Panel data models

Tutorial 2

Stanislav Avdeev

Goal for today's tutorial

1. Understand the panel structure of the data
2. Explore differences between pooled OLS, fixed and random effects estimators
3. Interpret the variation in the data
4. Make proper inferences using panel data models

Panel data

- Panel data is when you observe the same individual over multiple time periods
  - "individual" could be a person, a company, a state, 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
- We assume that we observe each individual the same number of times, i.e. a *balanced* panel (so we have $N \times T$ observations)
  - you can use these estimators with unbalanced panels too, it just gets a little more complex

Panel data

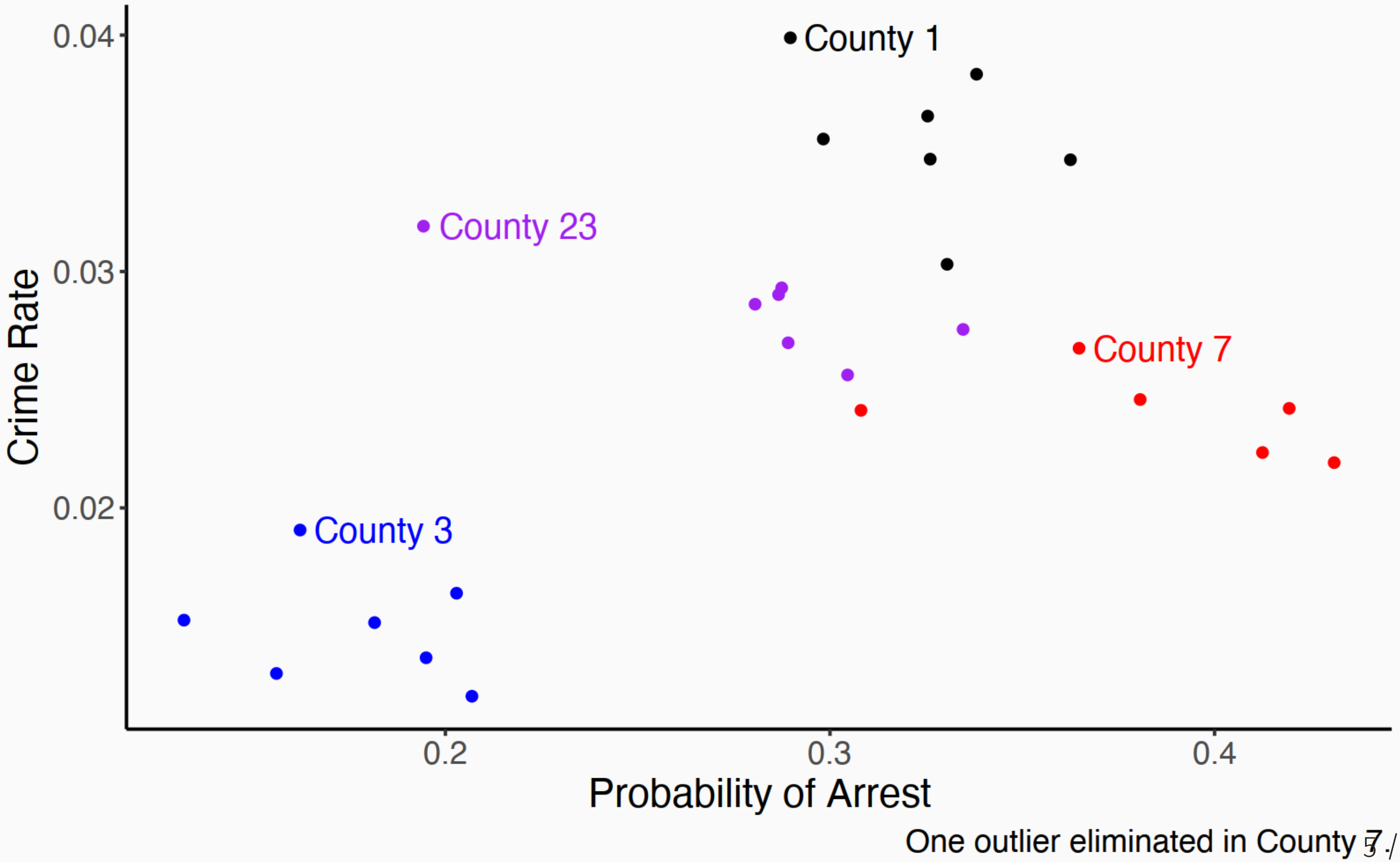
- Let's use a dataset from `wooldridge` package on crime data
  - you can use a lot of datasets from different packages, such as `wooldridge` which contains datasets from "Introductory Econometrics: A Modern Approach" by Wooldridge J.M.
- Here's what a panel data set looks like - a variable for individual (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
3	81	0.0163921	0.202899
3	82	0.0190651	0.162218
3	83	0.0151492	0.181586

