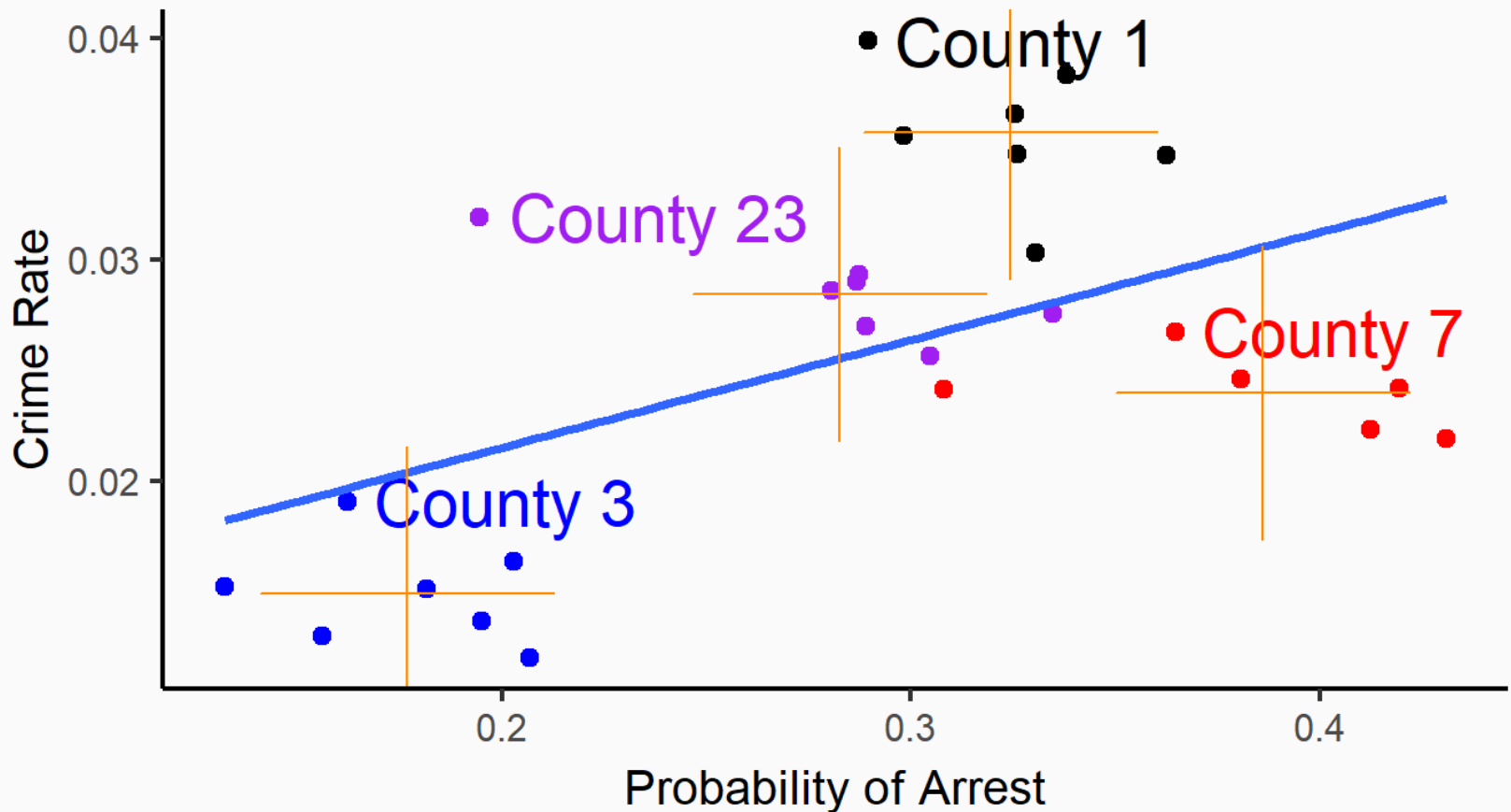
Between and Within

- We can clearly see that *between counties* there's a strong positive relationship

- But if you look *within* counties, the relationship seems weakly negative

- Which would make sense - if you think your chances of getting arrested are high, that should be a deterrent to crime

- But what are we actually doing here? Let's think about the data-generating process!

- What goes into the probability of arrest and the crime rate? Lots of stuff!

Between and Within

- For each of these variables we can ask if they vary *between counties* and/or *within counties*

- Lots of stuff like geography, landmarks, the quality of the schools only varies *between counties*, but not that much over the years

- The number of police on the streets, the poverty rate, and the probability of arrest, vary both *between* and *within* counties from year to year

- So county fixed effects sucks up all the variation for things that do not vary within counties

- That means even if we cannot measure some variable, if it only varies between counties, we can control for it!

