Uma introdução à regressão com dados de painel[1]

Rafael Mesquita - Universidade Federal de Pernambuco
Antônio Alves Tôrres Fernandes - Universidade Federal de Pernambuco
Dalson Britto Figueiredo Filho - Universidade Federal de Pernambuco

Resumo

Apesar da crescente oferta de dados em formato de painel, ainda são raros os estudos no Brasil que combinam as dimensões transversal e longitudinal na mesma análise. Por exemplo, em uma amostra de mais de 7 mil artigos publicados entre 2000 e 2018 em periódicos de CPRI, apenas 45 casos citavam técnicas específicas para lidar com observações de unidades espaciais (países, estados, pessoas) repetidas em intervalos regulares do tempo (anos, meses, dias). Diante dos benefícios inferenciais que tal abordagem pode proporcionar e da escassez de pesquisas sobre o tema, este artigo apresenta uma introdução à regressão de painel. Metodologicamente, sintetizamos as principais recomendações da literatura e mostramos a implementação no *R Statistical* com o pacote *plm*, indo desde a seleção de modelos até o tratamento dos dados e apresentação de resultados. Para aumentar o potencial pedagógico do trabalho, disponibilizamos os materiais de replicação, incluindo dados originais e scripts computacionais. Com este artigo esperamos difundir a utilização de análises longitudinais na pesquisa empírica em CPRI no Brasil.

Palavras-chave: Dados em painel; TSCS; CPRI no Brasil; Regressão linear; Metodologia política.

1 Esse trabalho se beneficiou dos comentários dos membros do grupo de Métodos de Pesquisa do Departamento de Ciência Política da Universidade Federal de Pernambuco (DCP – UFPE). Uma versão preliminar foi apresentada sob o título "Regressão com dados de painel: balanço da utilização na CPRI brasileira e estratégias para maior difusão do método", na seção de "Ensino e Pesquisa em Ciência Política e Relações Internacionais" do 12º encontro da Associação Brasileira de Ciência Política, João Pessoa, Paraíba (2020). Agradecemos os comentários e sugestões dos participantes, em particular os apontamentos da professora Elia Alves e Jakson Aquino. Parte do conteúdo advém da tese de doutorado de um dos autores (Mesquita, 2018). Materiais de replicação estão disponíveis em: <https://osf.io/5yx7g/?view_only=ac1691cced8549238d6d6e0a9d2b7f7b>. Agradecemos a Ivan Fernandes (UFABC) e Gustavo Fernandes (FGV) pela disponibilização do banco de dados utilizado no trabalho.

Abstract

In spite of the growing offer of data organized in panel format, there are still few studies in Brazil combining cross-sectional and longitudinal dimensions. For instance, out of a sample of over 7 thousand articles published between 2000 and 2018 in Brazilian Political Science and International Relations journals, only 45 articles cited specific techniques to handle observations comprising spatial units (countries, states, persons) repeated in regular time intervals (years, months, days). Given the inferential gains this approach can afford and the scarcity of research on the topic, this article presents an introduction to panel regression. Methodologically, we summarize the main recommendations of the literature and show how to implement them via *R Statistical* and *plm* package, going from model selection to data treatment and presentation of results. To improve the pedagogical potential of this work, we share replication materials, including the raw data and computational scripts. With this article, we aim to broaden the use of longitudinal analysis in empirical research in Political Science and International Relations in Brazil.

Keywords: Panel Data; TSCS; Political Science and International Relations in Brazil; Linear Regression; Political Methodology.

1. Introdução

Muitos dados em Ciência Política e Relações Internacionais (CPRI) estão dispostos em formato de painel, ou seja, observações de unidades espaciais mensuradas em intervalos regulares do tempo (Beck, 2001). A notação *Times-Series Cross-Section* (TSCS) é recorrente na literatura e revela que a base de dados combina uma dimensão espacial (países, estados, municípios, pessoas, etc.) e outra temporal (anos, meses, dias, etc.) (Beck, 2008). Para explicar, a expressão *time-series* informa que os casos estão dispostos ao longo do tempo enquanto a denominação *cross-section* indica que a base é formada por múltiplas observações transversais (Gujarati, 2004).

Para determinadas perguntas de pesquisa, essa combinação de informações temporais e espaciais é de alta importância para obtenção de respostas mais confiáveis do que seria produzido pela utilização de uma única abordagem. Se quiséssemos, por exemplo, verificar se o crescimento econômico afeta o resultado das eleições nos municípios brasileiros (Fernandes e Fernandes 2017), poderíamos de início fazer apenas um estudo *cross-section* comparando vários municípios na última eleição. Esse desenho de pesquisa, porém, será limitado, pois sabemos que nos últimos anos o crescimento econômico em praticamente todo o território nacional foi baixo, de modo que aprenderemos apenas sobre os efeitos da contração econômica e não do crescimento. Outras circunstâncias peculiares do período recente com impacto eleitoral (e.g.: denúncias de corrupção) tampouco poderão ser controladas se tiverem abrangência nacional. Contudo, superar essas dificuldades é possível se reunirmos dados não somente da última eleição, mas de repetidas eleições para todos os municípios de interesse.

Além disso, dados em painel são valiosos pois permitem ganhos importantes em termos de mensuração, controle de *confounders*, aprimoramento de inferências causais e tamanho da amostra (Baltagi, 2005; Fortin-Rittberger, 2013). Obviamente que essas vantagens vêm com custos. Por exemplo, a estrutura do termo de erro requer procedimentos específicos para que os resultados sejam consistentes e válidos (Worrall, 2008). Apesar de propício aos fenômenos estudados pela CPRI, o método continua pouco utilizado: em uma amostra de mais de 7 mil artigos publicados entre 2000 e 2018 em periódicos de CPRI, apenas 45 casos citavam técnicas específicas de regressão de painel.

Assim, diante dos benefícios analíticos de tal abordagem e da sua limitada utilização na área de CPRI, o presente artigo apresenta soluções didáticas sobre o que fazer e não fazer com dados de painel (Beck e Katz, 1995). Nosso público alvo são estudantes de graduação e pós-graduação em fases iniciais de treinamento. Essa contribuição pedagógica se torna tanto mais importante dada a ausência de material didático especialmente voltado para CPRI.[2] A maior parte dos manuais ou é direcionado a outras disciplinas, como economia (Wooldridge, 2002; Kennedy, 2003; Stock e Watson, 2004), ou apresenta um grau proibitivo de complexidade.[3] Pensando nisso, incluímos também uma seção de perguntas e respostas com dúvidas frequentes sobre o assunto, além dos scripts computacionais que permitem que os leitores adaptem as rotinas computacionais a seus respectivos interesses de pesquisa.

O artigo está organizado da seguinte forma: a primeira seção apresenta os resultados de uma análise bibliométrica sobre o uso da regressão de painel em CPRI; a segunda fornece uma introdução à lógica de análise de dados longitudinais; a terceira seção replica os dados de Fernandes e Fernandes (2017) para mostrar o passo a

2 Para trabalhos em português com o mesmo espírito didático, consultar Fávero (2013) e Marques (2000). Para um tutorial em inglês com ênfase computacional no pacote *plm*, ver Henningsen e Henningsen (2019).

3 Ver por exemplo o grau de dificuldade do exame final do curso de econometria para dados de painel do professor William Greene <http://people.stern.nyu.edu/wgreene/Econometrics/PanelDataEconometricsFinalExam.pdf>

passo da implementação computacional da regressão de painel; e a última seção sumariza nossas recomendações sobre como difundir análises longitudinais na pesquisa empírica em CPRI no Brasil.

2. Balanço do uso de painel de dados na CPRI brasileira

Quão populares são dados em painel na CPRI brasileira? Para responder essa pergunta, examinamos 7.764 artigos publicados entre 2000 e 2018 em 17 periódicos hospedados na *Scientific Electronic Library Online Platform* (Scielo)[4] (ver Tabela 1 nos Anexos). Semelhante a Medeiros et al. (2016), utilizamos análise automatizada de conteúdo para detectar palavras-chave no texto integral dos artigos. A partir dos pacotes *rscielo* (Meireles et al., 2019) e *quanteda* (Benoit et al., 2018), contabilizamos a frequência de expressões ligadas aos métodos de painel de dados. De toda a amostra, apenas 45, ou seja 0,58%, mencionaram expressões ligadas a dados em painel.[5] Embora o intervalo se inicie em 2000, somente em 2008 surgem artigos mencionando esses termos, o que significa que a utilização dessa técnica em CPRI é bastante recente.

Comparamos a frequência de dados em painel com o número de citações ao Método de Mínimos Quadrados Ordinários (MQO, em inglês *Ordinary Least Squares, OLS*), tendo em vista ser este último um dos métodos mais difundidos em CPRI (Krueger e Lewis-Beck, 2008).[6] Para cada artigo mencionando dados em painel (110 citações em 45 artigos) existem outros cinco que usam MQO (758 citações em 263 artigos). A Figura 1 ilustra essa comparação.

4 O *rscielo* automatiza a coleta de informações do Scielo. No momento da redação deste artigo, o pacote identifica um total de 374 periódicos, referentes a todas as áreas do conhecimento. Destes, selecionamos 65 que receberam conceito Qualis Capes no campo de CPRI em 2016. Como periódicos podem ser pontuados em mais de uma área, alguns destes títulos, apesar de avaliados no campo de CPRI, possuíam uma área de concentração distinta, como economia, educação ou saúde pública. Inspecionando a lista de periódicos e os sites de cada revista, retivemos 17 títulos cujo foco é declaradamente CPRI. Deste subconjunto, coletamos o texto de 8.829 artigos ao todo. Após eliminação de duplicatas e falsos-positivos, chegou-se a 8.789. Filtrando em seguida apenas publicações entre 2000 e 2018, obtivemos a quantidade final de 7.764.

5 Termos utilizados para o dicionário de dados em painel: "panel data", "dados em painel", "painel de dados", "tscs", "*time series cross section*".

6 Termos utilizados para o dicionário de MQO: "*ols*", "mmq", "mqo", "*ordinary least squares*", "mínimos quadrados", "*linear regression*", "regressão linear", "*regression analysis*", "análise de regressão".

Figura 1 – Prevalência de MQO/OLS vs. painel de dados

Fonte: elaborado pelos autores. O gráfico à esquerda contabiliza ocorrências das palavras-chaves ligadas a MQO e dados em painel e o da direita conta o total de artigos. As linhas azuis mostram os totais para dados em painel (PD) e as vermelhas para MQO (OLS).

Apesar da tendência de crescimento, a popularidade dos métodos de dados em painel hoje é equivalente à de MQO há uma década. O uso de dados em painel é, ademais, fortemente concentrado: 42% dos artigos foram publicados na Revista de Administração Pública. Das 17 revistas, seis não contabilizaram sequer uma menção ao método. Esses resultados apontam para uma baixa difusão dos dados em painel na produção nacional em CPRI.

No restante do artigo, introduziremos os conceitos essenciais para regressão de painel a partir da revisão bibliográfica dos principais manuais e mostraremos o passo a passo da implementação computacional a partir da replicação de Fernandes e Fernandes (2017).

3. Entendendo dados longitudinais

Esta seção apresenta os fundamentos da análise de dados longitudinais. Descrevemos primeiramente as diferentes formas de tabular informações ao longo do tempo. Discutimos ainda os principais tipos de painel (curto x longo; balanceado x desbalanceado) (seção 2.1). Em seguida, explicamos a notação do modelo de regressão de painel e as quatro principais modalidades de estimação (2.2). Por fim, apresentamos as vantagens e desvantagens de cada abordagem (2.3 e 2.4).

Dados longitudinais podem ser compreendidos a partir de uma comparação com dados transversais (Menard,

2002). Em um desenho transversal (*cross-section*), as informações sobre casos e variáveis são coletadas em um ponto específico do tempo. Por exemplo, taxa de homicídios (variável) para todas as unidades da federação (casos) em 2013 (tempo). Por outro lado, dados longitudinais são coletados em pelo menos dois períodos diferentes de tempo, o que permite observar a mudança dos valores em cada caso para cada variável de interesse.[7] Por exemplo, a taxa de homicídios (variável) para todas as unidades da federação (casos) entre 2000 e 2010 (tempo). A Figura 2 ilustra a diferença entre dados transversais, séries temporais, e sua combinação na forma de painel. Como exemplo, utilizamos a taxa de homicídios por 100 mil habitantes de três estados brasileiros.

Figura 2 – Estrutura dos dados

Transversal / *Cross-Section*

UF	Tempo	Taxa de homicídios	Média Geral	Desvio Entre Casos
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

Série Temporal / *Time-Series*

UF	Tempo	Taxa de homicídios	Média Individual	Desvio Intra-Caso
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

Dados em Painel / *Time-Series Cross-Section*

UF	Tempo	Taxa de homicídios	Média Geral	Média Individual	Desvio Geral	Desvio Intra-Caso	Desvio Entre Casos
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

Fonte: elaboração própria, com base em dados de Cerqueira (2017).

7 Menard (2002) identifica quatro desenhos de pesquisa longitudinal: (1) *total population designs*; (2) *repeated cross-sectional designs*; (3) *revolving panel designs* e (4) *longitudinal panel designs*. Wooldridge (2013, p.448) também diferencia entre "*independently pooled cross section*" e "*panel data*"/"*longitudinal data*"; no primeiro caso, trata-se de amostras aleatórias de indivíduos diferentes em sucessivos momentos no tempo (e.g.: amostragem do censo), ao passo que no último os sujeitos observados permanecem os mesmos ao longo do tempo.

Em nosso exemplo, a variável *UF* representa a unidade de análise transversal. A variável *Tempo* indica a dimensão longitudinal das informações, no caso, o ano. X_{it} representa o valor da variável de interesse para o caso *i* no tempo *t*. Por exemplo, o Acre teve uma taxa de homicídios de 22 em 2011. As demais colunas apresentam medidas agregadas derivadas de X_{it}; embora estas normalmente não figurem nas planilhas de dados propriamente, nós as reproduzimos para fins didáticos.

A média geral (por vezes chamada *grand mean*) é a soma total dos casos (237,7) dividida pelo tamanho da amostra (9), ou seja, 26,4. O desvio geral é a diferença de cada observação da média geral. Por exemplo, como em 2011 a Bahia teve taxa 39,4, ela estava 13 pontos acima da média geral (39,4 - 26,4). O desvio entre os casos é a diferença entre a média de cada caso e a média geral. Por exemplo, Santa Catarina dista 12,5 - 26,4 = -13,9 pontos da média geral. Por fim, o desvio intra-caso representa a diferença entre cada observação e a média individual daquele estado, de modo que, em 2013, Santa Catarina esteve 11,9 - 12,5 = -0,6 pontos abaixo da sua média histórica.

Esta configuração em que cada linha da planilha representa uma combinação única de caso no tempo é também conhecida como formato "*long*" e é o arranjo padrão para a maior parte dos programas estatísticos. Em diversos cenários, pode haver variação na disponibilidade de informações transversais ou longitudinais. Como dados em painel são o produto da combinação de ambas, resulta que diferentes tipos de painéis podem ser formados, a depender do número de casos ou de unidades temporais disponíveis, como explicamos a seguir.

3.1 Tipos de painel: curtos x longos, balanceados x não-balanceados

Painéis assumem duas principais configurações. Há um painel curto, "*stacked*" ou "*cross-section dominant*" quando o número de casos supera a quantidade de períodos temporais, ou seja, N>T. Por sua vez, dizemos que o painel é longo ("*temporally dominant*") quando o número de períodos temporais é maior do que a quantidade de casos (T>N). Alguns autores convencionaram chamar apenas os primeiros de painel de dados e os últimos de dados *Time-Series Cross-Section* (TSCS) (Beck, 2001; Fortin-Rittberger, 2013).

Para cada configuração do painel (curto ou longo), assume-se diferentes expectativas em relação à distribuição dos dados. Beck (2001) argumenta que, para dados em painel (N>T), as unidades selecionadas são encaradas como amostras de uma população, observadas em um curto intervalo de tempo. Dessa forma, as inferências não são específicas às unidades e podem ser generalizáveis. Já em TSCS (T>N), unidades são julgadas como fixas e observadas por um longo período de tempo. Desse modo, as inferências são orientadas às unidades e menos generalizáveis. O exemplo tradicional da primeira são pesquisas de censo, enquanto que da última são comparações macroeconômicas entre países.[8]

Outra característica importante diz respeito à disponibilidade da informação. Temos o painel balanceado ou equilibrado quando há informação para todos os casos em todos os períodos de tempo. Por outro lado, dizemos que o painel é não-balanceado ou desequilibrado quando alguns casos estão ausentes em determinados períodos de tempo. Uma complicação adicional, que pode afetar a consistência das estimativas, diz respeito à natureza do desequilíbrio do painel. A ausência de dados pode ser aleatória, o que tende a prejudicar a eficiência dos coeficientes (erro padrão, p-valor, intervalos de confiança). Todavia, o desbalanceamento pode ser relacionado com alguma variável de interesse

8 Semelhantemente, as propriedades assintóticas de TSCS supõem que N é fixo e T pode crescer até o infinito. Já as de dados em painel, que T é fixo e N pode crescer ao infinito (Fortin-Rittberger, 2013). Daí deduz-se que, quanto maior N em termos proporcionais, melhor os dados se adequam a modelos de dados em painel e, inversamente, quanto maior o T, melhor a adequação aos modelos TSCS.

do estudo. Por exemplo, pode ser mais fácil encontrar informações detalhadas para países mais ricos do que nações mais pobres. Quando a ausência de informação é sistemática por qualquer motivo, as estimativas tendem a apresentar viés. De acordo com Baltagi (2005), painéis desequilibrados tendem a ser mais frequentes do que painéis completos, o que reforça a importância de descrever exatamente o processo de coleta, tratamento e análise dos dados.

Explicamos agora como dados em painel podem ser utilizados para análise de regressão.

3.2 Notação e tipos de regressão de painel

Para os propósitos do artigo, é importante apresentar a notação clássica do modelo de regressão linear, assumindo dados apenas transversais:

$$Y_i = \alpha + \beta X_i + \varepsilon_i \qquad (1)$$

Em que Y representa a variável dependente, α indica o intercepto, β representa a variação observada em Y quando a variável independente, X, aumenta em uma unidade, e ε indica a natureza estocástica do modelo. O subscrito *i* indica que as observações são indexadas por caso. Continuando com o exemplo da Figura 2 (taxa de homicídios por UF ao longo do tempo), essa equação pode representar o efeito do desemprego (X) sobre a taxa de homicídios (Y). A regressão de painel é uma extensão do modelo anterior, em que:

$$Y_{it} = \alpha + \beta X_{it} + \varepsilon_{it} \qquad (2)$$

Os subscritos *i* e *t* informam que as observações são indexadas, respectivamente, por caso e tempo. Nosso modelo se torna mais poderoso, pois acumulamos informações sobre a relação entre desemprego e homicídios para várias UF e anos. Como prosseguir, entretanto, com a regressão? Descrevemos abaixo as quatro abordagens tradicionais para estimar modelos de painel.

Pooled OLS

Ao reunir todas as observações em uma mesma regressão linear ("*pooling*"), tem-se *Pooled OLS* (POLS). Essa abordagem pressupõe que todas as unidades podem ser contempladas pela mesma reta de regressão. Mesmo que as unidades *i* apresentem diferenças umas das outras, considera-se que o conjunto de variáveis independentes já carrega toda a informação importante. Ou seja, os principais fatores que diferenciam entre as observações já estão explicitados no modelo, de modo que não é necessário controlar por outros fatores não-observados. Em nosso exemplo, seria como se a taxa de homicídios, qualquer que fosse a UF, fosse principalmente determinada pelo desemprego, e não houvesse outras peculiaridades nas UF afetando o resultado: se a Bahia tivesse o nível de desemprego de Santa Catarina, teria a mesma taxa de homicídios desta. Como heterogeneidades individuais não são modeladas, temos a Equação (2), em que há apenas um intercepto (α) para toda a população e todo o mais que não é observável é alocado como ruído no fator de erro (ε).

Se houver fatores não-observados em ε que, entretanto, não forem aleatórios, mas sistemáticos, POLS já não será apropriado. Se tivermos elementos não-observados correlacionados com X, temos um caso tradicional de viés de variável omitida e a estimativa do parâmetro β ($\hat{\beta}$) será inconsistente.[9] Como as observações se repetem no tempo, a independência das observações está em risco. Suponha que para os três estados do exemplo haja algum fator que não podemos observar e que, contudo, impacta na taxa de homicídios, digamos a influência do

9 Se o fenômeno estudado é melhor descrito por um conjunto de variáveis e alguma dessas é deixada de fora da equação, o efeito das variáveis preservadas no modelo será ou sobre ou subestimado. Essa distância sistemática para mais ou para menos no parâmetro caracteriza um viés.

crime organizado. Se a influência do crime organizado for correlacionada com desemprego, haverá viés; se for específica em cada estado (e.g.: a influência criminosa no Acre é a mesma ao longo do tempo), os erros não serão independentes (ε_{it} será correlacionado com ε_{it-1}). Por essa razão, a primeira motivação para se afastar do modelo POLS é a necessidade de trabalhar a heterogeneidade oculta dentro do termo de erro de uma maneira mais sofisticada.

Os fatores não-observados se dividem em elementos que permanecem sempre os mesmos para cada unidade e elementos que variam segundo a unidade e o tempo. Então, o "erro composto" (ε_{it}) da Equação (2) pode ser destrinchado em duas partes: uma fixa para cada unidade (μ_i) e outra variável (ν_{it}).

$$Y_{it} = \alpha + \beta X_{it} + \mu_i + \nu_{it} \qquad (3)$$

Os erros persistentes às unidades (μ_i) também são conhecidos como heterogeneidade não-observada, efeito individual, ou efeito fixo do termo de erro, enquanto que ν_{it} é usualmente chamado de efeito variável ou idiossincrático. Diferente das outras três abordagens que serão apresentadas aqui, POLS não dispõe de estratégias para lidar com μ_i, pois não desagrega ε_{it}. Todas as demais especificações decompõem o fator de erro em fixo e variável.

Primeiras Diferenças / First Differences

Na abordagem de Primeiras Diferenças ou *First Differences* (FD), todas as variáveis são diferenciadas, isto é, subtraídas de seu valor no período temporal anterior. A subtração é eficiente em lidar sobretudo com problemas dinâmicos dos dados, como correlação serial e não-estacionariedade, já que, ao subtrair uma observação em *t* da observação em *t*-1, tudo o que não houver variado de um instante para o outro (inclusive a heterogeneidade não-observada) será apagado.

Por subtrair observações aos pares, FD traz alguns custos: reduz a variabilidade dos dados e assim aumenta os erros-padrão, sendo portanto mais recomendado para bases grandes. Ademais, a diferenciação remove todos fatores constantes ao longo do tempo ou variando monotonicamente (i.e.: sempre aumentando à mesma taxa todo ano), o que impossibilita o uso de variáveis como *dummies* para países, distâncias geográficas e variáveis de crescimento constante (Wooldridge, 2013, p. 462–463). O uso de FD é uma das soluções drásticas recomendadas por Beck (2008) diante de dependências temporais persistentes no modelo.

Efeitos Fixos / Fixed Effects

Fixed Effects (FE) ou Efeitos Fixos, assim como a FD, buscam livrar-se do componente fixo do erro (μ_i), porém, ao invés da diferenciação, utilizam a *within transformation*. Dá-se o nome de *within* a essa operação pois ela retém apenas o desvio intra-caso (ver Figura 2). Essa remoção do valor da média de cada observação ("*time demeaning*" ou "*group-mean centering*") terá como resultado a eliminação de α e μ_i, cujos valores para cada unidade não mudam com a passagem do tempo.[10] Portanto, assim como FD, o estimador FE omite fatores constantes no tempo.

Por lidar com o problema da heterogeneidade e ser bastante robusta, a especificação FE é vista por Croissant e Millo (2018, p.5) como um *benchmark* para dados em painel. Isso é especialmente o caso para disciplinas que lidam com dados observacionais, o que também se verificou em nosso levantamento bibliométrico: efeitos fixos foram o método mais recorrente para CPRI (26% dos artigos).

Um dos benefícios de FE em comparação com POLS é que, enquanto este postula apenas um

10 Os componentes α e μ_i são removidos para estimação consistente de β. Repare, não obstante, que alguns *softwares* estatísticos reportam valores para interceptos individuais (μ_i) ou populacionais (média de todos os μ_i). Vide Wooldridge (2002, seção 10.5.3) para detalhes.

intercepto para toda a amostra, aquele leva em consideração que cada unidade pode ter seu próprio intercepto. Em comparação com FE, POLS terá estimadores viesados, já que não lida com μ_i. Já sob FE, o modelo se torna menos viesado, porém também menos eficiente.

Computacionalmente, obtêm-se resultados semelhantes em aplicar um modelo FE e um POLS acrescido de *dummies* para todos os casos. Nesta situação, POLS geralmente atende pelo nome de *Least Square Dummy Variable* (LSDV). A Figura 3 ilustra os procedimentos de diferenciação e de subtração da média.