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

Between and within variation

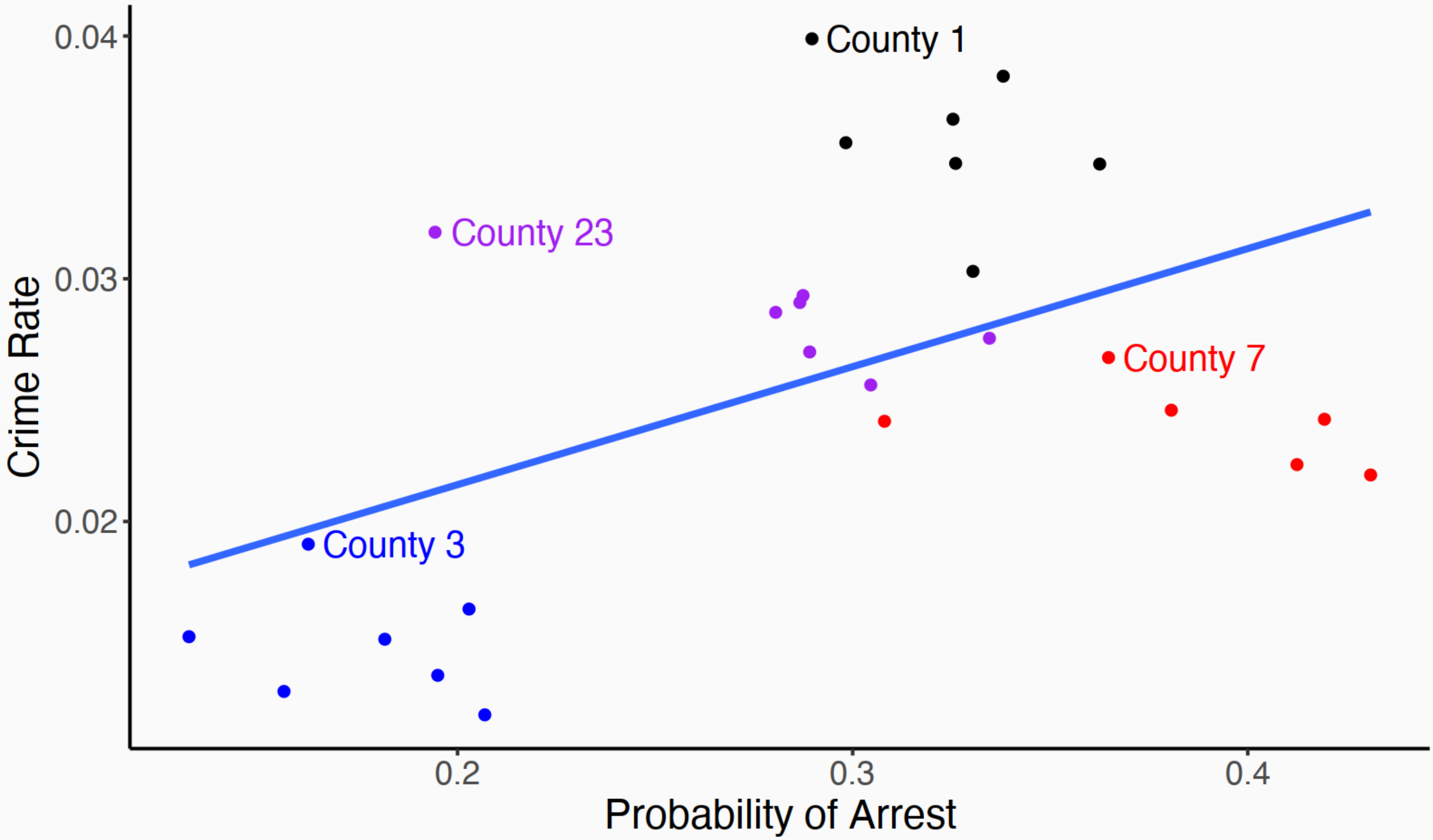
Let's pick a few counties and graph this out



One outlier eliminated in County 7.

