

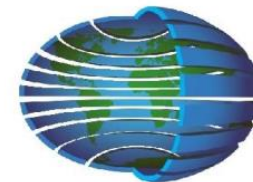
An Introduction to Panel Data Regression

Rafael Mesquita, Antônio Fernandes, Dalson B. Figueiredo Filho
Federal University of Pernambuco (UFPE)

2021



UNIVERSIDADE
FEDERAL
DE PERNAMBUCO



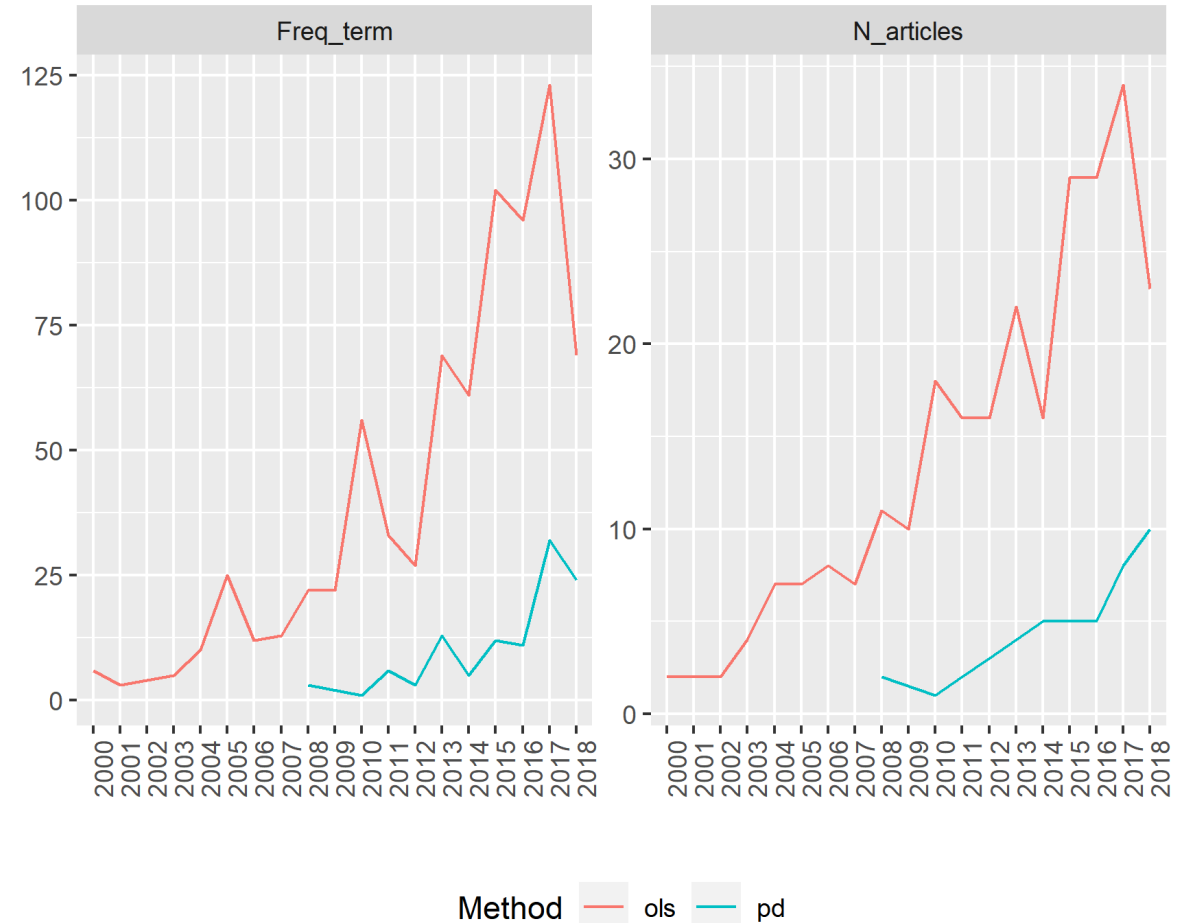
PPGCP
Programa de
Pós-Graduação em
Ciência Política

Supporting materials

- Article in *Política Hoje*
 - <https://periodicos.ufpe.br/revistas/politica hoje/article/view/246522/38645>
- OSF Repository
 - https://osf.io/5yx7g/?view_only=ac1691cced8549238d6d6e0a9d2b7f7b

The applicability of panel data for PIR

- Several PIR phenomena naturally fit the panel data format
 - E.g., Elections in cities; Trade between countries each year
- However, it is seldom used in Brazilian PIR
 - Only 0.6% in a sample of 7,7 thousand papers



Source: Mesquita et al. (2021)

The applicability of panel data for PIR

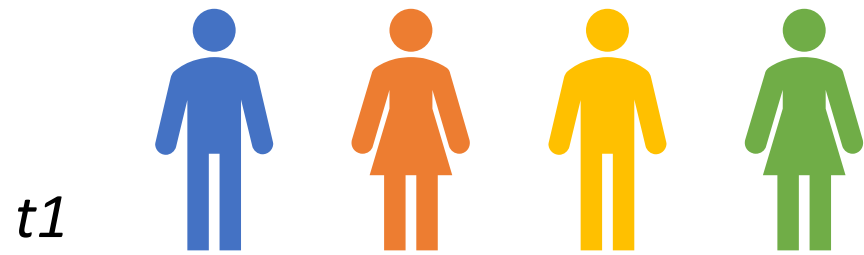
- Panel data offer several benefits, such as:
 - 1. Facilitates detection of causality in time
 - 2. Monitoring individual variation in the cases of interest
 - 3. Reduction of measurement errors
 - 4. Larger samples
 - 5. Ways of eliminating omitted variable bias (which could not be implemented with static cross-sectional data)

The structure of longitudinal data

The structure of longitudinal data

- Cross-sectional data have observations on 'i' subjects in a single time period (or ignoring the passage of time)
- Time series are observations of a single subject, repeated over 't' periods of time
- Panel data combine both: observe 'i' subjects over 't' periods of time

Cross-sectional data



t_2

t_3

Time series



Panel data



The layout of panel data

Cross-Section

UF	Time	Homicide Rate
i	t	X_i
Acre	2011	22,0
Bahia	2011	39,4
Santa Catarina	2011	12,8

Time-Series

UF	Time	Homicide Rate
i	t	X_t
Bahia	2011	39,4
Bahia	2012	43,4
Bahia	2013	37,8

Time-Series Cross-Section

UF	Time	Homicide Rate
i	t	X_{it}
Acre	2011	22
Acre	2012	27,4
Acre	2013	30,1
Bahia	2011	39,4
Bahia	2012	43,4
Bahia	2013	37,8
Santa Catarina	2011	12,8
Santa Catarina	2012	12,9
Santa Catarina	2013	11,9

Source:
Mesquita et al. (2021), based on
data from Cerqueira (2017)

The layout of panel data

UF	2011	2012	2013
Acre	22	27,4	30,1
Bahia	39,4	43,4	37,8
Santa Catarina	12,8	12,9	11,9

Wide format
- Each line, one state

UF	Time	Homicide Rate	Overall Mean
i	t	X_{it}	\bar{X}
Acre	2011	22	26,4
Acre	2012	27,4	26,4
Acre	2013	30,1	26,4
Bahia	2011	39,4	26,4
Bahia	2012	43,4	26,4
Bahia	2013	37,8	26,4
Santa Catarina	2011	12,8	26,4
Santa Catarina	2012	12,9	26,4
Santa Catarina	2013	11,9	26,4

Long format
- Each line an observation of state x year

Measures aggregated by time and unit

- The repetition allows the aggregation of values in different ways
 - E.g., Mean of a single state over time; Mean of all states in a given year.
- These measures are useful for descriptive statistics and some forms of estimation

Cross-Section

UF	Time	Homicide Rate	Overall Mean	Between Case Deviat.
i	t	X_i	\bar{X}	$X_i - \bar{X}$
Acre	2011	22,0	24,7	-2,7
Bahia	2011	39,4	24,7	14,7
Santa Catarina	2011	12,8	24,7	-11,9

Time-Series

UF	Time	Homicide Rate	Individual Mean	Intra-case Deviat.
i	t	X_t	\bar{X}	$X_t - \bar{X}$
Bahia	2011	39,4	40,2	-0,8
Bahia	2012	43,4	40,2	3,2
Bahia	2013	37,8	40,2	-2,4

Time-Series Cross-Section

UF	Time	Homicide Rate	Overall Mean	Individual Mean	Overall Deviat.	Intra-case Deviat.	Between Case Deviat.
i	t	X_{it}	\bar{X}	\bar{X}_i	$X_{it} - \bar{X}$	$X_{it} - \bar{X}_i$	$\bar{X}_i - \bar{X}$
Acre	2011	22	26,4	26,5	-4,4	-4,5	0,1
Acre	2012	27,4	26,4	26,5	1,0	0,9	0,1
Acre	2013	30,1	26,4	26,5	3,7	3,6	0,1
Bahia	2011	39,4	26,4	40,2	13,0	-0,8	13,8
Bahia	2012	43,4	26,4	40,2	17,0	3,2	13,8
Bahia	2013	37,8	26,4	40,2	11,4	-2,4	13,8
Santa Catarina	2011	12,8	26,4	12,5	-13,6	0,3	-13,9
Santa Catarina	2012	12,9	26,4	12,5	-13,5	0,4	-13,9
Santa Catarina	2013	11,9	26,4	12,5	-14,5	-0,6	-13,9

Types of panels

Types of panels: regarding the N/T ratio

$N > T$

- Short or cross-section dominant
- Typical example: census

$T > N$

- Long or temporally-dominant
- “Time-Series Cross-Section” (TSCS) for some authors (Beck)
- Typical example: country-year GDP data

Types de panels: regarding missing cases

Balanced

- All subjects are observed in all time periods
- Total observations = $N \times T$

UF	Time	Homicide Rate
<i>i</i>	<i>t</i>	X_{it}
Acre	2011	22
Acre	2012	27,4
Acre	2013	30,1
Bahia	2011	39,4
Bahia	2012	43,4
Bahia	2013	37,8
Santa Catarina	2011	12,8
Santa Catarina	2012	12,9
Santa Catarina	2013	11,9

Unbalanced

- Some subjects are not observed in some time frames
- Total observations < $N \times T$

UF	Time	Homicide Rate
<i>i</i>	<i>t</i>	X_{it}
Acre	2011	22
Acre	2012	
Acre	2013	30,1
Bahia	2011	39,4
Bahia	2012	43,4
Bahia	2013	37,8
Santa Catarina	2011	
Santa Catarina	2012	12,9
Santa Catarina	2013	11,9

The 4 main approaches

Pooled OLS

First
Differences

Fixed
Effects

Random
Effects

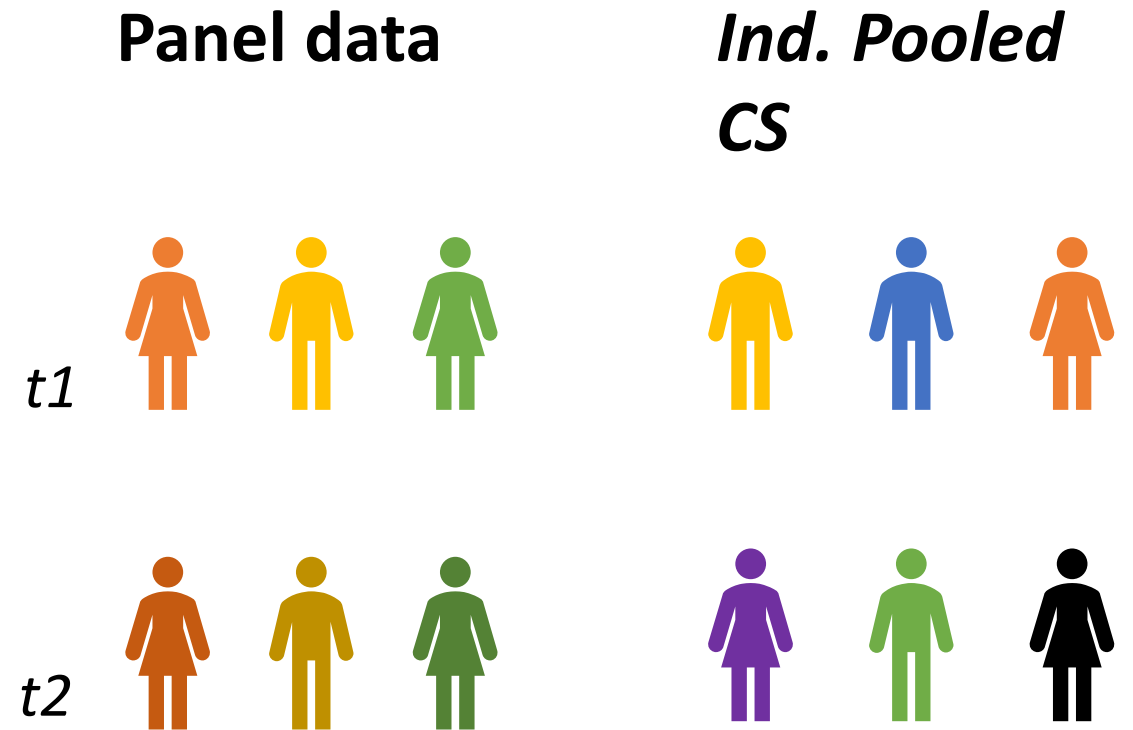
Pooled OLS

Pooled OLS

- If I have observations of several entities over time, is it appropriate to run a regular OLS model, the same way I would for cross-sectional data?
- Distinction between **panel data** and ***independently pooled cross sections***

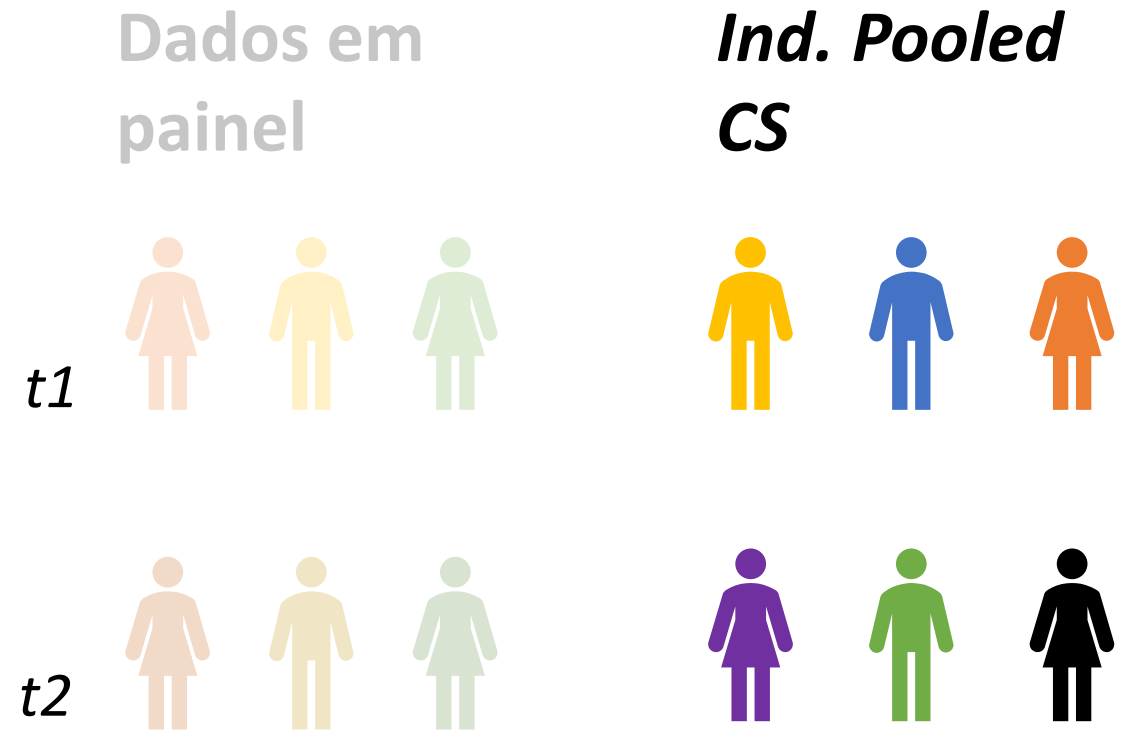
Independently pooled cross sections vs. Panel data

- Random samples of distinct cases at t_1 and t_2 may be pooled to increase the N
- Pooling cases does not threaten linear regression assumptions if there is randomness
- *Ind. pooled cross sections* are not exactly panel/longitudinal data because they do not observe the same units over time