Figura 3 - Ilustração dos procedimentos de diferenciação e *within transformation*

Primeiras Diferenças / First Differences

ID	Tempo	Taxa de homicídios	Taxa de desemprego
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

$$Y_{it-1} = \alpha + \beta X_{it-1} + \mu_i + \nu_{it-1}$$
$$- \left(Y_{it} = \alpha + \beta X_{it} + \mu_i + \nu_{it} \right)$$
$$\Delta Y_{it} = \beta \Delta X_{it} + \Delta \nu_{it}$$

ID	Tempo	Taxa de homicídios	Taxa de desemprego
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

Efeitos Fixos / Fixed Effects

ID	Tempo	Taxa de homicídios	Taxa de desemprego	Média Individual (homic.)	Média Individual (desemp.)
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

$$Y_{it} = \alpha + \beta X_{it} + \mu_i + \nu_{it}$$
$$- \left(\bar{Y}_i = \alpha + \beta \bar{X}_i + \mu_i + \bar{\nu}_i \right)$$
$$\ddot{Y}_i = \beta \ddot{X}_i + \ddot{\nu}_i$$

ID	Tempo	Desvio Intra-Caso (homic.)	Desvio Intra-Caso (desemp.)
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

Fonte: elaborado pelos autores, a partir de dados do Cerqueira et al. (2017) e IBGE (2017).

O esquema ilustra os efeitos da diferenciação e da *within transformation* apenas para as variáveis observadas: homicídios e desemprego. Mesmo assim, a subtração abrange também os fatores não-observados (indicados em itálico nas equações da imagem), resultando na supressão do intercepto geral (α) e do efeito individual do fator de erro (μ_i). Em nosso exemplo, significa que os modelos resultantes, ao estimarem o efeito do desemprego sobre a taxa de homicídios, eliminam a influência fixa do crime organizado em cada estado. Não só este, mas todo fator não-observado que for invariante no tempo será removido, e a estimação irá se basear apenas nos elementos que variam.

Efeitos Aleatórios / Random Effects

Como foi visto, uma deficiência de FD e FE é a impossibilidade de testar variáveis que não mudam com o tempo. Mesmo as que variam lentamente tendem a ser subestimadas (Clark e Linzer 2015). Todavia, há muitas variáveis de interesse para CPRI que são estáticas, como geografia, gênero, nacionalidade e outras. A especificação dos Efeitos Aleatórios ou *Random Effects* (RE) contorna essa dificuldade, permitindo a inclusão de variáveis que não mudam com a passagem do tempo.

Assim como para FE, supõe-se que o erro composto possui parte fixa e parte idiossincrática, ou seja, $\varepsilon_{it} = \mu_i + \nu_{it}$. Contudo, ao invés de eliminar μ_i por inteiro, supõe-se que ele não tenha correlação com X (Bell e Jones 2015, p. 142). Se o efeito fixo individual não tem correlação com as variáveis independentes, não é mais fonte de viés e portanto não precisa mais ser eliminado; e.g.: se acreditamos que crime organizado é independente do desemprego, não há motivo para removê-lo.

O segundo pressuposto distintivo de RE diz respeito aos efeitos individuais. Efeitos *fixos* supõem que o efeito de cada indivíduo pode ser calculado como um valor único. Por isso, FE calcula um intercepto para cada indivíduo na forma de um parâmetro. RE, em contraste, considera μ_i e ν_{it} como realizações *aleatórias* de uma distribuição. No exemplo envolvendo o crime organizado não-observado nas UFs, FE irá estimar esse impacto individual caso a caso, sem supor que há alguma relação entre o nível de crime organizado na Bahia e no Acre. Já RE imprime mais estrutura ao conjunto, pois supõe que os efeitos do crime tendem a uma distribuição normal, consideradas todas as UFs (Bell et al., 2019, p. 1061).

Alguns autores classificam RE como um meio-termo entre POLS e FE (Clark e Linzer 2015, p. 402). POLS não elimina μ_i e FE o apaga completamente. RE executa parcialmente este procedimento, pois subtrai das observações uma parcela da média intra-grupo ("*quasi-demeaning*"). A proporção dessa parcela é dada por θ (*theta*, no output do *plm*) que assume valores de 0 a 1. *Theta* pode ser interpretado como um diagnóstico de qual componente mais contribui com a variância de ε: se fixo (μ_i) ou idiossincrático (ν_{it}). Quando há pouca variância nos efeitos fixos ($\theta \to 0$) as estimativas de RE se aproximam de POLS e, no caso contrário, ($\theta \to 1$) de FE (Wooldridge, 2013, p. 494).

Em contraponto, a suposição que não há relação entre μ_i e X é difícil de sustentar em boa parte das situações reais. Outra diferença significativa é que estimadores RE sempre terão algum viés, podendo só ser eficientes. A escolha entre FE e RE é, portanto, um *tradeoff* entre entre eficiência (RE) e não-viesamento (FE).

A comparação direta entre RE e FE é recorrente na literatura. Disciplinas mais afeitas ao uso de experimentos randomizados tendem a preferir RE (Dieleman e Templin, 2014). Beck (2001, p. 284) argumenta que estudos observacionais preferem FE a RE, pois este último é mais recomendado quando se consideram as unidades como uma amostra de uma população mais ampla, e aquele quando se busca fazer inferências restritas às unidades. A ideia de "intercambialidade" captura essa noção de forma intuitiva: quando para um conjunto de observações podemos substituir um caso A por outro B sem grande perda de informação - ou seja, o *nome do caso* não importa - temos algo próximo de uma amostragem aleatória (Hausman 1978, p. 1262; King 2001, p. 498). A expectativa de RE de uma distribuição normal dos efeitos individuais se torna então cabível. Em alguns cenários de CPRI, essa intercambialidade é plausível (e.g.: deputados em um congresso), em outros não (e.g.: países membros do Conselho de Segurança da ONU).

Por fim, ao fazer uma análise empírica, é sempre informativo comparar lado a lado os resultados das quatro formas de regressão: POLS, FD, FE e RE. O contraste

aponta para os diferentes resultados obtidos a depender do que se faz com μ_i: POLS mantém o fator fixo totalmente inserido no componente de erro, RE mantém uma parcela, e FE e FD o eliminam completamente (Wooldridge, 2013, p. 494).

3.3 Benefícios dos dados em painel

O Quadro 1 sintetiza os principais benefícios técnicos e problemas operacionais associados aos dados em painel.

Quadro 1 – Benefícios e problemas dos dados em painel

Benefícios	Problemas
(1) Facilitar a detecção de causalidade; (2) Mensuração de variação individual; (3) Reduzir erros de mensuração; (4) Aumentar o tamanho da amostra; (5) Controlar por problemas de variáveis omitidas.	(1) Correlação serial dos resíduos; (2) Não-estacionariedade; (3) Heterogeneidade; (4) Heteroscedasticidade de painel; (5) Correlação contemporânea de erros; (6) Estruturas complexas de dependência.

Fonte: elaborado pelos autores, com base em Fortin-Rittberger (2013), Hsiao (2003) e Kennedy (2003).

A primeira vantagem dos painéis é facilitar a detecção de relações causais entre X e Y. Pearl (2000) aponta três pressupostos para identificar uma relação causal: (1) associação entre as variáveis (correlação); (2) precedência temporal e (3) não-espuriedade da relação. Diferente dos dados transversais, em que os valores de X e Y são mensurados no mesmo período, dados em painel permitem observar se a variação temporal de X está correlacionada com a variação temporal de Y. Dessa forma, além de controlar por eventuais variáveis espúrias, os dados longitudinais favorecem a satisfação do pressuposto da precedência temporal entre a variável independente e a dependente.

Além disso, a natureza longitudinal também favorece o estudo da mudança, já que conserva a mesma unidade de análise ao longo do tempo. Assim, é possível examinar a variação individual de diferentes indicadores entre os casos de interesse. Por exemplo, a análise agregada do número de homicídios em uma cidade é informativa. Já uma base de dados que informa o dia do óbito e bairro abre possibilidades adicionais. Será possível observar, por exemplo, se os finais de semana em certo bairro concentram a maior parte das ocorrências. Como aponta Fortin-Rittberger (2013), essa possibilidade de monitorar, no espaço e no tempo, onde e quando surgem certos fenômenos de interesse aproxima os estudos longitudinais, mesmo sendo observacionais, de uma lógica de experimento controlado.

Como terceira vantagem, temos medidas mais confiáveis sobre a mudança das variáveis. De acordo com Blalock (1979), o problema da mensuração é o principal desafio ao desenvolvimento das ciências sociais. Em particular, em CPRI, conceitos importantes, democracia por exemplo, não são diretamente observáveis. Uma das formas de examinar a confiabilidade de uma medida é mensurá-la várias vezes ao longo do tempo. Logo, os dados em painel ajudam a superar eventuais erros de mensuração.[11]

11 Erros podem ser sistemáticos ou aleatórios. O erro sistemático é prejudicial em análises descritivas pois as estimativas produzidas serão viesadas. Por outro lado, quando o erro sistemático atinge todos os casos com a mesma magnitude não afeta a consistência das estimativas em modelos de regressão. Por sua vez, o erro aleatório na variável dependente tende a reduzir a eficiência (maior erro padrão), mas não afeta a consistência das estimativas. Por fim, quando as variáveis independentes também apresentam erro aleatório, as estimativas tendem a ser inconsistentes e, em geral, subestimadas (King, Keohane e Verba, 1994).

Além dos ganhos em mensuração, outro trunfo dos dados em painel é aumentar o tamanho da amostra. Um dos problemas recorrentes em desenhos de pesquisa é a indeterminação, ou seja, "muitas variáveis, N pequeno" (Lijphart, 1971, p. 686). Quando temos poucas observações, é difícil sustentar que a relação causal proposta se explica por um motivo e não qualquer outro. Assim, a determinação de um desenho de pesquisa diz respeito à proporção entre variáveis e observações. Por isso, aumentar o número de casos é uma das recomendações costumeiras para solucionar a indeterminação (King, Keohane e Verba, 1994). Ademais, amostras maiores superam problemas de eficiência (menor erro padrão), multicolinearidade e facilitam a utilização de testes estatísticos mais robustos (Hsiao, 2003).

Por fim, dados em painel controlam os problemas de variáveis omitidas (Kennedy, 2003). A omissão de uma variável importante produzirá viés nas estimativas se ela estiver correlacionada com outras variáveis modeladas. A perspectiva puramente transversal não fornece muitas soluções para lidar com fatores não-observados. Frequentemente, somos simplesmente ignorantes sobre os *unobservables* que incidem sobre Y. Contudo, a partir do momento em que os casos são acompanhados ao longo do tempo, expedientes como FE, FD e RE podem ser usados para remover ou mitigar essas heterogeneidades individuais (Wooldridge 2002, seção 10.1).

Esses benefícios, entretanto, não vêm sem custo, como vemos a seguir.

3.4 Desafios dos dados em painel e como superá-los

Para fins didáticos, propomos uma breve recapitulação dos pressupostos da regressão linear de MQO como ponto de partida para abordar os problemas típicos das regressões de painel.[12] Seguindo Wooldridge (2013, p. 59, 119), os principais pressupostos para MQO podem ser resumidos em seis: (P.1) Há uma relação linear entre X e Y; (P.2) Ausência de autocorrelação (ou amostra aleatória);[13] (P.3) Ausência de multicolinearidade perfeita ("*full rank*");[14] (P.4) Exogeneidade de X (ou "média condicional zero"): o erro não deve ser uma função das variáveis independentes. Ou seja, ele é aleatório e nenhuma das variáveis independentes no modelo porta informação a respeito do seu valor;[15] (P.5) Homocedasticidade: há variância uniforme dos resíduos da regressão e estes não são correlacionados uns com os outros. Se por algum motivo em um segmento dos dados a variância dos resíduos muda em sintonia com os valores das variáveis explicativas, há heteroscedasticidade;[16] e (P.6) Distribuição normal dos erros.

Se as chamadas premissas Gauss-Markov (P.1-P.5) são respeitadas, pode-se dizer que os parâmetros estimados são os *Best Linear Unbiased Estimators* (ou BLUE) dos parâmetros populacionais. "*Best*" porque possuem a menor variância já que há (P.5) homocedasticidade (variância uniforme dos resíduos); "*Linear*" por conta da relação linear (P.1); e "*Unbiased*" porque não há viés no parâmetro, já que os regressores são exógenos e não afetam o erro (P.4).

Porém, modelos de painel facilmente podem infringir essas suposições. Por combinarem tanto séries temporais quanto comparações transversais, herdam os problemas de ambos os tipos de dados. Os problemas

12 A lista de pressupostos da MQO por vezes varia entre autores. Para outras listagens dos pressupostos da regressão linear e do teorema de Gauss-Markov, ver Berry (1993) e Larocca (2012).

13 Amostras aleatórias e ausência de autocorrelação podem parecer expectativas distintas, mas são equivalentes pela seguinte perspectiva: em ambos os casos, se tem que $Corr(\varepsilon i, \varepsilon j = 0)$, para $i \neq j$.

14 O principal dano provocado pela multicolinearidade é o aumento da variância dos coeficientes da regressão, o que aumenta a chance de encontrar resultados não-significativos.

15 A violação da exogeneidade equivale a um viés de variável omitida.

16 Em outras palavras, o modelo se encaixa melhor a certas porções dos dados e não a outras. Heteroscedasticidade não leva a viés ou inconsistência de estimadores MQO, porém por interferir na distribuição das variâncias impossibilita a construção de intervalos de confiança e testes de significância (Wooldridge, 2013, pp. 268–269).

podem ser divididos entre os dinâmicos e os espaciais. A literatura geralmente recomenda atentar em primeiro lugar para os problemas dinâmicos. Estes violam principalmente os requisitos da ausência de autocorrelação (P.2), exogeneidade (P.4) e homocedasticidade (P.5). Os principais são:

(1) **Correlação serial dos resíduos:** há correlação entre os resíduos no momento *t* e *t-1*. Trata-se de uma das características persistentes de painéis longos, pois, como Box et al. (2016) advertem, um aspecto intrínseco das séries temporais é que observações adjacentes são interdependentes. O principal problema gerado pela correlação dos erros é a inconsistência dos testes de significância (p-valor e intervalos de confiança).[17] Por esse motivo, a correlação serial deve ser detectada e solucionada antes de seguir adiante tratando dos problemas espaciais. Há diferentes ferramentas para detecção de processos temporais na estrutura de erros: os gráficos da função de autocorrelação (ACF) e autocorrelação parcial (PACF), os testes de Durbin-Watson, Breusch-Godfrey, Ljung-Box, entre outros. Em sua maioria, eles são direcionados a verificar processos autorregressivos de primeira ordem, ou AR(1), que são mais habituais.[18]

Pode-se testar correlação AR(1) nos erros de modelos POLS de forma simples e direta, fazendo uma regressão dos resíduos em *t* nos resíduos em *t*-1, sem intercepto, conforme $\varepsilon_{it} = \rho\varepsilon_{it-1} + \nu_{it}$. O resultado para o parâmetro ρ indicará se resíduos passados estão ou não correlacionados de forma significativa com os futuros (Wooldridge, 2002, seção 7.8.5; 2013, p. 417).

As especificações FE, RE e FD exigirão testes específicos. Cobrir todo o arsenal de testes existentes ultrapassa o escopo deste artigo (para uma lista, ver Croissant e Millo 2008, p. 22-28; 2018, seção 4.3), porém destacamos que a seleção de testes deve ser guiada pelo tipo de estimação utilizada no modelo (POLS, FE, RE ou FD), a proporção N/T, e o número de observações.

Uma vez que se perceba correlação serial, soluções recomendadas incluem adicionar a variável dependente em *lag*, inserir erros-padrões robustos, ou aplicar FD. Quanto à primeira alternativa: recomenda-se a inserção de uma *lagged dependent variable* (LDV), supondo que o processo seja AR(1) e após isso fazer novo teste para verificar se a autocorrelação foi eliminada (Beck, 2001). Tal solução, não obstante, foi criticada por Achen (2000, p. 14), afirmando que a LDV pode absorver todo o poder explicativo dos demais regressores e assim produzir viés.[19]

(2) **Não-Estacionariedade:** uma série temporal é estacionária quando está em equilíbrio estatístico, isto é, apresenta média e variância constantes ao longo do tempo. Ainda que os valores oscilem, retornarão à média. É não-estacionária quando a média e variância não são constantes, ou seja, a série não tende a retornar a uma média anterior após desvios. Dito de outra forma, a não-estacionariedade pode ser entendida como um efeito persistente de choques anteriores (Box et al., 2016, p. 7; Fortin-Rittberger, 2013). Desvios da estacionariedade podem advir do tipo de fenômeno monitorado, mas também podem resultar da frequência da mensuração: se as informações foram mensuradas por dia ou por mês, por exemplo, é possível encontrar efeitos sazonais e, para amostras com T grande, comportamento cíclicos. A inspeção visual da série temporal é uma das maneiras mais simples de identificar tendências ou sazonalidade. Diferentes testes oriundos da literatura sobre séries temporais

17 Tecnicamente, a autocorrelação afeta a eficiência das estimativas. Em particular, quando a autocorrelação for positiva, os erros padrão dos coeficientes serão subestimados, o que aumenta a chance de incorretamente rejeitar a hipótese nula (erro do tipo I).

18 Embora não seja habitual esperar processos mais complexos que AR(1) para dados em painel (Fortin-Rittberger, 2013), é possível que haja correlações de ordem superior, o que irá motivar remédios mais complexos (e.g.: a inclusão de variáveis com múltiplos *lags*).

19 Quando os regressores possuem alguma tendência, a introdução de uma LDV irá dominar a regressão. O coeficiente da LDV será viesado para cima e o dos demais regressores para baixo. Também não é recomendável quando o T é pequeno, pois sacrifica um período de análise (Achen, 2000).

podem ser empregados para verificar não-estacionaridade: ACF, PACF e testes de raiz unitária (e.g.: *Augmented Dickey-Fuller,* ADF). Mais recentemente, testes próprios ao contexto de dados em painel, "*Panel Unit Root Tests*", têm sido propostos (ver Croissant e Millo 2018, cap.8). A aplicação desses testes será dificultada se o painel tiver poucos períodos temporais. Se se verificar que a série é não-estacionária, soluções incluem utilizar FD ou, em casos de não-estacionariedade por tendência, incluir uma variável da passagem do tempo (*time trend*) (Fortin-Rittberger, 2013). Observemos agora os aspectos espaciais.

(3) **Heterogeneidade:** regressões supõem que os casos são homogêneos. Quaisquer características peculiares são explicadas pelas variáveis independentes, como denotado anteriormente através do conceito de "intercambialidade" (King, 2001, p. 498). Há heterogeneidade quando um ou alguns dos casos observados possuem características distintivas que, não tendo sido modeladas, irão alojar-se no termo de erro e assim provocar correlações entre ε_{it} e X_{it}. Heterogeneidade viola, desse modo, os pressupostos de exogeneidade (P.4) e homoscedasticidade (P.5). Se houver heterogeneidade, normalmente será indevido supor que há um só intercepto para toda a população, como nos modelos POLS. Se cada unidade possui características fixas impactantes, a ausência de interceptos próprios a cada uma levará a uma linha de regressão equivocada (Fortin-Rittberger, 2013). Por sua vez, a violação do pressuposto da homocedasticidade implica na produção de testes de significância inconsistentes, o que aumenta a chance de incorrer em inferências errôneas.

Como discutimos na seção 3.2, abandonar estimações POLS em favor de outras especificações é uma das formas de sanar a heterogeneidade. As últimas três dificuldades espaciais comprometem principalmente a exogeneidade (P.4), homocedasticidade (P.5) e a normalidade dos erros (P.6).

(4) **Heteroscedasticidade de painel:** os resíduos precisam ter variância constante em ambos os sentidos, isto é, entre unidades *i* e entre períodos temporais *t*. Há heteroscedasticidade de painel quando resíduos têm variância constante ao longo do tempo para cada unidade (*within unit*), mas inconstante entre as unidades (*across units*). Dito de outra forma, cada unidade terá sua própria variância de resíduo. Ela pode decorrer de má especificação do modelo ou quando um ou dois casos não se encaixarem bem na especificação.

(5) **Correlação contemporânea de erros:** o erro de um caso está correlacionado com o erro de outros casos para o mesmo momento no tempo. Se, por exemplo, um choque externo incide em um ano e afeta várias unidades ao mesmo tempo, os resíduos das unidades neste período terão variância diferente dos demais períodos.

(6) **Estruturas complexas de dependência:** há mais casos especiais que, de alguma forma, impõem dependência entre observações. Um dos mais explorados em estudos observacionais é a correlação espacial ou geográfica entre unidades. Ou seja, a contiguidade ou proximidade entre dois países os expõe a mais eventos em comum e, por não ser algo aleatório, pode levar a vieses (Fortin-Rittberger, 2013).

De modo geral, testes como o de Pesaran (2004) oferecem evidência se há dependências entre unidades tais como as descritas nos itens (4) a (6), embora haja variações especializadas para diferentes tipos de painel e de estrutura de dependência (ver Croissant e Millo 2008, p. 28-31; 2018, cap. 10).

Problemas espaciais em modelos POLS podem exigir uso de estimadores FE, RE ou FD. Também se recomenda, para qualquer estimador, a revisão do conjunto de variáveis, no intuito de contemplar novos fatores que possam abarcar a heterogeneidade entre as unidades; além da aplicação de erros-padrão robustos.

Concluindo, é útil lembrar que a probabilidade de ocorrência desses problemas segue o tipo de painel em questão. Painéis do tipo T>N tendem a sofrer mais de problemas decorrentes de heterocedasticidade e correlação serial de erros, ao passo que painéis onde N>T padecem frequentemente de heterogeneidade não-mensurada entre os casos.

A Figura 4 resume a lista de pressupostos e violações discutida nesta seção.

Figura 4 - Resumo dos pressupostos MQO e problemas dinâmicos e temporais

	Pressuposto	Formalização	Por que violá-lo é um problema?
P.1	Relação linear	$Y = \alpha + \beta X + \varepsilon$	Viés
P.2	Ausência de autocorrelação, Amostra aleatória	$Corr(\varepsilon i, \varepsilon j) = 0$ $i \neq j$	Viés
P.3	Ausência de colinearidade perfeita	$Corr(X1,X2) \neq \pm 1$	Impossível estimar $\hat{\beta}$; Viés; Infla Erro-Padrão($\hat{\beta}$), p-valor
P.4	Exogeneidade	$E(\varepsilon \mid x) = 0$	Viés
P.5	Homoscedasticidade	$Var(\varepsilon \mid x) = \sigma^2$	Invalida Erro-Padrão($\hat{\beta}$), p-valor
P.6	Distribuição normal dos erros	$\varepsilon \sim Normal(0, \sigma^2)$	Invalida Erro-Padrão($\hat{\beta}$), p-valor

Normalmente afetam:

	Problemas dinâmicos
1	Correlação serial dos resíduos
2	Não-estacionariedade
	Problemas espaciais
3	Heterogeneidade
4	Heteroscedasticidade de painel
5	Correlação contemporânea dos erros
6	Estruturas complexas de dependência

Fonte: elaborado pelos autores.

4. Aplicação

Nesta seção, reproduzimos os dados de Fernandes e Fernandes (2017) para ilustrar como se faz uma análise de painel. Embora seja costumeiro utilizar bases clássicas dos pacotes estatísticos para ensino (e.g.: "Grunfeld", "EmplUK", etc.), consideramos esta uma valiosa oportunidade para um *replication* que permita incentivar ao mesmo tempo o aprendizado e a cultura de transparência e reprodutibilidade nas ciências sociais (King, 1995).

O principal objetivo do artigo "A importância do crescimento econômico local na escolha do chefe do Executivo no Brasil" é analisar o impacto do crescimento econômico local sobre o percentual de votos recebidos pelo candidato incumbente nas eleições presidenciais e municipais brasileiras entre 2000 e 2010.

Para responder à pergunta de pesquisa, Fernandes e Fernandes (2017) mobilizam 17 variáveis independentes para explicar a variável dependente: a proporção de votos recebida pelo candidato mandatário. Conforme os autores: "Estimamos a relação por meio de técnicas de dados em painel que permitem analisar se as idiossincrasias não observáveis afetam ou não a estimação. Para isso, usamos os três modelos tradicionais de painel estático: *painel empilhado, efeitos aleatórios (EA) e efeitos fixos (EF)*" (Fernandes e Fernandes,

2017, p. 660). Para replicar os dados do trabalho, utilizamos o *software R*, *RStudio* e o pacote *plm* (Croissant e Millo, 2008). Para facilitar a compreensão do passo a passo computacional, reportaremos os scripts ao longo do texto.

Para os testes a seguir, é útil já ter preparado no ambiente do *RStudio* os quatro tipos de especificação a serem comparadas: POLS, FE, RE e FD. Assim, é necessário, em primeiro lugar, carregar o banco de dados e adequá-lo para execução dos modelos.

```
##### ABRINDO BANCO DE DADOS ############

library(haven) # pacote para ler dados do Stata
library(plm) # pacote para executar os modelos de dados em painel

BANCO <- read_dta("fernandes_2017.dta") # Arquivo da base de dados no diretório

BANCO <- pdata.frame(BANCO, index = c("codibge", "ano")) # aqui o banco é convertido
ao formato pdata.frame
# para executar os modelos. No 'index', São acrescentadas as dimensões espacial e
temporal dos dados.
# Nesse caso, 'codibge' representa a dimensão espacial (cód. do munic.) e 'ano' o
ano da eleição.
```

Aplicando o comando *pdim()* podemos ver que a base de dados possui N= 5.565 municípios, e T=2-6.[20] Há, ao todo, 28.945 observações. Ou seja, estamos diante de um painel curto (N>T) e não-balanceado. Ao transformar o banco em um objeto do tipo *pdata.frame*, é criado um índice contendo as dimensões espacial e temporal, o que viabiliza a estimação dos modelos.