Between variation

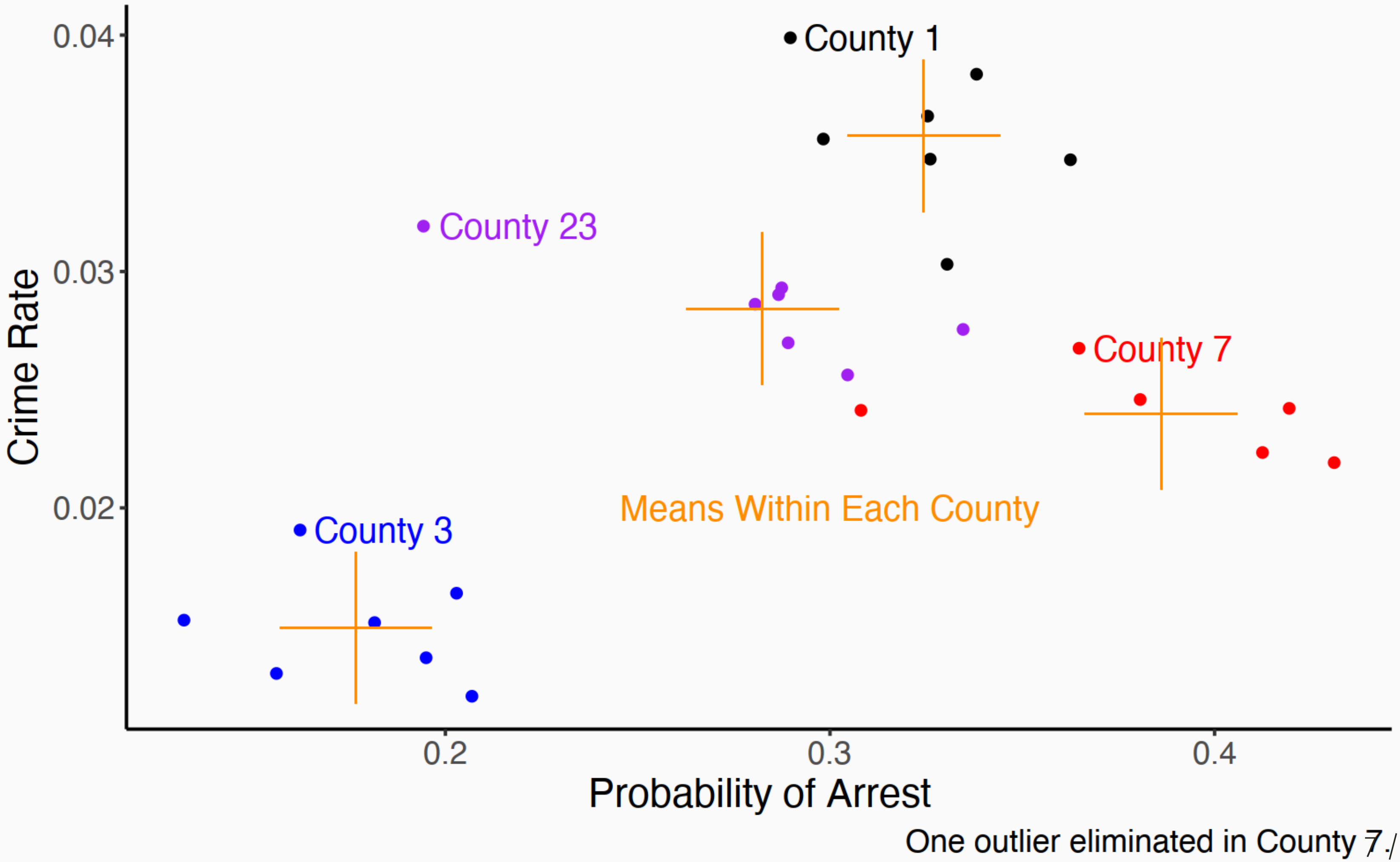
If we look at the **between** variation by using the **pooled** OLS estimator, we get this



One outlier eliminated in County 7.

