Static linear 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 contain information on the same individual over multiple time periods
  - "individual" could be a person, a company, a state, a country, etc. There are N individuals
  - "time period" could be a year, a month, a day, etc. There are T time periods
- 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 panel data estimators with unbalanced panels too, it just gets a little more complex

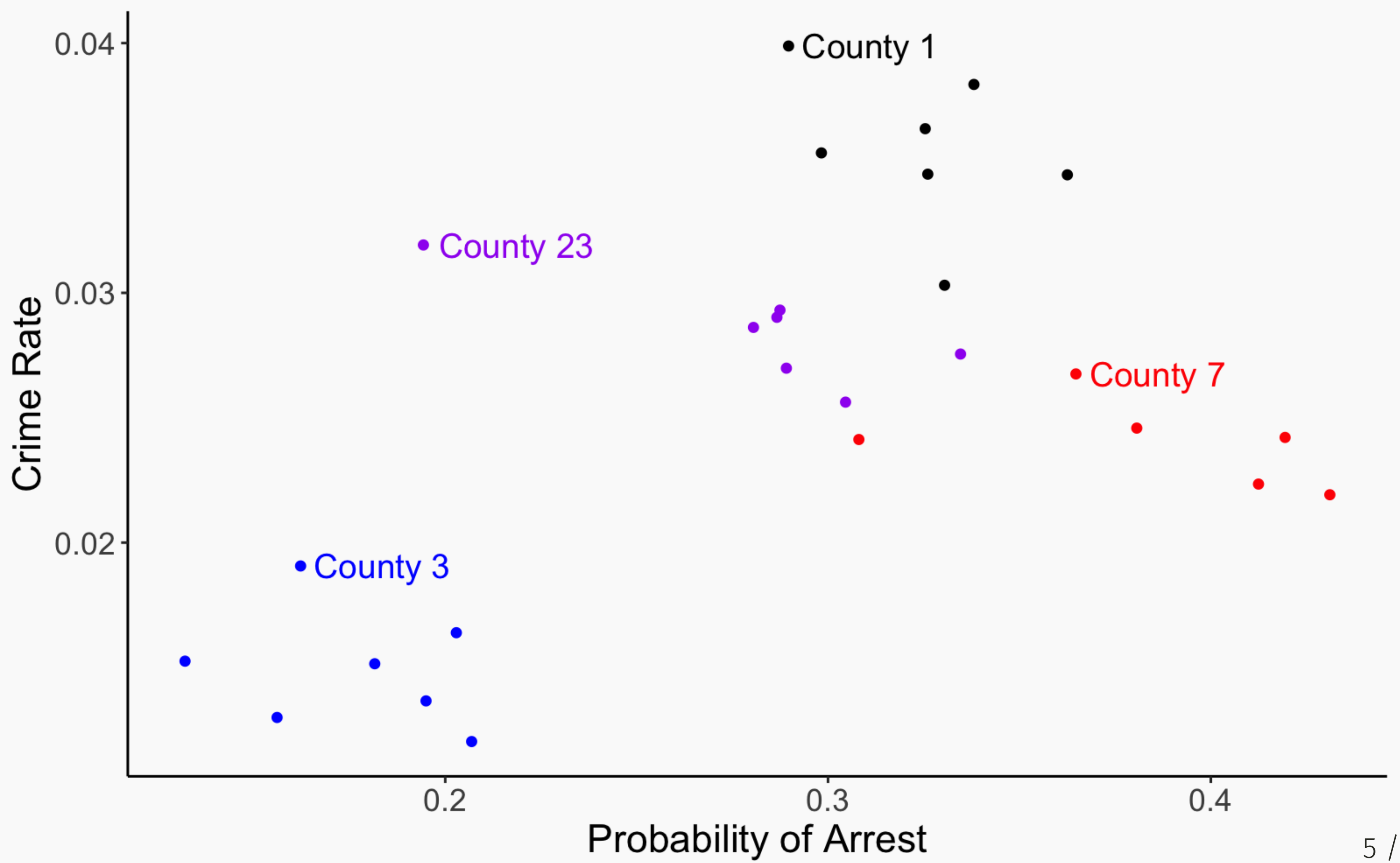
Panel data

- Let's use a data set from `wooldridge` package on crime data
  - you can use a lot of data sets from packages, such as `wooldridge` which contains data sets 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 different variables