20 Note que o período observado é "ano eleitoral". As observações começam em 2000 e vão até 2010, registrando dados a cada dois anos. Por esse motivo a base possui T=6 e não T=10.

```
##### FÓRMULA FERNANDES 2017 ##################

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

##### POOLED OLS ###############################

POLS <- plm(form, data = BANCO, model = "pooling")  # modelo pooled

##### EFEITOS FIXOS ############################

mode_fe <- plm(form, data = BANCO, model = "within") # O modelo de EF é executado
com o model = 'within'

##### EFEITOS ALEATÓRIOS #######################

mode_re <- plm(form, data = BANCO, model = "random") # o modelo de EA é executado
com o model = "random".

##### PRIMEIRAS DIFERENÇAS #####################

mode_fd <- plm(form, data = BANCO, model = "fd") # o modelo de FD é executado com o
model = "fd".
```

Executamos quatro modelos: POLS, FE, RE e FD. O modelo de POLS é executado por meio da especificação em *model* = "*pooling*" na função *plm*.[21] Do mesmo modo, os modelos FE, RE e FD também são especificados em *model*.

Qual dos modelos acima devemos utilizar? Para respondermos, é necessário verificarmos os pressupostos de cada um.

4.1 Pooled OLS?

A primeira e mais simples escolha é POLS, que aplica uma única linha de regressão, com o mesmo intercepto, para toda a população. Assim, para julgar a adequabilidade desta especificação, precisamos estimar o nível de homogeneidade de população, e se o nosso conjunto de regressores esgota ou não essa heterogeneidade. A relação entre votos em incumbentes e crescimento econômico (e outras variáveis de controle) é similar nos 5.565 municípios? Ou há variação significativa entre as unidades, mais do que foi possível explicar apenas pelas nossas variáveis? Para decidir, pode-se recorrer a (a) conhecimento contextual, (b) análise gráfica e (c) testes.

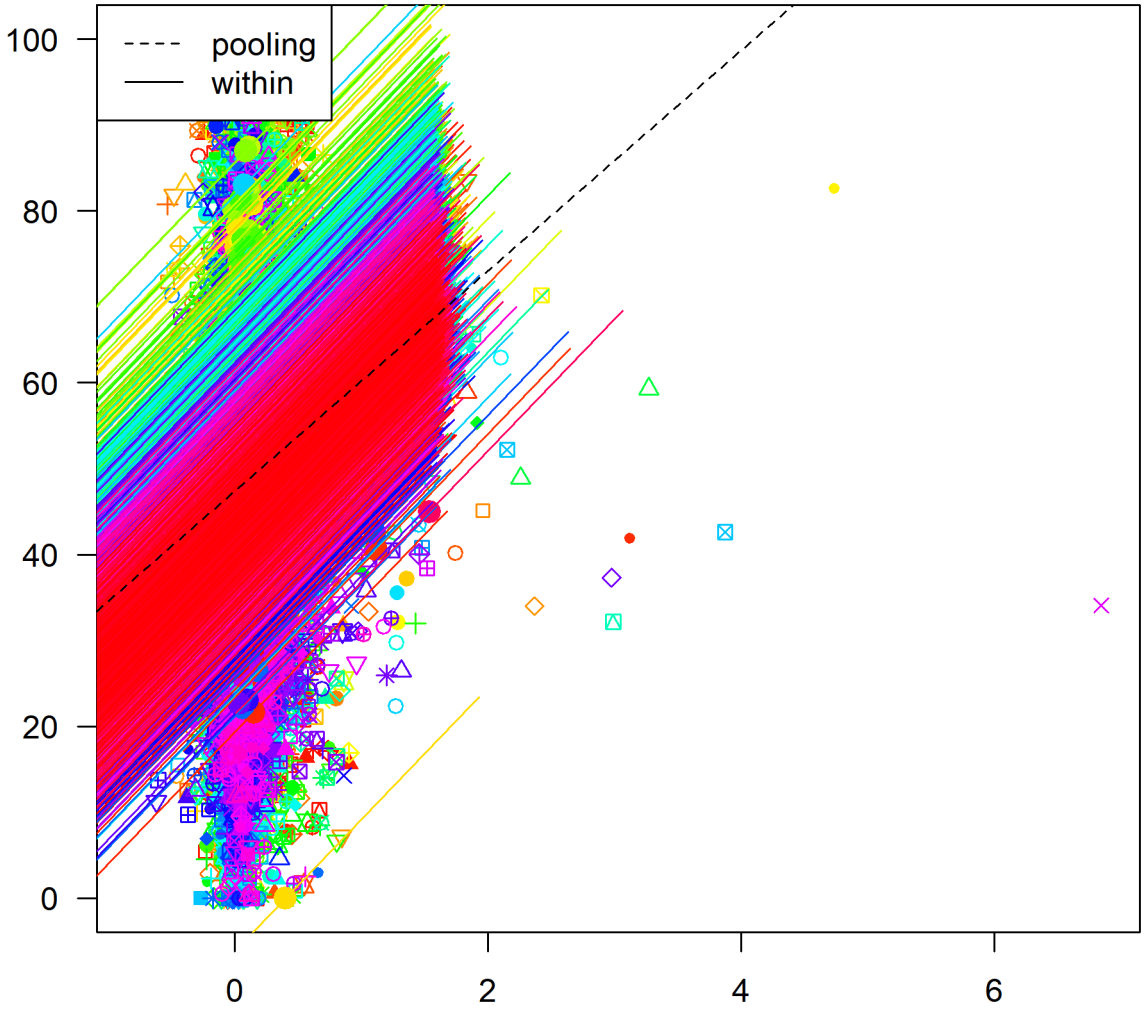
(a) O conhecimento de causa do pesquisador é o primeiro teste de plausibilidade. O quanto sabemos sobre o objeto estudado por vezes já é informativo sobre o grau de heterogeneidade ou homogeneidade entre unidades. Por exemplo, segundo a literatura, a dinâmica eleitoral na capital paulista é comparável ao que ocorre em Serra da Saudade (MG) que possui população de 781 habitantes? Assim, consultar a teoria sobre o assunto pode fornecer pistas sobre a plausibilidade de um *pooling*.

21 O modelo *pooled* também pode ser executado por meio da função *lm*. Entretanto, para execução dos testes para verificação de pressupostos é necessário que o banco de dados esteja no formato produzido via *plm*.

(**b**) Os gráficos são úteis para análise exploratória, principalmente quando N é pequeno. Para visualizar se podemos abarcar todos os municípios por uma única reta de regressão, podemos comparar, em um mesmo gráfico de dispersão, uma linha única aplicada à toda a população vs. linhas separadas para cada município. Exemplificamos a seguir com um gráfico bivariado com a principal variável independente, "crescimento econômico" (eixo horizontal), e a variável dependente "fração de votos no incumbente" (eixo vertical).

```
##### COMPARAÇÃO GRÁFICA INTERCEPTOS POLS VS FE #######

mode_eff <- plm(fracaovotos ~ cresc,
                data = BANCO, index = c("codibge", "ano", group = "codibge"),
                model = "within")
plot(mode_eff)
```

Figura 5 - Comparação Pooled OLS e FE para modelo bivariado fração de votos no incumbente x crescimento econômico



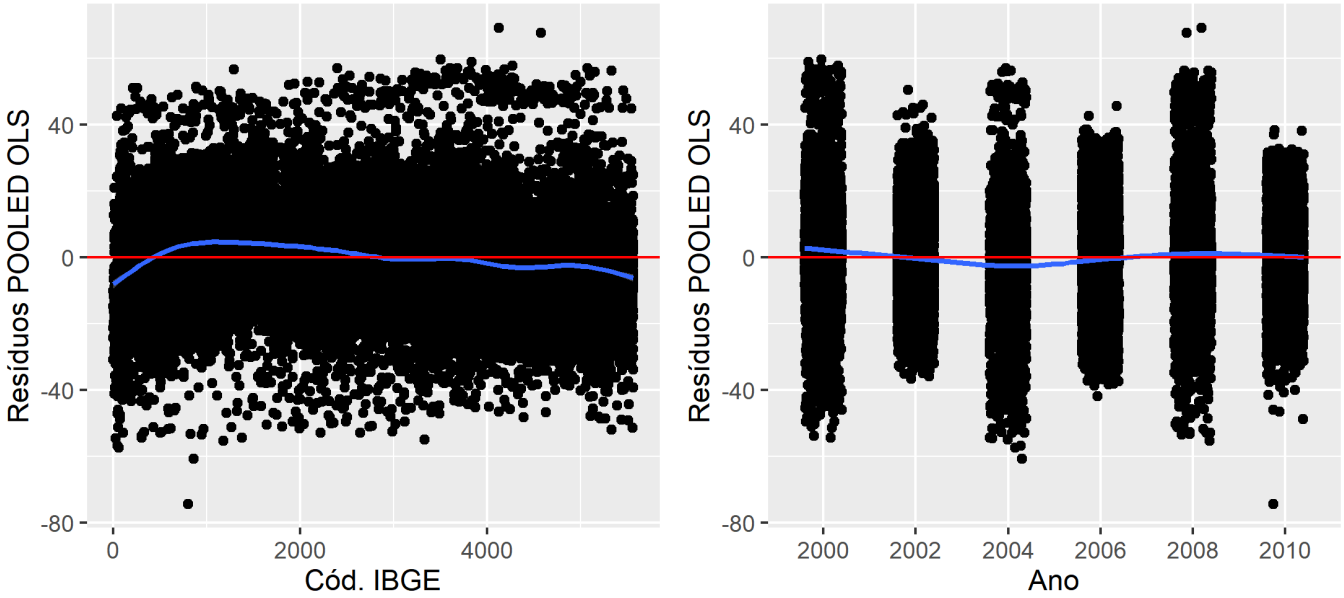
Fonte: elaborado pelos autores, baseado em dados de Fernandes e Fernandes (2017).

A solução *pooling* (linha pontilhada) propõe uma linha situada aproximadamente no meio das observações, ao passo que o modelo *within* (linhas sólidas coloridas) melhor representa a variação dos dados e indica que há larga diferença entre os interceptos de cada município.

Outra ferramenta de análise gráfica é o *plot* dos resíduos do modelo final. Como POLS pressupõe homogeneidade, seus resíduos devem ser semelhantemente aleatórios para todos os casos. Podemos agrupar os resíduos da regressão por município e por ano como forma de observar se a variância no termo de erro foi mais pronunciada para certos municípios ou períodos, como exemplificado adiante.

```
##### ANALISANDO OS RESIDUOS POOLED OLS ###########

library(broom)
library(ggplot2)

POLS_lm <- lm(fracaovotos ~ cresc+crescuf+crescbr+lpibreal+lpibuf+
               lpibbrasil+prefeitobasepresidente+persaude+lpop+leec+
               lheu+lses+laseps+ldesporc+ldespcor+linvest+ldespes,
             data = BANCO, na.action = "na.exclude")
# é necessario rodar o modelo POLS com lm para acrescentar
# os resíduos do modelo ao data.frame original

base <- augment(POLS_lm, data=BANCO)
# com o comando 'augment' do pacote broom, podemos juntar
# os resíduos do modelo com as variaveis do banco original.

                  #### Município ####

ggplot(base, aes(x=as.numeric(as.factor(codibge)), y=.resid))+
  geom_point()+
  geom_smooth()+
  geom_hline(yintercept = 0, color="red") + ylab("Resíduos POOLED OLS") +
  xlab("Cód. IBGE")
                  ####   Ano  ####

base$ano <- as.numeric(base$ano)

ggplot(base, aes(x=jitter(ano), y=.resid))+
  geom_point()+
  geom_smooth()+
  geom_hline(yintercept = 0, color="red") + ylab("Resíduos POOLED OLS") +
  xlab("Ano") + scale_x_discrete(limits=c("2000", "2002", "2004",
                                          "2006", "2008", "2010" ))
```

Figura 6 - Resíduos do modelo POLS agrupados por município e ano



Fonte: elaborado pelos autores, com base em Fernandes e Fernandes (2017).

Os resíduos dos municípios não são semelhantes: há algum fator sistematicamente diferenciando os municípios de código IBGE inferior a 3.000 dos demais com numeração mais alta. A variância entre anos eleitorais também é diferente. O motivo rápido se percebe: a variância é maior para anos de disputas municipais (2000, 2004 e 2008) e menor para as presidenciais (2002, 2006 e 2010). Esta situação ilustra o problema do *pooling* e da heterogeneidade, pois estamos tentando incluir num mesmo modelo preditivo eleições para prefeito e presidente. Esses resultados nos motivam a considerar a inclusão de *dummies* para cada ano ou tipo de eleição.

(c) o pacote *plm* permite implementar diferentes testes com o objetivo de verificar, sob a forma de hipótese nula/alternativa, se há efeitos individuais significativos. Podem ser aplicados o teste-F e o teste Breusch-Pagan LM. O teste-F compara se há diferenças significativas entre o POLS e o modelo FE, sendo a hipótese nula de que POLS é superior. Como podemos observar pelo resultado do p-valor (<0,05), o modelo FE é significativamente superior.

```
##### F-TEST POLS ################################

pFtest(mode_fe, POLS) # F-test com o modelo de efeitos fixos e Pooled
##
##  F test for individual effects
##
## data:  form
## F = 1.4214, df1 = 5548, df2 = 20786, p-value < 2.2e-16
## alternative hypothesis: significant effects
```

Já o teste Breusch-Pagan LM nos ajuda a escolher entre POLS e RE.[22] A hipótese nula do teste prevê que a variância específica por município é zero, o que recomendaria POLS. Conforme o resultado do teste, H0 é rejeitada e verifica-se então que POLS pode ser descartado (p-valor < 0,05).

```
##### BREUSCH-PAGAN TEST ########################

plmtest(POLS, type="bp", effect = "individual") # teste de Breusch-Pagan + efeitos
individuais
##
##  Lagrange Multiplier Test - (Breusch-Pagan) for unbalanced panels
##
## data:  form
## chisq = 107.15, df = 1, p-value < 2.2e-16
## alternative hypothesis: significant effects
```

Os comentários sobre efeitos individuais também se aplicam aos efeitos temporais. Seguindo a mesma lógica, trata-se de verificar se cada ano do estudo possui idiossincrasias específicas que fazem com que as observações naquele intervalo possuam alguma heterogeneidade peculiar. Isso corresponde à noção de choques no tempo. Assumindo tanto efeitos temporais como individuais, temos os chamados efeitos "*two ways*". Nos testes descritos acima, a presença desses efeitos pode ser testada modificando o argumento "*effect*", como exibido abaixo:

```
plmtest(POLS, type="bp", effect = "twoways") # teste de Breusch-Pagan + efeitos indi-
viduais e temporais
##
##  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
```

Diante da análise gráfica e do teste de Breusch-Pagan, sabemos que os modelos testados apresentam heterogeneidade espacial e também temporal. Assim, executamos novamente os modelos adicionando novas *dummies*. No estudo original, os autores adicionam ao modelo *dummies* para os anos e para diferenciar eleições municipais das presidenciais (Fernandes e Fernandes 2017, p.665). Embora na maioria das situações as *dummies* anuais sejam o caminho padrão para capturar heterogeneidade temporal, as especificidades desta base nos recomendam usar apenas a *dummy* de tipo de eleição.[23] Ela deverá aprimorar principalmente os modelos POLS e FD, enquanto que, para FE e RE, alteramos o efeito para "*two ways*", o que introduzirá efeitos por município e ano, implicando que *dummies* temporais como essa e variáveis como o PIB nacional possivelmente serão omitidas por colinearidade. Com a adição dessas variáveis, é possível controlar os efeitos heterogêneos identificados anteriormente.

22 O teste de Breusch-Pagan LM é um dos mais tradicionais, porém o comando *plmtest* pode calcular extensões do mesmo. Essas variações são úteis a depender das configurações dos dados em questão (*balanced vs unbalanced, short vs long*). Ver Croissant e Milo (2018, p. 86).

23 Para esses dados, as *dummies* para cada ano terão, muito provavelmente, colinearidade acentuada com algumas das demais variáveis, principalmente variáveis agregadas a nível nacional (como o PIB nacional, que terá um valor fixo a cada ano) e mesmo a nova *dummy* para eleições municipais.

```
##### MODELOS COM VARIÁVEL ELEIÇÃO ###############

# variavel para indicar se a eleição é pres (0) ou mun (1)
BANCO$tip_ele <- ifelse(BANCO$ano == 2000 |
                        BANCO$ano == 2004 |
                        BANCO$ano == 2008, 1, 0)

# formula com dummies para tipo de eleição
form_d <- as.numeric(fracaovotos) ~ cresc+crescuf+crescbr+lpibreal+lpibuf+
  lpibbrasil+prefeitobasepresidente+persaude+lpop+leec+
  lheu+lses+laseps+ldesporc+ldespcor+linvest+ldespes+
  tip_ele

##### POOLED OLS C DUMMIES ###############

POLS_dic <- plm(form_d, data = BANCO, model = "pooling")


##### FD C DUMMIES ########################

mode_fd_dic <- plm(form_d, data = BANCO, model = "fd")


##### FE TWO-WAYS #########################

mode_fe_2w <- plm(form_d, data = BANCO, model = "within", effect = "twoways")


##### RE TWO-WAYS #########################

mode_re_2w <- plm(form_d, data = BANCO, model = "random", effect = "twoways", random.
method = "walhus")
```

4.2 Efeitos fixos ou aleatórios?

Descartado POLS, podemos contemplar FE ou RE. Como já dissemos, nesta escolha é importante avaliar tanto a natureza dos dados quanto a intensidade da correlação entre μ_i e X. Para tanto, podemos recorrer ao (a) conhecimento contextual, (b) relevância ou indispensabilidade de certas variáveis de interesse, e (c) teste de Hausman.

(a) Como explicado na seção 3.2, o pesquisador deve responder à seguinte pergunta sobre a origem de seus dados: posso considerá-los como uma amostra aleatória de uma população mais ampla? Pensando em termos de exogeneidade, temos bons motivos para esperar que alguns municípios, simplesmente por ser, digamos, Salvador ou Sirinhaém, tenham peculiaridades não-observadas (e.g.: hegemonia histórica de certo partido) que se correlacionam com as variáveis selecionadas?

(b) Uma das limitações de FE (também de FD) é impedir o teste de elementos invariantes no tempo. Quando há interesse substantivo em testar variáveis deste tipo, o uso de RE pode ser contemplado, sempre levando em conta o *tradeoff* entre eficiência e viés. O modelo de Fernandes e Fernandes (2017) que replicamos não usa variáveis fixas no tempo, porém se desejássemos ir além e testar, por exemplo, *dummies* de controle para

as regiões brasileiras (Norte, Nordeste, Centro-Oeste, Sudeste e Sul), não poderíamos fazê-lo para FE.

(c) O comando *phtest* aplica o teste de especificação de Hausman (1978), comparando um modelo FE com RE, sob a hipótese nula de que ambos são consistentes. A hipótese nula será rejeitada quando se verificar correlação entre as heterogeneidades individuais não-observadas e as variáveis independentes. Em nosso caso, rejeita-se a hipótese nula que ambos FE e RE são consistentes. Sabemos portanto que há correlação entre a heterogeneidade individual não-observada e os regressores, de modo que apenas FE terá estimadores consistentes. É importante ter em mente que o teste de Hausman, mesmo sendo o padrão para discriminar entre FE e RE, pressupõe modelos bem especificados, livres de viés e homocedásticos.[24]

```
##### HAUSMAN TEST ##############################

phtest(mode_fe_2w, mode_re_2w) # teste de Hausman (modelo fixo e aleat.)
##
##  Hausman Test
##
## data:  form_d
## chisq = 851.98, df = 15, p-value < 2.2e-16
## alternative hypothesis: one model is inconsistent
phtest(form_d, data=BANCO, method="aux", vcov=vcovHC) # versão robusta do teste
##
##  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
```

4.3 Lidando com problemas temporais

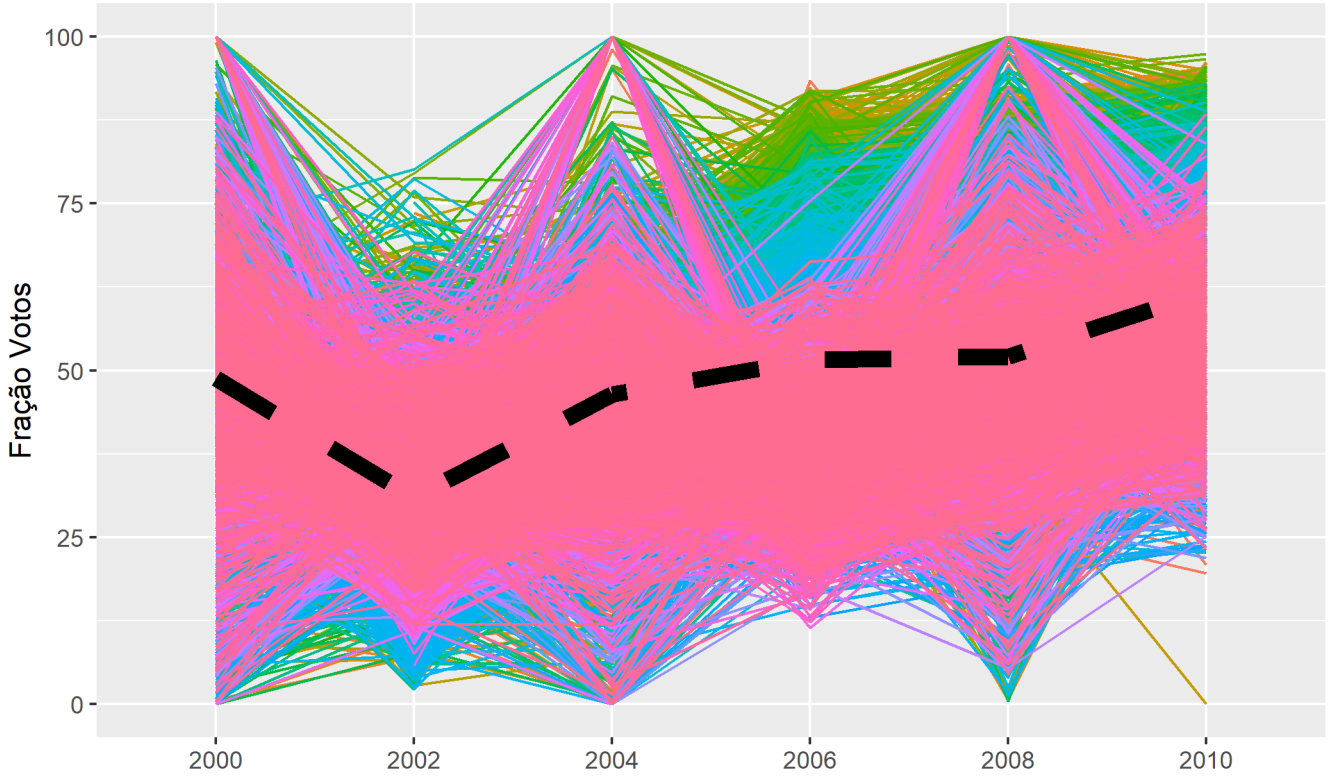
Como exposto na seção 3.4, os problemas temporais são tipicamente (a) não-estacionariedade e (b) correlação serial dos resíduos. O primeiro refere-se às variáveis brutas no banco de dados, e o segundo aos resíduos após a aplicação de modelos. Dada essa sequencialidade, deve-se primeiro verificar a estacionariedade.

(a) Painéis longos podem se beneficiar de testes como ACF, PACF, ADF ou *Panel Unit Root Tests* para verificar a estacionariedade. Como nosso painel é curto (T=6), esses testes seriam pouco informativos. Se ainda assim quisermos verificar de alguma maneira se os dados são estacionários ou não, podemos procurar de forma genérica por tendências não reversíveis de crescimento ou retração nas variáveis ao longo do tempo. Abaixo fazemos essa inspeção visual para a variável dependente "fração de votos", mostrando os valores por município e uma média geral de cada ano. Há alguma margem para esperarmos uma tendência de crescimento ao longo do tempo na fração de votos do incumbente.