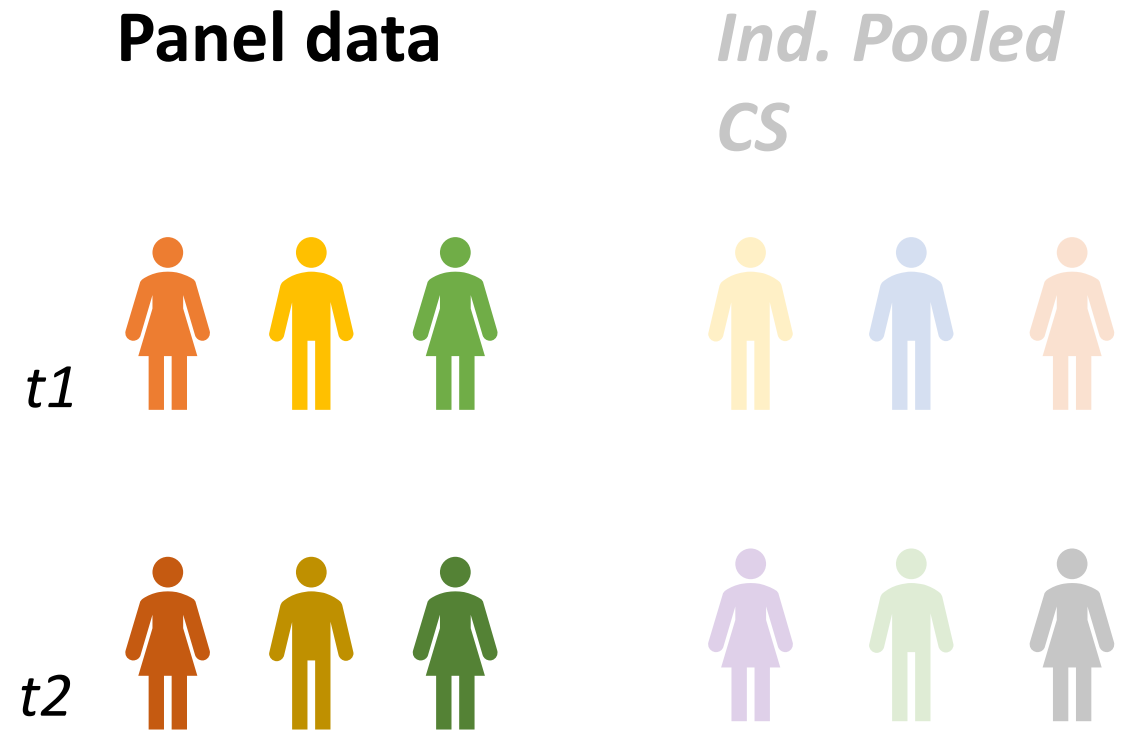
Independently pooled cross-sections

- Usually, a normal OLS regression can be used
- The differences between $t1$ and $t2$ can be captured satisfactorily by a dummy



Pooled OLS

- But what if we observe the same subjects over time? Is it possible to use a traditional OLS regression?
- Yes, if the set of IVs is appropriate and if it ensured that the **errors remain random and uncorrelated**

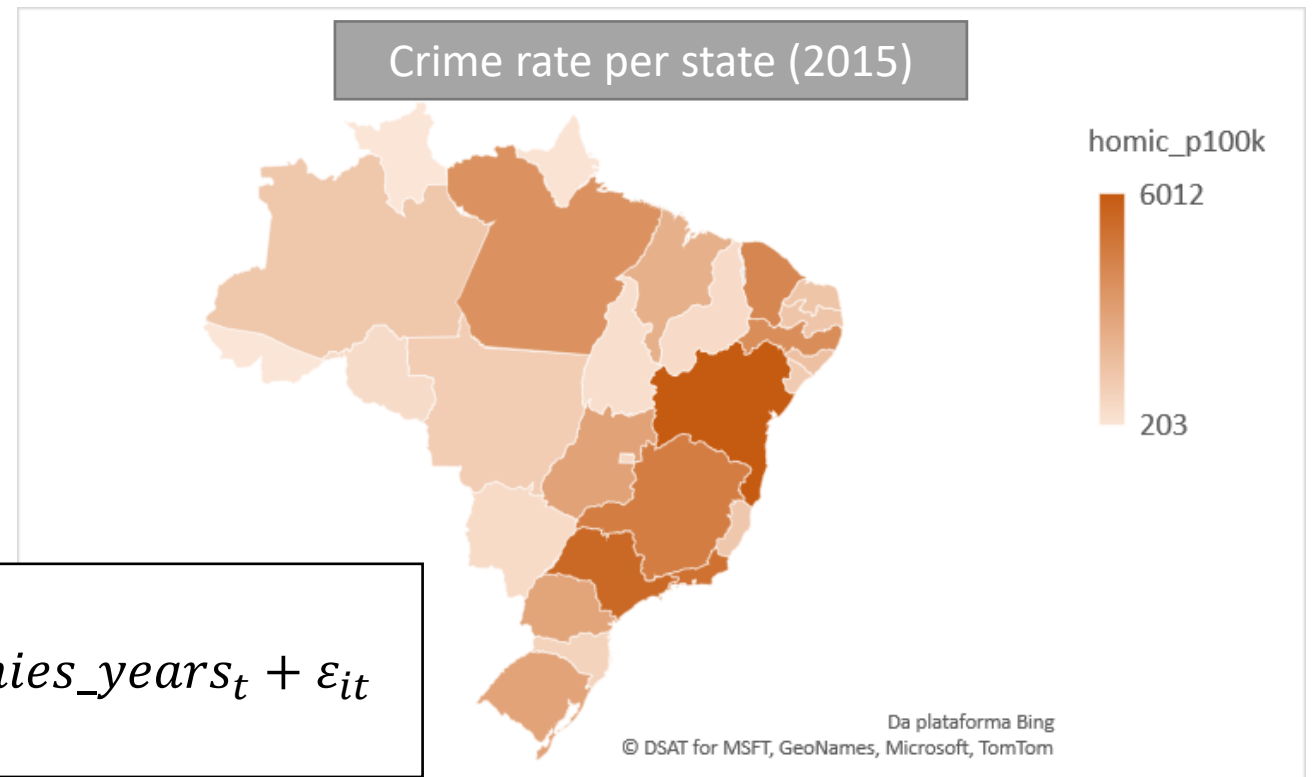


Example: Crime rate in Brazilian states

Explaining the crime rate in the 27 states and federal district between 2015 and 2017, given unemployment and size of population, using OLS

Model:

$$crime_{it} = \alpha + unempl_{it} + pop_{it} + dummies_years_t + \varepsilon_{it}$$



Data in OSF: https://osf.io/5yx7g/?view_only=ac1691cced8549238d6d6e0a9d2b7f7b

```

=====
Dependent variable:
-----
l_homic_p100k
-----
media_an_desoc_perc    0.077***
                        (0.015)

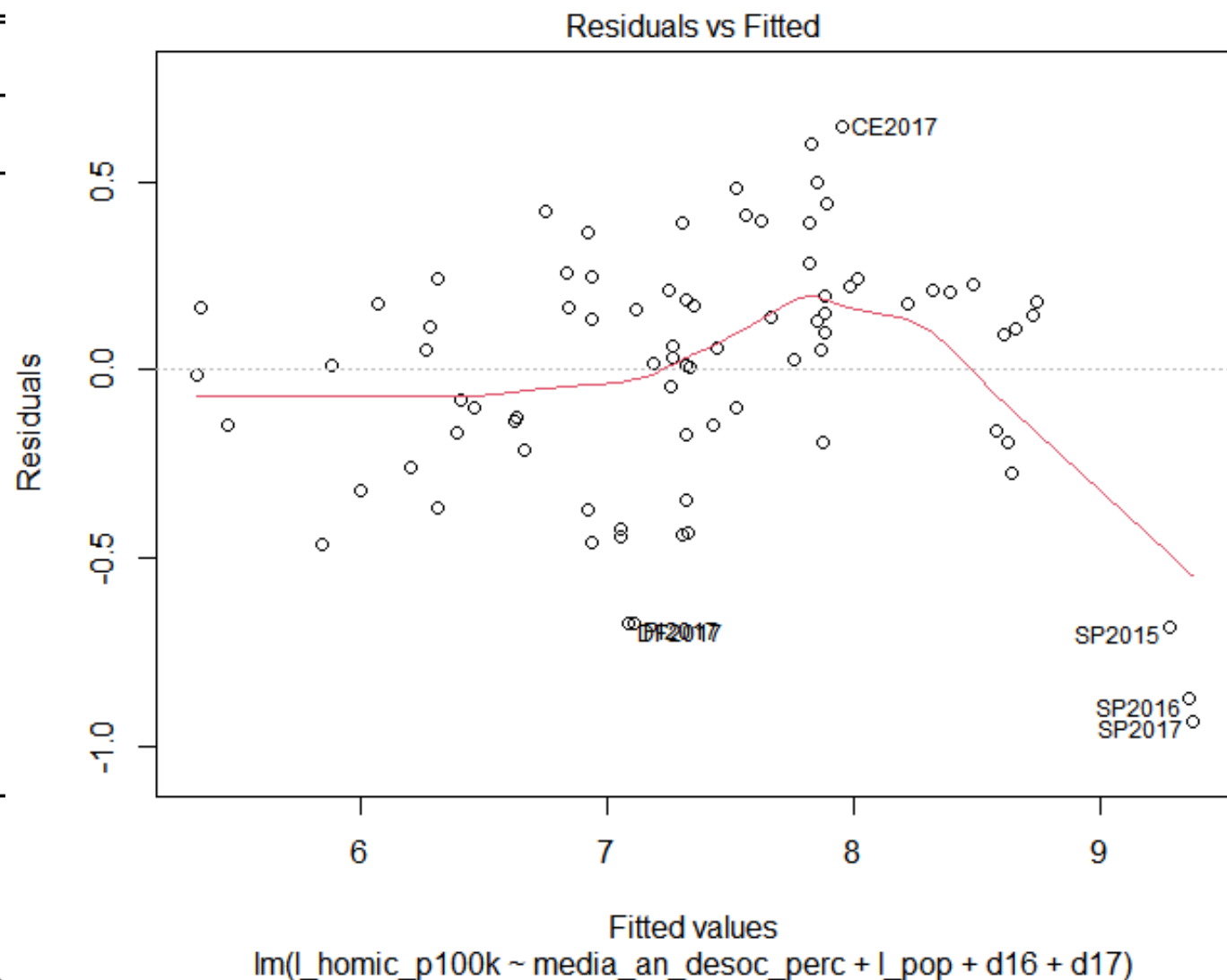
l_pop                  0.841***
                        (0.036)

d16                   -0.160
                        (0.101)

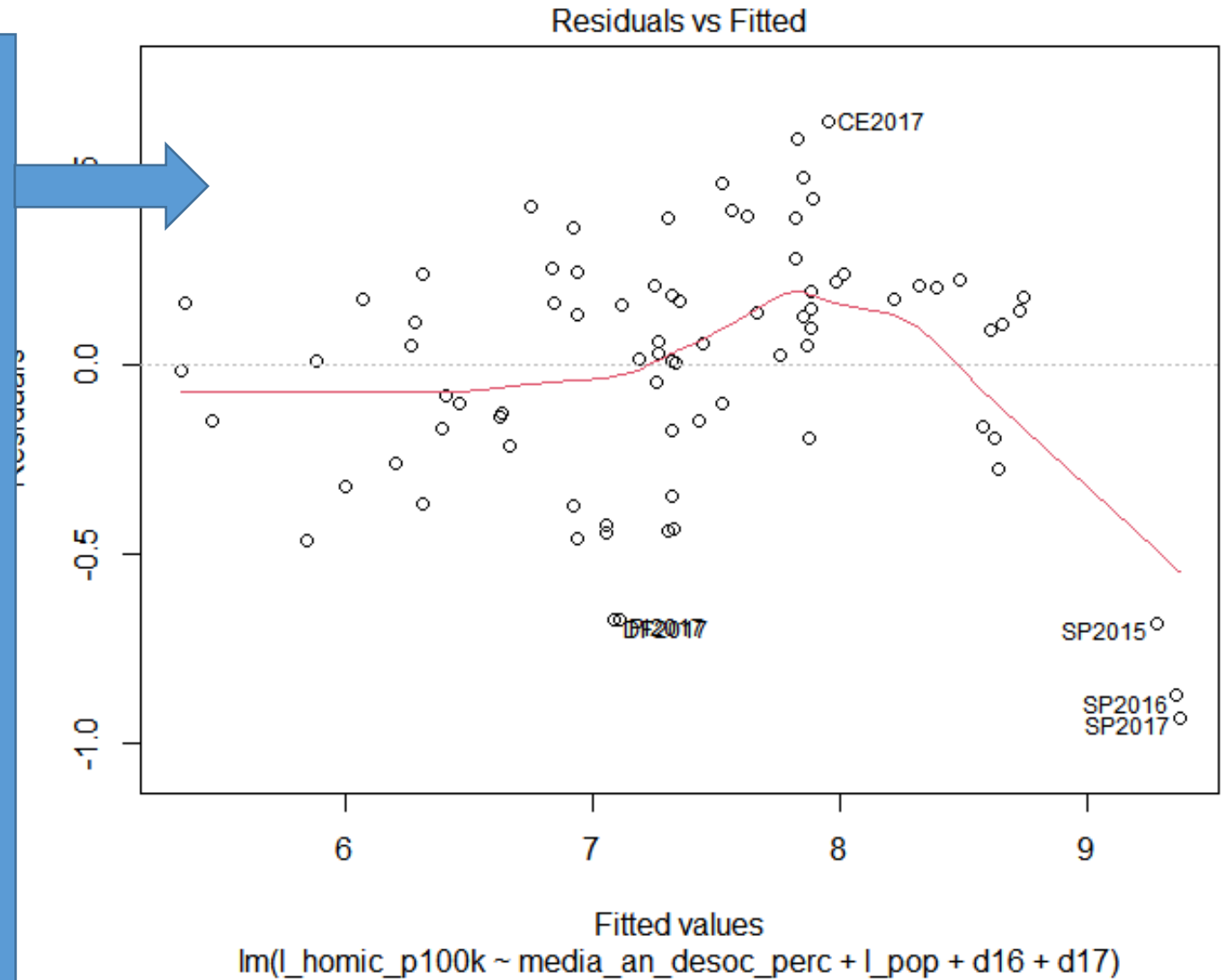
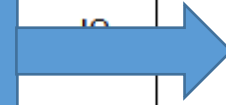
d17                   -0.239**
                        (0.110)

Constant              -6.247***
                        (0.571)
-----
Observations                81
R2                          0.883
Adjusted R2                 0.877
Residual Std. Error        0.335 (df = 76)
F Statistic                143.023*** (df = 4; 76)
=====
Note:          *p<0.1; **p<0.05; ***p<0.01

```



- The residuals vs. fitted diagnostics plot suggests that there may be something peculiar, unmodelled, occurring in DF, CE, and SP, leading their results to **systematically** diverge from the model's predictions
- This suggests that pooling is not recommended for these data

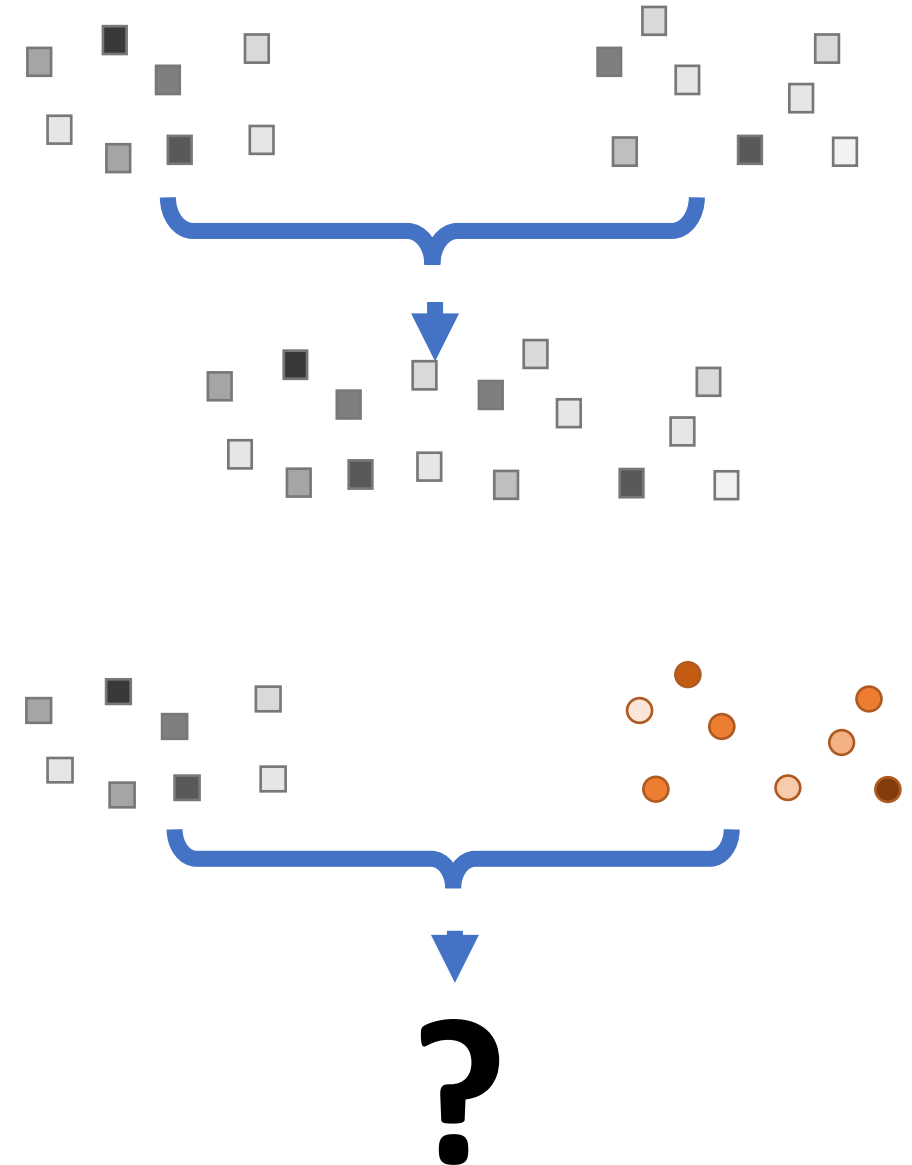


Note:

* $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

When can observations be pooled?