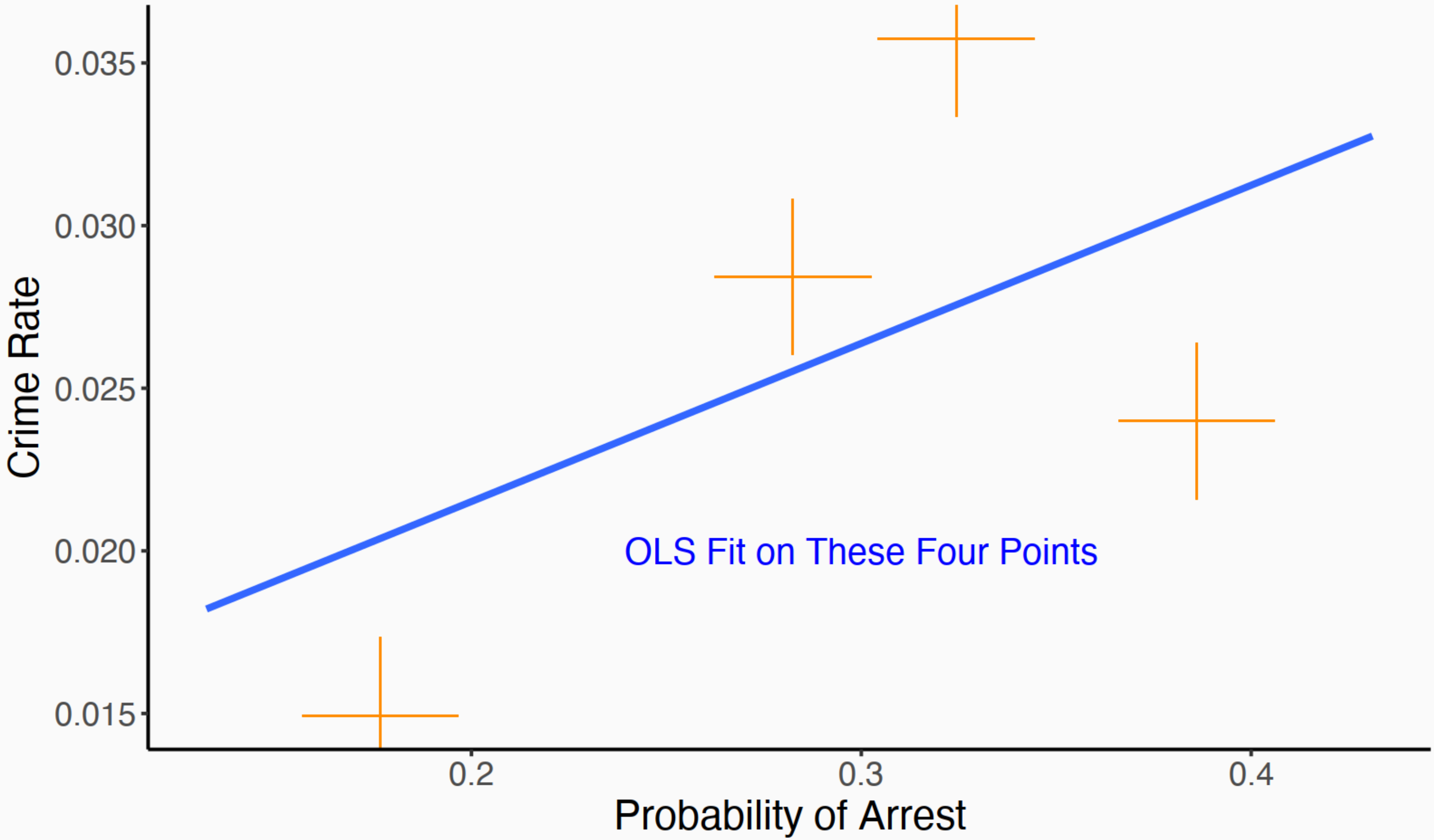
Between variation

Between variation looks at the relationship **between the means of each county**



Between variation

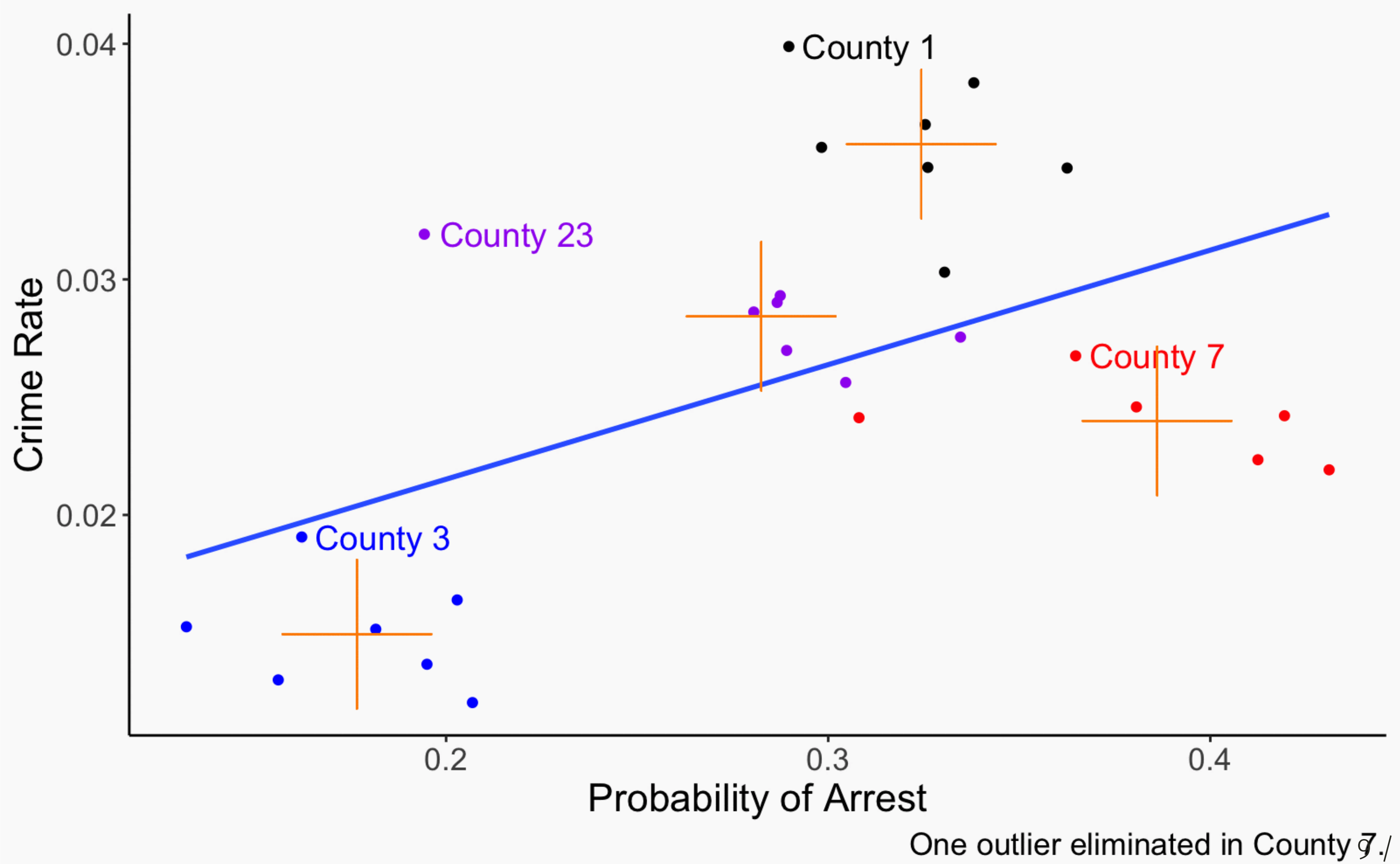
The individual year-to-year variation within county doesn't matter



One outlier eliminated in County 7.

Within variation

Within variation goes the other way: it looks at variation **within county from year-to-year**



Between and within variation

- We can clearly see that **between counties** there's a strong **positive** relationship
- But if you look **within** a given county, the relationship isn't that strong, and actually seems to be **negative**
    - which would make sense - if you think your chances of getting arrested are high, that should be a deterrent to crime
    - we are ignoring all differences between counties and looking only at differences within counties
- **Fixed effects** is sometimes also referred to as the **within** estimator

Panel data model

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

$$Y_{it} = \alpha + X'_{it}\beta + U_{it}$$

- What if there are individual-level components in the error term causing omitted variable bias?
    - X_{it} might be related to the variable which is not in the model and thus in the error term
- So we really have this then:

$$Y_{it} = \alpha + X'_{it}\beta + \eta_i + U_{it}$$

- If you think X_{it} η_i are **not** correlated (based on theory, previous research), you can use both FE and RE estimators
- If you think X_{it} η_i are correlated (based on theory, previous research), use FE estimator