24 Se os requisitos de ausência de *misspecification*, independência entre termo de erro e variáveis independentes, ausência de correlação serial e de heterocedasticidade não forem satisfeitos, recomenda-se, além inicialmente da revisão do modelo, o uso de formas robustas do teste. No *plm*, isso se opera com os argumentos *phtest(..., method="aux")* ou ainda *phtest(..., method="aux", vcov=vcovHC)* para matrizes variância-covariância customizadas.

```
##### VERIFICAR ESTACIONARIEDADE #############

library(dplyr)
##
## Attaching package: 'dplyr'
## The following objects are masked from 'package:plm':
##
##     between, lag, lead
## The following objects are masked from 'package:stats':
##
##     filter, lag
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
banco_est <- BANCO %>%
  group_by(ano) %>%
  mutate(mediafvotos = mean(fracaovotos, na.rm=T)) #banco com a media dos votos

ggplot(banco_est, aes(x=ano, y=fracaovotos, group=codibge, color=codibge))+
  geom_line() + geom_line(aes(x=ano, y=mediafvotos, group=1),
  linetype="dashed", size=3, color="black")+ ylab("Fração Votos") +
  xlab("") + guides(color="none")
```

Figura 7 - Fração de votos ao longo do tempo (municípios e média geral)



Fonte: elaborado pelos autores, com base em dados originais de Fernandes e Fernandes (2017).

Se há tendência, recomenda-se a inclusão de uma *time trend*, geralmente uma variável T que assume os valores de 1 para o primeiro ano, 2 para o segundo, e assim por diante.

(b) Até então nossa seleção entre os estimadores POLS, FE, RE e FD se pautou em quão bem cada um lidava com a heterogeneidade não-observada. Agora, o sucesso em eliminar correlações temporais também se torna importante para eleger um estimador. Como explicado em 2.4, há diversos testes para correlação serial que podem ser aplicados aos resíduos, a depender do tipo de painel (curto ou longo) e estimador empregado. O pacote *plm* oferece atualmente oito testes para correlação serial.[25] Como o teste de Hausman já nos aconselhou abandonar RE, nós utilizaremos apenas dois testes, voltados para FE e FD.

Para verificar a presença de correlação serial em modelos de efeito fixo, utilizamos a função *pwartest*. Como podemos observar no *output*, a hipótese alternativa do teste é de que existe correlação serial. O p-valor foi menor que 0,05, indicando a rejeição da hipótese nula (não existe correlação serial). Ou seja, conforme o resultado do teste, o modelo de efeitos fixos apresenta correlação serial.

```
##### TESTAR CORR. SERIAL PARA FE ########################

pwartest(mode_fe_2w) #pwartest para o modelo FE
##
##  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
```

Já o *pwfdtest* é utilizado para verificar a presença de correlação serial nos modelos FD. A hipótese nula é de que não existe correlação serial relacionada aos erros. Dado o p-valor (< 0,05), rejeitamos a hipótese nula. Ou seja, o modelo FD apresenta problemas de correlação serial.

```
##### TESTAR CORR. SERIAL PARA FD ######################

pwfdtest(mode_fd_dic) #pwfdtest para o modelo FD
##
##  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
```

Os resultados dos testes sugerem que tanto FE quanto FD abrigam correlação serial nos termos de erro. Da mesma forma, a inspeção dos dados brutos indicam tendência ascendente, ou seja, não-estacionariedade. Para remediar a situação, poderíamos contemplar a inclusão

25 Os seis testes não cobertos neste artigo são: teste genérico de Wooldridge para efeitos não-observados, aplicável a POLS (*pwtest*); teste de Breusch-Godfrey/Wooldridge para correlação serial, preferencialmente aplicável a RE, porém também a FEs longos (*pbgtest*); teste de Durbin-Watson, também preferencialmente a RE (*pdwtest*) e sua generalização por Bhargava, Narendranathan e Franzini (*pbnftest*); teste localmente robusto de Bera, Sosa-Escudero e Yoon para correlação serial ou efeitos aleatórios (*pbsytest*); teste de Baltagi e Li para correlação serial (*pbltest*).

de uma *time trend* (porém, como vimos em 2.2, essas não são estimáveis por FE ou FD), LDV ou emprego de erros-padrão robustos à correlação serial. Como foi visto anteriormente, problemas temporais tendem a ser mais preocupantes conforme T>N. Em nosso caso, tendo em vista que N é muito superior a T, e para manter a parcimônia do exemplo, vamos nos contentar em preservar as variáveis atuais e ao final empregar robustificação.

4.4 Lidando com problemas espaciais

Uma vez que já lidamos com a heterogeneidade não-observada, é de se esperar que os problemas espaciais restantes digam respeito à heteroscedasticidade de painel e à correlação contemporânea dos erros. A mesma inspeção visual aplicada na seção 4.1, agrupando resíduos por município, também informa esta decisão. A Figura 6 já mostrou que a variância dos resíduos entre municípios é sistemática.

Além da inspeção visual, podemos recorrer a testes específicos. O pacote *plm* oferece um comando geral para testar *cross-sectional dependence* (*pcdtest*), cabendo ao usuário escolher o tipo de teste, segundo as características do painel. Como temos T curto e N longo[26], o teste mais apropriado é o proposto por Pesaran (2004).

```
##### TESTAR HETEROSC. DE PAINEL  ##################################

pcdtest(mode_fe_2w, test = "cd") # modelo Efeitos Fixos
##
##  Pesaran CD test for cross-sectional dependence in panels
##
## data:  as.numeric(fracaovotos) ~ cresc + crescuf + crescbr + lpibreal +     lpi-
buf + lpibbrasil + prefeitobasepresidente + persaude +     lpop + leec + lheu + lses
+ laseps + ldesporc + ldespcor +     linvest + ldespes + tip_ele
## z = -0.66933, p-value = 0.5033
## alternative hypothesis: cross-sectional dependence
```

Os resultados mostram que nosso modelo FE *two ways* está livre de dependência contemporânea, tendo apenas problemas de correlação serial. Devemos aplicar erros-padrão robustos contra correlação serial. Podemos acessar diferentes opções de robustez no *plm*, a depender do tipo de painel, da forma de agrupamento dos dados e do tipo de problema a remediar (ver Croissant e Millo 2018, cap. 5; Henningsen e Henningsen 2019, pp. 382-384). Em nosso caso, como queremos lidar com a correlação serial de um FE com efeitos *two ways*, aplicamos a matriz *vcovDC* para estimar erros-padrão robustos:

```
##### LIDANDO COM CORR. SERIAL ############

library(lmtest)
mode_fe2w_rbst <- coeftest(mode_fe_2w,
                           vcov = vcovDC)
# adicionando erros robustos 'double cluster'
```

26 Breusch-Pagan LM, acionado pelo argumento test="lm" é recomendado para T longo e N curto; scaled-Breusch-Pagan LM, "sclm", para T e N longos; "rho" e "absrho" fornecem os valores do coeficiente de correlação entre resíduos para pares de observações (ver Croissant e Millo, 2018, seção 4.4).

4.5 Resultados

Apresentamos os resultados de todos os modelos na Tabela 1. Para permitir a comparação dos erros-padrão originais do modelo FE com os robustos, colocamos ambos em colunas separadas na tabela.

Sendo nosso interesse principal pedagógico e não substantivo, apresentamos apenas uma versão resumida da lista de variáveis. Assim, das mais de 20 variáveis do modelo original, conservamos 11 para a tabela.

Tabela 1 - Resultados da replicação

	Fração de Votos				
	POLS	**RE (2W)**	**FE (2W)**	**FE (2W, Robust.)**	**FD**
	(1)	**(2)**	**(3)**	**(4)**	**(5)**
Cresc. Munic.	1.683***	1.098***	1.148*	1.148	0.190
	(0.608)	(0.040)	(0.652)	(1.517)	(0.669)
Cresc. UF	0.095***	0.086***	-0.058*	-0.058	-0.144***
	(0.027)	(0.002)	(0.030)	(0.163)	(0.033)
Cresc. BR	3.660***	3.709***			4.231***
	(0.090)	(0.088)			(0.091)
Prefeito Base Presid.	1.211***	0.742***	1.289***	1.289***	1.136***
	(0.208)	(0.014)	(0.235)	(0.409)	(0.254)
Perc. Orç. Saúde	0.040***	0.121***	0.044*	0.044	0.058***
	(0.014)	(0.001)	(0.023)	(0.035)	(0.019)
Pop. Munic. (log)	0.302	-0.415***	-6.037***	-6.037**	-3.360*
	(0.258)	(0.018)	(1.449)	(2.609)	(1.824)
Desp. Educ. Cult. (log)	6.917***	6.461***	1.543***	1.543	0.162
	(0.358)	(0.025)	(0.525)	(1.284)	(0.553)
Desp. Invest. (log)	2.245***	1.983***	1.430***	1.430	1.644***
	(0.136)	(0.009)	(0.165)	(0.903)	(0.180)
Desp. Pessoal (log)	-1.424***	-0.335***	3.304***	3.304	1.167*
	(0.493)	(0.035)	(0.635)	(2.264)	(0.698)
Eleição Municip.	1.074***	0.828***			-0.915***
	(0.277)	(0.246)			(0.264)
Intercepto	102.026**	152.778***			-4.350***
	(43.701)	(37.792)			(0.373)
N Obs.	**26.352**	**26.352**	**26.352**		**20.803**
R2	**0.321**	**0.319**	**0.039**		**0.254**
R2 Ajust.	**0.320**	**0.318**	**-0.219**		**0.253**
Estat. F	**691.129*** (df = 18; 26333)**	**12.309.190*** **	**55.646*** (df = 15; 20783)**		**392.979*** (df = 18; 20784)**

Fonte: elaborado pelos autores, com base em Fernandes e Fernandes (2017). O valor de theta para o modelo de efeitos aleatórios two ways foi θ = 0,11.

Os testes de especificação nos conduzem a ter mais confiança nos resultados obtidos por meio de FE. Não obstante, a comparação de todos os modelos se revela instrutiva. Os resultados de POLS e RE tendem a estar mais próximos um do outro para muitos dos regressores, enquanto que as estimativas de FE e FD são mais semelhantes entre si. Isso é esperado pois, como vimos, os dois últimos removem a heterogeneidade individual de forma semelhante. Assim, ler os resultados de POLS/RE até FE/FD é uma maneira de verificar como se comportam as estimativas conforme vamos controlando mais e mais as idiossincrasias municipais; o "efeito líquido", por assim dizer, das variáveis, descontadas todas as peculiaridades dos casos.

Em algumas variáveis, a diferença se manifesta na troca de sinal. Por exemplo, os modelos POLS e RE afirmam que haveria efeitos positivos da taxa de crescimento econômico da UF sobre a fração de votos, ou negativos de despesas com pessoal. Temos um caso de viés pois, controlando pelas heterogeneidades de cada município por meio de FE e FD, vemos que seus impactos seriam na verdade o inverso.

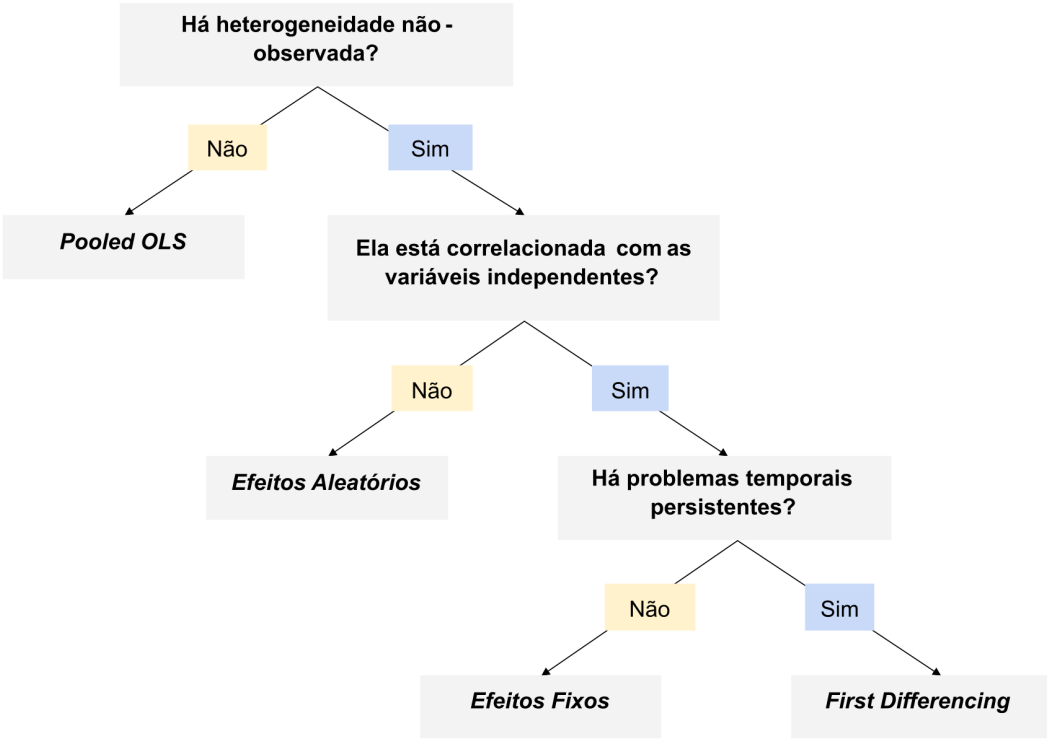
Para outros regressores, percebe-se uma mudança na força da relação. Às vezes a magnitude do efeito se torna mais débil (i.e.: despesa em educação e cultura, despesa em investimento), ocasionalmente mais forte (*dummy* para mesma base prefeito-presidente, população), e por vezes estável (percentual do orçamento de saúde). Ademais, enquanto POLS e RE tendem a acusar mais significância estatística, FD e principalmente FE com erros padrão robustos são mais conservadores. Notadamente, a principal variável de interesse do estudo, crescimento econômico municipal, não retém efeito significativo sobre os votos no incumbente.

Por fim, deve ser apontado que os interceptos POLS e RE apresentam valores irrealistas pelo fato de os dados não terem sido centralizados e que, como antecipamos, FE *two ways* omitiu *dummies* temporais (tipo de eleição) e aquelas fixas ao ano (PIB BR) por colinearidade.

4.6 Resumo do procedimento

Para concluir este tutorial, podemos representar o processo de escolha de um modelo como uma árvore de decisão similar à Figura 8.

Figura 8 - Como escolher o melhor estimador?



Fonte: elaborado pelos autores.[27]

Esquematicamente, primeiro se avalia em que medida existe heterogeneidade não-observada entre as entidades da base de dados. Isso pode ser feito a partir do conhecimento contextual do problema, da análise gráfica dos interceptos e pelo emprego de testes específicos (F e Breusch-Pagan LM). Uma vez descartada a opção POLS, devemos estimar se existe correlação entre a heterogeneidade não-observada e o conjunto de variáveis explicativas, por exemplo através do teste de Hausman. Inexistindo correlação, justifica-se o uso de RE. Por outro lado, na presença de correlação entre os erros e as variáveis independentes, podemos cogitar a utilização de FE ou FD. A opção entre esses dois últimos deve ser equilibrada por considerações de problemas temporais do modelo, bem como tamanho da base. Ao final do processo, testes para dependências temporais e espaciais são necessários para verificar a necessidade de erros-padrão robustos.

Este esquema é uma simplificação para fins didáticos e enfatiza as principais decisões que, se ignoradas, poderiam gerar viés ou ineficiência. Porém, como argumentamos ao longo do artigo, considerações substantivas também devem guiar essa escolha.

5. Conclusão

Este artigo apresentou uma introdução à lógica de dados de painel. Mostramos os principais benefícios advindos da análise de informações longitudinais e descrevemos os problemas mais frequentes que surgem quando a planilha de dados reúne, simultaneamente, dimensões de espaço e tempo. Depois disso, demonstramos como implementar uma análise de regressão com dados de painel no R. Por se tratar de um trabalho com finalidade pedagógica, compartilhamos ainda todos os dados e scripts de modo que alunos, professores e pesquisadores podem facilmente adaptar nossas rotinas computacionais aos seus interesses substantivos.

27 Este esquema é inspirado em um original mais complexo, que os autores creditam à professora Claudia Schwirplies (Uni HH).

No início deste artigo, apontamos como o uso de dados de painel na CPRI brasileira ainda é incipiente. Assim, acreditamos ser relevante explicar o que pode ser feito para mudar esse quadro. Entre as diferentes alternativas disponíveis, destacamos o seguinte: (1) reformular a estrutura curricular dos cursos de graduação e pós-graduação com o objetivo de incluir disciplinas específicas de cálculo, estatística e computação e ofertar disciplinas eletivas sobre como trabalhar com dados longitudinais, tal qual o curso do professor William Greene; (2) na impossibilidade de implementar o item 1, garantir a oferta de cursos especiais de verão, como o *Modeling Dynamics* ministrado pela professora Lorena Barberia (USP), *Modeling Dynamics in Space and Time* ministrado pelos professores Guy Whitten e Lorena Barberia (USP) e Análise de Dados em Painel por Igor Viveiros no curso MQ/UFMG; (3) diversificar, regionalmente, a oferta intermitente de cursos de análise de dados. Na impossibilidade de implementar o item 3, garantir apoio financeiro para que estudantes de mestrado e doutorado, principalmente do Norte e Nordeste, possam frequentar escolas de verão; (4) fomentar o aproveitamento de cursos on-line como créditos currículos válidos durante a pós-graduação; (5) edição, por parte das principais revistas de CPRI, de dossiês especiais sobre métodos de pesquisa, com especial ênfase para artigos intuitivos e pedagógicos, tal qual o volume 142 do Journal of Econometrics. Ainda, reservar, nas melhores revistas de CPRI, espaço para publicação de artigos com ênfase metodológica como Beck e Katz (1995); e (6) incentivar os pesquisadores que se dedicam à divulgação de métodos e técnicas de pesquisa em CPRI a partir de linhas específicas de financiamento em instituições de fomento (CAPES e CNPq).

Sabemos das dificuldades operacionais e logísticas que permeiam cada uma dessas iniciativas. No entanto, considerando os potenciais benefícios que podem ser incorporados em nossos modelos de análise e a limitada utilização de dados longitudinais na pesquisa empírica em CPRI, acreditamos que já passou da hora de levar o tempo a sério (Beck, Katz e Tucker, 1998; Boef e Keele, 2008).

REFERÊNCIAS BIBLIOGRÁFICAS

ACHEN, C. H. (2000), "Why lagged dependent variables can suppress the explanatory power of other independent variables". In: *annual meeting of the political methodology section of the American political science association, UCLA*.

BALTAGI, Badi H. (2005), *Econometric Analysis of Panel Data*. England, John Wiley & Sons Ltd.

BERRY, William D. (1993). *Understanding regression assumptions*. Thousand Oaks, Sage.

BECK, Nathaniel; KATZ, Jonathan N. (1995), "What to do (and not to do) with time-series cross-section data", *American political science review*, v. 89, n. 3: 634-647.

BECK, Nathaniel; KATZ, Jonathan N.; TUCKER, Richard. (1998), "Taking time seriously: Time-series-cross-section analysis with a binary dependent variable", *American Journal of Political Science*, v. 42, n. 4: 1260-1288.

BECK, Nathaniel. (2001), "TIME-SERIES–CROSS-SECTION DATA: What Have We Learned in the Past Few Years?", *Annual Review of Political Science*, v. 4, n. 1: 271–293.

BECK, Nathaniel. (2008), "Time-Series Cross-Section Methods", *in* J. Box-Steffensmeier; H. Brady; D. Collier (Eds.). *The Oxford Handbook of Political Methodology*. New York, OUP: 475–493.

BLALOCK, Hubert M. The presidential address: Measurement and conceptualization problems: The major obstacle to integrating theory and research. *American sociological review*, v. 44, n. 6, p. 881-894, 1979.

DE BOEF, Suzanna; KEELE, Luke. (2008), "Taking time seriously", *American Journal of Political Science*, v. 52, n. 1: 184-200.

BELL, Andrew; JONES, Kelvyn. (2015), "Explaining Fixed Effects: Random Effects Modeling of Time-Series Cross-Sectional and Panel Data", *Political Science Research and Methods*, v. 3, n. 1: 133–153.

BELL, Andrew; FAIRBROTHER, Malcolm; JONES, Kelvyn. (2019), "Fixed and random effects models: making an informed choice", *Quality & Quantity*, v. 53, n. 2: 1051-1074.

BENOIT, K. et al. (2018), "quanteda: An R package for the quantitative analysis of textual data", *Journal of Open Source Software*, v. 3, n. 30: 774.

BOX, G. E. P.; JENKINS, G. M.; REINSEL, G. C.; LJUNG, G. M. (2016), *Time series analysis: forecasting and control*. Hoboken/New Jersey, Wiley.

CLARK, Tom S.; LINZER, Drew A. (2015), "Should I Use Fixed or Random Effects?", *Political Science Research and Methods*, v. 3, n. 2: 399–408.

CROISSANT, Yves; MILLO, Giovanni. (2008), "Panel data econometrics in R: The plm package", *Journal of statistical software*, v. 27, n. 2: 1-43.

CROISSANT, Yves; MILLO, Giovanni. (2018), *Panel Data Econometrics with R*, Wiley.

CERQUEIRA, Daniel et al. (2017), *Atlas da violência 2017*.

DIELEMAN, J. L.; TEMPLIN, T. (2014), "Random-Effects, Fixed-Effects and the within-between Specification for Clustered Data in Observational Health Studies: A Simulation Study", *PLoS ONE*, v. 9, n. 10: e110257.

FÁVERO, L. P. L. (2013), "Dados em painel em contabilidade e finanças: teoria e aplicação", *BBR-Brazilian Business Review*, v. 10, n. 1: 131-156.

FERNANDES, Ivan Filipe de Almeida Lopes; FERNANDES, Gustavo Andrey de Almeida Lopes. (2017), "A importância do crescimento econômico local na escolha do chefe do Executivo no Brasil", *Revista de Administração Pública*, v. 51, n. 4: 653-688.

FINKEL, S. E. (1995), *Causal analysis with panel data* (No. 105). Sage.

FORTIN-RITTBERGER, J. (2013), "Time-Series Cross-Section", *in* H. Best; C. Wolf (Eds.), *The SAGE Handbook of Regression Analysis and Causal Inference*. London, SAGE Publications: 387–408.

GUJARATI, Damodar N. *Basic Econometrics*. McGraw-Hill Companies. New York, 2004.

HAUSMAN, Jerry A. (1978), "Specification tests in econometrics", *Econometrica: Journal of the econometric society*, v. 46, n. 6: 1251-1271.

HENNINGSEN, Arne; HENNINGSEN, Géraldine. (2019), "Analysis of Panel Data Using R", *in* M. Tsionas (Ed.) *Panel Data Econometrics: Theory*. London, Elsevier: 345-396.

HSIAO, C. (2003), *Panel data analysis*, Cambridge University Press.

IBGE. (2017), Pesquisa Nacional por Amostra de Domicílios Contínua. Disponível em: https://www.ibge.gov.br/estatisticas/sociais/trabalho/17270-pnad-continua.html?=&t=o-que-e. Acesso em 30-06-2020

KING, Gary. (2001). "Proper Nouns and Methodological Propriety: Pooling Dyads in International Relations Data", *International Organization*, v. 55, n. 2: 497-507.

KING, Gary. (1995), "Replication, replication", *PS: Political Science and Politics*, v. 28, n. 3: 444-452.

KING, Gary; KEOHANE, Robert O.; VERBA, Sidney. (1994), *Designing social inquiry: Scientific inference in qualitative research*. Princeton University Press.

KENNEDY, Peter. (2003), *A guide to econometrics*. MIT Press.

KRUEGER, James S.; LEWIS-BECK, Michael S. (2008), "Is OLS dead?", *The Political Methodologist*, v. 15, n. 2: 2-4.

LAROCCA, R. (2012). "Gauss–Markov Theorem" in SALKIND, N. J. (ed) *Encyclopedia of Research Design*. Thousand Oaks, Sage, 529-533.

LIJPHART, A. (1971), "Comparative Politics and the Comparative Method", *American Political Science Review*, v. 65, n. 3: 682–693.

MARQUES, Luís D. (2000), "Modelos Dinâmicos com Dados em Painel: revisão de literatura". *Working paper*. Disponível em http://wps.fep.up.pt/wps/wp100.pdf. Acesso em 23-06-2020.

MENARD, Scott. (2002), *Longitudinal research*. Thousand Oaks, SAGE Publications.

MEDEIROS, M.; BARNABE, I.; ALBUQUERQUE, R.; LIMA, R. (2016), "What does the field of International Relations look like in South America?", *Revista Brasileira de Política Internacional*, v. 59, n. 1: e004

MEIRELES, Fernando; SILVA, Denisson; BARBOSA, Rogerio. (2019), *rscielo: A Scraper for Scientific Journals Hosted on Scielo*. R package version 1.0.0. https://CRAN.R-project.org/package=rscielo.

MESQUITA, Rafael. (2018) *Liderança regional em perspectiva comparada: Brasil e Turquia*. Tese de doutorado, UFPE. Disponível em https://repositorio.ufpe.br/handle/123456789/32942. Acesso em 29-05-2021

MUMMOLO, J.; PETERSON, E. (2018). "Improving the interpretation of fixed effects regression results", *Political Science Research and Methods*, v. 6, n. 4: 829-835.

PESARAN, M. H. (2004), "General diagnostic tests for cross section dependence in panels", *CESifo Working Paper Series*, 1229.

PEARL, J. (2000), *Causality: Models, reasoning, and inference*. New York, Cambridge University Press.

SKRONDAL, A., RABE-HESKETH, S. (2008), "Multilevel and Related Models for Longitudinal Data", *in* Leeuw J.., Meijer E. (Eds.) *Handbook of Multilevel Analysis*, New York, Springer: 275-300.