Between and Within

- Now the task of identifying ProbArrest $\rightarrow$ CrimeRate becomes much simpler!

- If we control for County, that cuts out tons of omitted variables

- Conveniently, we can control for County just like it was any other variable!

- And when we do, we automatically *control for all variables that only have between variation*, whatever they are, even if we can't measure them directly or didn't think about them

- *All that's left is the within variation*

Implementation

Removing Between Variation

- Okay so that's the concept

- Remove all the between variation so that all that's left is within variation

- And in the process control for any variables that are made up only of between variation

- How can we actually do this? And what's really going on?

- Let's first talk about the regression model itself that this implies

- Then let's actually do the thing. There are two main ways: *de-meaning* and *binary variables* (they give the same result, for balanced panels anyway)

Estimation vs. Design

- To be clear, this is *exactly 0% different* from what we've done before in terms of controlling for stuff

- And in fact we're about to do the *exact same thing we did before* by just adding a categorical control variable for `county` or whatever

- (and in fact the "within" thing holds with other categorical controls - a categorical control for education isolates variation "within education levels")

- The difference is the *reason we're doing it*. It's fixed effects because *a categorical control for individual controls for a lot of stuff*, and we think closes a *lot* of back doors for us, not just one, and not just ones we can measure

The Model

The it subscript says this variable varies over individual i and time t

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \varepsilon_{it}$$

- What if there are individual-level components in the error term causing omitted variable bias?
- X_{it} is related to LocalStuff which is not in the model and thus in the error term!
- Regular ol' omitted variable bias. If we don't adjust for the individual effect, we get a biased $\hat{\beta}_1$
- (this bias is called "pooling bias" although it's really just a form of omitted variable bias)
- We really have this then:

$$Y_{it} = \beta_0 + \beta_1 X_{it} + (\alpha_i + \varepsilon_{it})$$

De-meaning

- Let's do de-meaning first, since it's most closely and obviously related to the "removing between variation" explanation we've been going for
- The process here is simple!

1. For each variable X_{it}, Y_{it}, etc., get the mean value of that variable for each individual $\bar{X}_i, \bar{Y}_i$
2. Subtract out that mean to get residuals $(X_{it} - \bar{X}_i), (Y_{it} - \bar{Y}_i)$
3. Work with those residuals

- That's it!

How does this work?

- That α_i term gets absorbed
- The residuals are, by construction, no longer related to the α_i, so it no longer goes in the residuals!

$$(Y_{it} - \bar{Y}_i) = \beta_0 + \beta_1(X_{it} - \bar{X}_i) + \varepsilon_{it}$$

Try it!

- We can use `group_by` to get means-within-groups and subtract them out

```r
# library(tidyverse)
data(crime4, package = 'wooldridge')
crime4 ← crime4 %>%
  # Filter to the data points from our graph
  filter(county %in% c(1,3,7, 23),
         prbarr < .5) %>%
  group_by(county) %>%
  mutate(demeaned_crime = crmrte - mean(crmrte,na.rm=TRUE),
         demeaned_prob = prbarr - mean(prbarr,na.rm=TRUE))
```

- *Note: I'm subsetting the data to just a few counties for pedagogical reasons.*

And Regress!

```
pooled_ols ← feols(crmrte ~ prbarr, data = crime4)
de_mean ← feols(demeaned_crime ~ demeaned_prob, data = crime4)
kable(etable(pooled_ols, de_mean,
  dict=c('demeaned_prob'="prbarr","demeaned_crime"="crime")),
  format="markdown")
```

	pooled_ols	de_mean
Dependent Var.:	crmrte	crime
Constant	0.0118* (0.0050)	1.41e-18 (0.0004)
prbarr	0.0486** (0.0167)	-0.0305* (0.0117)
_	_	_
S.E. type	IID	IID
Observations	27	27
R2	0.25308	0.21445
Adj. R2	0.22321	0.18303

Interpreting a Within Relationship

- How can we interpret that slope of -0.03?

- This is all *within variation* so our interpretation must be *within-county*

- So, "comparing a county in year A where its arrest probability is 1 (100 percentage points) higher than it is in year B, we expect the number of crimes per person to drop by .03"

- Or if we think we've causally identified it (and use a more realistic scale), "raising the arrest probability by 1 percentage point in a county reduces the number of crimes per person in that county by .0003".

- We're basically "controlling for county" (and will do that explicitly in a moment)

- So your interpretation should think of it in that way

  - *holding county constant → comparing two observations with the same value of county at different points in time*

Least Squares Dummy Variables

- De-meaning the data isn't the only way to do it!

- You can also use the least squares dummy variable (LSDV) method

- We just treat "individual" like the categorical variable it is and add it as a control!