- Assumes that two or more samples may be combined in a single pool of observations
- The original linear model is still good enough to describe all of the systematic variation
- **That is, introducing more observations does not add any relevant non-systematic variation**
- Usually, this the result of random samples (independently pooled cross sections) or well-specified models whose IVs exhaust the systematic variation adequately
- **However, a new set of observations that brings with it important heterogeneity, which is not captured by the IVs, will hinder pooling**

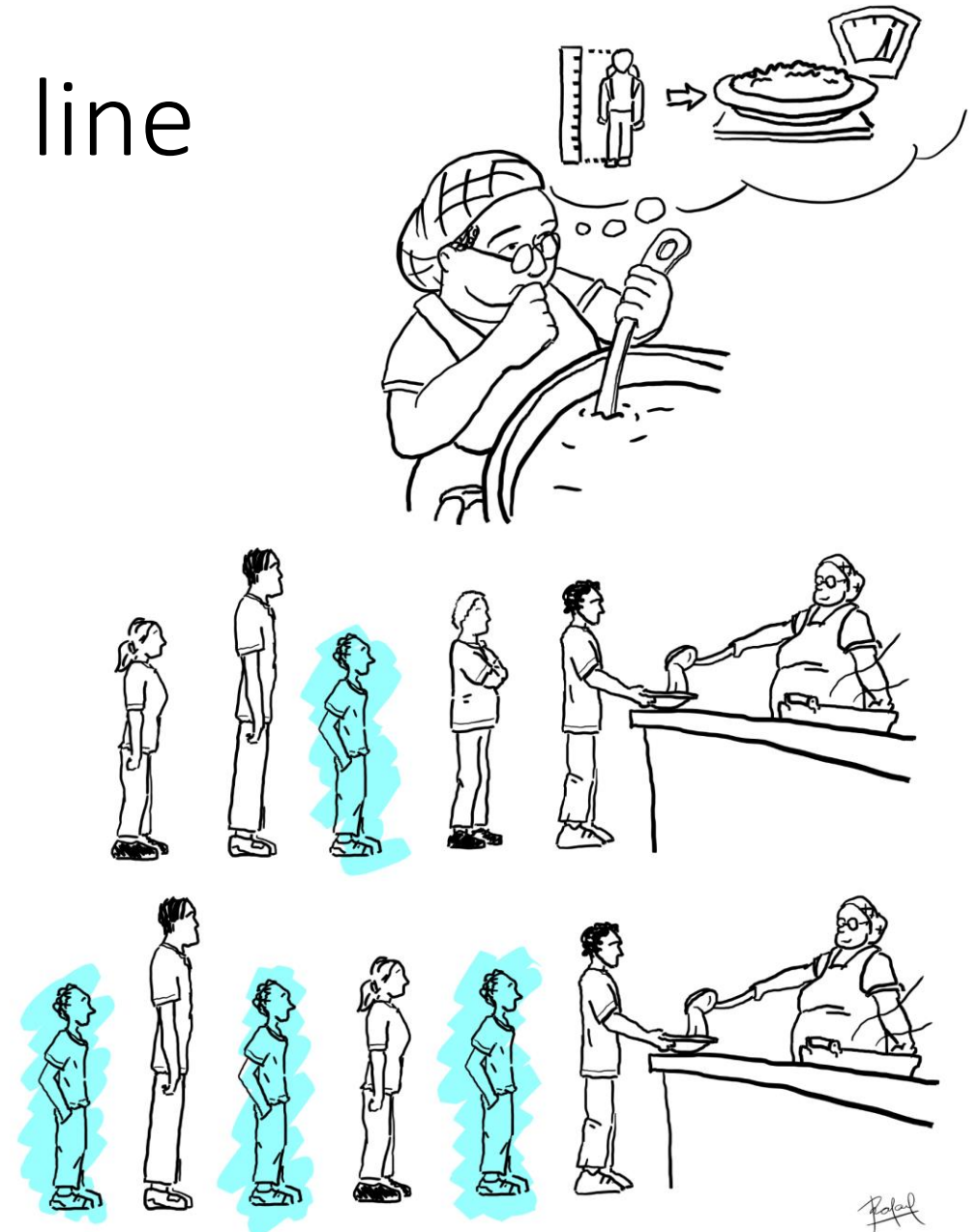


When can observations be pooled?

- Intuitively, we think that observing **the same subject** leads us to have a **more homogenous** database
- When in fact, repetition can lead to **more heterogeneity** and thus discourage pooling
- It all depends on how satisfactorily the IVs explain the variation in the data
- Example: “the school lunch line”

Example: the school lunch line

- A lunch lady may have a “mental model” predicting **the amount of food a child will eat (Y)** as a function of an observed variable, **the height of the child (X)**, plus **random factors (ϵ)**
- But there may be **a child** in line who, although being short, eats much more than their height suggests, due to some unobserved factor (e.g., gluttony)
- If that child only goes through the line once, the lunch lady’s “mental model” doesn’t fail in serving the class, because the “unmodelled” appetite of the gluttonous student will be compensated by another student’s moderation further down the line (i.e., a student who, even though tall, eats little).
- But if the gluttonous child joins the line **repeatedly**, the “predicted” food for the class will end sooner! This is because the unmodelled heterogeneity that was dissipated before in the population and could be treated as “noise” becomes a more persistent factor



Unobserved heterogeneity

Breaking down the “composite error”

$$Y_{it} = \alpha_i + \beta X_{it} + \varepsilon_{it}$$



$$\varepsilon_{it} = \mu_i + \nu_{it}$$



**Unit effect, fixed effect, or
unobserved heterogeneity**
Fixed for each spatial unit,
invariable over time



**Variable or
idiosyncratic effect**
Variable for time and
units

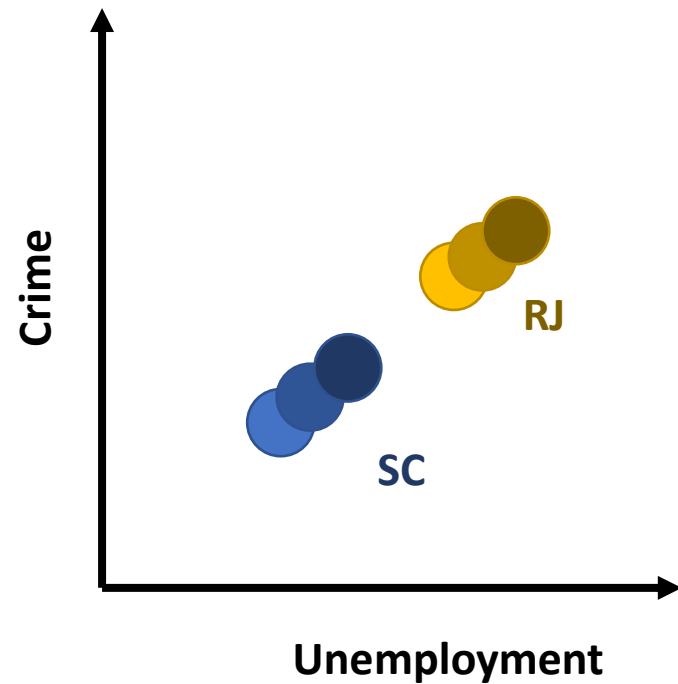
Unobserved heterogeneity

- Individual unobserved heterogeneity (μ_i) encapsulates all peculiar factors related to the spatial unit, not captured by X, that influence the result of Y
 - E.g., Why these crime rates if unemployment is low? “Because *it’s Rio de Janeiro*”
 - Crime rate = unemployment + strength of organized crime
- Why do states differ in their crime rates? If it is only due to distinct levels of unemployment, units are not heterogenous.
 - E.g., If RJ had the unemployment rate of SC, it would also have its crime rate



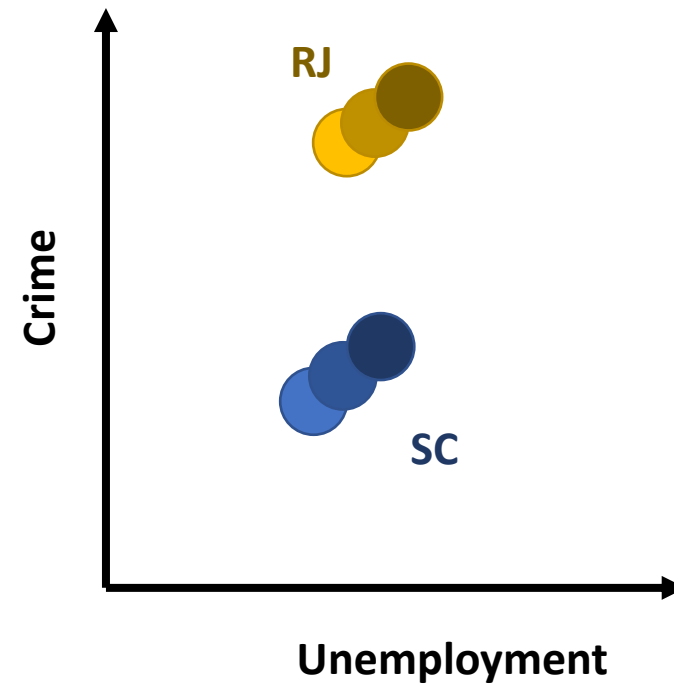
Unobserved heterogeneity

Without heterogeneity



Higher
unemployment,
more crime

With heterogeneity



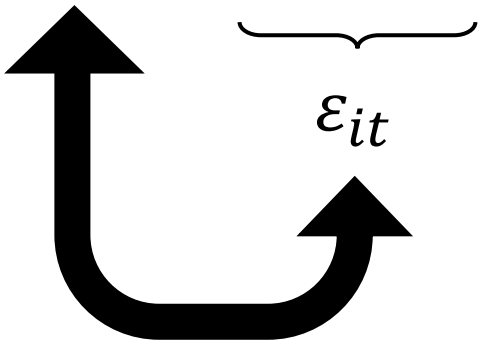
Same level of
unemployment,
but distinct
crime rates

Unobserved heterogeneity

- If the unobserved heterogeneity is negligible, POLS is appropriate
- If there is heterogeneity, its presence violates classic linear regression assumptions

How heterogeneity violates assumptions

- **Exogeneity/zero conditional mean assumption**
- **$E(\varepsilon | \mathbf{x}) = 0$**
- If μ_i is correlated with X_{it} then we have a case of omitted variable bias

$$Y_{it} = \alpha_i + \beta X_{it} + \underbrace{\mu_i + v_{it}}_{\varepsilon_{it}}$$


A diagram illustrating the relationship between the error term and the regression equation. A large, thick, black U-shaped arrow starts from the bottom right and points upwards to the left, ending at the α_i term in the equation. A smaller, thick, black arrow starts from the ε_{it} label and points upwards to the $\mu_i + v_{it}$ term in the equation. A horizontal curly brace is positioned above the $\mu_i + v_{it}$ term, with the label ε_{it} centered below it.

How heterogeneity violates assumptions

- **Random sample/absence of autocorrelation assumption:**
- **$\text{Corr}(\varepsilon_s, \varepsilon_j) = 0$ for $s \neq j$**
- If μ_i is constant for every unit i , errors ε at $t1$ will be correlated with the errors at $t2$, since μ_i will certainly be present in both observations

$$Y_{it} = \alpha_i + \beta X_{it} + \mu_i + v_{it}$$

$$t1 \quad Y_{i1} = \alpha_i + \beta X_{i1} + \underline{\mu_i} + v_{i1}$$

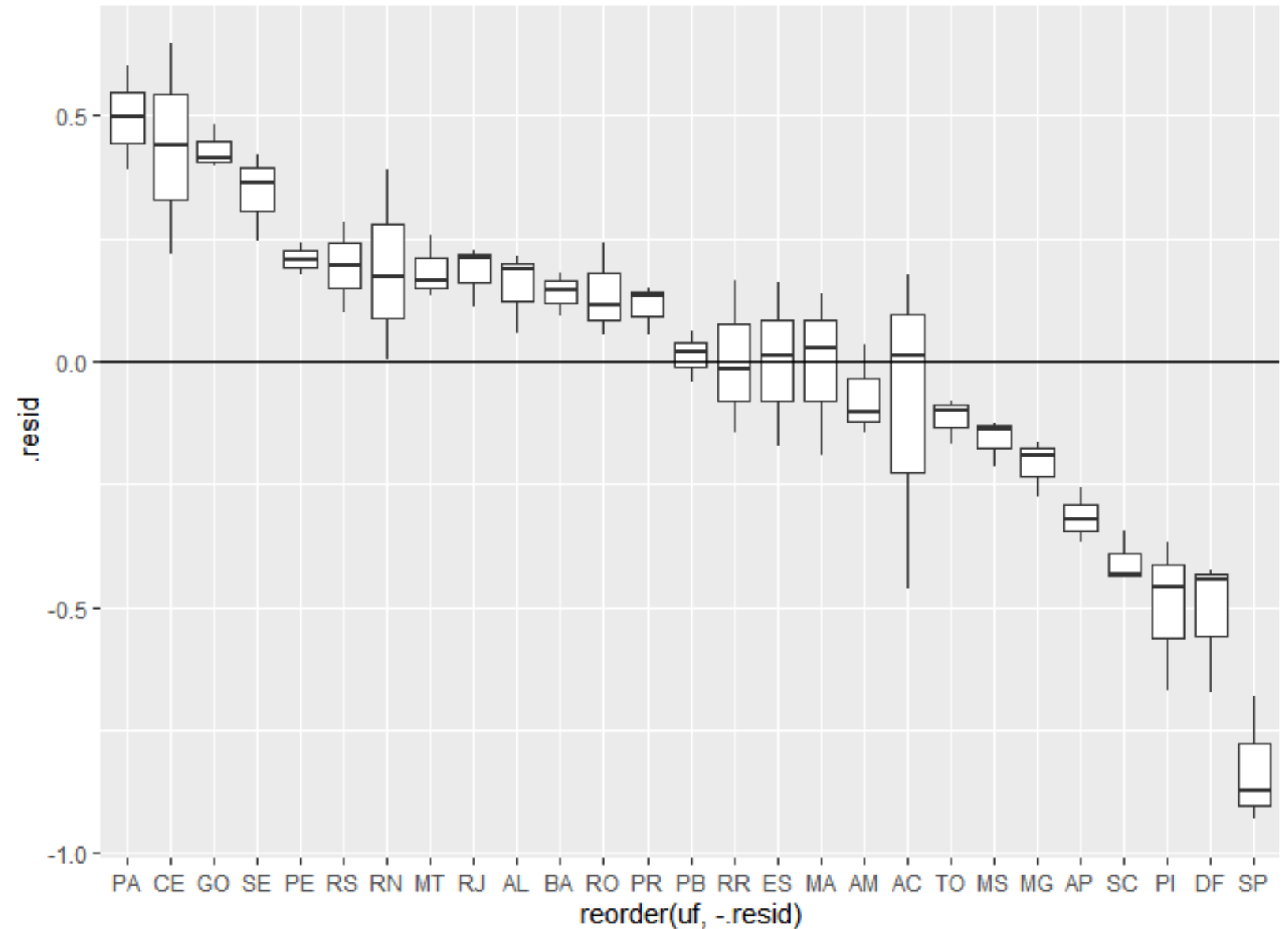
$$t2 \quad Y_{i2} = \alpha_i + \beta X_{i2} + \underline{\mu_i} + v_{i2}$$

Diagnosing heterogeneity (I)

Diagnosing heterogeneity (I)

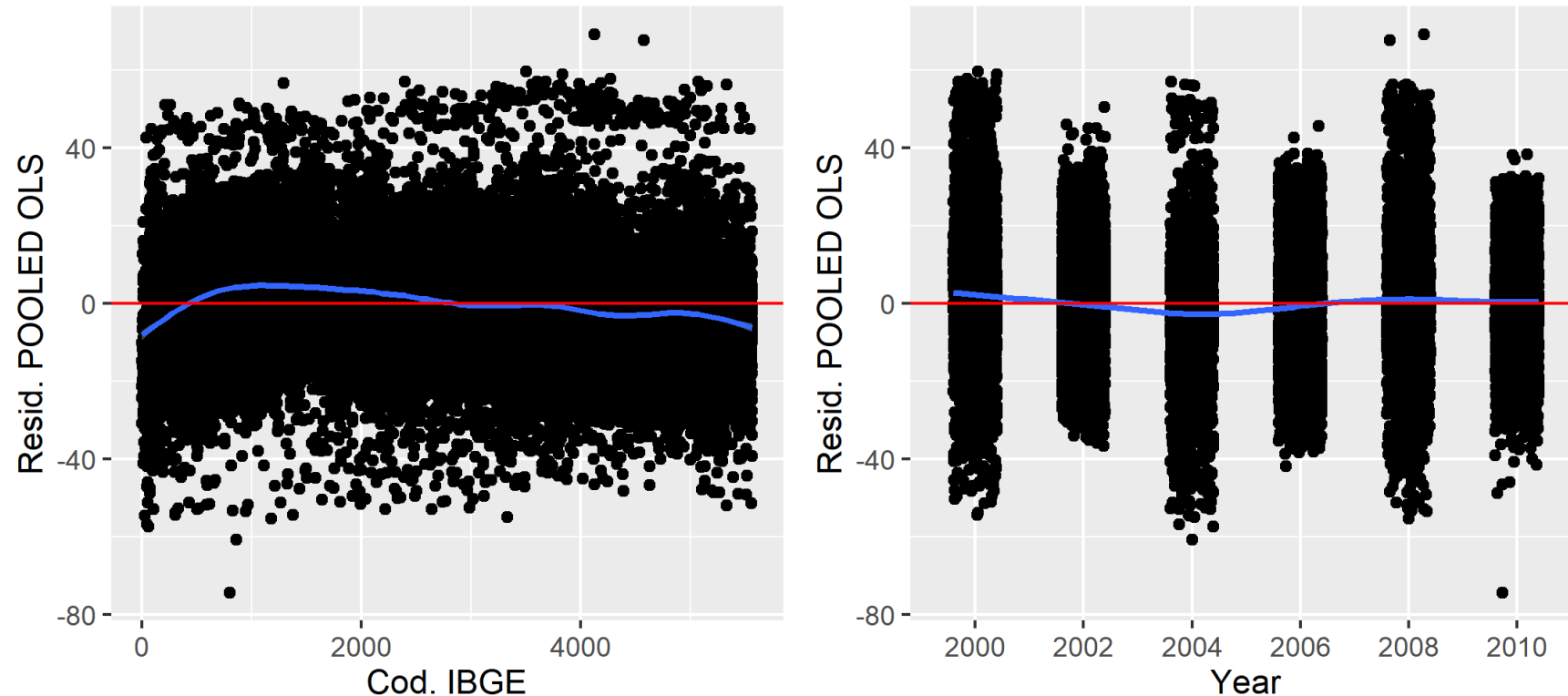
- (1) Substantive knowledge
 - E.g., Can we expect that the crime rate in RJ is caused by the same determinants that in SC?
- (2) Graphic analysis
 - Grouping residuals by spatial unit or by year

Residuals for crime
rate Pooled OLS
model, grouped by
state



Data by Fernandes and Fernandes (2017) for “votes for incumbent” in local and national elections between 2000 and 2010, for over 5 thousand municipalities.