STOCK, James H.; WATSON, Mark W. (2004), *Econometria*. Pearson.

VERBEEK, Marno; NIJMAN, Theo. (1992), "Testing for selectivity bias in panel data models", *International Economic Review*, v. 33, n. 3: 681-703.

WOOLDRIDGE, J. M. (2002), *Econometric Analysis of Cross Section and Panel Data*, Cambridge, MIT Press.

WOOLDRIDGE, J. M. (2013), *Introductory econometrics: a modern approach*. Australia, South-Western Cengage Learning.

WORRALL, John L. (2008), "An introduction to pooling cross-sectional and time series data", *in* S. Menard (Ed.) *Handbook of longitudinal research: Design, measurement, and analysis*, Elsevier: 233-248.

Perguntas e Respostas

1. **Devo usar efeitos fixos ou aleatórios?**

Em geral, efeitos aleatórios podem introduzir viés, isto é, aumentar a chance de subestimar ou superestimar sistematicamente o parâmetro de interesse. Porém, esse método tende a reduzir a variância dos coeficientes, o que significa estimativas mais eficientes (Kennedy, 2003). Assim, modelos com efeitos aleatórios tendem a produzir estimativas mais consistentes quando a amostra é pequena e quando a correlação entre a variável independente e os efeitos das observações é relativamente baixa (Clark e Linzer, 2015). Por sua vez, efeitos fixos tendem a produzir estimativas não-viesadas, principalmente em amostras grandes (Mummolo e Peterson, 2018). Uma desvantagem dos efeitos fixos é a redução do grau de liberdade do modelo por conta da inclusão de *dummies*. Se, por outro lado, a amostra é suficientemente grande, isso não será um problema na estimação dos coeficientes. Para mais informações: Wooldridge (2002) e Clark e Linzer (2015).

2. Qual é a diferença entre um painel equilibrado e um painel desequilibrado?

Stock e Watson (2004) definem painel equilibrado como um conjunto de dados em que não existem casos ausentes (*missing cases*), ou seja, temos informação para todas as unidades observadas (N) em todos os períodos do tempo (T). O desequilíbrio aparece quando, por qualquer motivo, ocorrem descontinuidades na matriz de dados, seja na falta de observações, seja pela ausência de unidades de tempo. Wooldridge (2002) alerta que, de forma geral, o tratamento estatístico será similar para ambos os tipos de painel. Em todos os casos, o pesquisador deve informar a razão do desequilíbrio e descrever detalhadamente os critérios de seleção da amostra. Para uma introdução sobre o assunto, ver Mayer (2010)[1]. Para uma discussão detalhada das especificidades do painel desequilibrado, Wooldridge (2002) e Baltagi (2005). Para um tratamento mais avançado sobre viés de seleção em dados de painel, ver Verbeek e Nijman (1992).

3. Painéis curtos versus painéis longos?

A amplitude do painel depende da razão entre entidades (N) e unidades de tempo (T). No painel curto, temos muitas entidades observadas durante um período relativamente curto de tempo. Imagine um conjunto de dados com informações detalhadas para todos os municípios brasileiros em 2000 e 2010. Sem dúvida, é muita informação, mas como a quantidade de entidades (mais de 5 mil municípios) é muito superior ao número de períodos de tempo (10 anos), estamos diante de um painel curto em que há dominância da dimensão transversal. O painel longo ocorre quando a oferta de dados transversais é significativamente menor do que a amplitude temporal das informações. Por exemplo, imagine uma base de dados das cinco regiões do país (N, NE, CO, SE e S) reunidas anualmente entre 1950 e 2020. Essa composição em que T>N nos aproxima de uma tendência *time series*, exigindo técnicas específicas. Para mais informações, ver Cameron e Trivedi (2005)[2] e Stock e Watson (2004).

1 Ver: <https://homepage.univie.ac.at/robert.kunst/pan2010_pres_mayer.pdf>.

2 Ver: <http://public.econ.duke.edu/~vjh3/e262p_07S/readings/Cameron_&_Trivedi_Microeconometrics_Chapters_2.pdf>.

4. Como evitar viés de variável omitida com dados de painel?

Como em outros desenhos de pesquisa, é preciso estar a par da literatura específica da área para conhecer a especificação teórica dos modelos. Muitas vezes, podemos achar uma variável importante em alguma base de dados já existente. Tipicamente, FE e FD oferecem boas salvaguardas contra variáveis omitidas quando os *unobservables* importantes são fixos às unidades. Mesmo assim, seu uso não substitui conhecimento abalizado da literatura. Lembrando que a inclusão de uma variável irrelevante no modelo explicativo apenas prejudicará a eficiência dos coeficientes. O erro padrão será maior do que deveria e isso reduz a confiabilidade do testes de significância, mas $\hat{\beta}$ ainda fornecerá uma estimativa consistente de β, ou seja, a inclusão de uma variável irrelevante não causa viés. Por outro lado, a exclusão de uma variável teoricamente importante tende a produzir estimativas viesadas, principalmente quando a variável omitida está correlacionada com as demais variáveis incluídas no modelo.

5. O que é e para que serve o teste de Hausman?

O objetivo do teste de Hausman é avaliar se existe correlação entre os erros do modelo e as variáveis independentes. No contexto de dados de painel, o teste é empregado para orientar a decisão entre efeitos fixos e efeitos aleatórios. A hipótese nula sustenta que os efeitos aleatórios são mais adequados. Operacionalmente, devemos estimar os dois modelos (com efeitos fixos e aleatórios) e depois comparar as estimativas. Um resultado significativo (p-valor<0,05) indica que a hipótese nula deve ser rejeitada, ou seja, devemos optar por efeitos fixos. Por outro lado, um resultado não significativo (p-valor>0,05) implica na não-rejeição da hipótese nula, o que significa a adoção do modelo com efeitos aleatórios. O teste de Hausman geralmente vem acompanhado de uma distribuição de chi-quadrado com teste de significância, além das estimativas dos coeficientes. Para mais, ver Hausman (1978).

6. O que é a autocorrelação serial e por que eu devo me preocupar com ela?

Embora a autocorrelação possa ocorrer no espaço, a autocorrelação que nos interessa aqui é aquela que ocorre no tempo, também chamada de correlação serial. O dicionário de Estatística de Cambridge define correlação serial como a dependência de mensurações da mesma entidade em estudos longitudinais. Em geral, quanto maior a proximidade temporal das mensurações, maior a magnitude da correlação. Por outro lado, quanto mais distante as mensurações no tempo, em média, menor o grau de correlação entre os valores da série. Essa estrutura de associação dos valores no tempo viola o pressuposto de independência dos erros no modelo de regressão. O principal problema da autocorrelação serial é a produção de testes de significância inconsistentes, ou seja, não poderemos confiar no p-valor e nos intervalos de confiança. É por esse motivo que a análise de dados longitudinais sempre deve vir acompanhada de estimativas do nível de autocorrelação da série.

7. Como aprender mais sobre análise de dados de painel?

Em Econometria, sugerimos Kennedy (2003), Wooldridge (2002, 2013), Baltagi (2005) e Gujarati (2004). Hsiao (2003) apresenta uma das mais completas introdução ao tema. O artigo "*What to do and (not to do with) with Time-Series Cross-Section Data*" de Beck e Katz (1995) é o trabalho mais amplamente citado em Ciência Política (7.414

citações no *google scholar* em junho de 2021). Fortin-Rittberger (2013) também apresenta um trabalho introdutório. Finkel (1995) enfoca as possibilidades causais a partir de dados longitudinais. Em dúvida sobre efeitos fixos ou aleatórios? Ver Bell e Jones (2015). Mummolo e Peterson (2018) reproduzem vários estudos que empregam dados de painel e demonstram como melhorar a interpretação dos coeficientes estimados utilizados efeitos fixos. Veja a programação do curso de verão da IPSA/USP: <http://summerschool.fflch.usp.br/tracks-and-courses/>.

8. **TSCS x dados de painel?**

Beck (2008) diferencia dados de painel dos dados TSCS. Para ele, os métodos que funcionam bem em painéis não necessariamente funcionam em dados TSCS e vice-versa. Na pesquisa empírica em CPRI, a combinação mais frequente de casos e tempo se materializa em bancos de dados que compilam informações sobre países (N) ao longo de anos (T) (Beck, 2001). Assim, como os dados do tipo TSCS consistem em exatamente os mesmos casos ao longo do tempo, espera-se maior correlação temporal do que para dados de painel. TSCS consiste na coleta de informações sobre unidades observacionais (indivíduos, firmas, unidades federativas, países) dispersas ao longo do tempo (dia, mês, ano). Quando maior for o T em relação ao N, maiores serão as preocupações sobre a estrutura serial dos dados. Por outro lado, quando existem muitos casos (N) para poucos períodos de tempo (T), menos graves serão problemas de autocorrelação.

9. Como recuperar os interceptos de cada unidade quando uso FE ou RE no plm?

Use os comandos fixef() e ranef(). Os valores correspondem ao desvio em comparação ao intercepto geral.

10. Qual a diferença entre utilizar o *plm* ou os pacotes *lme4* e *nlme*?

Dados em painel não são a única forma de lidar com mensurações repetidas. Outra abordagem consiste no uso de modelos lineares mistos, também conhecidos como modelos hierárquicos ou modelos multiníveis, que são executados por pacotes como *lme4* e *nlme.* Enquanto que a perspectiva de dados em painel é principalmente econométrica, em outros campos como saúde, biologia e educação, a abordagem multinível goza de uma tradição mais longa por terem dados tipicamente hierarquizados (e.g.: alunos em salas, salas em escolas). Dados em painel podem ser encarados como uma estrutura aninhada de dois níveis: há indivíduos *i* (nível superior) observados em ocasiões sucessivas *t* (nível inferior). As especificações FE e RE, por admitirem interceptos variáveis por indivíduo, possuem uma lógica próxima aos modelos hierárquicos em que há interceptos aleatórios (*random intercepts*). Além dos interceptos, é possível trabalhar com dados em painel prevendo coeficientes variáveis segundo indivíduo ou período ($\beta_{□□}$) (exemplos em Wooldridge 2013, p. 451-453; Croissant e Millo 2018, seção 8.2), embora essa possibilidade seja mais amplamente explorada nos modelos hierárquicos (*random intercepts* + *random slopes*). Quando temos configurações de dados em que o aninhamento não é no tempo, ou é mais complexo que apenas dois níveis (ex.: municípios em estados, estados em regiões), o uso de modelos hierárquicos é recomendado. Do ponto de vista operacional, o ecossistema de pacotes do R é mais bem desenvolvido em torno do *lme4/nlme* do que do *plm*, de modo que há maior facilidade em integrar os produtos das análises do *lme4/nlme* em outros pacotes gráficos, de testes e de apresentação de resultados. Para mais, ver Skrondal e Rabe-Hesketh (2008) e Croissant e Millo (2008, pp. 33-39).

Anexos

Tabela 1 – Total de artigos coletados por periódico da área de CPRI (2000-2018)

Periódico CPRI	Qualis (2016)	Total de artigos	% Total
Estudos Avançados	B1	896	12
Revista de Administração Pública	A1	726	9
Sociologias	B1	573	7
Revista Brasileira de Ciências Sociais	A2	559	7
Tempo Social	B1	533	7
Revista de Sociologia e Política	A1	518	7
Caderno CRH	A2	503	6
Sociedade e Estado	B1	483	6
Dados - Revista de Ciências Sociais	A1	469	6
Lua Nova: Revista de Cultura e Política	A2	450	6
Novos estudos CEBRAP	A2	420	5
Revista Brasileira de Política Internacional	A1	403	5
Opinião Pública	A1	340	4
Contexto Internacional	A2	323	4
Revista Brasileira de Ciência Política	B1	235	3
Civitas - Revista de Ciências Sociais	B2	183	2
Brazilian Political Science Review	A2	150	2
Total: 17 periódicos		7.764	100

Fonte: elaborado pelos autores

An introduction to panel data regression[1]

Rafael Mesquita - Federal University of Pernambuco
Antônio Alves Tôrres Fernandes - Federal University of Pernambuco
Rafael Mesquita - Federal University of Pernambuco

Resumo

Apesar da crescente oferta de dados em formato de painel, ainda são raros os estudos no Brasil que combinam as dimensões transversal e longitudinal na mesma análise. Por exemplo, em uma amostra de mais de 7 mil artigos publicados entre 2000 e 2018 em periódicos de CPRI, apenas 45 casos citavam técnicas específicas para lidar com observações de unidades espaciais (países, estados, pessoas) repetidas em intervalos regulares do tempo (anos, meses, dias). Diante dos benefícios inferenciais que tal abordagem pode proporcionar e da escassez de pesquisas sobre o tema, este artigo apresenta uma introdução à regressão de painel. Metodologicamente, sintetizamos as principais recomendações da literatura e mostramos a implementação no *R Statistical* com o pacote *plm*, indo desde a seleção de modelos até o tratamento dos dados e apresentação de resultados. Para aumentar o potencial pedagógico do trabalho, disponibilizamos os materiais de replicação, incluindo dados originais e scripts computacionais. Com este artigo esperamos difundir a utilização de análises longitudinais na pesquisa empírica em CPRI no Brasil

Palavras-chave: Dados em painel; TSCS; CPRI no Brasil; Regressão linear; Metodologia política.

1 This article received comments from the Research Methods group from the Political Science Department of the Federal University of Pernambuco (DCP – UFPE). A preliminary version was presented under the title "*Regressão com dados de painel: balanço da utilização na CPRI brasileira e estratégias para maior difusão do método*", in section "*Ensino e Pesquisa em Ciência Política e Relações Internacionais*" of the 12th Conference by the Brazilian Association of Political Science, João Pessoa, Paraíba (2020). We are grateful for the participants' comments and suggestions, in particular to professors Elia Alves and Jakson Aquino. Content partially from Mesquita (2018). Replication materials available at: <https://osf.io/5yx7g/?view_only=ac1691cced8549238d6d6e0a9d2b7f7b>. We are grateful to Ivan Fernandes (UFABC) and Gustavo Fernandes (FGV) for allowing the use of their database.

1. Introdução

O

Abstract

In spite of the growing offer of data organized in panel format, there are still few studies in Brazil combining cross-sectional and longitudinal dimensions. For instance, out of a sample of over 7 thousand articles published between 2000 and 2018 in Brazilian Politicand International Relations journals, only 45 articles cited specific techniques to handle observations comprising spatial units (countries, states, persons) repeated in regular time intervals (years, months, days). Given the inferential gains this approach can afford and the scarcity of research on the topic, this article presents an introduction to panel regression. Methodologically, we summarize the main recommendations of the literature and show how to implement them via *R Statistical* and *plm* package, going from model selection to data treatment and presentation of results. To improve the pedagogical potential of this work, we share replication materials, including the raw data and computational scripts. With this article, we aim to broaden the use of longitudinal analysis in empirical research in Political Science and International Relations in Brazil.

Keywords: Panel Data; TSCS; Political Science and International Relations in Brazil; Linear Regression; Political Methodology.

1. Introduction

A great variety of data in Politics and International Relations (PIR) are measured in panel form, that is, observations of spatial units measured at regular time intervals (Beck, 2001). The notation *Times-Series Cross-Section* (TSCS) is recurrent in the literature and reveals that a database combines spatial (countries, states, municipalities, individuals, etc.) and temporal dimensions (years, months, days, etc.) (Beck, 2008). The expression *time-series* means that the cases are laid out over time, and *cross-section* indicates that a database is formed by multiple transversal observations (Gujarati, 2004).

For certain research questions, that combination of temporal and spatial information can be very important to achieve more reliable answers than what could be produced by using a single approach. For example, if we wanted to verify if economic growth affects election results in Brazilian municipalities (Fernandes and Fernandes 2017), we could initially just conduct a cross-sectional study comparing several municipalities in the last election. That research design, however, would be limited, since we know that in the last few years economic growth in almost the whole national territory was low, meaning we will only know the effects of negative growth rather than positive. In addition, particular circumstances of the recent period (e.g., corruption accusations) with electoral impact could not be controlled for if they are nation-wide. However, overcoming these difficulties is possible if we gather data not only from the last election, but from repeated elections for all municipalities of interest.

In addition, panel data are valuable because they allow for important gains when it comes to measurement, controlling confounders, improvement of causal inferences, and sample size (Baltagi, 2005; Fortin-Rittberger, 2013). Obviously, these advantages come with costs. For example, the structure of the error term requires specific procedures for the results to be consistent and valid (Worrall, 2008). Despite being suitable to the phenomena studied by PIR, the method is still under-utilized: in a sample of over seven thousand published papers between 2000 and 2018 in PIR journals, only 45 cited specific panel regression techniques.

Thus, faced with the analytical benefits of this approach and its limited use in PIR fields, this paper shows didactic solutions on what to do and what not do with panel data (Beck and Katz, 1995). Our target audiences are undergraduate and graduate students in their initial stages of training. This pedagogical contribution becomes more important given the absence of didactic material especially focused on PIR.[2] Most manuals are either dedicated towards other disciplines, such as economics (Wooldridge, 2002; Kennedy, 2003; Stock and Watson, 2004), or have a prohibitive degree of complexity.[3] With this in mind, we have also included a question and answer section with frequently asked questions on the topic, in addition to the computational scripts that enable readers to adapt them to their respective research interests.

The article is organized as follows: the next section presents the results of a bibliometric analysis on the use of panel regression in PIR; section 3 gives an introduction the logic of longitudinal data analysis; section 4 replicates the data from Fernandes and Fernandes (2017), to demonstrate a computational implementation of a panel regression step-by-step; and the final section summarizes our recommendations on how to disseminate longitudinal analysis in PIR empirical research in Brazil.

2 For works in Portuguese with the same didactic goal, see Fávero (2013) and Marques (2000). For a tutorial in English with computational emphasis on the *plm* package, see Henningsen and Henningsen (2019).

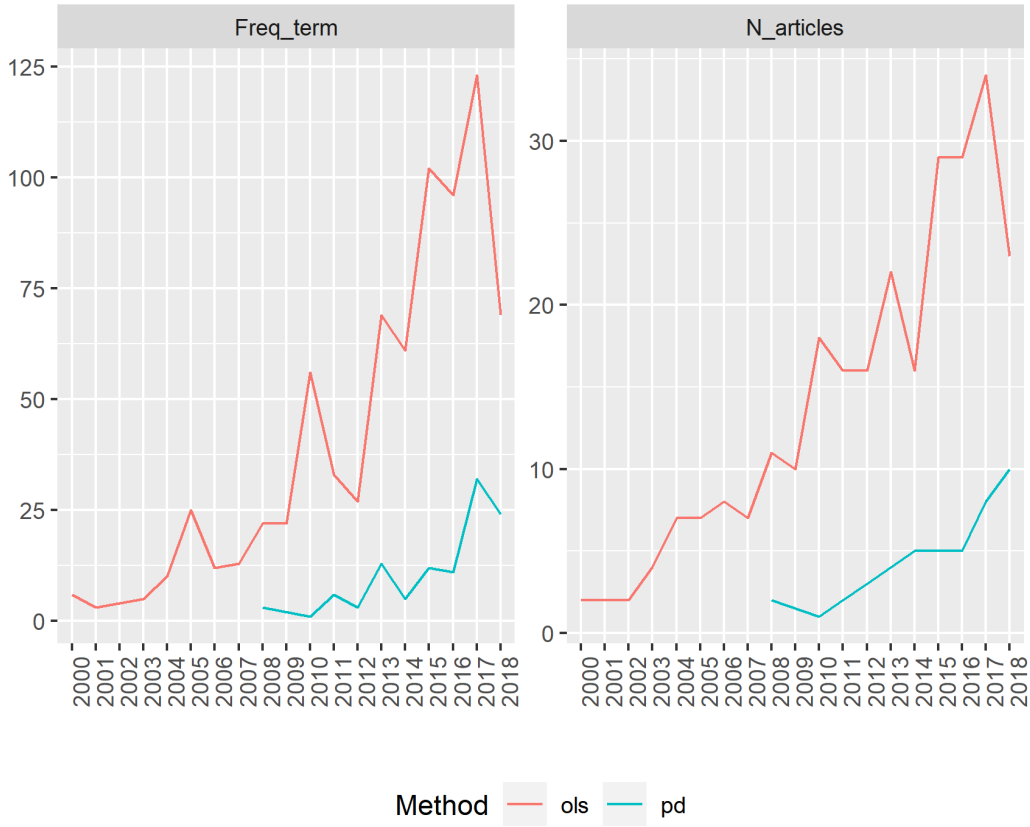
3 See, for example, the degree of difficulty in the final exam for the panel data econometrics course by professor William Greene <http://people.stern.nyu.edu/wgreene/Econometrics/PanelDataEconometricsFinalExam.pdf>

2. An assessment of the use of panel data in Brazilian PIR

How popular are panel data in Brazil PIR? To answer that question, we examined 7,764 articles published between 2000 and 2018 in 17 journals hosted at the Scientific Electronic Library Online Platform (Scielo)[4] (see Table 1 in the Appendix). Similarly to Medeiros et al. (2016), we used automated content analysis to detect keywords in the articles' full text. Using the *rscielo* (Meireles et al., 2019) and *quanteda* (Benoit et al., 2018) packages, we counted the frequency of expressions related to panel data methods. Out of the whole sample, only 45, that is, 0.58%, mentioned expressions related to panel data.[5] Although the interval begins in 2000, only in 2008 did papers with these terms come up, which means the use of this technique in PIR is very recent.

We compared the frequency of panel data with the number of citations for Ordinary Least Squares (OLS), given that this is one of the most disseminated methods in PIR (Krueger and Lewis-Beck, 2008).[6] For each article mentioning panel data (110 citations in 45 articles), there were another five using OLS (758 citations in 263 articles). Figure 1 illustrates that comparison.

Figure 1 – Prevalence of OLS vs. panel data



Source: the authors. The graph to the left counts the occurrence of keywords related to OLS and panel data and the one to the right the total of articles. The blue lines show the total for panel data (PD) and red for OLS.

4 The *rscielo* package automates information gathering from Scielo. As of the writing of this paper, the package identifies a total of 374 journals in all fields of knowledge. Of these, we selected 65 that had been evaluated by Qualis Capes in the PIR field in 2016. Since journals can be evaluated in more than one field, some of them, despite having been evaluated in the field of PIR, had a different concentration area, such as economics, education, or public health. Upon inspecting the list of journals and websites for each, we kept 17 whose focus was unmistakeably PIR. Of this subset, we collected the text of 8,829 papers in total. After removing duplicates and false positives, we arrived at 8,789. Further filtering to publications to only between 2000 and 2018, we reached the final number of 7,764.

5 Terms used for the panel data dictionary: "panel data", "*dados em painel*", "*painel de dados*", "tscs", "time series cross section".

6 Terms used for the OLS dictionary: "ols", "mmq", "mqo", "ordinary least squares", "*mínimos quadrados*", "linear regression", "*regressão linear*", "regression analysis", "*análise de regressão*".

Despite the growth trend, the popularity of panel data methods is today equivalent to OLS' a decade ago. Furthermore, the use of panel data is highly concentrated: 42% of articles were published in the Brazilian Journal of Public Administration (*Revista de Administração Pública*). Of the 17 journals, six did not mention the method at all. These results indicate low dissemination of panel data in national publications in PIR.

Over the remainder of the article, we introduce the essential concepts for panel regression using a bibliographical review of the key manuals and demonstrate a computational step-by-step by replicating Fernandes and Fernandes (2017).

3. Understanding longitudinal data

This section presents the foundations of longitudinal data analysis. First, we describe the different ways to tabulate information over time. We also discuss the main types of panel (short x long; balanced x unbalanced) (section 3.1). Next, we explain the notation of the panel regression model and the four main modalities of estimation (3.2). Lastly, we show the advantages and disadvantages of each approach (3.3 and 3.4).

Longitudinal data can be understood by comparison with cross-sectional data (Menard, 2002). In a cross-sectional design, the information on cases and variables are collected at a specific point in time. For instance, the homicide rate (variable) for all states in the country (cases) in 2013 (time). On the other hand, longitudinal data are collected for at least two different periods of time, which enables the observation of the change in value for each case, for each variable of interest.[7] For example, the homicide rate (variable), for all states in the country (cases), between 2000 and 2010 (time). Figure 2 shows the difference between cross-sectional data, temporal series, and their combination in panel form. As an example, we used the homicide rate per 100 thousand inhabitants in three Brazilian states.

7 Menard (2002) points out four longitudinal research designs: (1) total population designs; (2) repeated cross-sectional designs; (3) revolving panel designs; and (4) longitudinal panel designs. Wooldridge (2013, p. 448) also differentiates between "independently pooled cross section" and "panel data" or "longitudinal data"; the former uses random samples of different individuals in successive points in time (e.g., census sampling), while, in the latter, the subjects remain the same during the time frame.

Figure 2 – Data structure

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

Source: the authors, based on Cerqueira (2017).

In our example, the *UF* variable represents the cross-sectional unit of analysis (Brazilian state). The *Time* variable indicates the longitudinal dimension of information, here, the year. X_{it} represents the value of the variable of interest of case i in time t. For example, Acre had a homicide rate of 22 in 2011. The remaining columns show aggregate measures derived from X_{it}. Although these do not usually appear in databases, we reproduced them for teaching purposes.