Panel data model: simulation

- Let's simulate a panel dataset

```r
set.seed(7)
df <- tibble(id = sort(rep(1:600, 10)),
             time = rep(1:10, 600),
             x1 = rnorm(6000),
             # fixed variable within individual, e.g. gender
             x2 = ifelse(id %% 2 == 0, 1, 0),
             y = id + time + 2*x1 + 50*x2 + rnorm(6000))
```

id	time	x1	x2	y
1	1	2.2872472	0	6.4522242
1	2	-1.1967717	0	-0.3436617
1	3	-0.6942925	0	2.3772643
2	1	0.3569862	1	53.0468280
2	2	2.7167518	1	59.0964204
2	3	2.2814519	1	60.9361494

Panel data model: simulation

```r
# The true effect is 2
library(plm) # package to estimate FE and RE models
pooled <- plm(y ~ x1 + x2, model = "pooling", df) # or lm(y ~ x1 + x2, df)
fixed  <- plm(y ~ x1 + x2, model = "within", index = c("id", "time"), df)
random <- plm(y ~ x1 + x2, model = "random", index = c("id", "time"), df)
```

	Model 1	**Model 2**	**Model 3**
x1	1.278	1.900***	1.900***
	(2.235)	(0.043)	(0.043)
x2	51.049***		51.033***
	(4.474)		(14.176)
Num.Obs.	6000	6000	6000

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

- Pooled OLS estimates are off as it doesn't take into account the panel structure of data
- FE and RE estimators provide unbiased estimates
- FE estimator doesn't produce estimates of X_2 as it's not varying **within** individual

Panel data model: simulation

- Let's introduce the correlation between individual effects and individual characteristics

$$\mathrm{cov}(X_i, \eta_i) \neq 0$$

```r
set.seed(7)
df <- tibble(id = sort(rep(1:600, 10)),
             time = rep(1:10, 600),
             # add a correlated individual effect in x1
             x1 = rnorm(6000) + 0.1*id,
             x2 = ifelse(id %% 2 == 0, 1, 0),
             y = id + time + 2*x1 + 50*x2 + rnorm(6000))
```

Panel data model: simulation

```r
# The true effect is 2
pooled_corr ← plm(y ~ x1 + x2, model = "pooling", df)
fixed_corr  ← plm(y ~ x1 + x2, model = "within", index = c("id", "time"), df)
random_corr ← plm(y ~ x1 + x2, model = "random", index = c("id", "time"), df)
```

	Model 1	**Model 2**	**Model 3**
x1	11.969***	1.900***	11.720***
	(0.008)	(0.043)	(0.023)
x2	49.768***		49.799***
	(0.272)		(0.791)
Num.Obs.	6000	6000	6000

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

- Pooled OLS and RE estimates are off since $\mathbf{cov}(X_i, \eta_i) \neq 0$
- FE still provide unbiased estimates since η_i are eliminated
- How does FE estimator eliminate η_i?

Estimation: de-meaning approach

- To estimate FE model, we need to remove between variation so that all that's left is within variation
- There are two main ways
    - **de-meaning**
    - **binary variables**
- They give the same result (for balanced panels anyway)
- Let's do de-meaning first, since it's most closely and obviously related to the "removing between variation" explanation
    - for each variable X_{it}, Y_{it}, etc., get the mean value of that variable for each individual $\bar{X}_i, \bar{Y}_i$
    - subtract out that mean to get residuals $(X_{it} - \bar{X}_i), (Y_{it} - \bar{Y}_i)$
    - work with those residuals
- α and η_u terms get absorbed
- The residuals are, by construction, no longer related to the η_i

$$Y_{it} - \bar{Y}_i = (X_{it} - \bar{X}_i)'\beta + (U_{it} - \bar{U}_i)$$

Estimation: LSDV approach

- De-meaning the data is not the only way to do it
    - and sometimes it can make the standard errors wonky, since they don't recognize that you've estimated those means
- You can also use the **least squares dummy variable** - LSDV (another word for "binary variable") method
    - we just treat "individual" like the categorical variable it is and add it as a control

Estimation: empirical example

- Let's get back to the crime dataset
- To demean the data, we can use `group_by` to get means-within-groups and subtract them out

```
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(mean_crime = mean(crmrte),
         mean_prob = mean(prbarr)) %>%
  mutate(demeaned_crime = crmrte - mean_crime,
         demeaned_prob = prbarr - mean_prob)
```