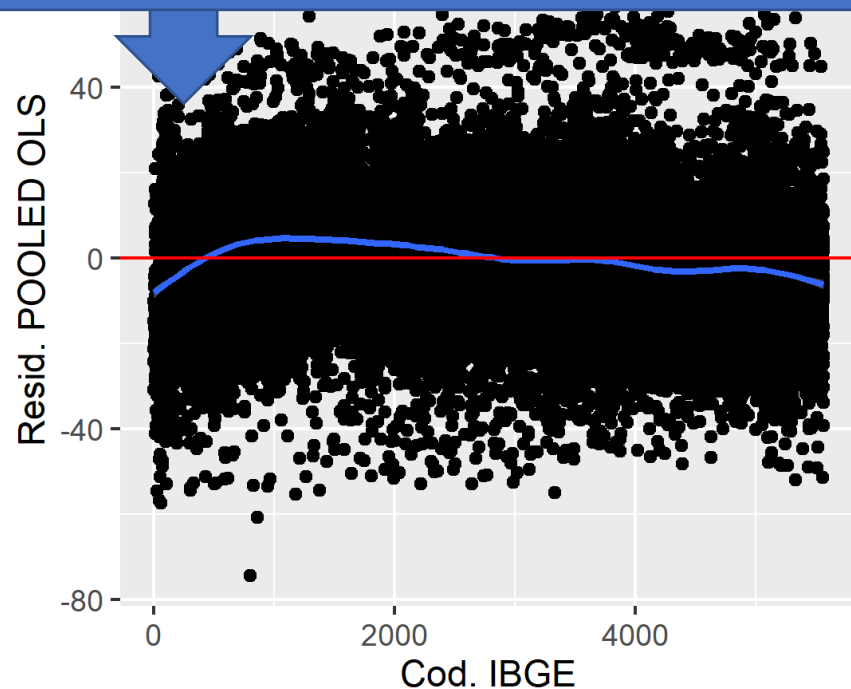
Residuals grouped by municipality and year/election



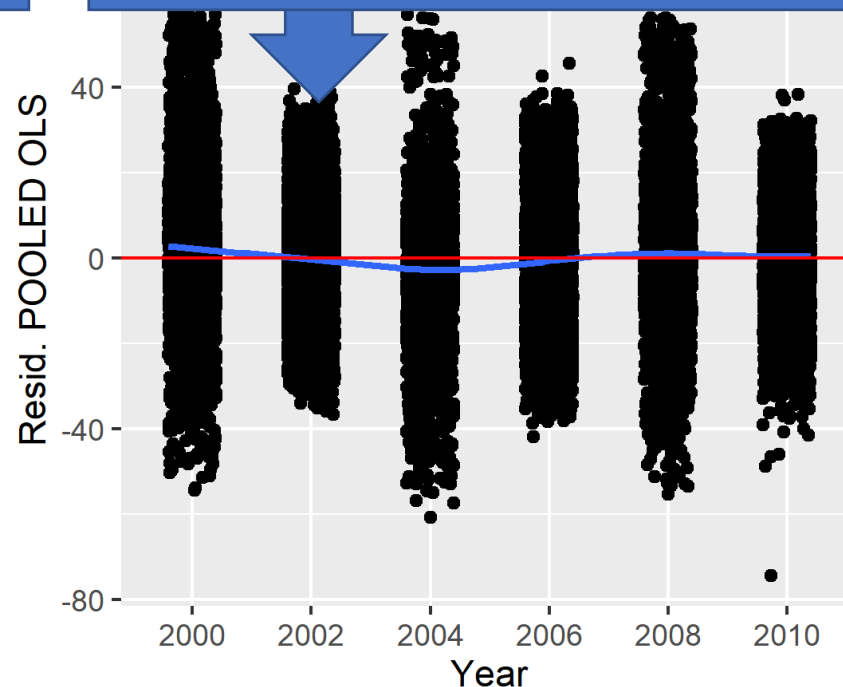
Source: Mesquita et al. (2021), based on data by Fernandes and Fernandes (2017)

Data in the OSF: https://osf.io/5yx7g/?view_only=ac1691cced8549238d6d6e0a9d2b7f7b

There is something that systematically differentiates the predictions for municipalities with IBGE code between 500 and 3000 from those with code >4000



There is something that systematically differentiates the predictions for majoritarian elections in 2002, 2006, and 2010 (national?)



Source: Mesquita et al. (2021), based on data by Fernandes and Fernandes (2017)

Data in the OSF: https://osf.io/5yx7g/?view_only=ac1691cced8549238d6d6e0a9d2b7f7b

Eliminating heterogeneity: First Differences

First Differences

- First differences (FD) are a way of removing individual fixed effects (μ_i) and enabling an OLS regression free of violations
- Since μ_i is the same for every i over time, it is possible to subtract (“differentiate”) for every i each observation from its previous one; thus, elements invariant in time are removed, leaving only varying ones

First Differences

$$Y_{it} = \alpha_i + \beta X_{it} + \mu_i + v_{it}$$

$$Y_{i1} = \alpha_i + \beta X_{i1} + \mu_i + v_{i1} \quad t1$$

$$- \left[Y_{i2} = \alpha_i + \beta X_{i2} + \mu_i + v_{i2} \right] \quad t2$$

$$\Delta Y_i = \beta \Delta X_i + \Delta v_i$$

ID	Time	Homic. rate	Unempl. rate
i	t	Y_{it}	X_{it}
Acre	2011	22	5,4
Acre	2012	27,4	8
Acre	2013	30,1	9,6
Acre	2014	29,4	9,8
Bahia	2011	39,4	10,5
Bahia	2012	43,4	10,2
Bahia	2013	37,8	9,9
Bahia	2014	40	10
Santa Catarina	2011	12,8	3,6
Santa Catarina	2012	12,9	3,1
Santa Catarina	2013	11,9	3,4
Santa Catarina	2014	13,5	3,1



ID	Time	Homic. rate	Unempl. rate
i	t	ΔY_{it}	ΔX_{it}
Acre	d1	5,4	2,6
Acre	d2	-0,7	0,2
Bahia	d1	4	-0,3
Bahia	d2	2,2	0,1
Santa Catarina	d1	0,1	-0,5
Santa Catarina	d2	1,6	-0,3

Hence, ideally T is even. If T is odd, 1 cross section is discarded

Obs.: Actually, we only differentiate observable elements in the database (X and Y). However, the unobserved ones (μ_i) are also removed when the regression is run.

Source:
Mesquita et al. (2021)

First Differences

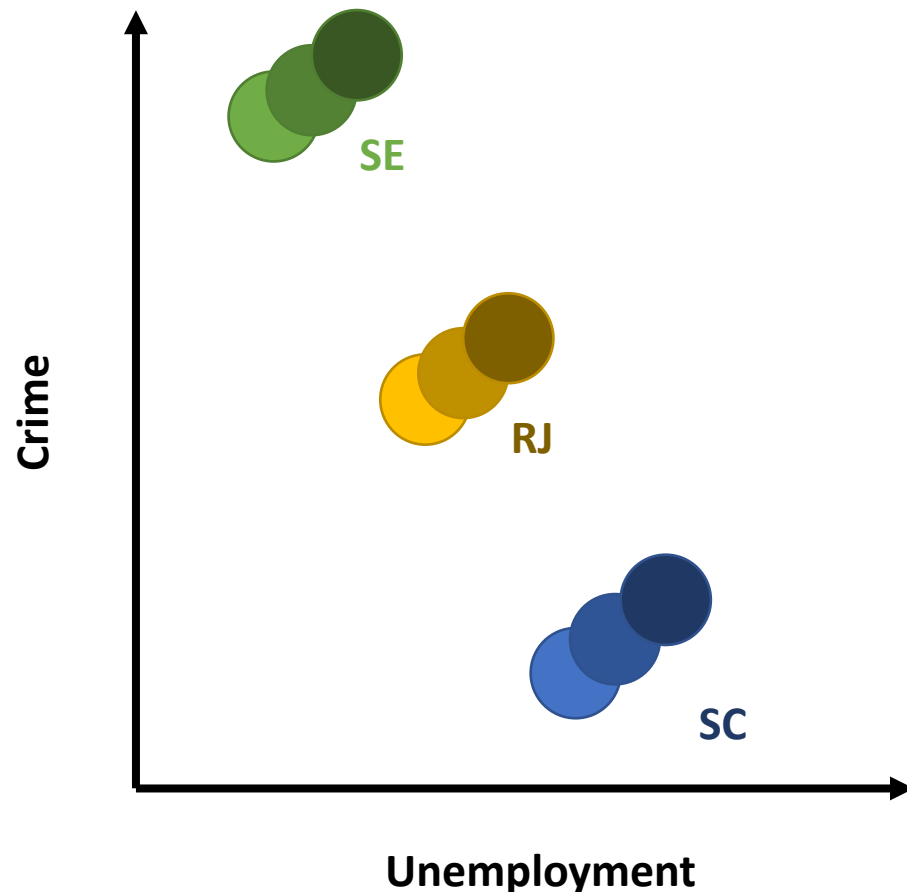
- **Benefits of FD**
- It is expected that unobserved heterogeneity is a problem for several types of data. FD benefits from repeated measures to eliminate this source of bias
- It is a way of remedying autocorrelation

First Differences

- **Limitations of FD**
- Everything that is invariable in time (or varies monotonically) is differentiated (e.g., geography, gender)
- Reduction in the number of observations by half
- Increase in the standard errors
- Consequently, FD is recommended as a drastic solution for cases of severe autocorrelation (Beck 2008) and when there is a reasonably large number of observations

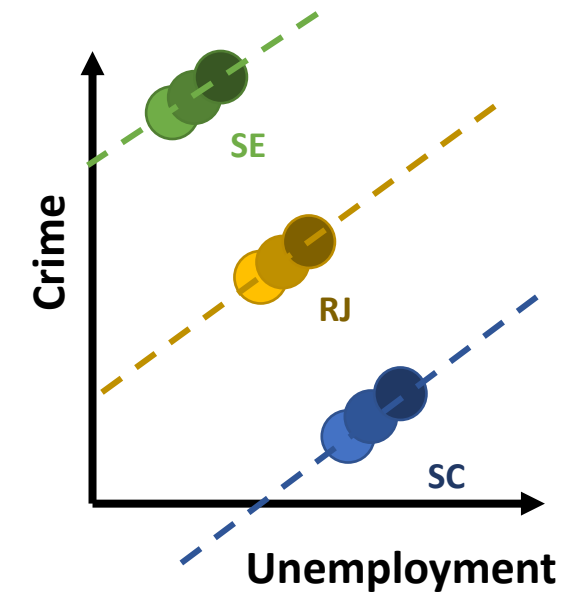
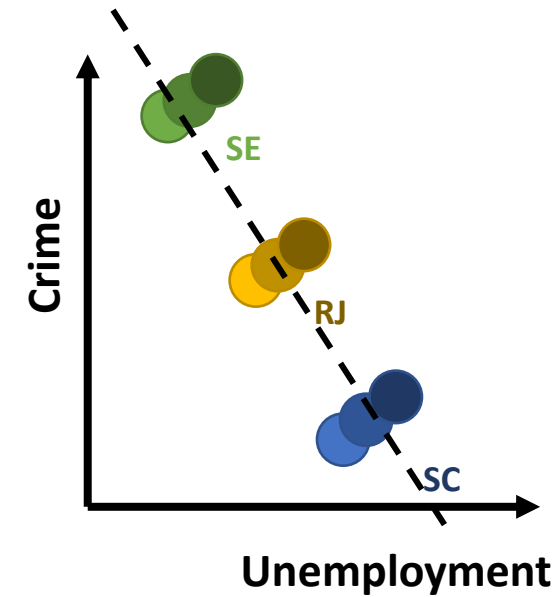
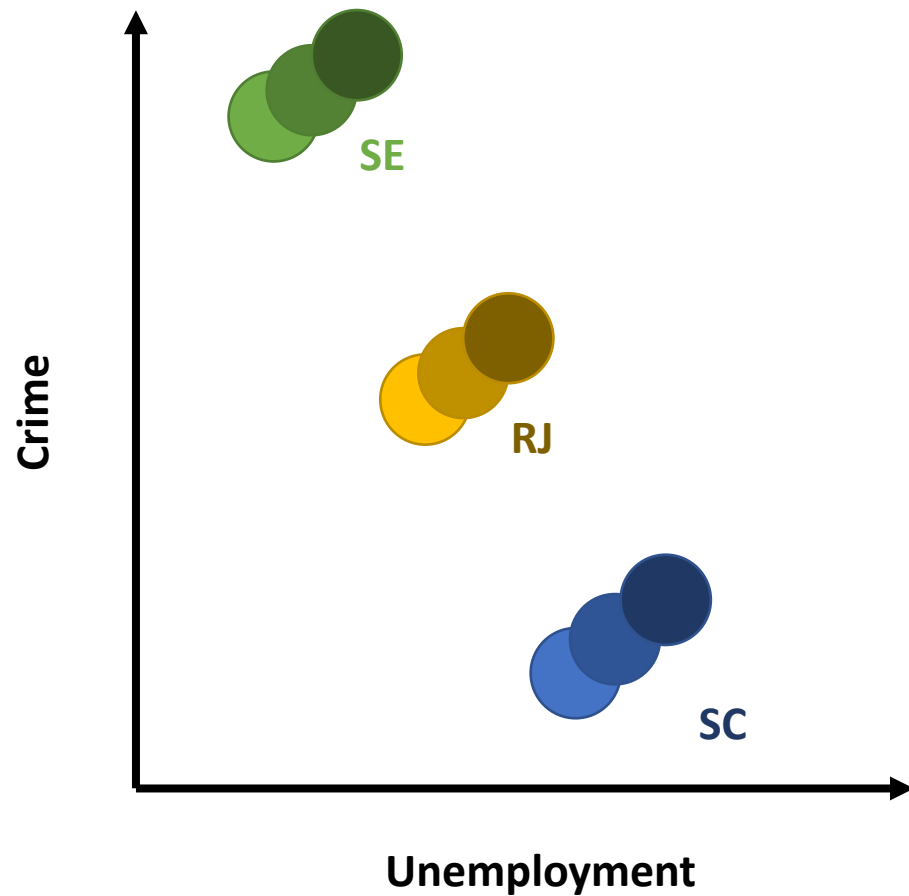
Fixed Effects

One or several regression lines?



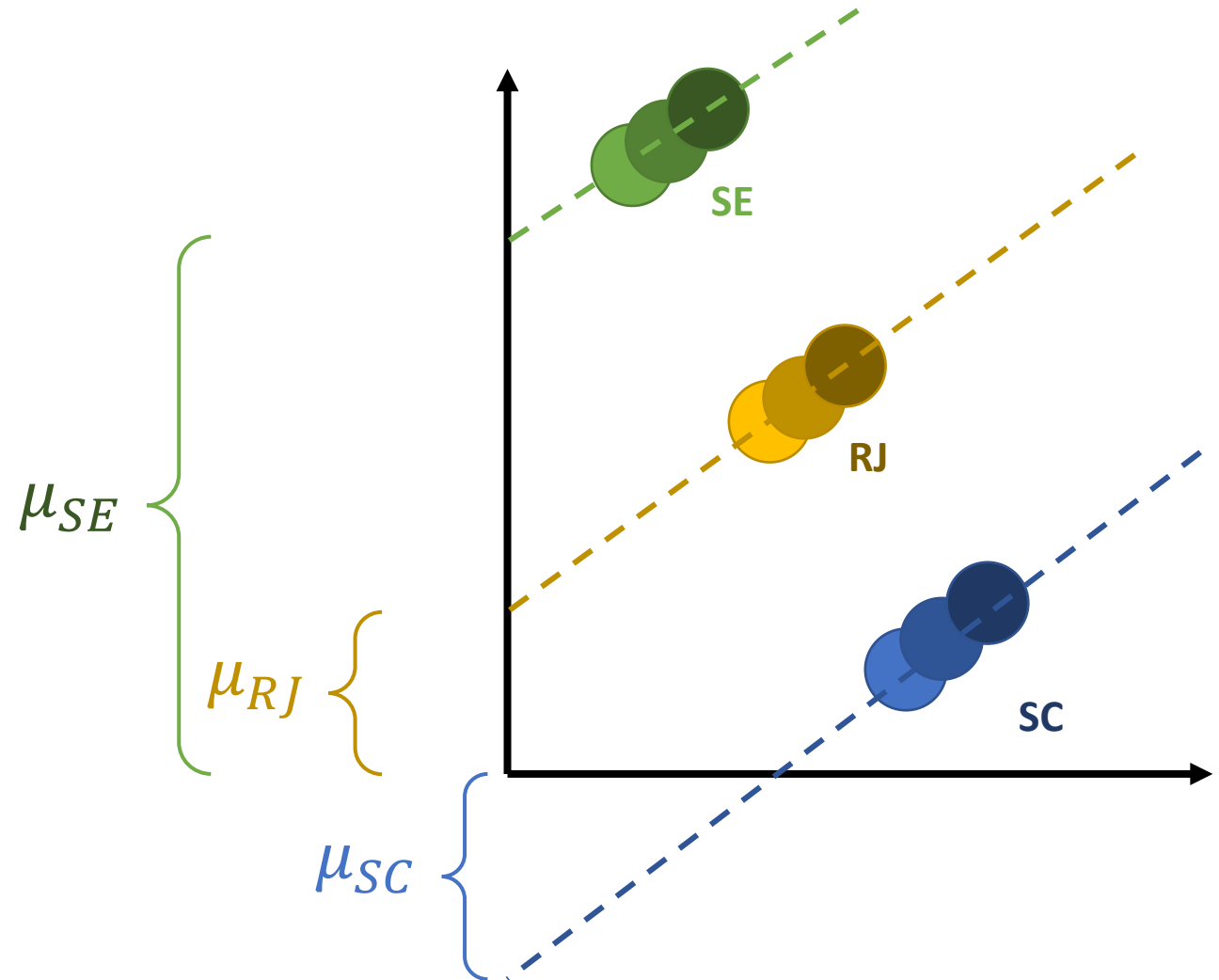
- Another way of understanding the relationship between unmodelled heterogeneity and pooling is to think if it is possible to draw a single regression line that adequately encompasses all cases
- In the example to the left, that single line for the whole population of cases would predict a **negative** relationship between unemployment and crime

One or several regression lines?