The overall mean (sometimes called grand mean) is the total sum of cases (237.7) divided by the size of the sample (9), that is, 26.4. The overall deviation is the difference between each observation from the overall mean. For instance, since in 2011 Bahia had a rate of 39.4, it was 13 points above the overall mean (39.4 – 26.4). the deviation between the cases is the difference between the mean of each case and the overall mean. For example, Santa Catarina has a distance of 12.5 – 26.4 = -13.9 points from the overall mean. Lastly, the within-case deviation represents the difference between each observation and the individual mean of that state, meaning that, in 2013, Santa Catarina was 11.9 – 12.5 = -0.6 points below its historical average.

This configuration where each line of the dataset represents a unique combination of case in time is also known as long format and it is the standard arrangement for most statistical software. In several scenarios, there may be variation in the availability of cross-sectional or longitudinal data. Since panel data are the product of the combination of both, this means that different types of panels may be formed, depending on the number of cases or temporal units available, as we explain next.

3.1 Types of panel: short x long, balanced x unbalanced

Panels have two main configurations. There is a short panel, "stacked" or "cross-section dominant", where the number of cases is higher than the amount of time periods, that is, N>T. On the other hand, we say the panel is

long ("temporally dominant") where the number of time periods is higher than the number of cases (T>N). Some authors have standardized referring only to the former as panel data and the latter as Time-Series Cross-Section (TSCS) (Beck, 2001; Fortin-Rittberger, 2013).

For each panel configuration (short or long) different expectations are assumed regarding the distribution of the data. Beck (2001) argues that, for panel data (N>T), the selected units are seen as samples of a population, observed over a short time interval. Thus, inferences are not specific to the units and may be generalizable. For TSCS (T>N), units are seen as fixed and observed for a long period of time. Therefore, inferences are about the units and less generalizable. The traditional examples of the former are census surveys and, for the latter, macroeconomic comparisons between countries.[8]

Another important feature is the availability of information. We have a balanced panel when there is information for all cases in all time periods. On the other hand, we say the panel is unbalanced when some of these cases are absent for certain time periods. An additional complication, which can affect the consistency of estimations, is the nature of the panel's imbalance. The absence of data may be random, which tends to harm the efficiency of the coefficients (standard error, p-value, confidence intervals). However, the imbalance may be related to some variable of interest to the study. For example, it may be easier to find detailed information for richer countries than poorer nations. When the absence of information is systematic for any reason, the estimations tend to be biased. According to Batalgi (2005), unbalanced panels tend to be more frequent than balanced ones, which reinforces the need to describe with precision the process of data collection, treatment, and analysis.

8 Similarly, the asymptotic properties of TSCS presuppose that N is fixed and T can grow to infinity. For panel data, T is fixed, and N can grow to infinity (Fortin-Rittberger, 2013). From this it is deduced that the larger the N, proportionally speaking, the better the data fits to panel data models and, inversely, the larger the T, the better the fit to TSCS models.

We now explain how panel data may be used in regression analysis.

3.2 Notation and types of panel regression

For the purposes of this article, it is important to show the classic notation of the linear regression model, using only cross-sectional data:

$$Y_i = \alpha + \beta X_i + \varepsilon_i \qquad (1)$$

In which Y represents the dependent variable, α indicates the intercept, β represents the variation observed in Y when the independent variable, X, increases in one unit, and ε indicates the stochastic nature of the model. The i in subscript indicates that the observations are indexed by case. Continuing with the example in Figure 2 (homicide rate by state over time), this equation can represent the effect of unemployment (X) on the homicide rate (Y). A panel regression is an extension of the previous model, in which:

$$Y_{it} = \alpha + \beta X_{it} + \varepsilon_{it} \qquad (2)$$

The i and t in subscript inform that the observations are indexed, respectively, by case and time. Our model becomes more powerful, because we accumulate information on the relationship between unemployment and homicide for several states and years. How to move ahead, then, with the regression? We describe below the four traditional approaches to estimate panel models.

Pooled OLS

By gathering all observations in the same linear regression ("pooling"), we have a Pooled OLS (POLS). This approach assumes that all units can be covered by the same regression line. Even if the units of i show differ-

ences between each other, it is considered that the set of independent variables already carries all important information. That is, the main factors that differentiate observations are already included in the model, making it unnecessary to control for other unobserved factors. In our example, it would be as if the homicide rate, in whatever state, was mainly determined by unemployment, and there were no other peculiarities in the states affecting the outcome: if Bahia had the unemployment level of Santa Catarina, it would have the same homicide rate as that state. Since individual heterogeneities are not modelled, we have Equation (2), where there is only one intercept (α) for the whole population and all else that is not observable is allocated as noise in the error term (ε).

However, if there are unobserved factors in ε which are not random but systematic, POLS is no longer appropriate. If we have unobserved elements that are correlated with X, we have a traditional case of omitted variable bias, and the parameter estimate β ($\hat{\beta}$) will be inconsistent.[9] Since observations are repeated over time, the independence of observations is in jeopardy. Suppose that for the three states in the example there is some factor we cannot observe that nonetheless has an impact on the homicide rate, for instance, the influence of organized crime. If the influence of organized crime is correlated with unemployment there will be a bias; if it is specific in each state (e.g., the criminal influence in Acre is the same over time), the errors will not be independent (ε_{it} will be correlated with ε_{it-1}). Because of this, the first motivation to not use the POLS model is the need to work with the heterogeneity hidden in the error term in a more sophisticated way.

The unobserved factors are divided into elements that remain the same for each unit and elements that vary according to unit and time. Therefore, the "composite error" (ε_{it}) of Equation (2) can be separated in two parts: one fixed for each unit (μ_i) and another one variable (ν_{it}).

$$Y_{it} = \alpha + \beta X_{it} + \mu_i + \nu_{it} \qquad (3)$$

The errors persistent to the units (μ_i) are also known as unobserved heterogeneity, unit effect, or fixed effect of the error term, while ν_{it} is usually called variable or idiosyncratic effect. Unlike the other three approaches presented here, POLS does not have strategies to deal with μ_i, because it does not disaggregate ε_{it}. All remaining specifications break down the error factor into fixed and variable.

First Differences

In the First Differences (FD) approach, all variables are differentiated, that is, subtracted from their value in the previous time period. The subtraction is especially efficient in handling dynamic issues with the data, such as serial correlations and non-stationarity, given that, when subtracting an observation at *t* from the observation at *t*-1, everything that has not varied from one moment to the next (including the unobserved heterogeneity) will be erased.

Because it subtracts observations by pairs, FD has costs: it reduces the data's variability and thus increases the standard errors, being, therefore, more suited for large datasets. Furthermore, the differentiation removes all factors constant over time or that vary monotonically (i.e., always increasing at the same rate every year), which hinders the use of dummy variables for countries, geographical distances, and constant growth variables (Wooldridge, 2013, p. 462–463). The use of FD is one of the drastic solutions recommended by Beck (2008) when faced with persistent temporal dependences in the model.

9 If the phenomenon under analysis is better described by a set of variables and one of them and left out of the equation, the effect of the variables preserved in the model will be over- or underestimated. That systematic upward or downward distance in the parameter characterizes a bias.

Fixed Effects

Fixed Effects (FE), as FD, seek to remove the fixed component in the error (μ_i). However, rather than differentiation, they use a "within transformation". The operation is called "within" because it only retains the within-case deviation (see Figure 2). That removal of the value of each observation's mean ("time demeaning" or "group-mean centering") will result in the elimination of α and μ_i, whose values for each unit do not change with the passage of time.[10] Therefore, like FD, the FE estimator omits factors constant in time.

Because it handles the heterogeneity problem and is very robust, the FE specification is seen by Croissant and Millo (2018, p. 5) as a benchmark for panel data. That is the case especially for disciplines that deal with observational data, which was also found in our bibliometric research: fixed effects were the most used method in PIR (26% of articles).

One of the benefits of FE in comparison to POLS is that, while the latter postulates only one intercept for the whole sample, the former takes into account that each unit can have its own intercept. In comparison to FE, POLS will have biased estimators, given that it does not deal with μ_i. Under FE, the model is less biased, although also less efficient.

Computationally, similar results occur when using an FE model and a POLS model with dummies for all cases. In this situation, POLS is usually referred to as Least Square Dummy Variable (LSDV). Figure 3 illustrates the differentiation and mean subtraction procedures.

Figure 3 – Illustration of the differentiation and within transformation procedures

First Differences

ID	Time	Homicide Rate	Unemploym. 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

$$Y_{it-1} = \alpha + \beta X_{it-1} + \mu_i + \nu_{it-1}$$
$$- \left[Y_{it} = \alpha + \beta X_{it} + \mu_i + \nu_{it} \right]$$
$$\Delta Y_{it} = \beta \Delta X_{it} + \Delta \nu_{it}$$

ID	Time	Homicide Rate	Unemploym. 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

Fixed Effects

ID	Time	Homicide Rate	Unemploym. Rate	Individual Mean (homic.)	Individual 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

$$Y_{it} = \alpha + \beta X_{it} + \mu_i + \nu_{it}$$
$$- \left[\bar{Y}_i = \alpha + \beta \bar{X}_i + \mu_i + \bar{\nu}_i \right]$$
$$\ddot{Y}_i = \beta \ddot{X}_i + \ddot{\nu}_i$$

ID	Time	Intra-case Deviat. (homic.)	Intra-case Deviat. (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

Source: the authors, based on data from Cerqueira et al. (2017) and IBGE (2017).

The figure illustrates the effects of differentiation and within transformation only for the observed variables: homicides and unemployment. Even so, the subtraction also encompasses unobserved factors (indicated in italics in the figure's equations), resulting in the suppression of the overall intercept (α) and the individual effect of the error term (μ_i). In our example, it means that the resulting models, when estimating the effect

10 The α and μ_i components are removed for a consistent estimation of β. Nevertheless, note that some statistical packages report values for individual (μ_i) or population (mean of all μ_i) intercepts. See Wooldridge (2002, section 10.5.3) for details.

of unemployment on the homicide rate, eliminate the fixed influence of organized crime in each state. Not only this, but every unobserved factor that is invariant will be removed and the estimation will only be based on the elements that vary.

Random Effects

As seen previously, a deficiency in FD and FE is the impossibility of testing variables that do not change with time. Even the ones that vary slowly tend to be underestimated (Clark and Linzer 2015). Nevertheless, there are many variables of interest for PIR that are static, such as geography, gender, nationality, among others. The Random Effects (RE) specification bypasses that issue, enabling the inclusion of variables that do not change with the passage of time.

As with FE, it is assumed that the composite error has fixed and idiosyncratic parts, that is, $\varepsilon_{it} = \mu_i + \nu_{it}$. However, instead of eliminating μ_i completely, it is assumed that it has no correlation with X (Bell and Jones 2015, p. 142). If the unit fixed effect is not correlated with the independent variables, it is no longer a source of bias and, consequently, no longer needs to be eliminated (e.g., if we believe that organized crime is independent from unemployment, there is no reason to remove it).

The second distinctive assumption of RE is about unit effects. *Fixed* effects assume that the effect of each unit can be calculated as a unique value. Therefore, FE calculates an intercept for each individual in the form of a parameter. In contrast, RE considers μ_i and ν_{it} as *random* realizations in a distribution. In the example involving unobserved organized crime in the states, FE will estimate that individual impact case-by-case, without assuming that there is a relationship between the level of organized crime in Bahia and Acre. RE gives more structure to the set, because it assumes that the effects of crime tend towards a normal distribution, all states considered (Bell et al., 2019, p. 1061).

Some authors classify RE as a middle ground between POLS and FE (Clark and Linzer 2015, p. 402). POLS does not eliminate μ_i and FE erases it completely. RE partially executes this procedure, since it subtracts a portion from the observations of the within-group mean ("quasi-demeaning"). The proportion of that fraction is given by θ (*theta*, in the *plm* output), which takes on values between 0 and 1. *Theta* can be interpreted as a diagnostic of which component most contributes to the variance of ε: if fixed (μ_i) or idiosyncratic (ν_{it}). When there is little variance in the fixed effects ($\theta \to 0$) the RE estimates approach POLS and, in the opposite scenario ($\theta \to 1$), they approach FE (Wooldridge, 2013, p. 494).

On the other hand, the assumption that there is no relationship between μ_i and X is hard to maintain in most real situations. Another significant difference is that RE estimators will always have some bias and can only be efficient. The choice between FE and RE is, therefore, a trade-off between efficiency (RE) and unbiasedness (FE).

The direct comparison between RE and FE is recurrent in the literature. Disciplines more accustomed with randomized experiments tend to prefer RE (Dieleman and Templin, 2014). Beck (2001, p. 284) argues that observational studies prefer FE to RE since the latter is more recommended when the units are considered to be a sample of a larger population, and the former when the aim is to make inferences restricted to those units. The idea of "interchangeability" captures that notion intuitively: when, for a set of observations, we can replace case A for case B without great loss of information – that is, the *name of the case* does not matter – then we have something close to a random sample (Hausman 1978, p. 1262; King 2001, p. 498). The

RE expectation of a normal distribution of individual effects thus becomes acceptable. In some PIR scenarios, that interchangeability is plausible (e.g., representatives in a legislature), but not in others (e.g., member-states in the UN Security Council).

Lastly, when conducting an empirical analysis, it is always informative to make a side-by-side comparison of the results from the four types of regression: POLS, FD, FE, and RE. The contrast indicates the different results obtained depending on what is done with μ_i: POLS keeps the fixed factors completed inserted in the error component, RE maintains a portion, and FE and FD remove it completely (Wooldridge, 2013, p. 494).

3.3 Benefits of panel data

Box 1 summarizes the main technical benefits and operational issues associated with panel data.

Box 1 – Benefits and issues of panel data

Benefits	Issues
(1) Facilitates the detection of causality; (2) Measures individual variation; (3) Reduces measurement errors; (4) Increases the size of the sample; (5) Controls omitted variables issues.	(1) Serial correlation of the residuals; (2) Non-stationarity; (3) Heterogeneity; (4) Panel heteroskedasticity; (5) Contemporaneous correlation of the errors; (6) Complex dependence structures.

Source: the authors, based on Fortin-Rittberger (2013), Hsiao (2003), and Kennedy (2003).

The first advantage of panel data is to facilitate the detection of causal relationships between X and Y. Pearl (2000) highlights three assumptions to identify a causal relationship: (1) association between the variables (correlation); (2) temporal precedence; and (3) nonspuriousness of the relationship. Unlike cross-sectional data, in which values of X and Y are measured in the same time frame, panel data enable one to observe if the temporal variation of X is correlated with the temporal variation of Y. Thus, in addition to controlling for possible spurious variables, longitudinal data favor the assumption of temporal precedence between the independent and dependent variable.

Furthermore, the longitudinal nature also favors the study of change since it maintains the same unit of analysis over time. Therefore, it is possible to examine the individual variation of different indicators among the cases of interest. For example, the aggregate analysis of the number of homicides in a city is informative. But a database that informs the day of death and neighbourhood brings additional possibilities. It would be possible to observe, for instance, if the majority of occurrences are concentrated in the weekends of a certain neighbourhood. As Fortin-Rittberger (2013) highlights, that possibility to monitor, in space and time, where and when certain phenomena of interest arise brings longitudinal studies, though being observational, closer to the logic of a controlled experiment.

As a third advantage, we have more reliable measures on the change of the variables. According to Blalock (1979), measuring is the main challenge to the development of the social sciences. In particular, in PIR, important concepts, such as democracy, are not directly observable. One of the ways to examine the reliability of a measure is to measure it on several occasions over

time. Consequently, panel data help to overcome possible measuring errors. [11]

In addition to measuring gains, another benefit of panel data is increasing the sample size. One of the recurring problems in research design is indeterminacy, that is, "many variables, small N" (Lijphart, 1971, p. 686). When we have a few observations, it is difficult to argue that the causal relationship proposed is explained by one reason and not any other. Thus, determinacy of a research design relates to the proportion of variables and observations. For this reason, increasing the number of cases is a common recommendation to solve indeterminacy (King, Keohane and Verba, 1994). Furthermore, larger samples solve efficiency issues (smaller standard error), multicollinearity, and make using more robust statistical tests easier (Hsiao, 2003).

Lastly, panel data control the problem of omitted variables (Kennedy, 2003). The omission of an important variable will produce bias in the estimates if it is correlated with other modelled variables. A purely cross-sectional perspective does not offer many solutions to deal with unobserved factors. Frequently, we are simply ignorant on the unobservables that impact Y. However, from the moment that cases are studied over time, tools such as FE, FD, and RE may be used to remove or mitigate these individual heterogeneities (Wooldridge 2002, section 10.1).

Nevertheless, these benefits do not come without a cost, as we show next.

11 Errors can be systematic or random. Systematic error is harmful in descriptive analyses because the estimates produced will be biased. On the other hand, when the systematic error impacts all cases with the same magnitude, it does not affect the consistency of the estimates in regression models. Conversely, random error in the dependent variable tends to reduce efficiency (higher standard error) but does not affect consistency of the estimates. Lastly, when independent variables also present random error, the estimates tend to be inconsistent and, generally, underestimated (King, Keohane and Verba, 1994).

3.4 Challenges with panel data and how to overcome them

For pedagogical purposes, we propose a brief summary of the assumptions of an OLS linear regression as a starting point to address the typical issues of panel regressions.[12] According to Wooldridge (2013, p. 59, 119), the main assumptions for OLS can be summed up into six: (A.1) There is a linear relationship between X and Y; (A.2) Absence of autocorrelation (or random sample); [13] (A.3) Absence of perfect multicollinearity ("full rank"); [14] (A.4) Exogeneity of X (or "zero conditional mean") – the error must not be a function of the independent variables, that is, it is random and none of the independent variables in the model carries information regarding its value;[15] (A.5) Homoskedasticity – there is uniform variance in the regression residuals and these are not correlated to each other – if, for some reason, in a portion of the data, the variance of residuals changes according to the values of the explanatory variables, there is heteroskedasticity;[16] (A.6) Normal distribution of the errors.

If the so-called Gauss-Markov assumptions (A.1-A.5) are respected, it can be said that the parameters estimated are the *Best Linear Unbiased Estimators* (or BLUE) of the population's parameters. "Best" because they have the smallest variance (A.5) since there is homoskedasticity (uniform variance of residuals); "Linear" because of the linear relationship (A.1); and

12 The list of OLS assumptions sometimes varies among authors. For other lists of linear regression assumptions and the Gauss-Markov theorem, see Berry (1993) and Larocca (2012).

13 Random samples and the absence of autocorrelation may seem like distinct expectations, but they are equivalent once we consider that, in both cases, we have $\mathrm{Corr}(\varepsilon i, \varepsilon j = 0)$, for $i \neq j$.

14 The main damage caused by multicollinearity is the increase in the variance of the regression coefficients, which increases the chance of finding non-significant results.

15 The violation of exogeneity equals an omitted variable bias.

16 In other words, the model fits better for certain sections of the data and not others. Heteroskedasticity does not lead to bias or inconsistency of OLS estimators, but because it interferes in the distribution of variances, it hinders the construction of confidence intervals and significance tests (Wooldridge, 2013, pp. 268–269).

"Unbiased" because there is no bias in the parameter, since the regressors are exogenous and do not affect the error (A.4).

However, panel models can easily violate those assumptions. Because they combine temporal series and cross-sectional comparisons, they inherit the issues of both types of data. These problems can be divided into dynamic and spatial. The literature usually recommends attention to dynamic issues first. These especially violate the requirements of absence of autocorrelation (A.2), exogeneity (A.4), and homoskedasticity (A.5). The main ones are:

(1) **Serial correlation of the residuals:** there is correlation between the residuals at moments *t* and *t-1*. This is a persistent characteristic of long panels given that, as Box et al. (2016) caution, an intrinsic aspect of temporal series is that adjacent observations are dependent. The main issue generated by the correlation of errors is an inconsistency of significance tests (p-value and confidence intervals).[17] For this reason, the serial correlation must be detected and solved before moving on to spatial issues. There are different tools for detecting temporal processes in the error structure: the plots of the autocorrelation (ACF) and partial autocorrelation functions (PACF), the Durbin-Watson, Breusch-Godfrey, and Ljung-Box tests, among others. Most of them are directed towards checking first-order autoregressive processes, or AR(1), which are more common.[18]

It is possible to test an AR(1) correlation in the errors of POLS models in a simple and direct way – by running a regression of the residuals at *t* on those at *t-1*, without an intercept, as in $\varepsilon_{it} = \rho\varepsilon_{it-1} + \nu_{it}$. The result for parameter ρ will indicate whether or not the residuals from the past are significantly correlated with the residuals from the future (Wooldridge, 2002, section 7.8.5; 2013, p. 417).

The FE, RE, and FD specifications will require specific tests. To cover the whole arsenal of existing tests goes beyond the scope of this article (for a list, see Croissant and Millo 2008, p. 22-28; 2018, section 4.3). However, we highlight that the selection of tests must be guided by the type of estimation used in the model (POLS, FE, RE, or FD), the N/T proportion, and the number of observations.

Once a serial correlation is found, recommended solutions include adding the dependent variable lagged, inserting robust standard errors, or using FD. Regarding the first alternative: the inclusion of a lagged dependent variable (LDV) is recommended, under the assumption that the process is AR(1) and, afterwards, running a new test to check if the autocorrelation has been eliminated (Beck, 2001). However, this solution has been criticized by Achen (2000, p. 14), who states that the LDV may absorb the whole explanatory power of the remaining regressors and thus produce bias.[19]

(2) **Non-stationarity:** a temporal series is stationary when it is in statistical equilibrium, that is, it has constant mean and variance over time. Although values oscillate, they will return to the mean. It is non-stationary when the mean and variance are not constant, that is, when the series does not tend towards a prior mean after deviations. In other words, non-stationarity may be understood as a persistent effect of prior shocks (Box et al., 2016, p. 7;

17 Technically, autocorrelation affects the efficiency of estimates. In particular, if the autocorrelation is positive, the standard error of the coefficients will be underestimated, which increases the chances of incorrectly rejecting the null hypothesis (type I error).

18 Although it is not common to expect more complex processes than AR(1) for panel data (Fortin-Rittberger, 2013), it is possible that there are higher-order correlations, which will motivate more complex solutions (e.g., the inclusion of variables with multiple lags).

19 When the regressors present some trend, the introduction of an LDV will dominate the regression. The LDV's coefficient will be biased upwards and the remaining regressors, downwards. It is also not recommended when the T is small, because it sacrifices a period of analysis (Achen, 2000).

Fortin-Rittberger, 2013). Deviations of stationarity may come from the type of phenomenon monitored but may also result from the measuring frequency: if information was measured by day or by month, for example, it is possible to find seasonal effects and, for samples with a large T, cyclical behavior. The visual inspection of the time series is one of the simplest ways of identifying trends or seasonality. Different tests in the literature on time series may be employed to verify non-stationarity: ACF, PACF, and unit root tests (e.g., Augmented Dickey-Fuller, ADF). More recently, tests specific to the context of panel data, "Panel Unit Root Tests", have been proposed (see Croissant and Millo 2018, chapter 8). The application of these tests will be hindered if the panel has few time periods. If the series is not stationary, solutions include using FD or, in cases of non-stationarity by trend, including a time index or trend variable (Fortin-Rittberger, 2013). Now we turn to spatial aspects.

(3) **Heterogeneity:** regressions assume that cases are homogenous. All particular characteristics are explained by the independent variables, as previously explained with the concept of "interchangeability" (King, 2001, p. 498). Heterogeneity is present when one or some of the cases observed have distinct characteristics that, having not been modelled, will end up in the error term and thus generate correlations between $\varepsilon_{\square\square}$ and X_{it}. Consequently, heterogeneity violates the assumptions of exogeneity (A.4) and homoskedasticity (A.5). If heterogeneity is present, it will usually be inaccurate to assume that there is only one intercept for the whole population, as in POLS models. If every unit has impactful fixed characteristics, the absence of unit-specific intercepts will lead to a mistaken regression line (Fortin-Rittberger, 2013). On the other hand, violating the homoskedasticity assumption implies producing inconsistent significance tests, which increases the chances of mistaken inferences.

As discussed in section 3.2, abandoning POLS estimates in favor of other specifications is one way of solving heterogeneity. The last three spatial difficulties mainly compromise exogeneity (A.4), homoskedasticity (A.5), and normality of the errors (A.6).

(4) **Panel heteroskedasticity:** the residuals must have constant variance in both directions, that is, between units *i* and time periods *t*. Panel heteroskedasticity is present when residuals present constant variance over time within each unit, but inconstant across units. In other words, each unit will have its own residual variance. It may happen due to the model's poor specification or when one or two cases do not fit well within the specification.

(5) **Contemporaneous correlation of the errors:** a case's error is correlated with the error of others for the same moment in time. For example, if an external shock occurs in a year and affects several units at the same time, the units' residuals for this period will present a different variance from other periods.

(6) **Complex dependence structures:** there are more special cases that, somehow, impose dependence between observations. One of the most explored in observational studies is the spatial or geographical correlation between units. That is, contiguity or proximity between two countries exposes them to more events in common and, because this is not random, may lead to biases.