Estimation: empirical example

- To use least squares dummy variable, we only need to add FE as categorical variables

```
pooling ← lm(crmrte ~ prbarr, data = crime4)
lsdv    ← lm(crmrte ~ prbarr + factor(county), data = crime4)
de_mean ← lm(demeaned_crime ~ demeaned_prob, data = crime4)
```

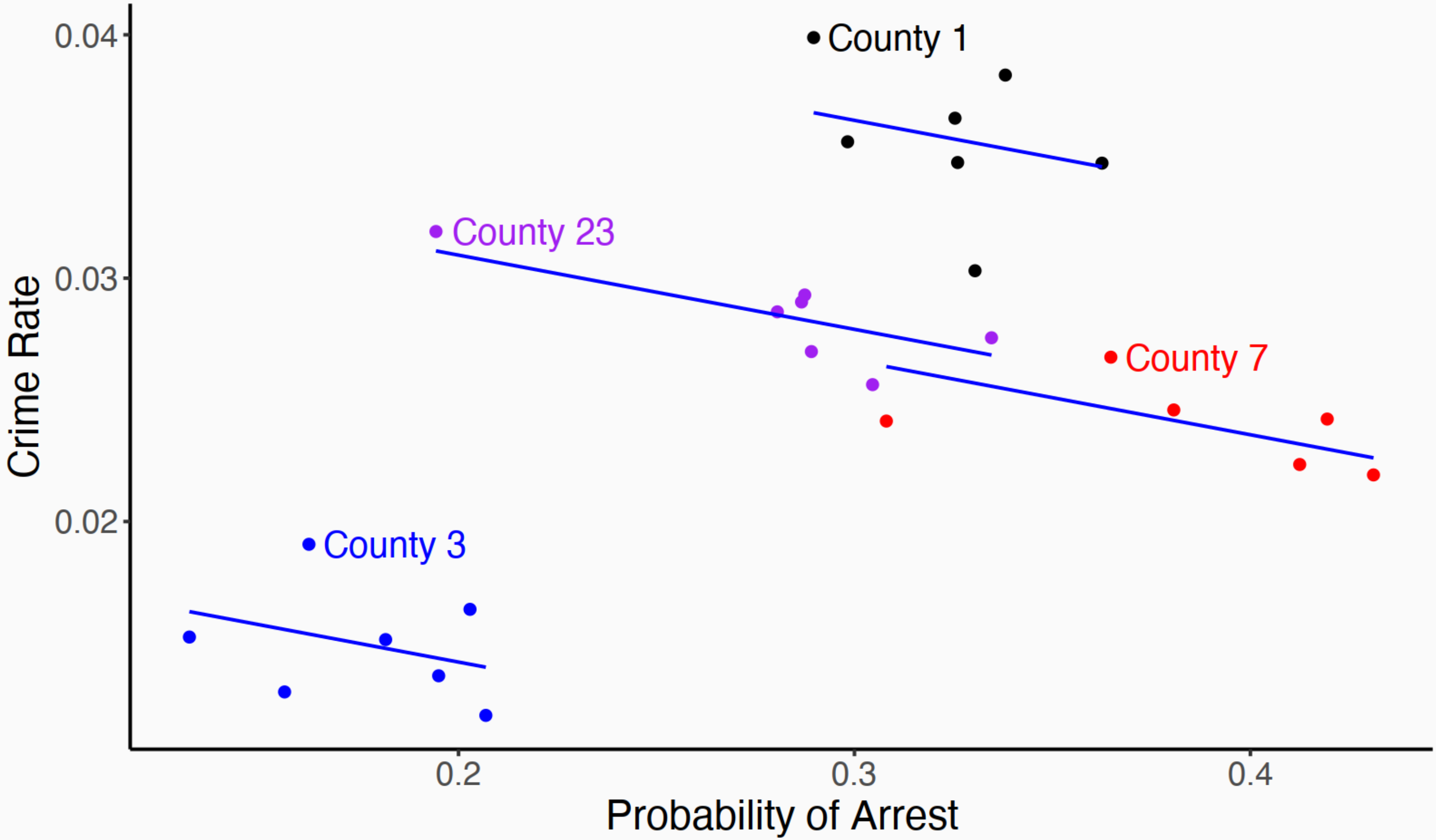
	Model 1	Model 2	Model 3
prbarr	0.049**	-0.030*	
	(0.017)	(0.012)	
demeaned_prob			-0.030*
			(0.012)
Num.Obs.	27	27	27