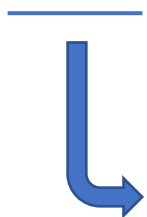
One or several regression lines?

- Each state should have its own intercept
- But the angle is the same, so the β is the same for all units



Recap.: breaking down the “composite error”

$$Y_{it} = \alpha_i + \beta X_{it} + \varepsilon_{it}$$



$$\varepsilon_{it} = \mu_i + \nu_{it}$$



**Fixed effect or unobserved
heterogeneity**



**Variable or idiosyncratic
effect**

$$Y_{it} = \alpha_i + \beta X_{it} + \mu_i + \nu_{it}$$

Standard model

$$Y_{it} = \alpha_i + \beta X_{it} + \mu_i + v_{it}$$

- Similarly to FD, FE also tries to erase individual heterogeneity (μ_i)

Subtracting from unit mean

$$\begin{array}{r} Y_{it} = \alpha_i + \beta X_{it} + \mu_i + v_{it} \\ - \left[\bar{Y}_i = \alpha_i + \beta \bar{X}_i + \mu_i + \bar{v}_i \right] \\ \hline \ddot{Y}_i = \beta \ddot{X}_i + \ddot{v}_i \end{array}$$

- “Time demeaning”, “group-mean centering” or “within transformation”:
subtracting each observation it from the mean for individual i for the whole of t

ID	Time	Homic. rate	Unempl. rate	Indiv. Mean (homic.)	Indiv. Mean (unempl.)
i	t	Y_{it}	X_{it}	\bar{Y}_i	\bar{X}_i
Acre	2011	22	5,4	27,2	8,2
Acre	2012	27,4	8	27,2	8,2
Acre	2013	30,1	9,6	27,2	8,2
Acre	2014	29,4	9,8	27,2	8,2
Bahia	2011	39,4	10,5	40,2	10,2
Bahia	2012	43,4	10,2	40,2	10,2
Bahia	2013	37,8	9,9	40,2	10,2
Bahia	2014	40	10	40,2	10,2
Santa Catarina	2011	12,8	3,6	12,8	3,3
Santa Catarina	2012	12,9	3,1	12,8	3,3
Santa Catarina	2013	11,9	3,4	12,8	3,3
Santa Catarina	2014	13,5	3,1	12,8	3,3

ID	Time	Intra-case dev. (homic.)	Intra-case dev. (unempl.)
i	t	$Y_{it} - \bar{Y}_i$	$X_{it} - \bar{X}_i$
Acre	2011	-5,2	-2,8
Acre	2012	0,2	-0,2
Acre	2013	2,9	1,4
Acre	2014	2,2	1,6
Bahia	2011	-0,8	0,4
Bahia	2012	3,3	0,0
Bahia	2013	-2,4	-0,3
Bahia	2014	-0,1	-0,2
Santa Catarina	2011	0,0	0,3
Santa Catarina	2012	0,1	-0,2
Santa Catarina	2013	-0,9	0,1
Santa Catarina	2014	0,7	-0,2



Within transformation

Time fixed effects and two-ways effects

- Instead of effects for each individual i , it is also possible to assume fixed effects for each time period t (e.g., exogenous shock, such as a financial crisis)

$$Y_{it} = \alpha_i + \beta X_{it} + \eta_t + v_{it}$$

- Combining fixed effects for individuals and time periods, we have the two-ways FE

$$Y_{it} = \alpha_i + \beta X_{it} + \mu_i + \eta_t + v_{it}$$

Advantages and disadvantages of FE

- **Advantages**
 - Intuitive (identical to POLS with dummies for all units i : Least Squares Dummy Variable, LSDV)
 - Eliminates unmodelled individual heterogeneity
 - Most popular approach for observational data
 - Allows visualization of the individual intercept of the units i
- **Disadvantages**
 - Similarly to FD, variables that do not vary with time are not computed (e.g., geography, gender)
 - Less parsimonious
 - Smaller efficiency and precision given that for each new unit i a new intercept is calculated, and only the within-unit variability is used (which is usually smaller than across-unit variability)
 - Sensitivity to the sample

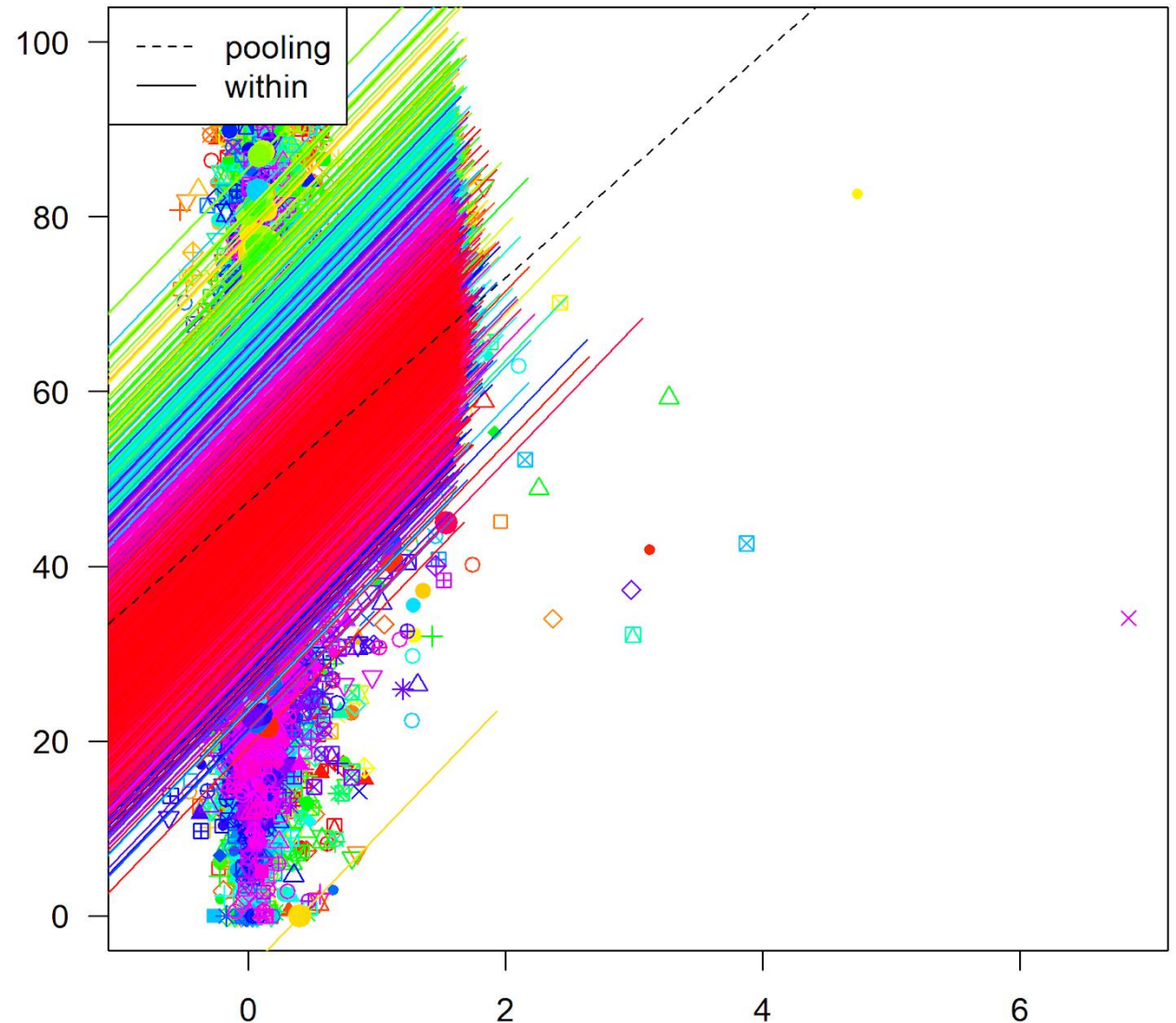
How do I know if I should use FE?

Diagnosing heterogeneity (II)

- (1) Substantive knowledge
- (2) Graphic analysis
- (3) Tests

Graphic analysis

- Goal is to check if it is preferable to cover the data by a single regression line and intercept (POLS) vs. several lines and individual intercepts
- *Pooling vs. within* plots are useful for bivariate relationships $X \sim Y$



Tests

There are two tests to recommend abandoning POLS:

F-Test

- POLS vs. FE
- F-test to check if individual effects are jointly significant
- Null hypothesis:
 - $\mu_i = \dots \mu_N = 0$
(if individual effects equal zero)
- Alternative hypothesis:
 - $\mu_i \neq \dots \mu_N \neq 0$

Breusch-Pagan LM

(not to be confused with the Breusch-Pagan test for heteroskedasticity!)

- POLS vs. RE
- Test to check if the variance of the individual effects is zero
- Null hypothesis:
 - $\sigma_{\mu_i}^2 = 0$
(variance of individual effects equals zero)
- Alternative hypothesis :
 - $\sigma_{\mu_i}^2 \neq 0$

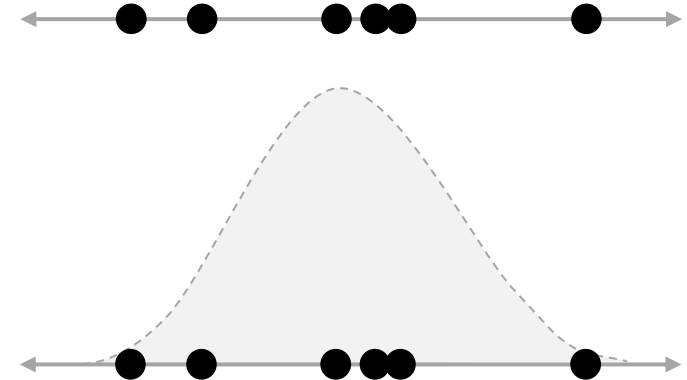
Tests

- Rejection of H_0 suggests that there are indeed significant fixed individual effects
- They can also test for time or two-ways effects

Random Effects

Randomness

- Individual heterogeneity (μ_i) can be treated like an isolated parameter (“case-by-case basis”) or as **random realizations** belonging to the same distribution
- Such singularity vs. generality of individual effects depends on the nature of the data
 - E.g., Exporting countries in the world economy (very peculiar and heterogenous cases) vs. Survey respondents (more similar and homogenous cases)



Randomness

- The RE approach assumes that individual effects μ_i come from the same **normal distribution**, with mean and variance calculated from the collected data.

$$\mu_i \sim N(0, \sigma_\mu^2)$$

Recap.: Omitted variable bias

- We eliminate the fixed effect (μ_i) via FD or FE because it is a form of omitted variable bias
- Remembering that: the omission of a variable “Z” generates bias if:
 - a. Z is correlated with Y; or
 - b. Z is correlated with X.
- Therefore, there will be **no** risk of bias if:
 - a. **The omitted variable Z is not correlated with Y; or**
 - b. **The omitted variable Z is not correlated with X**

RE

- RE assumes that there is no relationship between the individual effect μ and X , and therefore the effect does not need to be removed via FE or FD
- *But if we assume that μ and X are independent, why not just use Pooled OLS?*
 - Answer: The fixed effect μ is not considered a source of bias in the RE approach, but μ may still generate serial correlation in the errors; therefore, it is necessary to deal with it (refer to slides “How heterogeneity violates assumptions”)

Quasi-demeaning

- RE proposes “quasi-demeaning” the data: observations are subtracted from a *portion* of the intra-group mean

$$Y_{it} - \theta \bar{Y}_i = \alpha_i(1 - \theta) + \beta(X_{it} - \theta \bar{X}_i) + (\varepsilon_{it} - \theta \bar{\varepsilon}_i)$$

- The size of the fraction is given by θ , which is estimated:

$$\hat{\theta} = 1 - \sqrt{\frac{1}{1 + T\left(\frac{\sigma_{\mu}^2}{\hat{\sigma}_{\nu}^2}\right)}}$$

(in Wooldrige 2013, p.493)

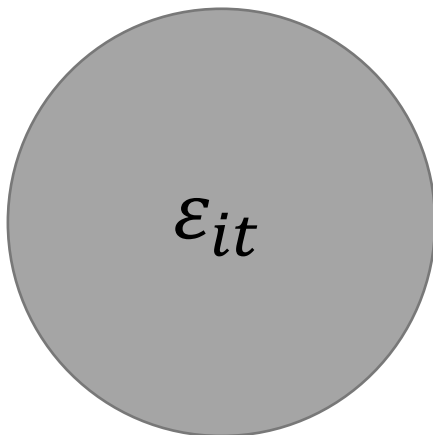
Theta

- θ varies between 0 and 1
- It may be interpreted as diagnostic of which component most contributed to the variance of the “composite error” (ε_{it}): if the fixed (μ_i) or idiosyncratic (ν_{it}) effect
- RE as a “middle-ground” between POLS and FE
 - When there is little variance in the fixed effects (μ_i), then $\theta \rightarrow 0$ and RE estimates will come closer to POLS
 - When there is a lot of variance in the fixed effects, then $\theta \rightarrow 1$ and RE will have results closer to FE

Therefore, the four approaches basically differ on what to do regarding unobserved heterogeneity

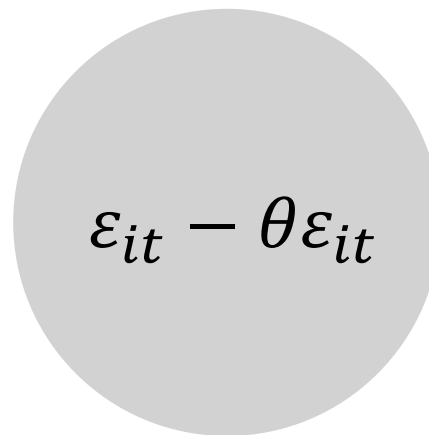
POLS

- Nothing is done about heterogeneity


$$\varepsilon_{it}$$

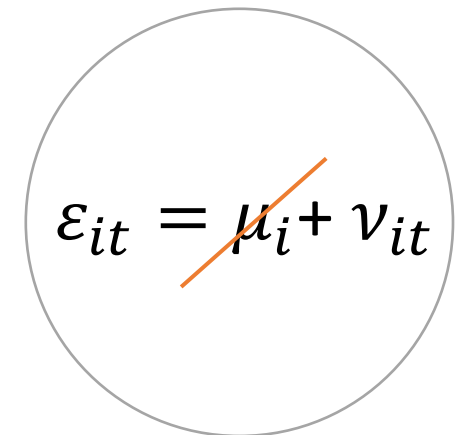
RE

- Heterogeneity is partially eliminated


$$\varepsilon_{it} - \theta \varepsilon_{it}$$

FE/FD

- Heterogeneity is completely eliminated


$$\varepsilon_{it} = \cancel{\mu_i} + \nu_{it}$$

Advantages and disadvantages of RE

- **Advantages**

- Given $X_i - \theta X_i \neq 0$ (if $\theta \neq 1$), it is possible to insert X with time constant values (e.g., geography, gender, etc.)
- More efficient (smaller standard errors)

- **Disadvantages**

- There will always be bias: the larger the $\text{cov}(X, \mu_i)$, the more severe it will be

FE or RE?