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
3	81	0.0163921	0.202899
3	82	0.0190651	0.162218
3	83	0.0151492	0.181586
3	84	0.0136621	0.194986

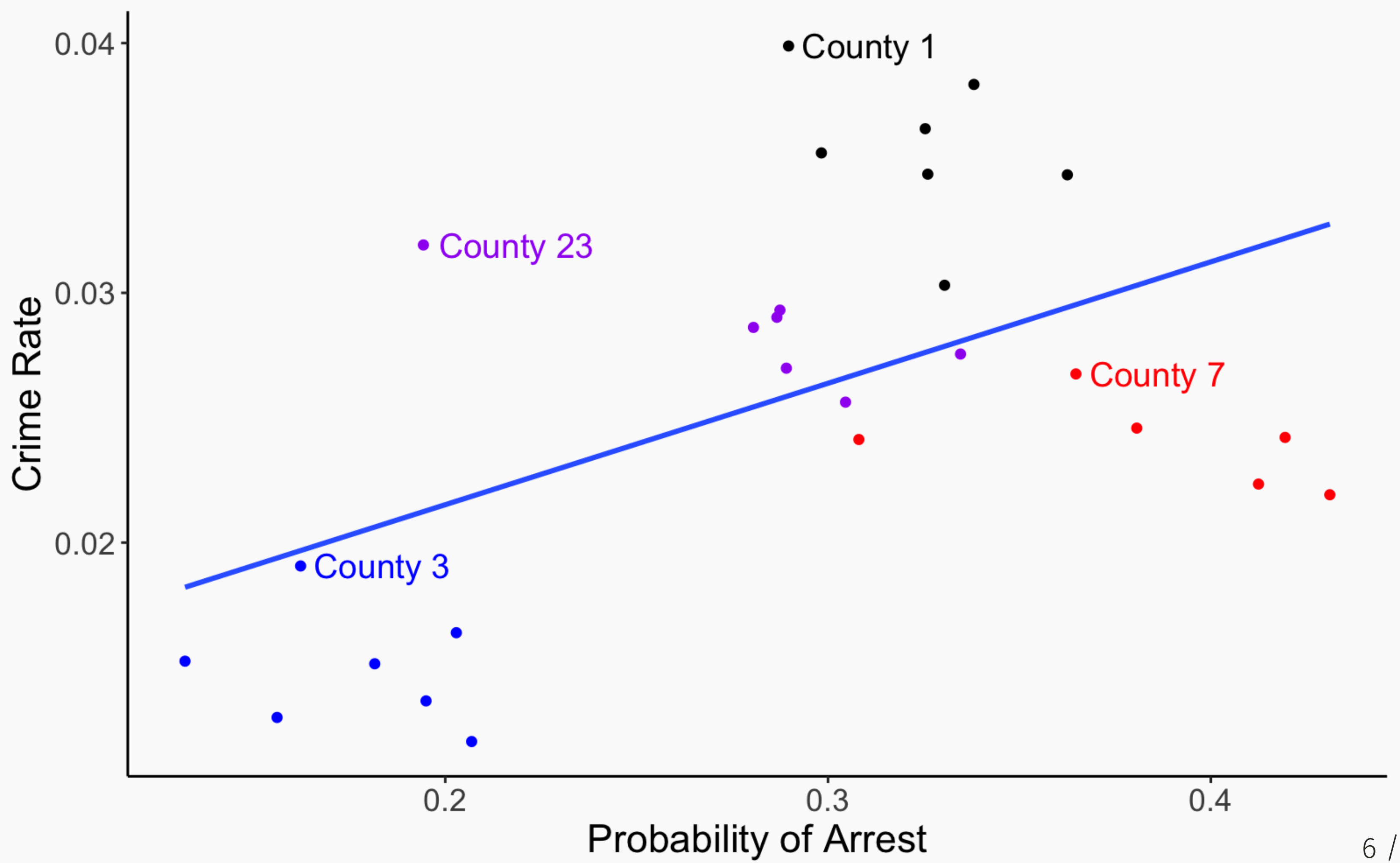