Generally, tests such as Pesaran's (2004) offer evidence on whether there is dependence between units such as the ones described in items (4) to (6), although there are specialized variations for different types of panel and dependence structure (see Croissant & Millo 2008, p. 28-31; 2018, chapter 10).

Spatial issues in POLS models may require the use of FE, RE, or FD estimators. It is also recommended, for

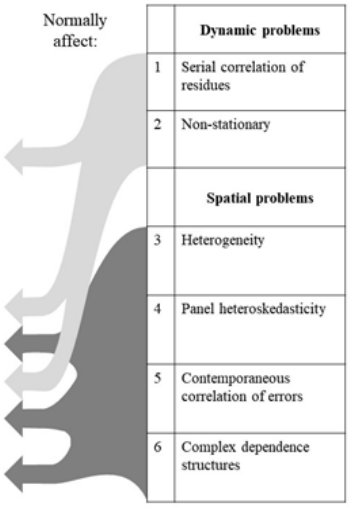
whichever estimator, reviewing the set of variables, in order to consider new factors that may capture the heterogeneity between the units; in addition to using robust standard errors.

In conclusion, it is useful remembering that the likelihood of these issues occurring depends on the panel type. T>N panels tend to have more problems with heteroskedasticity and serial correlation of errors, while N>T panels frequently have issues with non-measured heterogeneity between the cases.

Figure 4 summarizes the list of assumptions and violations discussed in this section.

Figure 4 – Summary of OLS assumptions and dynamic and temporal issues

	Assumption	Formal	Why is it a problem to violate it?
A.1	Linear relation	$Y = \alpha + \beta X + \varepsilon$	Bias
A.2	No autocorrelation, Random sample	$Corr(\varepsilon i, \varepsilon j) = 0$ $i \neq j$	Bias
A.3	No perfect collinearity	$Corr(X1,X2) \neq \pm 1$	Impossible to estimate $\hat{\beta}$; Bias; Inflated SE($\hat{\beta}$), p-value
A.4	Exogeneity	$E(\varepsilon \mid x) = 0$	Bias
A.5	Homoskedasticity	$Var(\varepsilon \mid x) = \sigma^2$	Invalidates SE($\hat{\beta}$), p-value
A.6	Normal distribution of errors	$\varepsilon \sim Normal(0, \sigma^2)$	Invalidates SE($\hat{\beta}$), p-value



Source: the authors.

4. Application

In this section, to illustrate how to conduct a panel analysis, we replicate data by Fernandes and Fernandes (2017). Although it is common to use classic databases from statistical packages for teaching (e.g., "Grunfeld", "EmplUK", etc.), we consider this a valuable opportunity for replication that simultaneously incentivizes learning and a transparency and replicability culture in the social sciences (King, 1995).

The main goal of the article "*A importância do crescimento econômico local na escolha do chefe do Executivo no Brasil*" ("The importance of local economic growth in the choice for president and mayor Brazil") is to analyze the impact of local economic growth on the incumbent's vote share in Brazilian presidential and municipal elections between 2000 and 2010.

To answer the research question, Fernandes and Fernandes (2017) mobilize 17 independent variables to explain the dependent variable: the vote share received by the incumbent. According to the authors: "We estimated the relationship using panel data techniques that enabled us to analyze whether unobservable idiosyncrasies affect the estimation or not. To so do, we used the three traditional models of static panel: pooled OLS, random effects (RE), and fixed effects (FE)" (Fernandes & Fernandes, 2017, p. 660).[20] To replicate the article's data, we used R, RStudio, and the *plm* package (Croissant & Millo, 2008). To make understanding the computational step-by-step easier, we report the scripts throughout the text.[21]

20 In the original: "*Estimamos a relação por meio de técnicas de dados em painel que permitem analisar se as idiossincrasias não observáveis afetam ou não a estimação. Para isso, usamos os três modelos tradicionais de painel estático: painel empilhado, efeitos aleatórios (EA) e efeitos fixos (EF)*" (Fernandes & Fernandes, 2017, p. 660).

21 Note that the example dataset at OSF has its variables named in Portuguese. Translations into English in the scripts reproduced here are merely for clarity.

For the following tests, it is useful to already have in the RStudio environment the four types of specifications that will be compared: POLS, FE, RE, and FD. First, it is necessary to load the database and fit it to execute the models.

```
##### 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(DATA, 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.
```

By applying the *pdim()* command, we can see that the database has N= 5,565 municipalities and e T=2-6.[22] In total, there are 28,945 observations. That is, we have a short (N>T) and unbalanced panel. By transforming the database into a *pdata.frame* object, an index containing the spatial and time dimensions is created, enabling the estimation of the models.

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

form <- as.numeric(voteshare) ~ growth+growth_state+growth_country+log_gdp+log_gdp_state+
  log_gdp_country+mayor_base_presid+health_exp+log_pop+leec+
  lheu+lses+laseps+ldesporc+ldespcor+linvest+ldespers

##### 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".
```

22 The period under observation is "electoral year". The observations begin in 2000 and go up to 2010, registering data every two years. This is why the database has T=6 and not T=10.

We executed four models: POLS, FE, RE, and FD. The POLS model is executed through the specification *model = "pooling"* in the *plm* function.[23] Similarly, the FE, RE, and FD models are specified in *model*.

Which of the above models should we use? To answer that, we need to check the assumptions of each one.

4.1 Pooled OLS?

The first and simplest choice is POLS, which applies a single regression line, with the same intercept, for the whole population. Therefore, to judge this specification's adequacy, we need to assess the population's degree of homogeneity and whether or not our set of regressor exhausts that heterogeneity. Is the relationship between incumbent votes and economic growth (and other control variables) similar across the 5,565 municipalities? Or is there significant variation among the units, more than it was possible to explain with only our variables? To decide, we can turn to (a) contextual knowledge, (b) graphical analysis, and (c) testing.

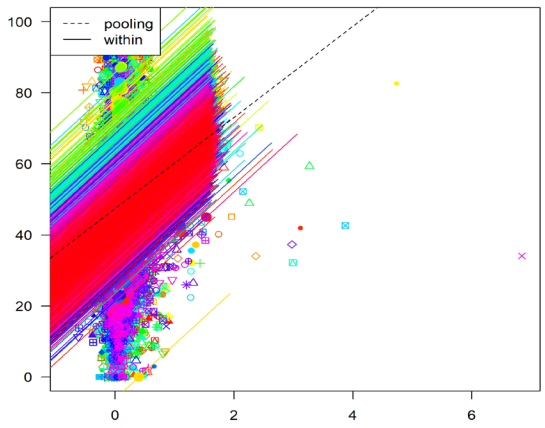
(a) Researcher knowledge on the subject matter is the first plausibility test. What we know about the object studied is often informative enough on the degree of heterogeneity or homogeneity across units. For example, according to the literature, is the electoral dynamics in São Paulo's capital is comparable to what happens in Serra da Saudade (in the state of Minas Gerais), population 781? Consequently, consulting theory on a subject may offer clues on the plausibility of pooling.

(b) Graphs are useful for exploratory analysis, especially when the N is small. To visualize if we can incorporate all municipalities in a single regression line, we can compare, in the same scatter plot, a single line applied to the whole population vs. separate lines for each municipality. The following example shows a bivariate plot with the main independent variable, "economic growth" (horizontal axis), and the dependent variable "incumbent vote share" (vertical axis).

```
##### GRAPHIC COMPARISON OF POLS VS FE INTERCEPTS #######

mode_fee <- plm(voteshare ~ growth,
                data = DATA, index = c("codibge", "year", group = "codibge"),
                model = "within")
plot(mode_fee)
```

23 The pooled model can also be executed with the *lm* function. However, to execute the tests for assumption check it is necessary that the database is in the format produced by *plm*.

Figure 5 – Pooled OLS and FE comparison for the bivariate model of incumbent vote share x economic growth



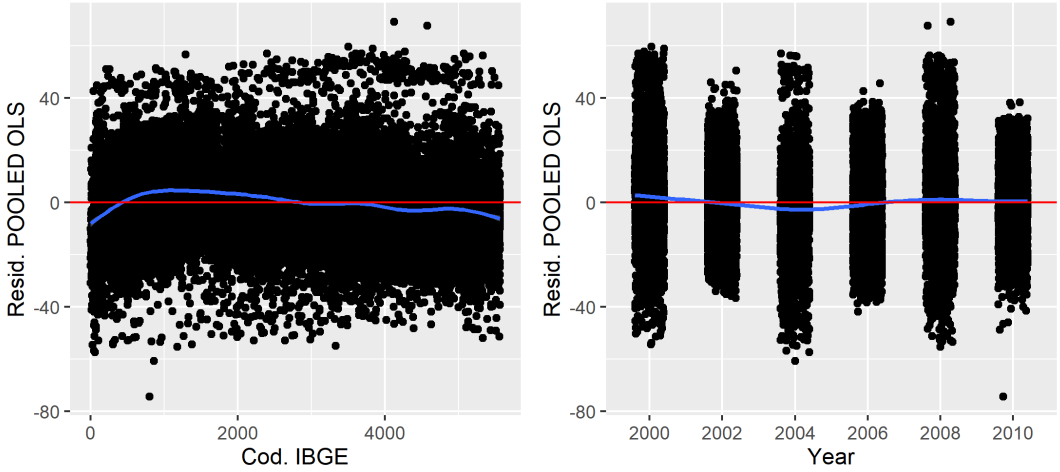
Source: the authors, based on data by Fernandes and Fernandes (2017).

The pooling solution (dotted line) proposes a line approximately in the middle of the observations, while the within model (solid colorful lines) better represents the data's variation and indicates that there is a large difference between the intercepts of each municipality.

Another graphic analysis tool is the residuals plot of the final model. Since POLS assumes homogeneity, its residuals must be similarly random for all cases. We can group the regression's residuals by municipality and year as a way to observe if the variance in the error term was more pronounced for certain municipalities or periods, as shown below.

```
##### POOLED OLS RESIDUALS ANALYSIS ###########

library(broom)
library(ggplot2)

POLS_lm <- lm(voteshare ~ growth+growth_state+growth_country+log_gdp+log_gdp_state+
  log_gdp_country+mayor_base_presid+health_exp+log_pop+leec+
  lheu+lses+laseps+ldesporc+ldespcor+linvest+ldespers,
            data = DATA, na.action = "na.exclude")
# running the POLS model with lm is required to add
# the model's residuals to the original data.frame

base <- augment(POLS_lm, data=DATA)
# with the 'augment' command in the broom package, we can join
# the model's residuals with the original database's variables.

                  #### Municipality ####

ggplot(base, aes(x=as.numeric(as.factor(codibge)), y=.resid))+
  geom_point()+
  geom_smooth()+
  geom_hline(yintercept = 0, color="red") + ylab(" Resid. POOLED OLS") +
  xlab("Cod. IBGE")
                  ####   Year  ####

base$year <- as.numeric(base$year)

ggplot(base, aes(x=jitter(year), y=.resid))+
  geom_point()+
  geom_smooth()+
  geom_hline(yintercept = 0, color="red") + ylab("Resid. POOLED OLS") +
  xlab("Year") + scale_x_discrete(limits=c("2000", "2002", "2004",
                                           "2006", "2008", "2010" ))
```

Figure 6 – POLS model residuals grouped by municipality and year



Source: the authors, based on Fernandes and Fernandes (2017).

The municipalities' residuals are not similar: there is something systematically differentiating the municipalities with IBGE codes lower than 3,000 from those with higher code values. The variance between electoral years is also different. The reason can be seen quickly: the variance is larger for municipal election years (2000, 2004, and 2008) and smaller for presidential ones (2002, 2006, and 2010). This illustrates the issue with pooling and heterogeneity, since we are trying to include mayoral and presidential elections in the same predictive model. These results encourage us to consider the inclusion of dummies for each year or election type.

(c) The *plm* package enables the implementation of different tests with the goal of checking, in a null/alternative hypothesis way, if there are significant individual effects. Both the F-test and the Breusch-Pagan LM test can be applied. The F-test checks if there are significant differences between the POLS and FE models, with the null hypothesis being that POLS is superior. As can be seen by the result of the p-value (<0.05), the FE model is significantly superior.

```
##### 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
```

The Breusch-Pagan LM test helps us to choose between POLS and RE.[24] The test's null hypothesis predicts that the municipality-specific variance is zero, which would recommend POLS. According to the test's result, H0 is rejected, and POLS can then be ruled out (p-value < 0.05).

```
##### 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
```

The comments on individual effects also apply to temporal effects. According to the same logic, the goal is to check if each year in the study has specific idiosyncrasies that make observations in that interval have some peculiar heterogeneity. This corresponds to the notion of shocks in time. When assuming both temporal and individual effects, we have the so-called "two-ways" effects. In the tests described above, the presence of these effects can be tested by modifying the "effect" argument, as shown below:

24 The Breusch-Pagan LM test is one of the more traditional ones, but the *plmtest* command can calculate its extensions. These variations are useful depending on the configuration of the data in question (balanced vs. unbalanced, short vs. long). See Croissant and Milo (2018, p. 86).

```
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
```

Given the results of the graphic analysis and the Breusch-Pagan test, we know that the models tested present spatial and temporal heterogeneity. Thus, we executed them again adding new dummies. In the original study, the authors added to the model dummies for the years and to differentiate municipal and presidential elections (Fernandes and Fernandes 2017, p. 665). Although in most situations annual dummies are the standard approach to capture temporal heterogeneity, the specificity of this database recommends that we use only the election type dummy.[25] It should enhance mainly the POLS and FD models. Meanwhile, for FE and RE, we altered the effect for "two-ways", which will introduce effects by municipality and years, implicating in time dummies such as this and variables such as the national GDP possibly being omitted due to collinearity. With the addition of these variables, it is possible to control the heterogenous effects identified previously.

25 For these data, dummies for each year will very likely have accentuated collinearity with some of the variables, especially variables aggregated at the national level (such as the national GDP, which will have a fixed value every year) and even the new dummy for municipal elections.

```
##### MODELS WITH THE ELECTION VARIABLE ###############

# variable to indicate if the election is pres (0) or munic (1)
DATA$type_ele <- ifelse(DATA$year == 2000 |
                        DATA$year == 2004 |
                        DATA$year == 2008, 1, 0)

# formula with election type dummy
form_d <- as.numeric(voteshare) ~ growth+growth_state+growth_country+log_gdp+log_
gdp_state+
  log_gdp_country+mayor_base_presid+health_exp+log_pop+leec+
  lheu+lses+laseps+ldesporc+ldespcor+linvest+ldespers
  type_ele

##### POOLED OLS W DUMMIES ###############

POLS_dic <- plm(form_d, data = DATA, model = "pooling")


##### FD W DUMMIES ########################

mode_fd_dic <- plm(form_d, data = DATA, model = "fd")


##### FE TWO-WAYS #########################

mode_fe_2w <- plm(form_d, data = DATA, model = "within", effect = "twoways")


##### RE TWO-WAYS #########################

mode_re_2w <- plm(form_d, data = DATA, model = "random", effect = "twoways", random.
method = "walhus")
```

4.2 Fixed or random effects?

With POLS ruled out, we can consider FE or RE. As we have pointed out, for this choice it is important to evaluate both the nature of the data as well as the intensity of the correlation between $_i$ and X. To do so, we can turn to (a) contextual knowledge, (b) relevance or requirement of certain variables of interest, and (c) the Hausman test.

(a) As explained in section 3.2, the researcher must answer the following question on the data's origin: can I consider them a random sample of a broader population? Thinking in exogeneity terms, do we have good reason to expect that some municipalities, just by virtue of being, let's say, Salvador or Sirinhaém, have unobserved peculiarities (e.g., a certain party's historic hegemony) that correlate with the selected variables?

(b) One of FE's limitations (also FD's) is not allowing testing elements that are invariant in time. Where there is substantive interest in testing this type of variable, RE can be considered, always taking into account the trade-off between efficiency and bias. The model by Fernandes and Fernandes (2017) does not use variables fixed in time, but if we wished to go further and test, for instance, control dummies for Brazilian regions

(North, Northeast, Midwest, Southeast, and South), we could not use FE.

(c) The *phtest* command runs the specification test by Hausman (1978), comparing FE and RE models, with the null hypothesis that both are consistent. The null hypothesis will be rejected if the correlation between unobserved individual heterogeneities and independent variables is found. In our case, the null hypothesis that both RE and FE are consistent is rejected. Therefore, we know that the unobserved individual heterogeneity and the regressors are correlated, meaning only FE will have consistent estimators. It is crucial to keep in mind that the Hausman test, albeit the standard one to differentiate between FE and RE, assumes models are well-specified, free of bias, and homoscedastic.[26]

```
##### HAUSMAN TEST ##############################

phtest(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
phtest(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
```

4.3 Handling temporal issues

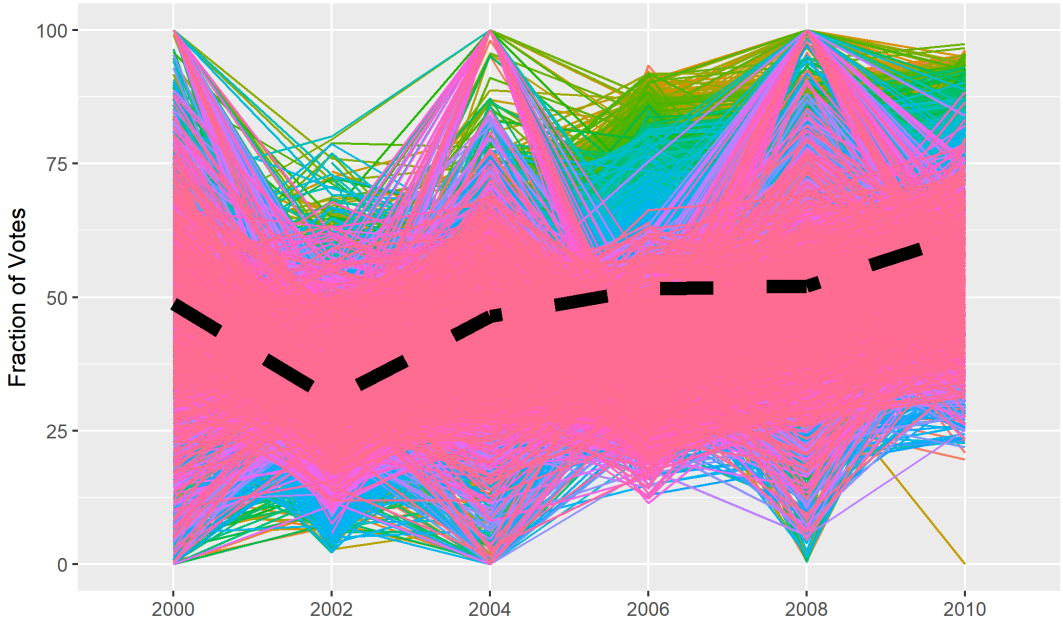
As described in section 3.4, temporal issues are usually (a) non-stationarity and (b) serial correlation of the residuals. The former concerns the raw variables in the database and the latter, the residuals after running the models. Given this sequence, one must first check for stationarity.

(a) Long panels may benefit from tests such as ACF, PACF, ADF, or Panel Unit Root Tests to check stationarity. Since our panel is short (T=6), these tests would not be very informative. If we still want to check if our data are stationary or not, we can look for overall trends towards irreversible growth or retraction in the variables over time. We have done this visual inspection below for the dependent variable "vote share", showing values by municipality and an overall mean for each year. We can have some expectation for a growth trend over time in the incumbent's vote share.

26 If the requirements of absence of misspecification, independence between the error term and independent variables, absence of serial correlation, and heteroskedasticity are not met, it is recommended, in addition to a model revision, the use of robust forms of the test. In *plm*, this is done with the arguments *phtest(..., method="aux")* or yet *phtest(..., method="aux", vcov=vcovHC)* for customized variance-covariance matrices.

```
##### CHECKING STATIONARITY #############

library(dplyr)
##
## Attaching package: 'dplyr'
## The following objects are masked from 'package:plm':
##
##     between, lag, lead
## The following objects are masked from 'package:stats':
##
##     filter, lag
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
data_est <- DATA %>%
  group_by(year) %>%
  mutate(meanvotes = mean(voteshare, na.rm=T)) #database with the vote mean

ggplot(data_est, aes(x=year, y=voteshare, group=codibge, color=codibge))+
  geom_line() + geom_line(aes(x=year, y=meanvotes, group=1),
  linetype="dashed", size=3, color="black")+ ylab("Fraction of Votes") +
  xlab("") + guides(color="none")
```

Figure 7 – Vote share over time (municipalities and overall mean)



Source: the authors, based on original data by Fernandes and Fernandes (2017).

If there is a trend, the inclusion of a time trend is recommended, usually a T variable that takes on the values of 1 for the first year, 2 for the second, and so on.

(b) So far, our selection between the POLS, FE, RE, and FD estimators has been based on how well each approach handled unobserved heterogeneity. Now, success in eliminating temporal correlations also becomes important in choosing an estimator. As explained in 3.4, there are several tests for serial correlation that can be run on residuals, depending on which type of panel (short or long) and the estimator employed. The *plm* package cur-

rently offers eight tests for serial correlation.[27] Since the Hausman test has already told us to abandon RE, we will run only two tests, focusing on FE and FD.

To check the presence of serial correlation on fixed effects models, we used the *pwartest* function. We can see in the output that the test's alternative hypothesis is that there is serial correlation. The p-value was lower than 0.05, indicating the rejection of the null hypothesis (there is no serial correlation). That is, according to the test results, the fixed effects model presents serial correlation.

```
##### 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
```

The *pwfdtest* function is used to check the presence of serial correlation in FD models. The null hypothesis is that there is no serial correlation related to the errors. Given the p-value (< 0.05), we reject the null hypothesis. That is, the FD model presents serial correlation issues.

```
##### 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
```

The tests' results suggest that both FE and FD harbor serial correlation in the error terms. Similarly, the inspection of the raw data indicates an ascending trend, that is, non-stationarity. To remedy the situation, we could consider including a time trend (however, as we saw in 3.2, these cannot be estimated by FE or FD), LDV or employing standard errors robust against serial correlation. As seen previously, temporal issues tend to be more worrisome according to T>N. In our case, given that N is much bigger than T, and to maintain the example's parsimony, we will preserve the current variables and use robust standard errors at the end.

4.4 Handling spatial issues

Having dealt with unobserved heterogeneity, it is expected that remaining spatial issues will be about panel heteroskedasticity and contemporaneous correlation of the errors. The same visual inspection used in section 4.1, grouping residuals by municipality, also informs this decision. Figure 6 has already demon-

27 The six tests not covered in this article are: the generic Wooldridge test for unobserved effects, applicable to POLS (*pwtest*); the Breusch-Godfrey/Wooldridge test for serial correlation, preferably run on RE, but also long FEs (*pbgtest*); the Durbin-Watson test, also preferable for RE (*pdwtest*) and its generalization by Bhargava, Narendranathan, and Franzini (*pbnftest*); the locally-robust test by Bera, Sosa-Escudero, and Yoon for serial correlation or random effects (*pbsytest*); Baltagi and Li test for serial correlation (*pbltest*).

strated that the variance of the residuals between municipalities is systematic.

In addition to the visual inspection, we can turn to specific tests. The *plm* package offers a general command to test cross-sectional dependence (*pcdtest*), wherein the user must choose the type of test according to the panel's characteristics. Since we have a short T and a long N,[28] the most appropriate test is the one proposed by Pesaran (2004).