- Choice must be based on:
- **(1) Nature of the data**
 - Interchangeability: does the *name* of the cases matter? This can reveal if they should be considered random realizations or not
 - Observational (FE) vs. experimental (RE)
- **(2) Substantive interest in time invariant variables**
- **(3) Bias [β] vs. Inefficiency [standard error(β), p-value]**
 - FE: - bias, + inefficiency
 - RE: + bias, - inefficiency
- **(4) Hausman test**

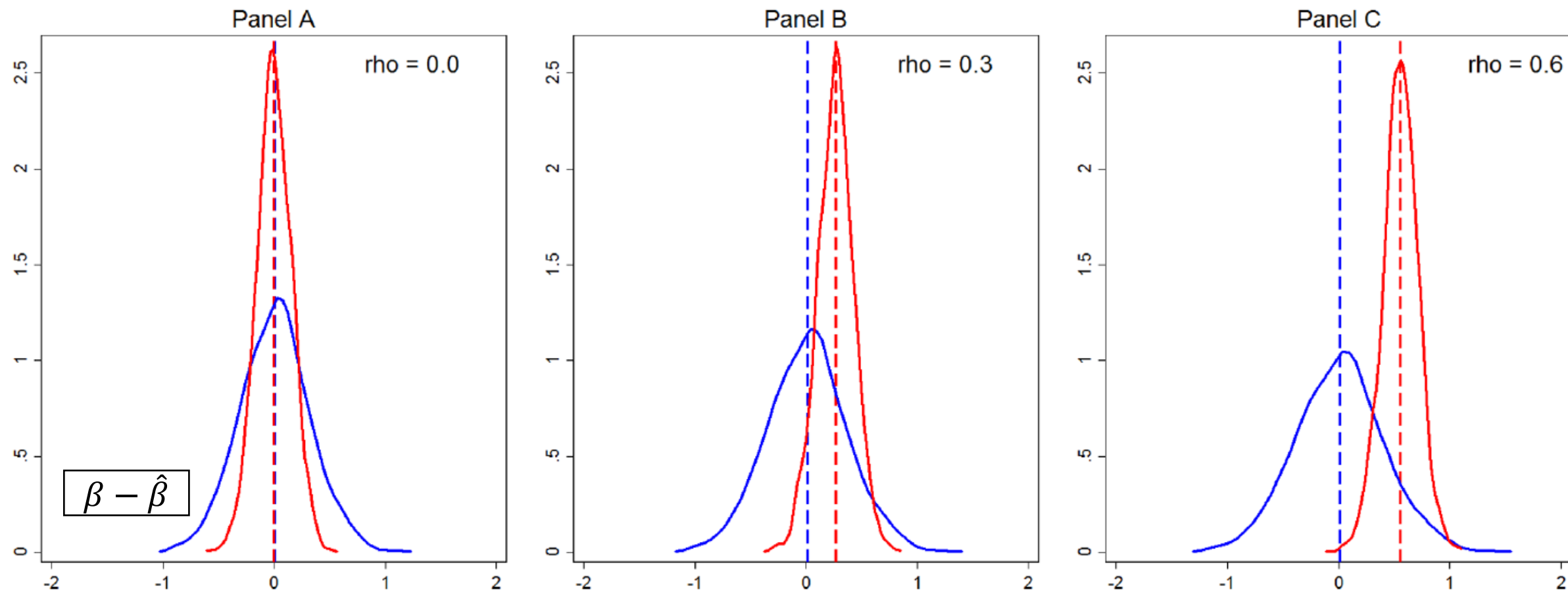


Figure 2. Distribution of errors. Red lines show the distribution of errors from RE estimation, while the blue lines show the distribution of errors from FE estimation. Each panel shows the correlation between the explanatory variable and the group-level effect set to a different value of ρ (0.0, 0.3, 0.6), increasing from left to right. Simulation based on the correct specification of a model with 50 groups and 10 observations per group. 50% of the variation of the outcome variable is explained by residuals, while only 10% of the variation in the explanatory variable is within-groups. doi:10.1371/journal.pone.0110257.g002

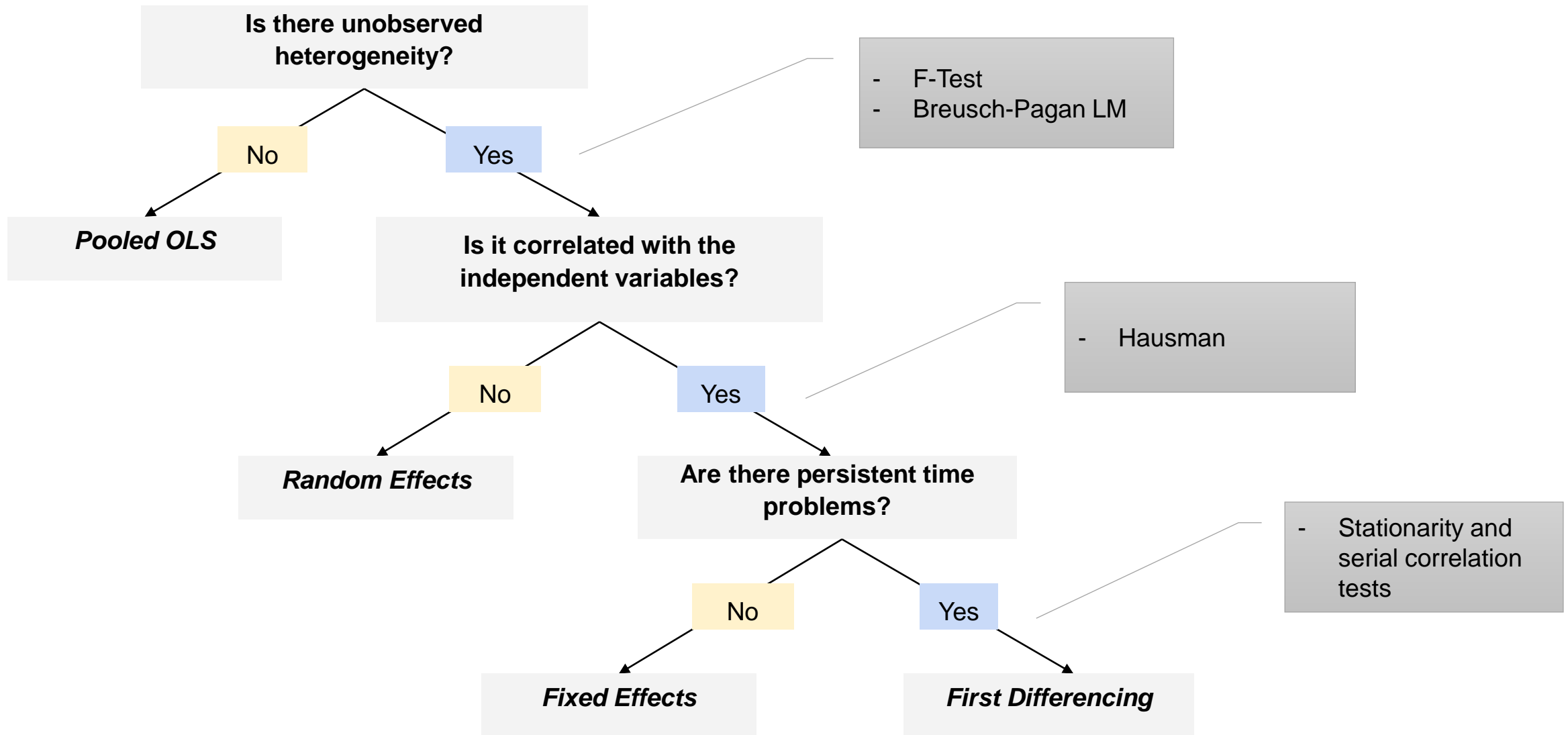
Source: Dieleman and Templin (2014, p. 5)

Hausman Test (1978)

- **H0: $\hat{\beta}_{fe} = \hat{\beta}_{re}$** (the 2 estimators are consistent)
 - Under H0, RE is preferred for being more efficient
- It is a test for well-specified models. It is assumed:
 - 1. That there is no misspecification
 - 2. That the idiosyncratic error v_{it} is independent of X_{it} and μ_i
 - 3. That the idiosyncratic error does not have serial correlation nor heteroskedasticity
 - 4. Usually, large samples
- If there are violations: there is a robust version of the test

Choosing an estimator

POLS, RE, FE, FD



Source: Mesquita et al. (2021)

Application:

What is the impact of local economic growth in the vote share of the incumbent in Brazilian presidential and municipal elections between 2000 and 2010?

Fernandes and Fernandes (2017)

Data https://osf.io/5yx7g/?view_only=ac1691cced8549238d6d6e0a9d2b7f7b

R package *plm* (Croissant and Milo 2008)

Database and model

```
##### OPENING DATABASE #####

library(haven) # package to read Stata data
library(plm) # package to execute the panel data models

DATA <- read_dta("fernandes_2017.dta") # Database file in the directory

DATA <- pdata.frame(BANCO, index = c("codibge", "year")) # here the database is
converted to the pdata.frame format
# to execute the models. In 'index', the data's spatial and temporal dimensions are
added.
# In this case, 'codibge' represents the spatial dimension (municipal code with the
Brazilian Institute of Geography and Statistics) and 'year' the election year.

##### FERNANDES FORMULA 2017 #####

form <- as.numeric(voteshare) ~ cresc+crescuf+crescbr+lpibreal+lpibuf+
  lpibbrasil+prefeitobasepresidente+persaude+lpop+leec+
  lheu+lse+laseps+ldesporc+ldespcor+linvest+ldespes
```

Create the four models: POLS, FE, RE, and FD

```
##### POOLED OLS #####
```

```
POLS <- plm(form, data = DATA, model = "pooling") # pooled model
```

```
##### FIXED EFFECTS #####
```

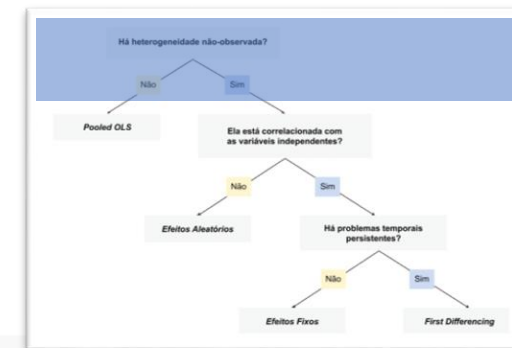
```
mode_fe <- plm(form, data = DATA, model = "within") # The FE model is executed with  
the model = 'within'
```

```
##### RANDOM EFFECTS #####
```

```
mode_re <- plm(form, data = DATA, model = "random") # The RE model is executed with  
the model = "random".
```

```
##### FIRST DIFFERENCES #####
```

```
mode_fd <- plm(form, data = DATA, model = "fd") # The FD model is executed with the  
model = "fd".
```



Is there unobserved heterogeneity?

- **F-test**
- FE is preferable to POLS
- **BP**
- RE is preferable to POLS

POLS F-TEST

```
pFtest(mode_fe, POLS) # F-test with the fixed effects and Pooled models
```

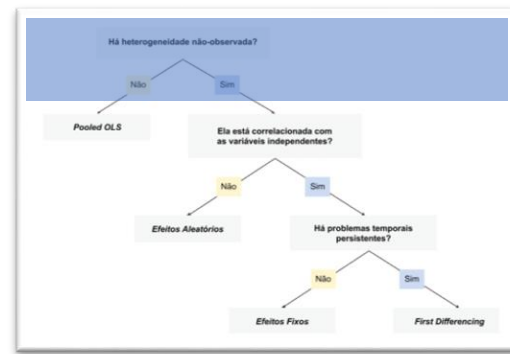
```
##
## F test for individual effects
##
## data: form
## F = 1.4214, df1 = 5548, df2 = 20786, p-value < 2.2e-16
## alternative hypothesis: significant effects
```

BREUSCH-PAGAN TEST

```
plmtest(POLS, type="bp", effect = "individual") # Breusch-Pagan test + individual effects
```

```
##
## Lagrange Multiplier Test - (Breusch-Pagan) for unbalanced panels
##
## data: form
## chisq = 107.15, df = 1, p-value < 2.2e-16
## alternative hypothesis: significant effects
```

Is there unobserved heterogeneity?

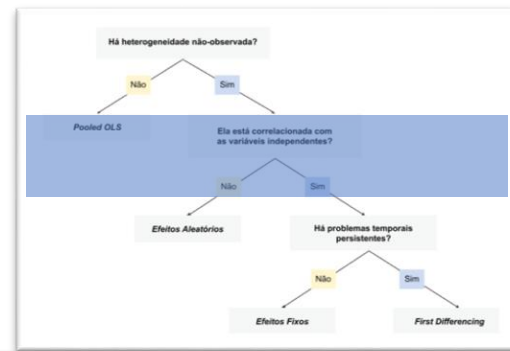


- **BP**
- Not only is there individual heterogeneity, but also temporal
- “Two-ways” effects must be used

```
plmtest(POLS, type="bp", effect = "twoways") # Breusch-Pagan test + individual and time effects
```

```
##  
## Lagrange Multiplier Test - two-ways effects (Breusch-Pagan) for  
## unbalanced panels  
##  
## data: form  
## chisq = 2762.9, df = 2, p-value < 2.2e-16  
## alternative hypothesis: significant effects
```

Is it correlated with the IVs?



- Hausman
- The FE model is preferable to RE

HAUSMAN TEST

```
phptest(mode_fe_2w, mode_re_2w) # Hausman test (fixed and random models)
```

```
##
```

```
## Hausman Test
```

```
##
```

```
## data: form_d
```

```
## chisq = 851.98, df = 15, p-value < 2.2e-16
```

```
## alternative hypothesis: one model is inconsistent
```

```
phptest(form_d, data= DATA, method="aux", vcov=vcovHC) # robust version of the test
```

```
##
```

```
## Regression-based Hausman test, vcov: vcovHC
```

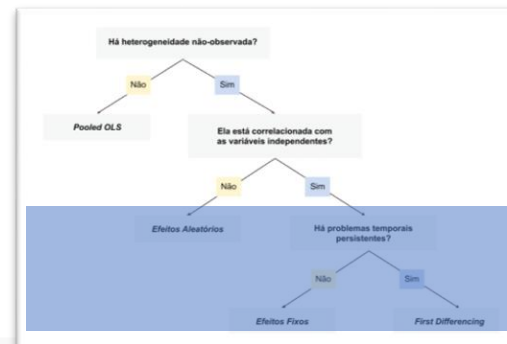
```
##
```

```
## data: form_d
```

```
## chisq = 810.83, df = 18, p-value < 2.2e-16
```

```
## alternative hypothesis: one model is inconsistent
```

Are there serial correlation issues?



- There is serial correlation in both the FE and the FD models
- Final decision: adopt FE or FD, with robust standard errors

TESTING SERIAL CORRELATION FOR FE

```
pwartest(mode_fe_2w) #pwartest for FE model
```

```
##
## Wooldridge's test for serial correlation in FE panels
##
## data: mode_fe_2w
## F = 71.527, df1 = 1, df2 = 20801, p-value < 2.2e-16
## alternative hypothesis: serial correlation
```

TESTING SERIAL CORRELATION FOR FD

```
pwfdtest(mode_fd_dic) #pwfdtest for FD model
```

```
##
## Wooldridge's first-difference test for serial correlation in panels
##
## data: mode_fd_dic
## F = 6004.9, df1 = 1, df2 = 15307, p-value < 2.2e-16
## alternative hypothesis: serial correlation in differenced errors
```

(Excerpt from the full table)

	Vote share				
	POLS	RE (2W)	FE (2W)	FE (2W, Robust.)	FD
Municipal growth	1.683 ^{***} (0.608)	1.098 ^{***} (0.040)	1.148 [*] (0.652)	1.148 (1.517)	0.190 (0.669)
State growth	0.095 ^{***} (0.027)	0.086 ^{***} (0.002)	-0.058 [*] (0.030)	-0.058 (0.163)	-0.144 ^{***} (0.033)
National growth	3.660 ^{***} (0.090)	3.709 ^{***} (0.088)			4.231 ^{***} (0.091)
Mayor Base Presid.	1.211 ^{***} (0.208)	0.742 ^{***} (0.014)	1.289 ^{***} (0.235)	1.289 ^{***} (0.409)	1.136 ^{***} (0.254)
% Budget Health	0.040 ^{***} (0.014)	0.121 ^{***} (0.001)	0.044 [*] (0.023)	0.044 (0.035)	0.058 ^{***} (0.019)
Expend. Educ. Cult. (log)	6.917 ^{***} (0.358)	6.461 ^{***} (0.025)	1.543 ^{***} (0.525)	1.543 (1.284)	0.162 (0.553)
Municipal Election	1.074 ^{***} (0.277)	0.828 ^{***} (0.246)			-0.915 ^{***} (0.264)
Intercepto	102.026 ^{**} (43.701)	152.778 ^{***} (37.792)			-4.350 ^{***} (0.373)
N Obs.	26,352	26,352	26,352		20,803
R ²	0.321 ^{***}	0.319 ^{***}	0.039 ^{***}		0.254 ^{***}
F Stat	691.129 ^{***} (df = 18; 26333)	12,309.190 ^{***}	55.646 ^{***} (df = 15; 20783)		392.979 ^{***} (df = 18; 20784)

*** p < 0,001; ** p < 0,01; *p < 0,05

- POLS coefficients close to a RE, and FE's close to FD
- Change in direction of some variables ("state growth") shows that unmodelled factors generated bias
- Change in magnitude and significance in others
- Variables fixed in the year for all municipalities ("national growth", *dummy* "municipal election") are not estimated by FE 2W
- From the left to the right of the table, the result of ignoring heterogeneity (POLS) and eliminating it entirely (FE/FE) can be seen

Extra:

Handling the spatial and temporal
problems of panel data models

Reviewing the assumptions of multiple regression analysis

Based on Wooldridge (2013)

	Assumption	Notation	Problem	Common causes	Possible solution
A 1	Linear relationship	$Y = \alpha + \beta X + \varepsilon$	Bias	Misspecification of the model	Revise model
A 2	Random sample (absence of autocorrelation)	$\text{Corr}(\varepsilon_i, \varepsilon_j) = 0$ $i \neq j$	Bias	Collection procedure; Nature of the data	Randomization; Revise model to eliminate correlation
A 3	Absence of perfect collinearity	$\text{Corr}(X_1, X_2) \neq \pm 1$	Impossible to estimate $\hat{\beta}$; Bias; Inflates $EP(\hat{\beta})$, p-value	X_1 and X_2 do not occur independently; Small N	Revise model; Increase N
A 4	Exogeneity	$E(\varepsilon x) = 0$	Bias	Omitted variable; Systematic measuring error; Truncation	Revise model
A 5	Homoskedasticity	$\text{Var}(\varepsilon x) = \sigma^2$	Invalidates $EP(\hat{\beta})$, p-value	Misspecification of the model; Systematic measuring error; Small N	Revise model; Increase N; Robustification
A 6	Normal distribution of the errors	$\varepsilon \sim \text{Normal}(0, \sigma^2)$	Invalidates $EP(\hat{\beta})$, p-value	Misspecification of the model; Systematic measuring error	(less important with a large N)

	Assumption
A 1	Linear relationship
A 2	Random sample (absence of autocorrelation)
A 3	Absence of perfect collinearity
A 4	Exogeneity
A 5	Homoskedasticity
A 6	Normal distribution of the errors