- Again, the regression approach is exactly the same as with any categorical control, but the *research design* reason for doing it is different

Let's do it!

```
lsdv ← feols(crmrte ~ prbarr + factor(county), data = crime4)
kable(etable(pooled_ols, de_mean, lsdv,
  keep = c('prbarr', 'demeaned_prob'),
  dict=c('demeaned_prob'="prbarr","demeaned_crime"="crmrte")), format="markdown")
```

	pooled_ols	de_mean	lsdv
Dependent Var.:	crmrte	crmrte	crmrte
prbarr	0.0486** (0.0167)	-0.0305* (0.0117)	-0.0305* (0.0124)
_	_	_	_
S.E. type	IID	IID	IID
Observations	27	27	27
R2	0.25308	0.21445	0.94114
Adj. R2	0.22321	0.18303	0.93044

- The result is the same (as it should be) except for the R^2 -- any guesses why they're so different?

Comparing LSDV and de-meaning

- Because de-meaning takes out the part explained by the fixed effects (α_i) *before* running the regression, while LSDV does it *in* the regression

- So the .94 is the portion of `crmrte` explained by `prbarr` *and* `county`, whereas the .21 is the "within - R^2 " - the portion of *the within variation* that's explained by `prbarr`

    - You can literally report the overall R^2 and the within R^2 side by side

- Neither is wrong (and the .94 isn't "better"), they're just measuring different things

Why LSDV?

- A benefit of the LSDV approach is that it calculates the fixed effects α_i for you
- We left those out of the table with the `coefs` argument of `export_summs` (we rarely want them) but here they are:

```
lsdv
```

```
## OLS estimation, Dep. Var.: crmrte
## Observations: 27
## Standard-errors: IID
##                   Estimate Std. Error   t value    Pr(>|t|)
## (Intercept)       0.045631   0.004116  11.08640 1.7906e-10 ***
## prbarr           -0.030491   0.012442  -2.45068 2.2674e-02 *
## factor(county)3  -0.025308   0.002165 -11.68996 6.5614e-11 ***
## factor(county)7  -0.009870   0.001418  -6.96313 5.4542e-07 ***
## factor(county)23 -0.008587   0.001258  -6.82651 7.3887e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 0.001933   Adj. R2: 0.930441
```

- Interpretation is exactly the same as with a categorical variable - we have an omitted county, and these show the difference relative to that omitted county

Why LSDV?

- LSDV makes clear what's happening by creating a separate intercept for each county
- Graphically, de-meaning moves all the points together, while LSDV moves the line up and down to meet the points
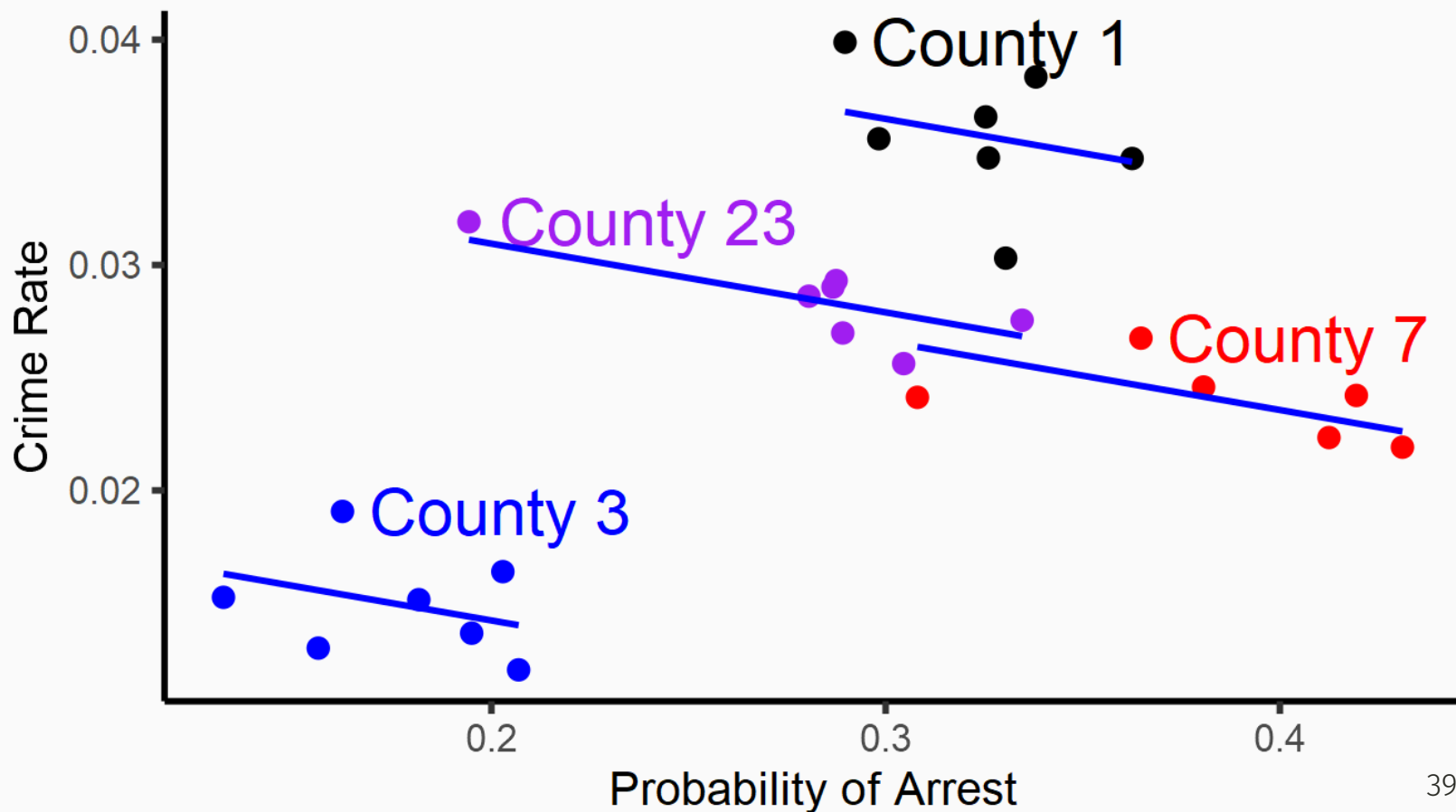
Why recover FEs?