```
##### TESTING PANEL HETEROSKEDASTICITY ############################

pcdtest(mode_fe_2w, test = "cd") # Fixed Effects model
##
##  Pesaran CD test for cross-sectional dependence in panels
##
## data:  as.numeric(voteshare) ~ cresc + crescuf + crescbr + lpibreal +     lpibuf
+ lpibbrasil + prefeitobasepresidente + persaude +     lpop + leec + lheu + lses +
laseps + ldesporc + ldespcor +     linvest + ldespes + tip_ele
## z = -0.66933, p-value = 0.5033
## alternative hypothesis: cross-sectional dependence
```

The results show that our FE two-ways model is free of contemporaneous dependence, having only serial correlation issues. We must use standard errors robust against serial correlation. We can access different robustness options in *plm*, depending on the type of panel, the way the data are grouped, and the type of problem to be fixed (see Croissant & Millo 2018, chapter 5; Henningsen & Henningsen 2019, pp. 382-384). In our case, because we want to deal with the serial correlation of an FE with two-ways effects, we run the *vcovDC* matrix to estimate robust standard errors:

```
##### HANDLING SERIAL CORRELATION #############

library(lmtest)
mode_fe2w_rbst <- coeftest(mode_fe_2w,
                           vcov = vcovDC)
# adding 'double cluster' robust errors
```

4.5 Results

We present the results of all models in Table 1. To enable the comparison of the original standard errors in the FE model with the robust ones, we put both in separate columns in the table.

Since our main interest is educational and not substantive, we show only a summarized list of variables. Therefore, of the more than 20 variables in the original model, we kept only 11 for the table.

28 Breusch-Pagan LM, activated by the argument test="lm", is recommended for long T and small N; scaled-Breusch-Pagan LM, "sclm", for long T and N; "rho" and "absrho" provide correlation coefficients between the residuals between pairs of observations (see Croissant & Millo, 2018, section 4.4).

Table 1 – Replication results

	Vote Share				
	POLS	RE (2W)	FE (2W)	FE (2W, Robust.)	FD
	(1)	(2)	(3)	(4)	(5)
Municipal growth	1.683***	1.098***	1.148*	1.148	0.190
	(0.608)	(0.040)	(0.652)	(1.517)	(0.669)
State growth	0.095***	0.086***	-0.058*	-0.058	-0.144***
	(0.027)	(0.002)	(0.030)	(0.163)	(0.033)
National growth	3.660***	3.709***			4.231***
	(0.090)	(0.088)			(0.091)
Mayor Base Presid.	1.211***	0.742***	1.289***	1.289***	1.136***
	(0.208)	(0.014)	(0.235)	(0.409)	(0.254)
% Budget Health	0.040***	0.121***	0.044*	0.044	0.058***
	(0.014)	(0.001)	(0.023)	(0.035)	(0.019)
Pop. Munic. (log)	0.302	-0.415***	-6.037***	-6.037**	-3.360*
	(0.258)	(0.018)	(1.449)	(2.609)	(1.824)
Expenditures Educ. Cult. (log)	6.917***	6.461***	1.543***	1.543	0.162
	(0.358)	(0.025)	(0.525)	(1.284)	(0.553)
Expenditures Invest. (log)	2.245***	1.983***	1.430***	1.430	1.644***
	(0.136)	(0.009)	(0.165)	(0.903)	(0.180)
Expenditures Personnel (log)	-1.424***	-0.335***	3.304***	3.304	1.167*
	(0.493)	(0.035)	(0.635)	(2.264)	(0.698)
Municipal Election	1.074***	0.828***			-0.915***
	(0.277)	(0.246)			(0.264)
Intercept	102.026**	152.778***			-4.350***
	(43.701)	(37.792)			(0.373)
N Obs.	26.352	26.352	26.352		20.803
R^2	0.321	0.319	0.039		0.254
Adjust. R^2	0.320	0.318	-0.219		0.253
F-Test	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

Source: the authors, based on Fernandes and Fernandes (2017). The value of theta for the random effects two-ways model was $\theta = 0.11$.

The specification tests lead us to be more confident in the results obtained via FE. Nevertheless, the comparison of all the models is informative. The POLS and FE results tend to be much closer to each other for many of the regressors, while the FE and FD estimates are more similar. This is expected since, as we saw, the latter two remove individual heterogeneity in a similar way. Thus, reading the POLS/RE and FE/FD results is a way of checking how the estimates behave as we control more and more the municipal idiosyncrasies; the "net effect",

as it were, of the variables, all the peculiarities of the cases being removed.

In some variables the difference is manifested in sign change. For example, the POLS and RE models state that there would be positive effects from the state's economic growth rate on the vote share, or negative ones with personnel expenditures. We have a case of bias given that, controlling for the heterogeneities of each municipality with FE and FD, we see that their impact is actually the inverse.

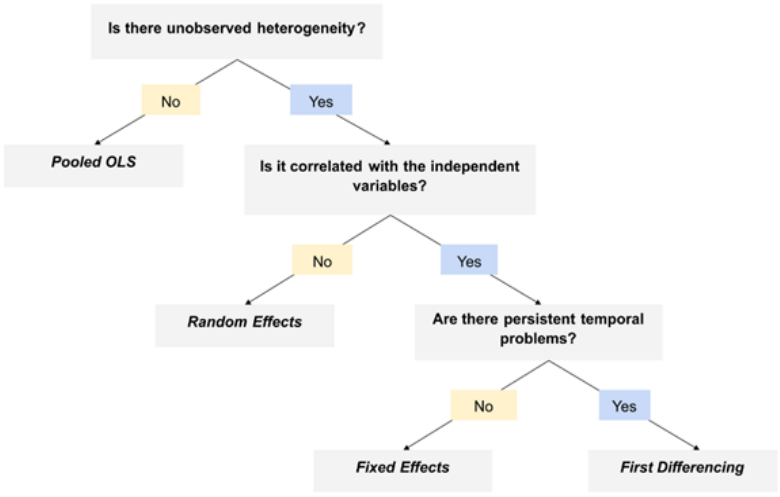
For other regressors, a change in the strength of the relationship is noticed. Sometimes, the effect becomes weaker (e.g., expenditures in education and culture, expenditures in investment), occasionally stronger (dummy for the same mayor-president base, population) and, every so often, stable (percentage of the health budget). Furthermore, while POLS and RE tend to indicate more statistical significance, FD and mainly FE with robust standard errors are more conservative. Notably, the main variable of interest in this study, municipal economic growth, does not retain significant effect on votes on the incumbent.

Lastly, it must be highlighted that the POLS and RE intercepts present unrealistic values due to the fact that the data were not centralized, and that, as we anticipated, FE two-ways omitted time dummies (election type) and those tied to the year (national GDP) due to collinearity.

4.6 Summary of the procedure

To conclude this tutorial, we can represent a model selection process as a decision tree similar to Figure 8.

Figure 8 – How to choose the best estimator?



Source: the authors.[29]

Schematically, first we evaluate to what extent there is unobserved heterogeneity among the database's entities. This can be done with contextual knowledge of the problem, graphic analysis of the intercepts, and by employing specific tests (F and Breusch-Pagan LM). Once POLS is ruled out, we must estimate if there is correlation between the unobserved heterogeneity and the set of explanatory variables, for example, using the Hausman test. If there is no correlation, the use of RE is justified. On the other hand, in the pres-

29 This tree is inspired by a more complex original, which the authors credit to professor Claudia Schwirplies (Uni HH).

ence of correlation between the errors and the independent variables, we can consider using FE or FD. The option between these must take into account the model's temporal issues, as well as the database's size. At the end of the process, tests for time and spatial dependencies are required to check for the need of robust standard errors.

This strategy is a simplification for educational purposes and emphasizes the main decisions that, if ignored, could generate bias or inefficiency. However, as we argue throughout the paper, substantive considerations must also guide that choice.

5. Conclusion

This article presented an introduction to the logic of panel data. We showed the main benefits that come from the analysis of longitudinal information and described the most frequent issues that arise when the dataset gathers, simultaneously, spatial and temporal dimensions. Subsequently, we demonstrated how to implement a regression analysis with panel data in R. Because it is a article with educational purposes, we have also shared all the data and scripts, so that students, professors, and researchers can easily adapt our computational routines to their substantive interests.

At the beginning of this article, we indicated how the use of panel data in Brazilian PIR is still incipient. Thus, we believe it is relevant to explain what can be done to change those circumstances. Among the different options available, we highlight the following: (1) restructure the curricula of undergraduate and graduate courses with the aim of including specific calculus, statistics, and computation classes and offer electives on how to work with longitudinal data, such as professor William Green's class[30]; (2) if item 1 cannot be implemented, ensure the offer of special summer courses such as "Modelling Dynamics" given by professor Lorena Barberia (USP), "Modelling Dynamics in Space and Time" given by professors Guy Whitten and Lorena Barberia (USP), and "Panel Data Analysis" by Igor Viveiros in the MQ/UFMG course;[31] (3) regionally diversify the intermittent offer of data analysis courses. Item 3 being impossible to implement, ensure financial support so that masters and PhD students, specially from the North and Northeast of Brazil, can attend summer schools; (4) encourage the acceptance of online courses as valid credits for graduate curricula; (5) special dossiers on research methods edited by the main PIR journals, with a special emphasis on intuitive and educational articles, such as volume 142 of the Journal of Econometrics.[32] In addition, set space aside in the best PIR journals for the publication of papers with methodological emphasis such as Beck and Katz (1995); and (6) incentivize researchers that are dedicated to the publication of research methods and techniques in PIR through specific financing via funding institutions (CAPES and CNPq).

We are aware of the operational and logistical difficulties that surround each of these initiatives. Nevertheless, considering the potential benefits that may be incorporated in our models of analysis and the limited use of longitudinal data in PIR empirical research, we believe it is about time we take time seriously (Beck, Katz, and Tucker, 1998; Boef and Keele, 2008).

30 http://people.stern.nyu.edu/wgreene/Econometrics/PanelDataNotes.htm

31 http://summerschool.fflch.usp.br/ipsa-usp-summer-school-2017-2/

32 https://www.sciencedirect.com/journal/journal-of-econometrics/vol/142/issue/2

BIBLIOGRAPHICAL REFERENCES

ACHEN, C. H. (2000), "Why lagged dependent variables can suppress the explanatory power of other independent variables". In: *annual meeting of the political methodology section of the American political science association, UCLA*.

BALTAGI, Badi H. (2005), *Econometric Analysis of Panel Data*. England, John Wiley & Sons Ltd.

BERRY, William D. (1993). *Understanding regression assumptions*. Thousand Oaks, Sage.

BECK, Nathaniel; KATZ, Jonathan N. (1995), "What to do (and not to do) with time-series cross-section data", *American political science review*, v. 89, n. 3: 634-647.

BECK, Nathaniel; KATZ, Jonathan N.; TUCKER, Richard. (1998), "Taking time seriously: Time-series-cross-section analysis with a binary dependent variable", *American Journal of Political Science*, v. 42, n. 4: 1260-1288.

BECK, Nathaniel. (2001), "TIME-SERIES–CROSS-SECTION DATA: What Have We Learned in the Past Few Years?", *Annual Review of Political Science*, v. 4, n. 1: 271–293.

BECK, Nathaniel. (2008), "Time-Series Cross-Section Methods", *in* J. Box-Steffensmeier; H. Brady; D. Collier (Eds.). *The Oxford Handbook of Political Methodology*. New York, OUP: 475–493.

BLALOCK, Hubert M. The presidential address: Measurement and conceptualization problems: The major obstacle to integrating theory and research. *American sociological review*, v. 44, n. 6, p. 881-894, 1979.

DE BOEF, Suzanna; KEELE, Luke. (2008), "Taking time seriously", *American Journal of Political Science*, v. 52, n. 1: 184-200.

BELL, Andrew; JONES, Kelvyn. (2015), "Explaining Fixed Effects: Random Effects Modeling of Time-Series Cross-Sectional and Panel Data", *Political Science Research and Methods*, v. 3, n. 1: 133–153.

BELL, Andrew; FAIRBROTHER, Malcolm; JONES, Kelvyn. (2019), "Fixed and random effects models: making an informed choice", *Quality & Quantity*, v. 53, n. 2: 1051-1074.

BENOIT, K. et al. (2018), "quanteda: An R package for the quantitative analysis of textual data", *Journal of Open Source Software*, v. 3, n. 30: 774.

BOX, G. E. P.; JENKINS, G. M.; REINSEL, G. C.; LJUNG, G. M. (2016), *Time series analysis: forecasting and control*. Hoboken/New Jersey, Wiley.

CLARK, Tom S.; LINZER, Drew A. (2015), "Should I Use Fixed or Random Effects?", *Political Science Research and Methods*, v. 3, n. 2: 399–408.

CROISSANT, Yves; MILLO, Giovanni. (2008), "Panel data econometrics in R: The plm package", *Journal of statistical software*, v. 27, n. 2: 1-43.

CROISSANT, Yves; MILLO, Giovanni. (2018), *Panel Data Econometrics with R*, Wiley.

CERQUEIRA, Daniel et al. (2017), *Atlas da violência 2017*.

DIELEMAN, J. L.; TEMPLIN, T. (2014), "Random-Effects, Fixed-Effects and the within-between Specification for Clustered Data in Observational Health Studies: A Simulation Study", *PLoS ONE*, v. 9, n. 10: e110257.

FÁVERO, L. P. L. (2013), "Dados em painel em contabilidade e finanças: teoria e aplicação", *BBR-Brazilian Business Review*, v. 10, n. 1: 131-156.

FERNANDES, Ivan Filipe de Almeida Lopes; FERNANDES, Gustavo Andrey de Almeida Lopes. (2017), "A importância do crescimento econômico local na escolha do chefe do Executivo no Brasil", *Revista de Administração Pública*, v. 51, n. 4: 653-688.

FINKEL, S. E. (1995), *Causal analysis with panel data* (No. 105). Sage.

FORTIN-RITTBERGER, J. (2013), "Time-Series Cross-Section", *in* H. Best; C. Wolf (Eds.), *The SAGE Handbook of Regression Analysis and Causal Inference*. London, SAGE Publications: 387–408.

GUJARATI, Damodar N. *Basic Econometrics*. McGraw-Hill Companies. New York, 2004.

HAUSMAN, Jerry A. (1978), "Specification tests in econometrics", *Econometrica: Journal of the econometric society*, v. 46, n. 6: 1251-1271.

HENNINGSEN, Arne; HENNINGSEN, Géraldine. (2019), "Analysis of Panel Data Using R", *in* M. Tsionas (Ed.) *Panel Data Econometrics: Theory*. London, Elsevier: 345-396.

HSIAO, C. (2003), *Panel data analysis*, Cambridge University Press.

IBGE. (2017), Pesquisa Nacional por Amostra de Domicílios Contínua. Disponível em: https://www.ibge.gov.br/estatisticas/sociais/trabalho/17270-pnad-continua.html?=&t=o-que-e. Acesso em 30-06-2020

KING, Gary. (2001). "Proper Nouns and Methodological Propriety: Pooling Dyads in International Relations Data", *International Organization*, v. 55, n. 2: 497-507.

KING, Gary. (1995), "Replication, replication", *PS: Political Science and Politics*, v. 28, n. 3: 444-452.

KING, Gary; KEOHANE, Robert O.; VERBA, Sidney. (1994), *Designing social inquiry: Scientific inference in qualitative research*. Princeton University Press.

KENNEDY, Peter. (2003), *A guide to econometrics*. MIT Press.

KRUEGER, James S.; LEWIS-BECK, Michael S. (2008), "Is OLS dead?", *The Political Methodologist*, v. 15, n. 2: 2-4.

LAROCCA, R. (2012). "Gauss–Markov Theorem" in SALKIND, N. J. (ed) *Encyclopedia of Research Design*. Thousand Oaks, Sage, 529-533.

LIJPHART, A. (1971), "Comparative Politics and the Comparative Method", *American Political Science Review*, v. 65, n. 3: 682–693.

MARQUES, Luís D. (2000), "Modelos Dinâmicos com Dados em Painel: revisão de literatura". *Working paper*. Available at http://wps.fep.up.pt/wps/wp100.pdf. Accessed 23-06-2020.

MENARD, Scott. (2002), *Longitudinal research*. Thousand Oaks, SAGE Publications.

MEDEIROS, M.; BARNABE, I.; ALBUQUERQUE, R.; LIMA, R. (2016), "What does the field of International Relations look like in South America?", *Revista Brasileira de Política Internacional*, v. 59, n. 1: e004

MEIRELES, Fernando; SILVA, Denisson; BARBOSA, Rogerio. (2019), *rscielo: A Scraper for Scientific Journals Hosted on Scielo*. R package version 1.0.0. https://CRAN.R-project.org/package=rscielo.

MESQUITA, Rafael. (2018) *Liderança regional em perspectiva comparada: Brasil e Turquia*. PhD Thesis, UFPE. Available at https://repositorio.ufpe.br/handle/123456789/32942. Accessed 29-05-2021

MUMMOLO, J.; PETERSON, E. (2018). "Improving the interpretation of fixed effects regression results", *Political Science Research and Methods*, v. 6, n. 4: 829-835.

PESARAN, M. H. (2004), "General diagnostic tests for cross section dependence in panels", *CESifo Working Paper Series*, 1229.

PEARL, J. (2000), *Causality: Models, reasoning, and inference.* New York, Cambridge University Press.

SKRONDAL, A., RABE-HESKETH, S. (2008), "Multilevel and Related Models for Longitudinal Data", *in* Leeuw J.., Meijer E. (Eds.) *Handbook of Multilevel Analysis*, New York, Springer: 275-300.

STOCK, James H.; WATSON, Mark W. (2004), *Econometria*. Pearson.

VERBEEK, Marno; NIJMAN, Theo. (1992), "Testing for selectivity bias in panel data models", *International Economic Review*, v. 33, n. 3: 681-703.

WOOLDRIDGE, J. M. (2002), *Econometric Analysis of Cross Section and Panel Data*, Cambridge, MIT Press.

WOOLDRIDGE, J. M. (2013), *Introductory econometrics: a modern approach*. Australia, South-Western Cengage Learning.

WORRALL, John L. (2008), "An introduction to pooling cross-sectional and time series data", *in* S. Menard (Ed.) *Handbook of longitudinal research: Design, measurement, and analysis*, Elsevier: 233-248.

Questions and Answers

1. **Should I use fixed or random effects?**

Generally, random effects can introduce bias, that is, increase the chance of systematically under- or overestimating the parameter of interest. However, this method tends to reduce the coefficients' variance, which means more efficient estimates (Kennedy, 2003). Therefore, models with random effects tend to produce more consistent estimates when the sample is small and when the correlation between the independent variable and the effects of the observations of relatively small (Clark and Linzer, 2015). On the other hand, fixed effects tend to produce unbiased estimates, especially for large samples (Mummolo and Peterson, 2018). A disadvantage of fixed effects is the reduction of the model's degree of freedom given the inclusion of dummies. However, if the sample is sufficiently large, this will not be a problem in estimating the coefficients. For more information: Wooldridge (2002) and Clark and Linzer (2015).

2. What is the difference between an unbalanced and a balanced panel?

Stock and Watson (2004) define a balanced panel as a dataset in which there are no missing cases, that is, we have information for all observed units (N) for all time periods (T). A panel is unbalanced when, for whatever reason, discontinuities in the data matrix occur, whether due to lack of observations or the absence of time units. Wooldridge (2002) warns that, overall, the statistical treatment will be the same for both panel types. In all cases, the research must inform the reason for the unbalanced panel and detail the criteria for sample selection. For an introduction on the subject, see Mayer (2010).[1] For a detailed discussion on the specificities of the unbalanced panel, Wooldridge (2002) and Baltagi (2005). For a more advanced treatment on selection bias in panel data, see Verbeek and Nijman (1992).

3. Short or long panels?

The panel's amplitude depends on the ratio between entities (N) and time units (T). In a short panel, we have many observed entities over a relatively short period of time. Imagine a dataset with detailed information for all Brazilian municipalities between 2000 and 2010. Undoubtedly, it is a lot of information, but since the number of entities (over five thousand municipalities) is much higher to the number of time frames (ten years), we have a short panel where the cross-sectional dimension is dominant. A long panel is present when the offer of cross-section data is significantly smaller than the temporal range of the information. For example, imagine a dataset of the country's five regions (N, NE, MW, SE, and S) gathered annually between 1950 and 2020. That composition, in which T>N, brings us closer to a time series trend, requiring specific techniques. For more information, see Cameron and Trivedi (2005)[2] and Stock and Watson (2004).

1 See: <https://homepage.univie.ac.at/robert.kunst/pan2010_pres_mayer.pdf>.

2 See: <http://public.econ.duke.edu/~vjh3/e262p_07S/readings/Cameron_&_Trivedi_Microeconometrics_Chapters_2.pdf>.

4. How does one avoid omitted variable bias with panel data?

As with other research designs, it is necessary to be up to date on the field's specific literature to know the theoretical specification of the models. Often we can find an important variable in a database that already exists. Typically, FE and FD offer good safeguards against omitted variables when the important unobservables are fixed to the units. Even then, their use does not substitute authoritative knowledge of the literature. It is important to remember that the inclusion of an irrelevant variable in the explanatory model will only harm the efficiency of the coefficients. The standard error will be higher than it should be and that reduces the reliability of the significance tests, but $\hat{\beta}$ will still provide a consistent estimate of β, that is, the inclusion of an irrelevant variable does not cause bias. On the other hand, the exclusion of a theoretically important variable tends to produce biased estimates, especially when the omitted variable is correlated with the remaining variables included in the model.

5. What is the Hausman test and what is for?

The goal of the Hausman test is to evaluate if there is correlation between the errors of the model and the independent variables. Within the context of panel data, the test is employed to guide the decision between fixed and random effects. The null hypothesis states that random effects are more suitable. Operationally, we must estimate both models (with fixed and random effects) and after comparing the estimates. A significant result (p-value<0.05) indicates that the null hypothesis must be rejected, that is, we must choose fixed effects. On the other hand, a non-significant result (p-value>0.05) implies not rejecting the null hypothesis, which means adopting the model with random effects. The Hausman test usually comes with a chi-square distribution with a significance test, in addition to the estimates of the coefficients. For more, see Hausman (1978).

6. What is serial correlation and why should I be concerned with it?

Although autocorrelation can occur in space, the one that interests us here is the one that happens in time, also known as serial correlation. The Cambridge Dictionary of Statistics defines serial correlation as the dependence between measures of the same entity in longitudinal studies. Usually, the closer the temporal proximity in the measurements, the stronger the magnitude of the correlation. On the other hand, the farther the measurements of time, on average, the weaker the degree of correlation between the values of the series. This association structure of the values in time violates the assumption of independence of the errors in the regression model. The main issue with serial correlation is production of inconsistent significance tests, that is, we will not be able to trust the p-value and the confidence intervals. That is the reason longitudinal data analysis must always be accompanied by estimates of the level of autocorrelation in the series.

7. How do I learn more about panel data analysis?

In Econometrics, we suggest Kennedy (2003), Wooldridge (2002, 2013), Baltagi (2005) and Gujarati (2004). Hsiao (2003) is one of the most complete introductions to the theme. The paper "What to do and (not to do with) with Time-Series Cross-Section Data" by Beck and Katz (1995) is the most widely cited paper in Political

Science (7,414 citations in Google Scholar in June 2021). Fortin-Rittberger (2013) is also an introductory paper. Finkel (1995) gives focus to causal possibilities using longitudinal data. In doubt between fixed or random effects? See Bell and Jones (2015). Mummolo and Peterson (2018) reproduces several studies that employ panel data and demonstrates how to improve the interpretation of the coefficients estimated using fixed effects. See the schedule the IPSA/USP summer school: <http://summerschool.fflch.usp.br/tracks-and-courses/>.

8. **TSCS x panel data?**

Beck (2008) differentiates panel data from TSCS data. According to him, the methods that work well for panels do not necessarily work well for TSCS data and vice-versa. In PIR empirical research, the most frequent combination of cases and time is seen in databases that compile information about countries (N) over years (T) (Beck, 2001). Thus, since data of the TSCS type consist of exactly the same cases over time, a stronger temporal correlation than for panel data is expected. TSCS consists of the collection of information on observational units (individuals, companies, federal units, countries) over time (day, month, year). The bigger the T is in relation to the N, the stronger are concerns over the serial structure of the data. On the other hand, when there are a lot of cases (N) for few time frames (T), autocorrelation problems will be less severe.

9. How do I recover each unit's intercepts when using FE or RE in *plm*?

Use the fixef() and ranef() commands. The values correspond to the deviation in comparison to the general intercept.

10. What is the difference between using the *plm* or the *lme4* and *nlme* packages?

Panel data are not the only way of handling repeated measures. Another approach consists of using mixed linear models, also known as hierarchical or multilevel models, which are executed by packages such as *lme4* and *nlme*. While the panel data perspective is mainly econometric, in other fields such as healthcare, biology, and education, the multilevel approach has a longer tradition as their data are typically hierarchical (e.g., students in classrooms, classrooms in schools). Panel data may be seen as a two-level nested structure: there are individuals *i* (upper level) observed at successive occasions *t* (lower level). Since FE and RE specifications admit intercepts varying by individual, they have a logic close to hierarchical models in which there are random intercepts. In addition to the intercepts, it is possible to work with panel data predicting variable coefficients according to individual or period ($\beta_{\square\square}$) (examples in Wooldridge 2013, p. 451-453; Croissant and Millo 2018, section 8.2), although that possibility is more broadly explored in hierarchical models (random intercepts + random slopes). When we have data configurations in which nesting is not in time or is more complex than in only two levels (e.g., municipalities in states, states in regions), the use of hierarchical models in recommended. From an operational point of view, the ecosystem of R packages is more well-developed around *lme4/nlme* than *plm*, which means it is easier to integrate products from *lme4/nlme* analyses into other graphic, testing, and result presentation packages. For more, see Skrondal and Rabe-Hesketh (2008) and Croissant and Millo (2008, pp. 33-39).

Appendix

Table 1 – Total number of articles collected per journal in the PIR field (2000-2018)

PIR Journal	Qualis (2016)	Total number of articles	% Total
Estudos Avançados	B1	896	12
Revista de Administração Pública	A1	726	9
Sociologias	B1	573	7
Revista Brasileira de Ciências Sociais	A2	559	7
Tempo Social	B1	533	7
Revista de Sociologia e Política	A1	518	7
Caderno CRH	A2	503	6
Sociedade e Estado	B1	483	6
Dados - Revista de Ciências Sociais	A1	469	6
Lua Nova: Revista de Cultura e Política	A2	450	6
Novos estudos CEBRAP	A2	420	5
Revista Brasileira de Política Internacional	A1	403	5
Opinião Pública	A1	340	4
Contexto Internacional	A2	323	4
Revista Brasileira de Ciência Política	B1	235	3
Civitas - Revista de Ciências Sociais	B2	183	2
Brazilian Political Science Review	A2	150	2
Total: 17 journals		7,764	100

Source: the authors