$\hat{\beta}$ is free of bias

$\hat{\beta}$ has the smallest variance

Gauss-Markov Assumptions

B est

L inear

U nbiased

E stimator

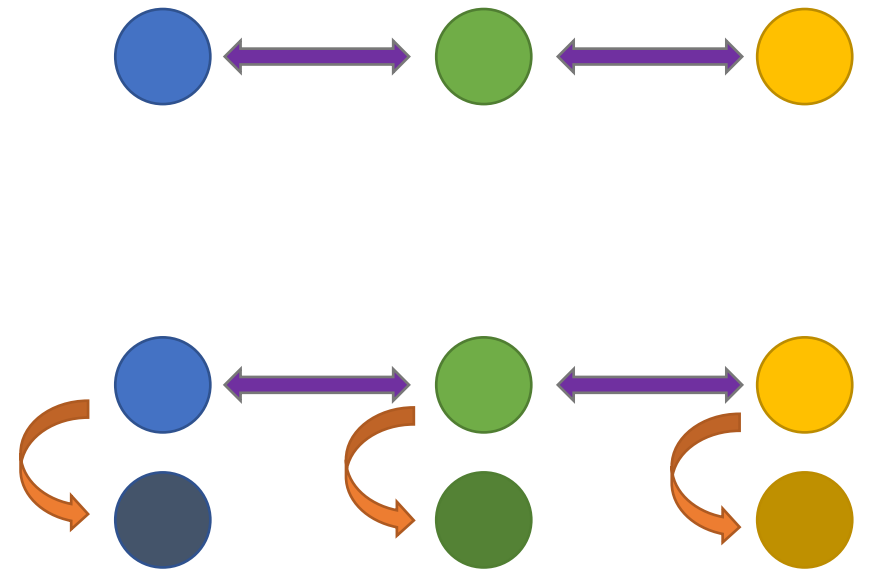
$\hat{\beta}$ has the smallest variance of any unbiased estimator

Classic linear regression assumptions

	Assumption	Notation	Why is it a problem to violate it?	Usually affect:		Dynamic issues
A 1	Linear relationship	$Y = \alpha + \beta X + \varepsilon$	Bias		1	Serial correlation of residuals
A 2	Random sample (absence of autocorrelation)	$\text{Corr}(\varepsilon_i, \varepsilon_j) = 0$ $i \neq j$	Bias		2	Non-stationarity
A 3	Absence of perfect collinearity	$\text{Corr}(X_1, X_2) \neq \pm 1$	Impossible to estimate $\hat{\beta}$; Bias; Inflates $EP(\hat{\beta})$, p-value			Spatial issues
A 4	Exogeneity	$E(\varepsilon x) = 0$	Bias		3	Heterogeneity
A 5	Homoskedasticity	$\text{Var}(\varepsilon x) = \sigma^2$	Invalidates $EP(\hat{\beta})$, p-value		4	Panel heteroskedasticity
A 6	Normal distribution of the errors	$\varepsilon \sim \text{Normal}(0, \sigma^2)$	Invalidates $EP(\hat{\beta})$, p-value		5	Contemporaneous correlation of the errors
					6	Complex dependence structures

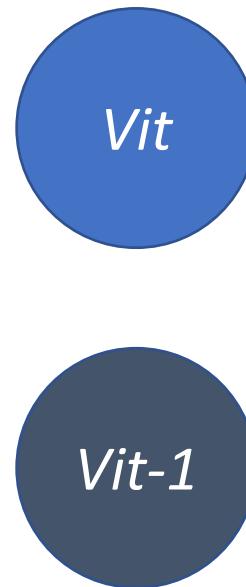
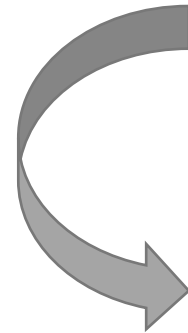
The problem with dependence

- In a simple cross-sectional regression, we just need to monitor 1 dimension to check the independence between observations: **spatial**
- In panel data, there are 2 dimensions: **spatial** and **temporal**



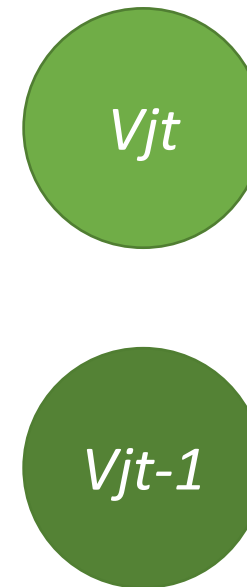
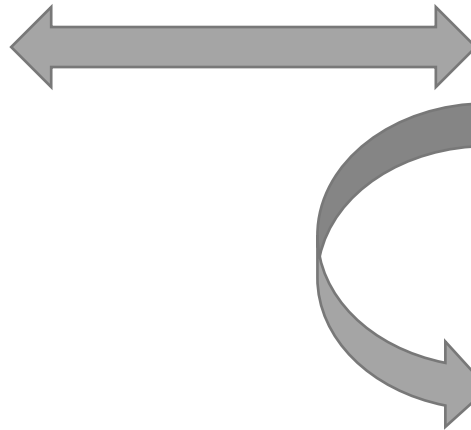
Temporal dependencies

Are the residuals of a unit i at t related to the ones of the same unit at $t-1$?



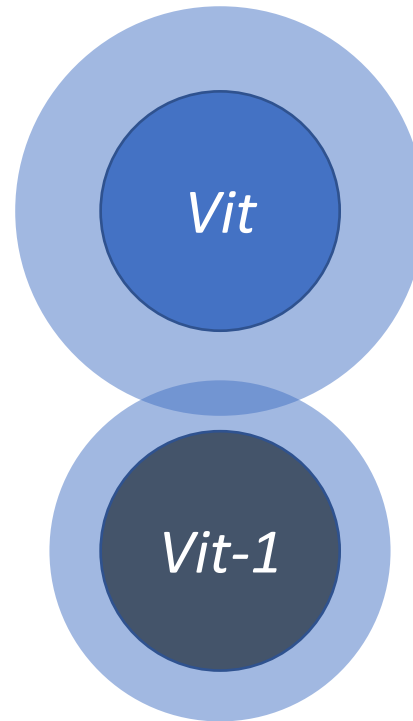
Spatial dependencies

Are the residuals of unit i at t related to the ones of another unit j at t ?



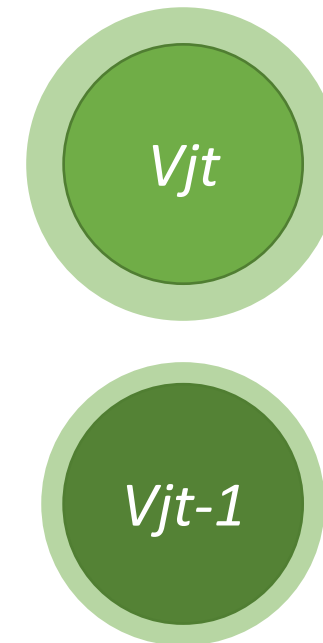
Temporal dependencies

Is the variance of moment t different from moment $t-1$?



Spatial dependencies

Is the variance of the residuals of unit i different from the variance of unit j ?



Temporal issues

1. Non-
stationarity

2.
Autocorrelation

Usually violate:

A 2 Absence of
autocorrelation

A 4 Exogeneity

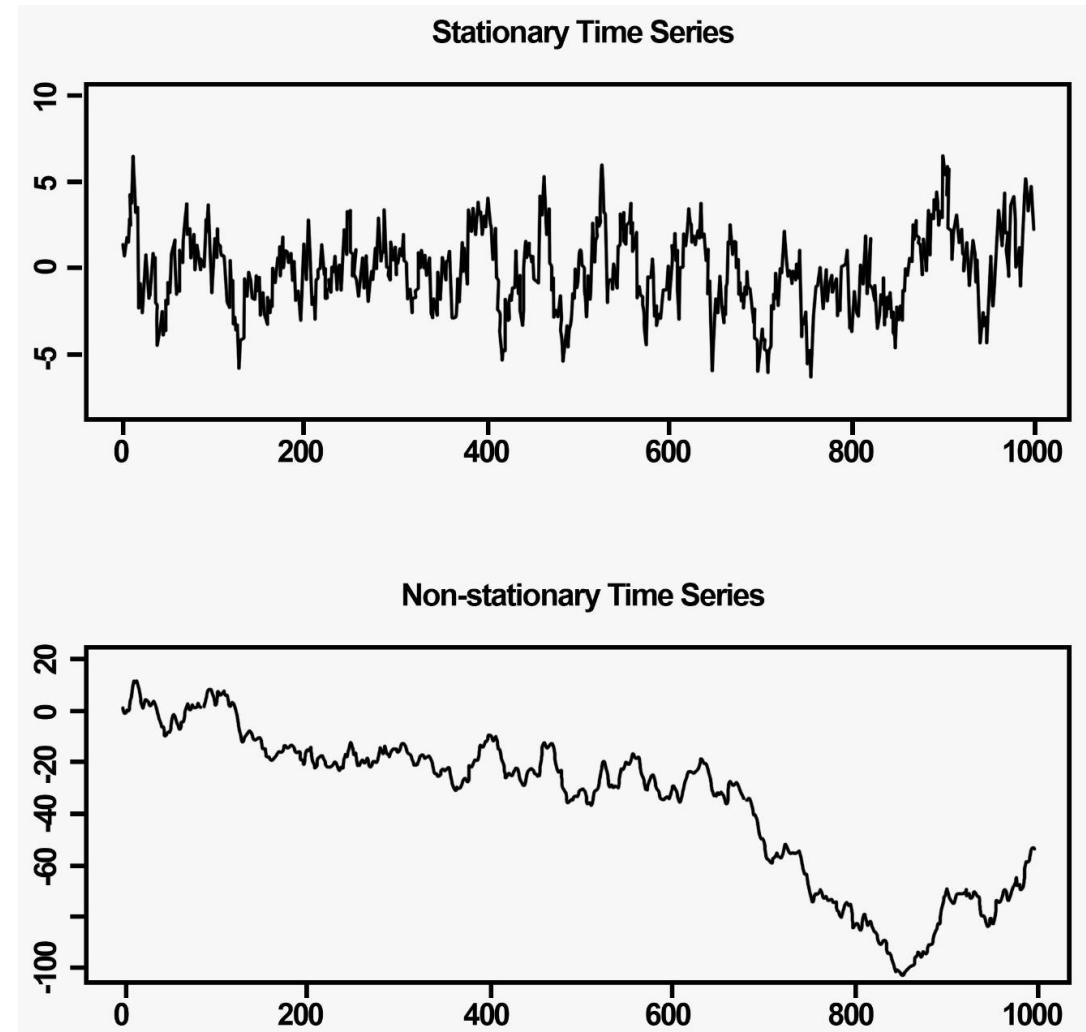
A 5
Homoskedasticity

Temporal issues

- **Initial considerations:**
- It is recommended to first address temporal issues and then spatial ones
- They will be more severe the larger T is in relation to N
- Time series techniques applicable to $T > \sim 20$

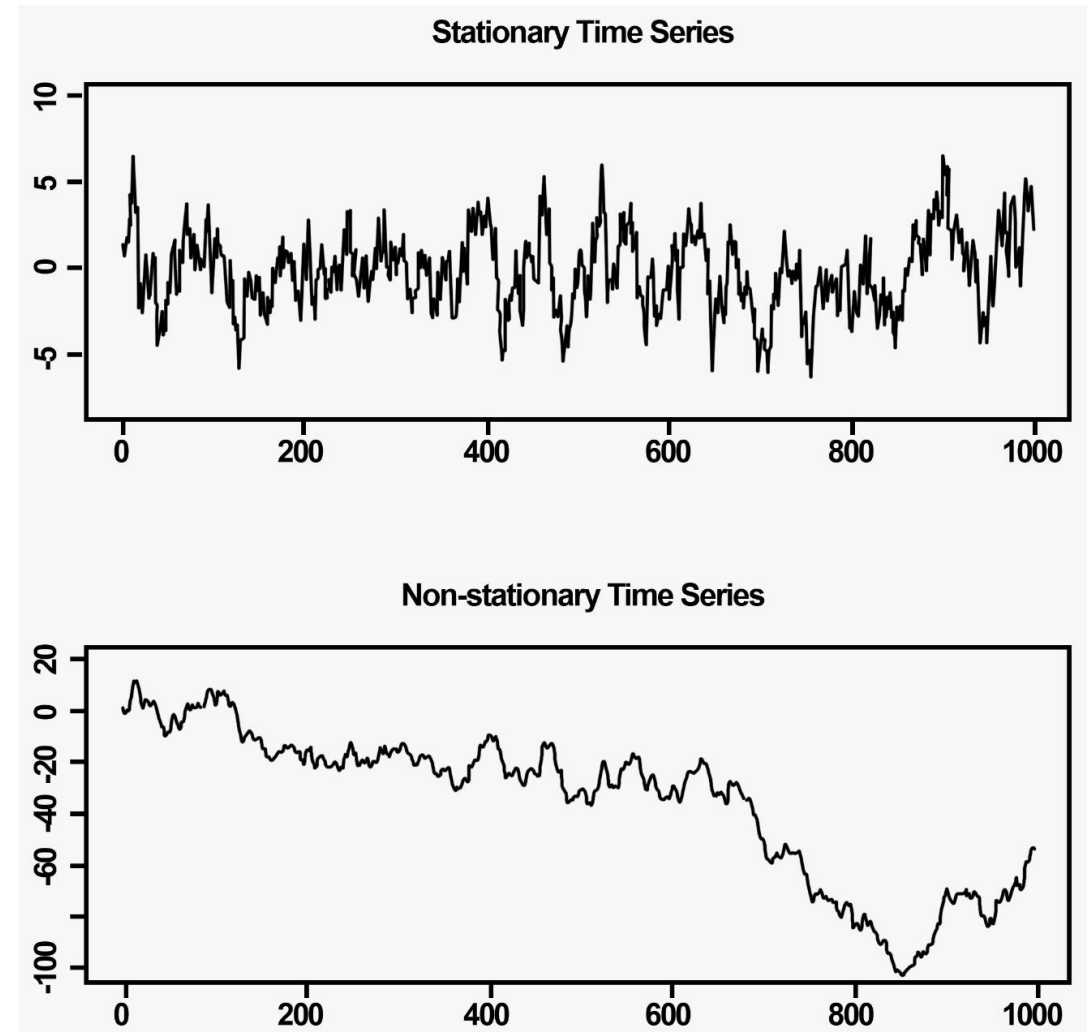
1. Non-stationarity: definition

- Stationary time series: mean and variance constant over time
 - Although values oscillate, they return to the mean
- Non-stationary: mean and variance are not constant
 - The series does not tend to return to a previous mean after deviations
 - Effects of persistent shocks
- **Ask yourself:**
 - *“Does the series return to the origin?”*
 - *“Is it similar at every segment?”*



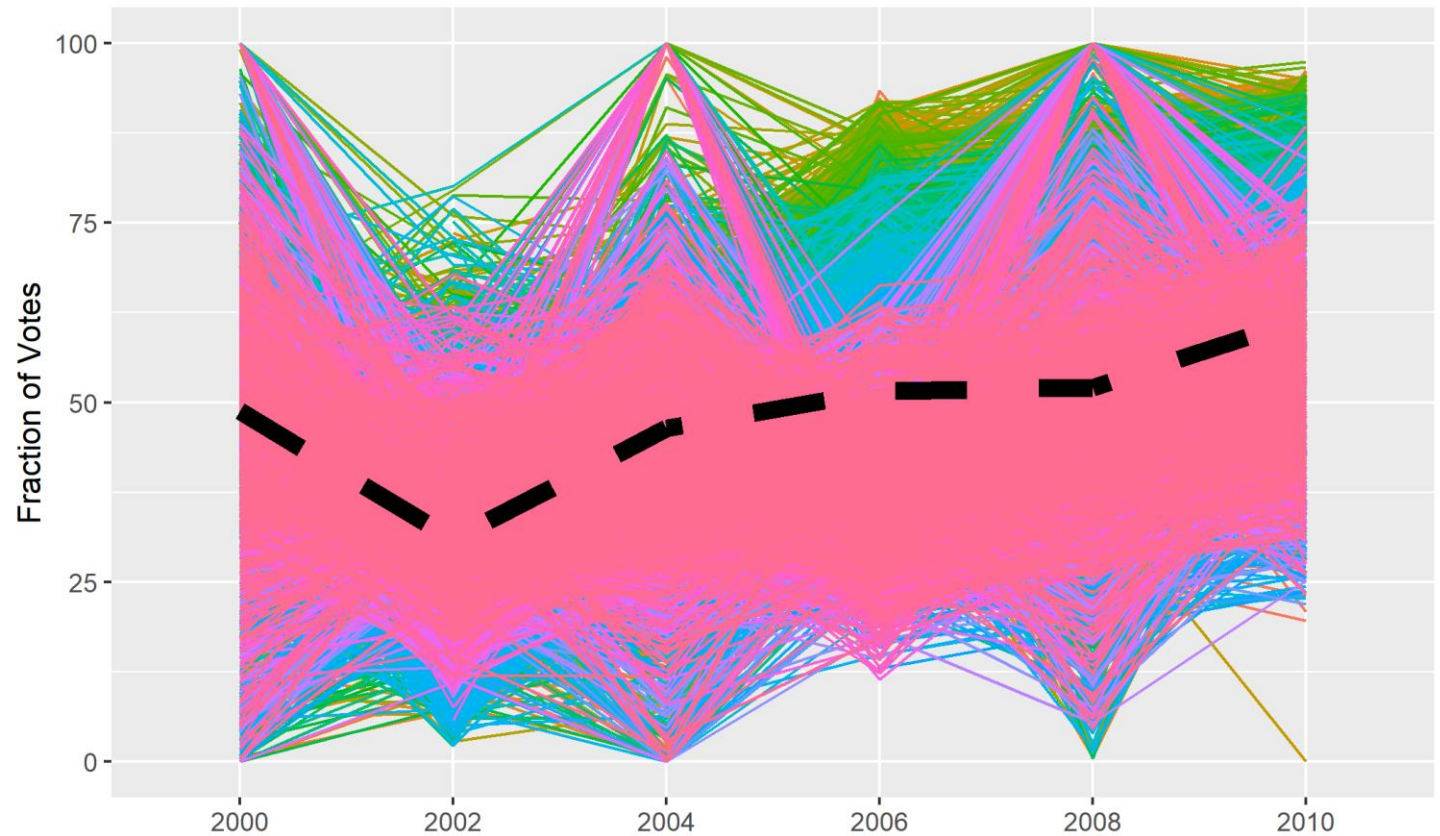
1. Non-stationarity: definition

- Why is it a problem?
 - Absence of autocorrelation
 - Exogeneity
 - Homoskedasticity
- Common causes of non-stationarity
 - Trends
 - Seasonality



1. Non-stationarity: diagnostics

- Diagnostics
 - 1. Visual inspection (time series plots)
 - 2. Unit root tests for longer series
 - 3. ACF and PACF plots, also for longer series



1. Non-stationarity: solutions

- Changing the operationalization of the variable (e.g., adjusting for inflation over time)
- Adding a time trend
 - Captures the effect of “the passage of time”
 - But are not estimated for FE or FD models (monotonic growth)
- Running FD

2. Serial correlation of the residuals: definition

- Occurs when the residual v_{it} is correlated with v_{it-1}
- Common in panel data, since observations are repeated
- Typically, autoregressive processes of the 1st order or ou “AR(1)”
- Why is it a problem?
 - Autocorrelation
 - Exogeneity
 - Distribution of the errors

2. Serial correlation of the residuals: diagnostics

- Diagnostics
 - 1. Regression of the residuals (for AR1)
 - 2. Tests

Diagnostics of serial correlation type AR(1)

- Regression of v_{it} in v_{it-1} (POLS) (Wooldridge)
- $\varepsilon_{it} = \rho\varepsilon_{it-1} + e_{it}$
- Check if ρ is significant

plm tests for serial correlation (usually: H_0 = there is no autocorr.)