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

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**
  - if we think we've causally identified it, "raising the arrest probability by 1 percentage point in a county reduces the number of crimes per person in that county by 0.0003"
  - we're basically **controlling for county**, i.e. comparing a county to itself at a different point in time
- A benefit of the LSDV approach is that it calculates the fixed effects α_i for you
  - interpretation is exactly the same as with a categorical variable - we have an omitted category (one county), and these show the difference relative to that omitted county
  - this also makes clear another element of what's happening. Just like with a categorical variable, the line is moving *up and down* to meet the counties
  - graphically, de-meaning moves all the points together in the middle to draw a line, while LSDV moves the line up and down to meet the points

Interpreting a within relationship



One outlier eliminated in County 7.

Panel data: estimation

- Applied researchers rarely do either of these, and rather will use a command specifically designed for the FE estimator
  - `feols` in **fixest**
  - `felm` in **lfe**
  - `plm` in **plm**
  - `lm_robust` in **estimatr**
- `feols` in **fixest** seems to be a better choice
  - it does all sorts of other neat stuff like fixed effects in nonlinear models like logit, regression tables, joint-test functions, and so on
  - it's very fast, and can be easily adjusted to do fixed effects with other regression methods like logit, or combined with instrumental variables
  - it clusters the standard errors by the first fixed effect by default

Panel data: estimation

Let's see at the output of `feols`

```r
library(fixest)
fe_plm ← plm(crmrte ~ prbarr, model = "within", index = "county", crime4)
fe_feols ← feols(crmrte ~ prbarr | county, crime4)
```

	Model 1	**Model 2**
prbarr	-0.030*	-0.030*
	(0.012)	(0.006)
Num.Obs.	27	27
Std.Errors		by: county

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

Fixed effects: limitations

1. Fixed effects 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. If there’s not a lot of within variation, fixed effects are going to be very noisy. Make sure there’s variation to study
5. The estimate pays the most attention to individuals with *lots of variation in treatment*

- 2 and 3 can be addressed by using the RE estimator instead (although you need to be certain that $\mathbf{cov}(X_i, \eta_i = 0)$
  - How can you check that?

Fixed or random effects?

- To decide between FE or RE estimators you can run the **Hausman test** where the null hypothesis is that the preferred model is the RE estimator vs. the alternative - the FE estimator
- It basically tests whether the errors are correlated with the regressors, the null hypothesis is they are not
  - under H_0: if $\mathbf{cov}(X_i, \eta_i = 0)$ both RE and FE estimators are consistent, but the RE estimator is more efficient
  - under H_1: if $\mathbf{cov}(X_i, \eta_i \neq 0)$ only FE estimator is consistent

Fixed or random effects?

- Let's apply it to two simulated datasets with and without correlated individual effects

```
phtest(fixed, random)
```

```
##
##     Hausman Test
##
## data:  y ~ x1 + x2
## chisq = 0.0018287, df = 1, p-value = 0.9659
## alternative hypothesis: one model is inconsistent
```

```
phtest(fixed_corr, random_corr)
```

```
##
##     Hausman Test
##
## data:  y ~ x1 + x2
## chisq = 71554, df = 1, p-value < 2.2e-16
## alternative hypothesis: one model is inconsistent
```

- As expected, we should use the RE estimator in the first model, and the FE estimator in the second model

Panel data: inference

- It’s common to cluster standard errors at the level of the fixed effects, since it seems likely that errors would be correlated over time (autocorrelated errors)
  - it is a default function in `feols` in **fixest**
- It’s possible to have more than one set of fixed effects
  - but interpretation gets tricky - think through what variation in X you’re looking at (we will discuss that in the 5^{th} tutorial on difference-in-differences design)

References

Books

- Huntington-Klein, N. The Effect: An Introduction to Research Design and Causality, Chapter 16: Fixed Effects
- Cunningham, S. Causal Inference: The Mixtape, Chapter 7: Panel Data

Slides

- Huntington-Klein, N. Econometrics Course Slides, Week 6: Within Variation and Fixed Effects
- Huntington-Klein, N. Causality Inference Course Slides, Lecture 8: Fixed Effects