Between and within variation

Let's pick a few counties and graph this out



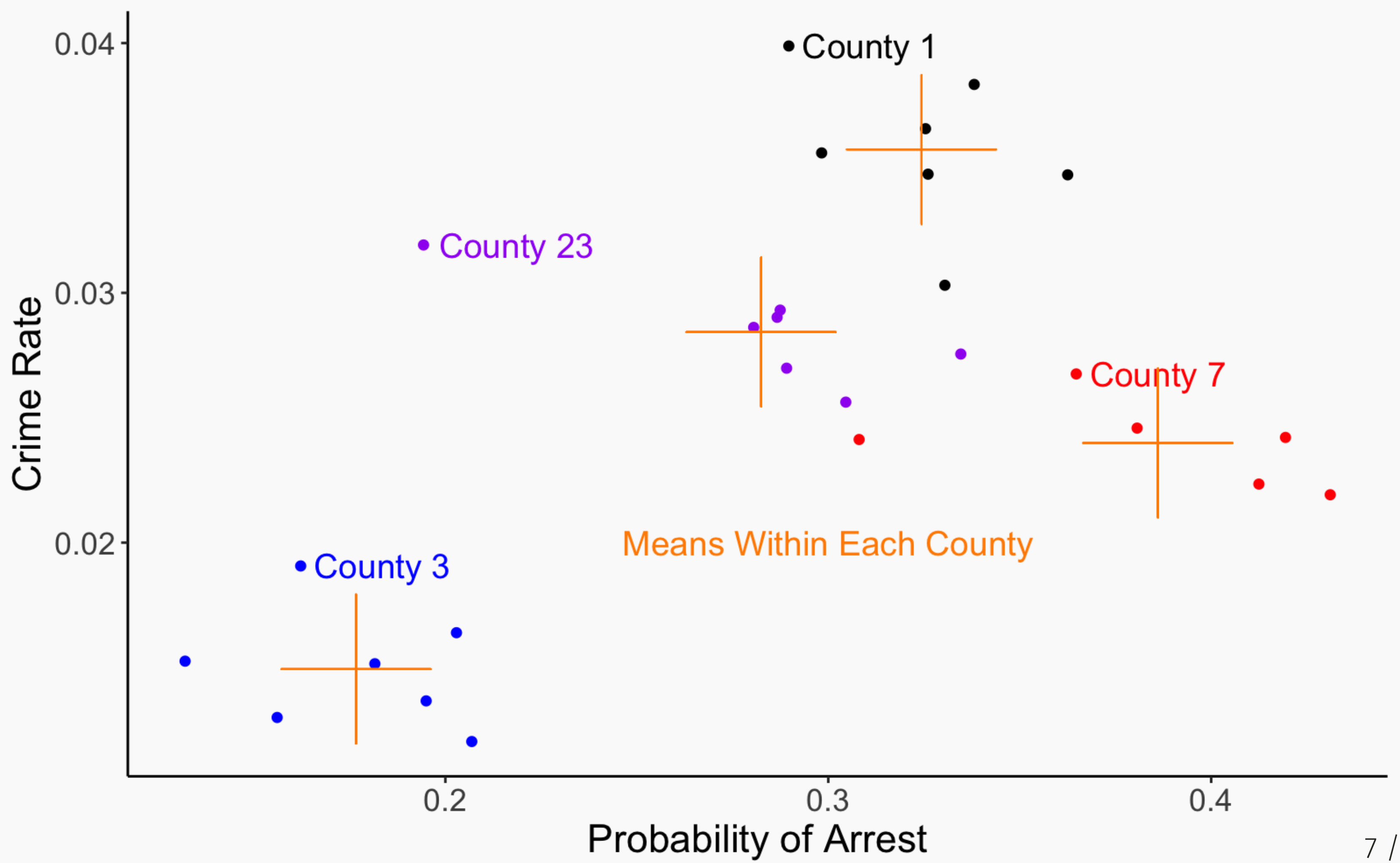
Between variation