POLS	<ul style="list-style-type: none">• Manual regression $\varepsilon_{it} \sim \varepsilon_{it-1}, \rho$• pwtest(pols) # Generic test by Wooldridge for the absence of unobserved effects in the residuals• pbsytest(pols) # Test for AR(1) or RE(1) (Bera, Sosa, and Yoon), AR(1) and RE(1) (Batalgi and Li)<ul style="list-style-type: none">• obs: <code>pbsytest(pols, test=c("ar", "re", "j"))</code>
RE	<ul style="list-style-type: none">• pbltest(formula, data=.) # Test by Batalgi and Li for AR(1)/MA(1)
FE	<ul style="list-style-type: none">• pwartest(fe) # Test by Wooldridge for AR(1) in FE
FD	<ul style="list-style-type: none">• pwfdtest(fd) # Test by Wooldridge for AR(1) in FD
General	<ul style="list-style-type: none">• pbgtest(model) # Breusch-Godfrey/Wooldridge test for serial correlation• pdwtest(model) # Durbin Watson test<ul style="list-style-type: none">• obs: BG and DW for FE require long T• pbnftest(model) # Generalization of DW (Bhargava, Narendranathan, and Franzini)<ul style="list-style-type: none">• obs: does not report a p-value. Usually:• DW: 0 – 2 positive autocorr.; 2 no autocorr.; 2 – 4 negative autocorr.

2. Serial correlation of the residuals: solutions

- Fixed time effects
- Introducing a lagged dependent variable (LDV)
 - However, it can diminish the effects of the other coefficients
- Robustification of the standard errors
- Run FD

Spatial issues

3. Heterogeneity

(refer to “unobserved heterogeneity” slides)

4. Panel heteroskedasticity

5. Contemporaneous correlation of the errors

6. Complex dependence structures

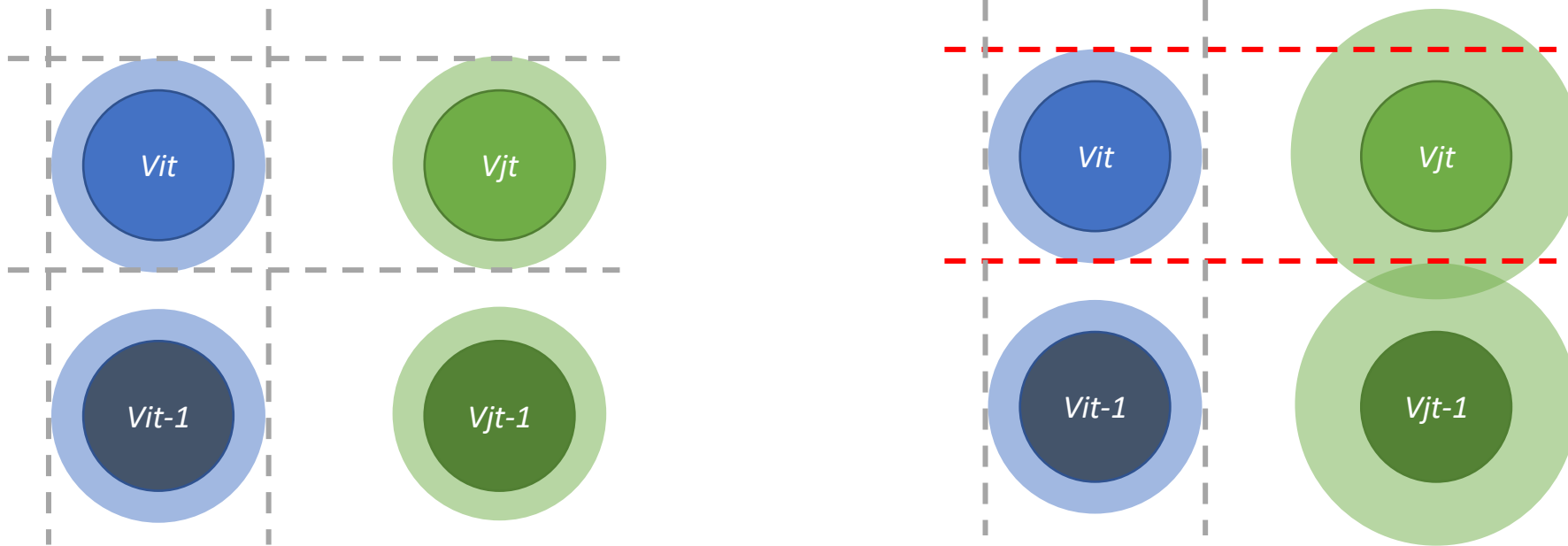
Usually violate:

A 4 Exogeneity

A 5
Homoskedasticity

A 6 Distribution of the errors

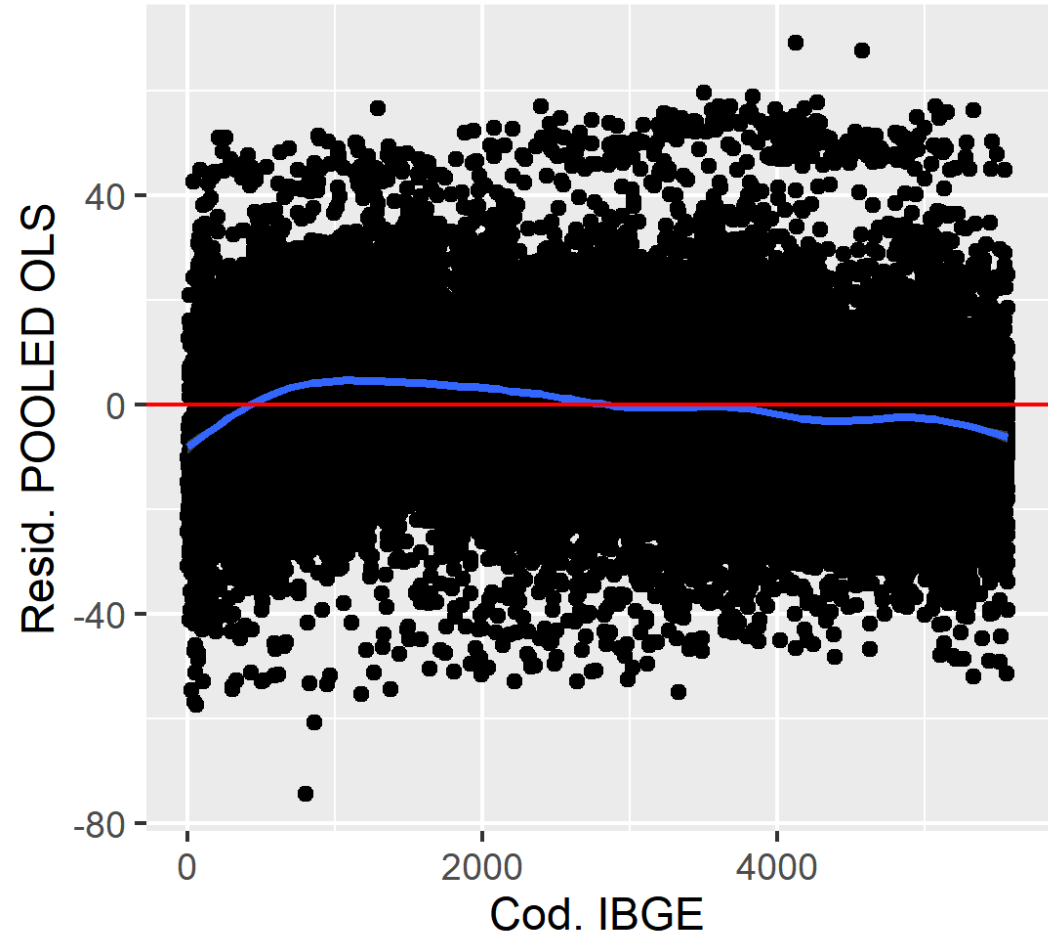
4. Panel heteroskedasticity: definition



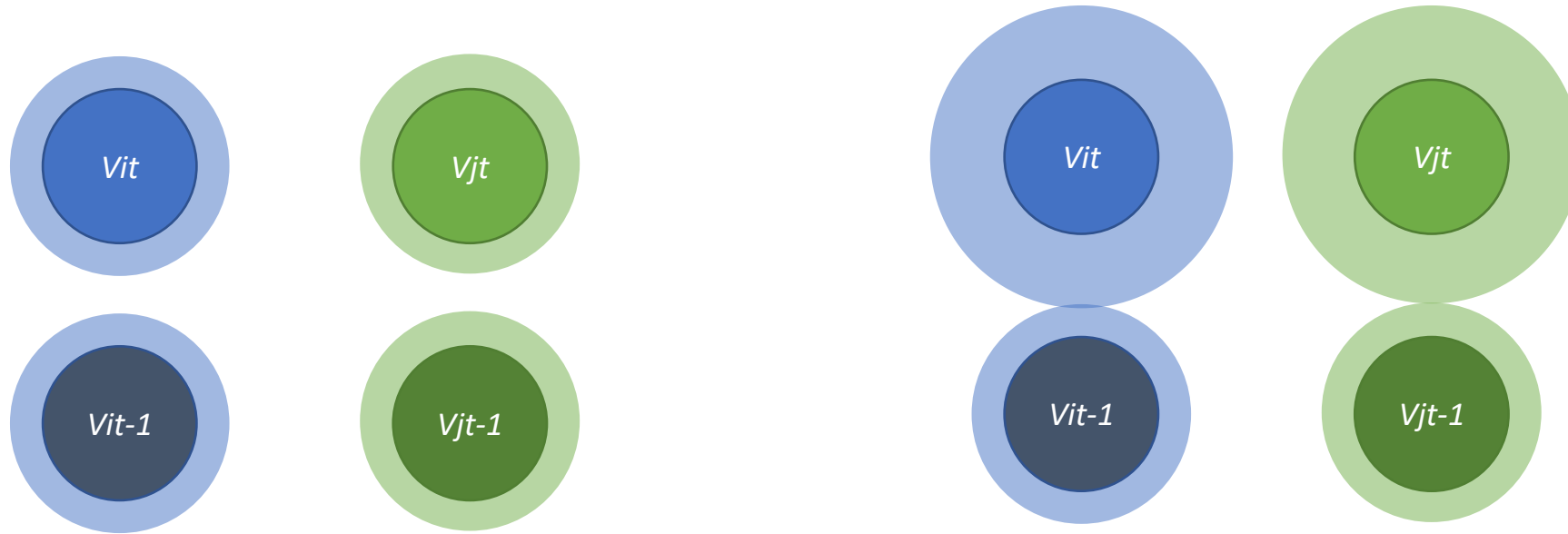
- Homoskedasticity requires that the variance in the errors be the same for the whole sample
- Panel heteroskedasticity occurs when residuals have constant variance over time *within units*, but inconstant *across units*

4. Panel heteroskedasticity: diagnostics

- Visual inspection:
 - Grouping residuals by spatial unit
- Tests



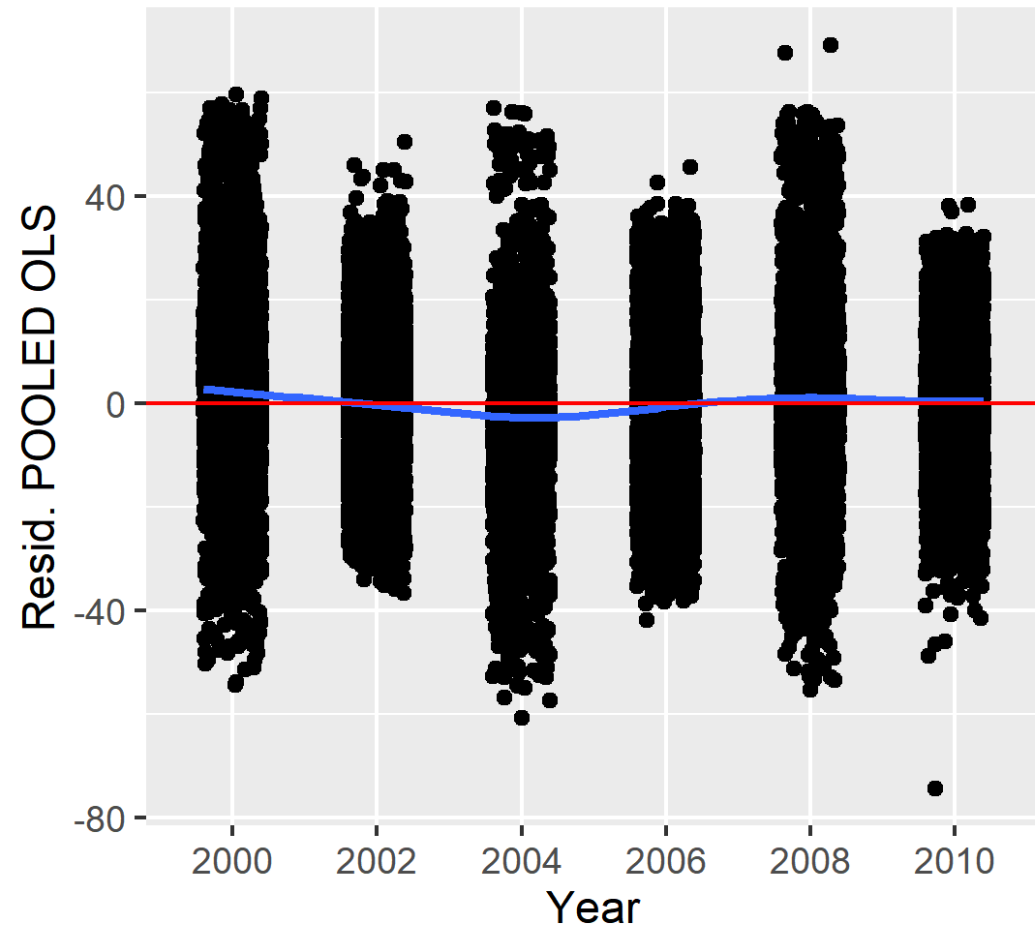
5. Contemporaneous correlation of the errors: definition



- The residual of a case is correlated with the residual of other cases for the same moment in time
- Exogenous shocks affect all units in the same way at instant t

5. Contemporaneous correlation of the errors: diagnostics

- Residuals grouped by year



5. Contemporaneous correlation of the errors: solutions

- Add dummies for the “shock years”
- Time or two-ways effects
- Robustification

6. Complex dependence structures

- There is some process of spatial diffusion (e.g., geographical units that are close will experience similar impact)

plm tests for cross-sectional dependence (usually: H_0 = there is not cross-sectional dependence)

Long T, short N	<ul style="list-style-type: none">• <code>pcdtest(model, test="lm")</code> # Breusch-Pagan LM test
Long T, long N	<ul style="list-style-type: none">• <code>pcdtest(model, test="sclm")</code> # scaled Breusch-Pagan LM
General	<ul style="list-style-type: none">• <code>pcdtest(model, test="cd")</code> # cross-sectional dependence Pesaran test<ul style="list-style-type: none">• obs.: standard test in <code>pcdtest()</code>• <code>pcdtest(model, test="rho")</code> <code>pcdtest(model, test="absrho")</code> # values of the correlation coefficients ρ between pairs of observations

Robustification

- Standard errors and p-value need to be made robust against
 - (1) spatial issues (heteroskedasticity, cross-sectional dependence)
 - (2) temporal issues (autocorrelation)

Variance x covariance matrices in the sandwich, plm packages for robustification

	Standard errors robust against...
vcovHC	<ul style="list-style-type: none">• Heteroskedasticity (same as cross-sectional data)
vcovNW	<ul style="list-style-type: none">• Heteroskedasticity and serial correlation
vcovSCC	<ul style="list-style-type: none">• Heteroskedasticity and serial correlation (for $T > 20$)
vcovDC	<ul style="list-style-type: none">• Double-clustering
vcovBK	<ul style="list-style-type: none">• Panel Corrected Standard Errors (PCSE)<ul style="list-style-type: none">• obs.: usually for POLS with LDV1

What to try first?



1. Improve data

2. Improve model

3. Robustification

Reference for these slides

MESQUITA, Rafael; FERNANDES, Antônio; FIGUEIREDO FILHO, Dalson B.;. Uma introdução à regressão com dados de painel. **Política Hoje**, Vol. 30, N. 1, 2021.