- We can recover the fixed effects from the LSDV regression and sometimes interpret them

- This is useful in a variety of research fields:

  - Industrial Organization: Recover unobservable firm-specific characteristics, which can be used to comprae relative welfare of different policy simulations, etc.
  - Criminology: Recover judge-specific effects, which can be used to measure the leniency of judges at trial
  - Labor economics: Decompose wages into firm-specific, worker-specific, and firm-worker specific components to evaluate monopsony power, importance of firm-specific human capital, etc.

- These are really cool applications, but they usually require administrative data that we can't get our hands on

- Still I want you to know about them

Why Not LSDV?

- LSDV is computationally expensive

- If there are a lot of individuals, or big data, or if you have many sets of fixed effects (yes you can do more than just individual - we'll get to that next time!), it can be very slow

- Most professionally made fixed-effects commands use de-meaning, but then adjust the standard errors properly

- (They also leave the fixed effects coefficients off the regression table by default)

Going Professional

- Applied researchers rarely do either of these, and rather will use a command specifically designed for fixed effects

- Like good ol' `feols()`! (what did you think the "fe" part stood for?)

- Note there are also functions in **fixest** that do fixed effects in non-linear models like logit, probit, or poisson regression (`feglm()` and `fepois()`)

- Plus, it clusters the standard errors by the first fixed effect by default, which we usually want!

Going Professional

```r
library(fixest)
pro ← feols(crmrte ~ prbarr | county, data = crime4)
kable(etable(de_mean, pro), format="markdown",
  dict=c('demeaned_prob'="prbarr","demeaned_crime"="crmrte"))
```

	de_mean	pro
Dependent Var.:	demeaned_crime	crmrte
Constant	1.41e-18 (0.0004)	
demeaned_prob	-0.0305* (0.0117)	
prbarr		-0.0305* (0.0064)
Fixed-Effects:	----------------	----------------
county	No	Yes
	_	_
S.E. type	IID	by: county
Observations	27	27
R2	0.21445	0.94114

Is this causal?

- After controlling for everything within county, is this causal?

Is this causal?

- After controlling for everything within county, is this causal?

- Probably not! Why not?

Is this causal?

- After controlling for everything within county, is this causal?

- Probably not! Why not?

Within variation persists

1. Within-county time variation: maybe crime and arrest probability moved together (e.g. a crime wave)

2. Reverse causality: maybe officers respond to crime rates by changing their arrest effort

3. Omitted variable bias: maybe poverty or population density is driving both crime and arrest probability and changes over time

Limits to Fixed Effects

- Okay! At this point we have the concept behind fixed effects, can execute them, and know what they're good for

- What aren't they good for?

1. They don't control for anything that has within variation
2. They control away *everything* that's between-only, so we can't see the effect of anything that's between-only ("effect of geography on crime rate?" Nope!)
3. Anything with only a *little* within variation will have most of its variation washed out too ("effect of population density on crime rate?" probably not)
4. The estimate pays the most attention to individuals with *lots of variation in treatment*

- 2 and 3 can be addressed by using "random effects" (see the The Effect chapter on Fixed Effects for more)