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



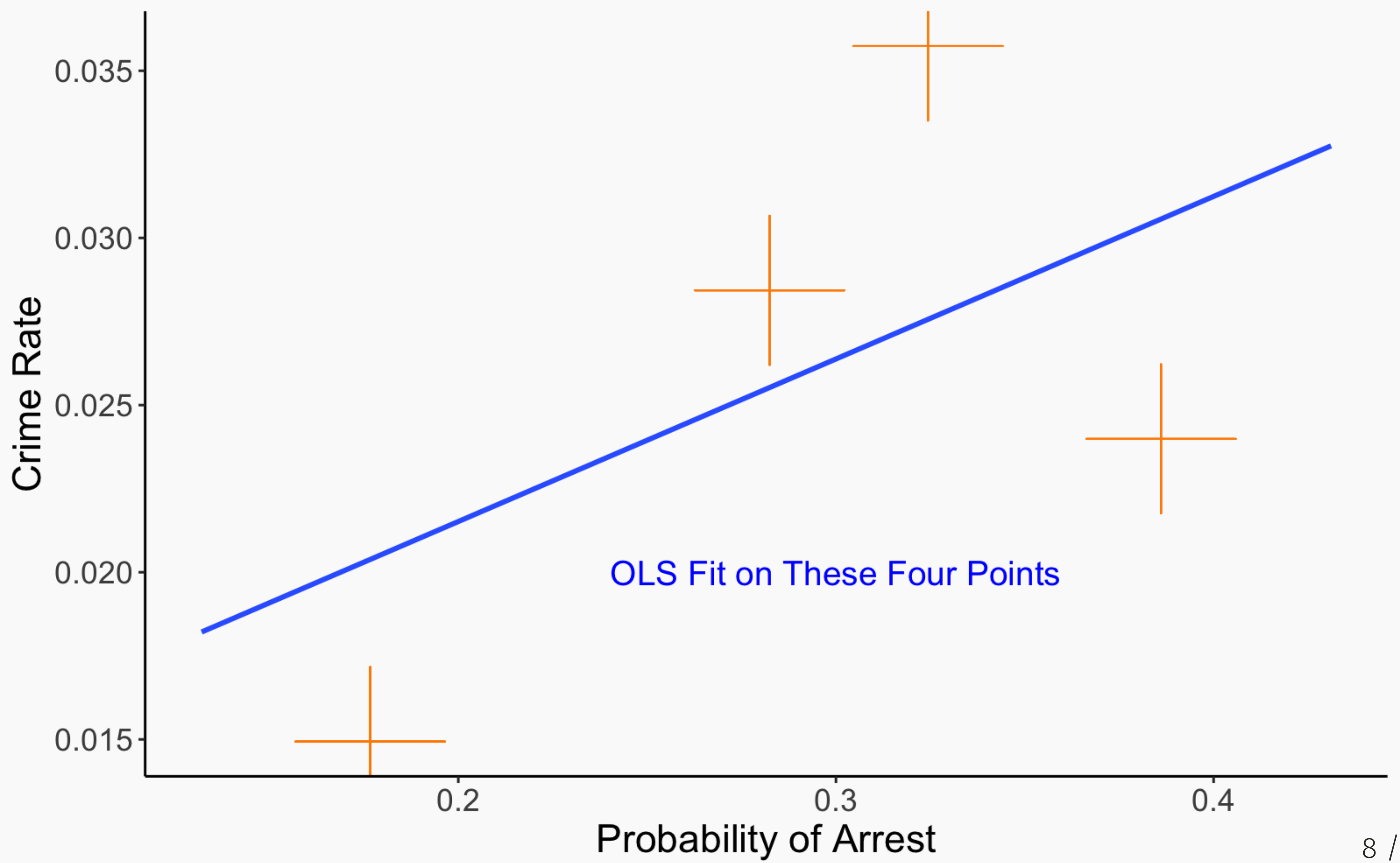
Between variation

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



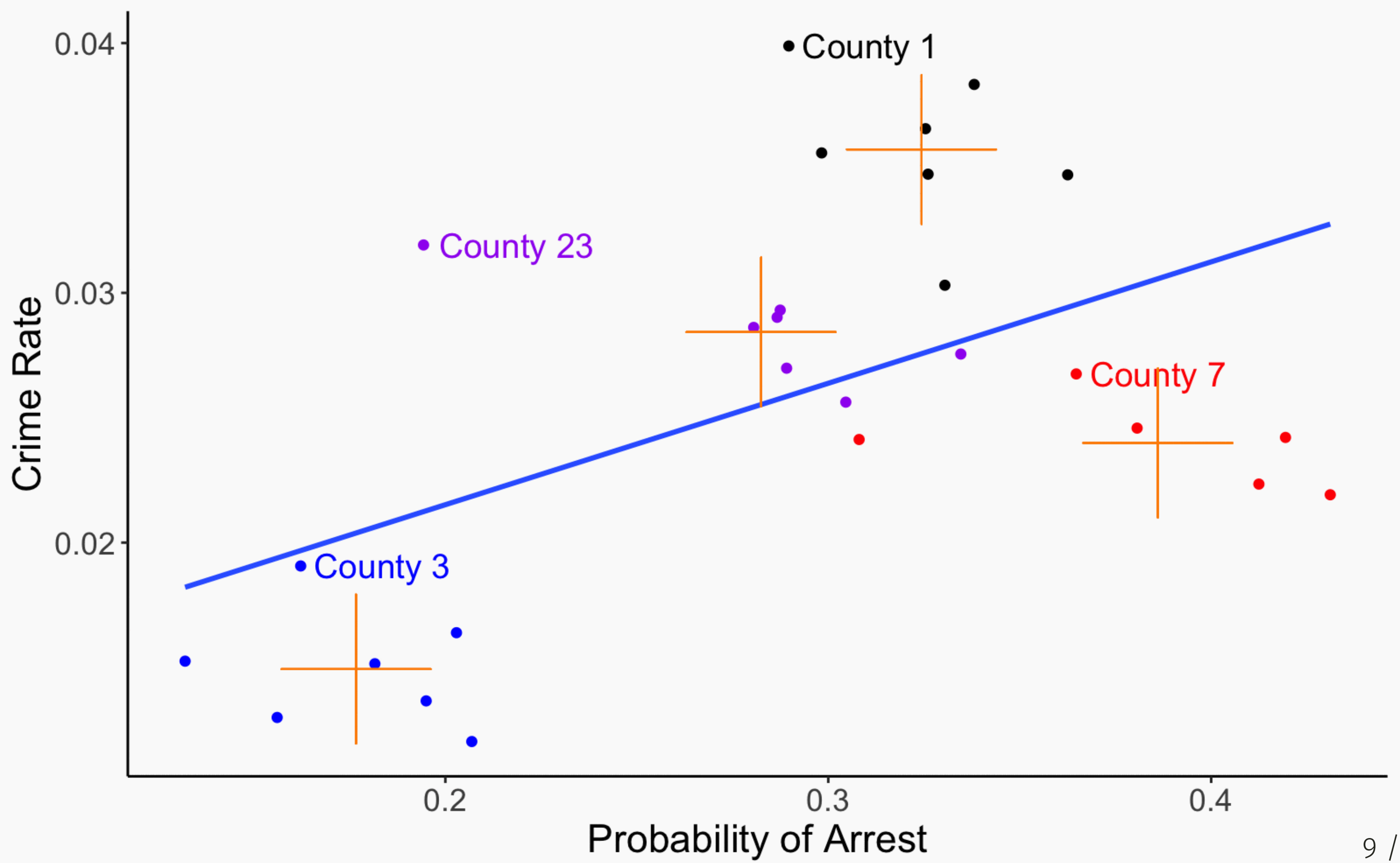
Between variation

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



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
- Thus, we have the following model

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

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

Panel data model: simulation

- Let's simulate a panel data set

```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, 0, 1),
             y = id + time + 2*x1 + 50*x2 + rnorm(6000))
```

id	time	x1	x2	y
1	1	2.2872472	1	56.452224
1	2	-1.1967717	1	49.656338
1	3	-0.6942925	1	52.377264
2	1	0.3569862	0	3.046828
2	2	2.7167518	0	9.096420
2	3	2.2814519	0	10.936149

Panel data model: simulation

```
# The true effect is 2
library(plm) # package to estimate FE and RE models (fixest is preferred for FE)
pooled ← plm(y ~ x1 + x2, model = "pooling", df) # or lm(y ~ x1 + x2, df)
random ← plm(y ~ x1 + x2, model = "random", index = c("id", "time"), df)
fixed  ← plm(y ~ x1 + x2, model = "within", 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	48.951***	48.967***	
	(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
- RE and FE 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 characteristics and individual effects

$$\text{corr}(X_{it}, \eta_i) \neq 0$$

```r
set.seed(7)
df <- tibble(id = sort(rep(1:600, 10)),
             time = rep(1:10, 600),
             x1 = rnorm(6000) + 0.05*id, # add a correlated individual effect
             x2 = ifelse(id %% 2 == 0, 1, 0),
             y = id + time + 2*x1 + 50*x2 + rnorm(6000))
```

id	time	x1	x2	y
1	1	2.3372472	0	6.5522242
1	2	-1.1467717	0	-0.2436617
1	3	-0.6442925	0	2.4772643
2	1	0.4569862	1	53.2468280
2	2	2.8167518	1	59.2964204
2	3	2.3814519	1	61.1361494

Panel data model: simulation

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

	Model 1	**Model 2**	**Model 3**
x1	21.743***	15.749***	1.900***
	(0.030)	(0.122)	(0.043)
x2	49.517***	49.976***	
	(0.522)	(2.553)	
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 $\mathbf{corr}(X_{it}, \eta_i) \neq 0$
- FE estimator still provides **unbiased** estimates since η_i are eliminated
  - how does the 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 that give the same results
  - **de-meaning**
  - **binary variables**
- Let's do de-meaning first, since it's closely related to the "removing between variation" explanation
  - start with a standard panel data model

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

  - for each variable get the mean value of that variable for each individual
  - subtract out that mean to get residuals

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

  - work with those residuals

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

- The residuals are, by construction, no longer related to the η_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 and add it as a control

Estimation: empirical example

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

```
data(crime4, package = 'wooldridge')
crime4 <- crime4 %>%
  filter(county %in% c(1, 3, 7, 23), # filter to the data points from our graph
         prbarr < .5) %>%
  group_by(county) %>%
  mutate(mean_crime = mean(crmrte),
         mean_prob = mean(prbarr)) %>%
  mutate(demean_crime = crmrte - mean_crime,
         demean_prob = prbarr - mean_prob)
```

county	year	crmrte	prbarr	mean_crime	mean_prob	demean_crime	demean_prob
1	81	0.0398849	0.289696	0.0357414	0.3243583	0.0041435	-0.0346623
1	82	0.0383449	0.338111	0.0357414	0.3243583	0.0026035	0.0137527
3	81	0.0163921	0.202899	0.0149364	0.1766691	0.0014557	0.0262299
3	82	0.0190651	0.162218	0.0149364	0.1766691	0.0041287	-0.0144511

Estimation: empirical example

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

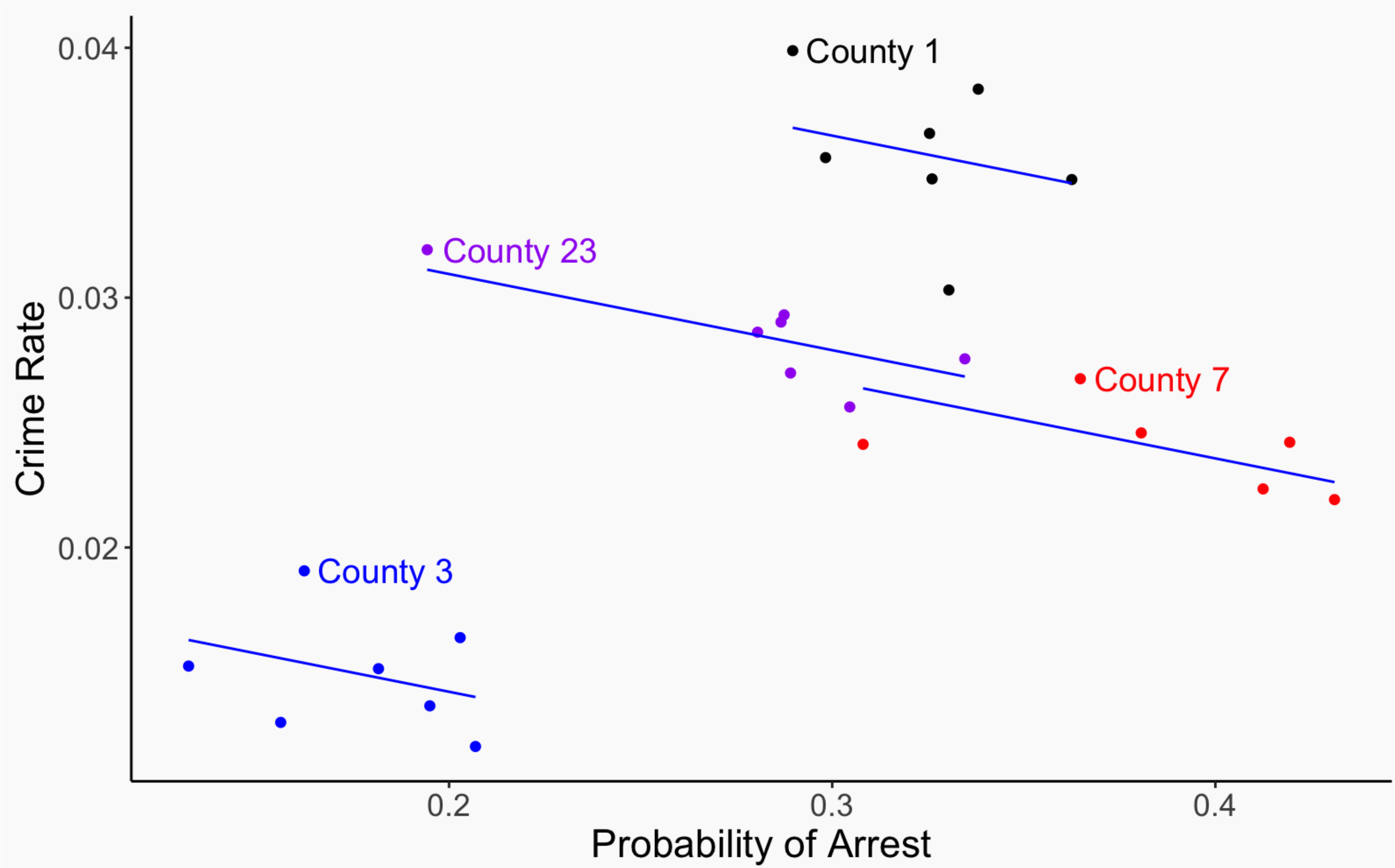
```
pooling ← lm(crmrte ~ prbarr, data = crime4)
lsdv    ← lm(crmrte ~ prbarr + factor(county), data = crime4)
de_mean ← lm(demean_crime ~ demean_prob, data = crime4)
```

	Model 1	Model 2	Model 3
prbarr	0.049**	-0.030*	
	(0.017)	(0.012)	
demean_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 a county**
  - if we think we've **causally** identified it then "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 different points in time
- A benefit of the LSDV approach is that it calculates the fixed effects 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
  - 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



Panel data: estimation

- Empirical 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 IV
    - it clusters the standard errors by the first fixed effect by default

Panel data: estimation

Let's look at the output of `plm` and `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 FE estimator 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

$$\mathrm{corr}(X_{it}, \eta_i) = 0$$

    - 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
- The Hausman test is a broad set of tests that compare the estimates in one model against the estimates in another and sees if they are different
- It basically tests whether the errors are correlated with the regressors
    - under H_0: $\mathbf{corr}(X_{it}, \eta_i) = 0$ and both RE and FE estimators are consistent, but the RE estimator is more efficient
    - under H_1: $\mathbf{corr}(X_{it}, \eta_i) \neq 0$ and only FE estimator is consistent
- FE estimator is almost always preferred to the RE estimator, except when you are quite sure that the right-hand-side variables X_{it} are unrelated to the individual effects η_i

Fixed or random effects

- Let's apply it to two simulated data sets 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 = 14621, 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

- One of the assumptions of the regression model is that the error terms are independent of each other
  - however, we might imagine that some of the left variation is shared across all individuals, making them correlated with each other
  - thus, not taking that into account would make the s.e. wrong
- Two conditions need to hold for clustering to be necessary
  - first, there needs to be **treatment effect heterogeneity**. That is, the treatment effect must be quite different for different individuals
- If that is true, there's a second condition
  - either **DGP** is clustered, meaning the individuals/groups in your data represent a **non-random sampling of the population**. For example, some groups are more likely to be included in your sample than others
  - or **treatment assignment mechanism** is clustered, meaning within individuals/groups your **treatment variable is assigned in a clustered way**. For example, if you belong to a certain group, you are more likely to get treatment
- So before clustering, think about whether both conditions are likely to be true (Abadie et al. 2017)

Panel data inference

- It's common to cluster s.e. at the level of the fixed effects, since it seems likely that errors would be correlated over time
  - clustered s.e. calculate the standard errors while allowing some level of correlation between the error terms
  - `feols` in `fixest` clusters by the first FE by default
- 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

Panel data inference: simulation

```r
set.seed(7)
df <- tibble(id = sort(rep(1:600, 10)),
             time = rep(1:10, 600),
             # we don't generate x2 as FE eliminates it anyway
             x1 = rnorm(6000),
             # Now the error term has two components:
             # 1. the individual cluster (5*id),
             # 2. the normal error term (rnorm(6000))
             y = id + time + 2*x1 + (5*id + rnorm(6000)))
```

id	time	x1	y
1	1	2.2872472	11.452224
1	2	-1.1967717	4.656338
1	3	-0.6942925	7.377264
2	1	0.3569862	13.046828
2	2	2.7167518	19.096420
2	3	2.2814519	20.936149

Panel data inference: simulation

```
# The true effect is 2
fe_clustered     ← feols(y ~ x1 | id, df)
fe_not_clustered ← feols(y ~ x1 | id, se = 'standard', df) # make s.e. i.i.d.
```

	Model 1	Model 2
x1	1.900***	1.900***
	(0.046)	(0.043)
Num.Obs.	6000	6000
Std.Errors	by: id	IID
+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001		

- Not accounting for clustering at the individual level leads to incorrect s.e.

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 8: Panel Data

Slides

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

Articles

- Abadie, A., Athey, S., Imbens, G. W., & Wooldridge, J. (2017). When Should You Adjust Standard Errors for Clustering? (No. w24003). National Bureau of Economic Research