UMA INTRODUÇÃO AOS MODELOS DE DADOS EM PAINEL: O QUE SÃO E COMO SE ESTIMAM EM STATA

Pedro Gomes Rodrigues
Investigador Integrado no Centro de Administração e Políticas Públicas (CAPP)
Professor Auxiliar no Instituto Superior de Ciências Sociais e Políticas (ISCSP),
Universidade de Lisboa

Resumo: Esta nota destina-se principalmente aos alunos com interesse, no geral, por métodos quantitativos e, em particular, pelos chamados modelos de dados em painel. Sendo de natureza pedagógica, é uma primeira introdução ao tema, com vista a habilitar o aluno à sua estimação e interpretação em Stata, o software de eleição entre microeconometristas. Para além de explicitar a diferença entre duas especificações alternativas, são apresentados os princípios comummente usados para guiar a escolha entre elas.

Palavras-chave: Dados em painel; Efeitos fixos (FE); Efeitos aleatórios (RE); Stata;

Abstract: This note is primarily aimed at students interested, in general, in quantitative methods and, in particular, in panel data models. Having a pedagogical nature, it is a first introduction to the topic, with the objective of equipping the student with what he/she needs to estimate and interpret them in Stata, the software that is preferred by microeconometricians. In addition to explaining the difference between two alternative specifications, we present the commonly-used principles to guide the choice between them.

Keywords: Panel data; Fixed effects (FE); Random effects (RE); Stata;

1. O que são dados em painel?

Dados em painel, também conhecidos como dados longitudinais, têm uma natureza multidimensional: registam-se séries temporais para um conjunto de entidades que podem ser, por exemplo, indivíduos, empresas, países ou cidades. Desta forma, acompanha-se a evolução ao longo do tempo do conjunto de entidades que queremos estudar (Durlauf & Blume, 2009; Asteriou & Hall, 2015).

Sendo Y_{it} a variável dependente ou a explicar, referente à entidade $i = 1, \ldots, M$ no período de tempo $t = 1, \ldots, T$ e sob a hipótese de linearidade, a sua equação escreve-se genericamente como:

$$(1)\ y_{it} = \beta_1 x_{1,it} + \cdots + \beta_K x_{K,it} + \alpha_i + \varepsilon_{it}.$$

Nesta especificação, há K variáveis independentes – os xs – (também conhecidas como variáveis explicativas ou regressores) e o resíduo, u_{it}, é a soma de duas componentes: α_i, constante ao longo do tempo, e ε_{it} que se assume como independente e identicamente distribuído (i.i.d.). Vale a pena frisar que α_i é como uma característica que assume um valor que é diferente entre entidades.

Antes de prosseguir, o leitor interessado poderá querer refrescar os seus conhecimentos sobre métodos quantitativos. Para tal, poderá consultar referências mais recentes como Wooldridge (2013), Stock & Watson (2018), ou Studenmund & Johnson (2017), complementados pelos vídeos do Professor Mark L. Burkey que no seu conjunto formam uma intuitiva introdução à econometria.[1]

A equação genérica (1) encapsula duas variantes, apelidados de 'efeitos fixos' (FE, ou *fixed effects*) e de 'efeitos aleatórios' (RE, ou *random effects*), que a seguir se detalham.

1.1. A especificação assumindo efeitos fixos (FE)

No caso FE, representa uma característica não observada, específica a cada entidade e invariante no tempo. Sendo a entidade de referência, i.e., aquela com a qual se fazem comparações na interpretação das diferentes ordenadas na origem, na variante FE a equação (1) toma a forma de um modelo de mínimos quadrados

[1] http://www.burkeyacademy.com/home/statistics-econometrics

com variáveis dicotómicas ou binárias (LSDV, ou *Least Squares Dummy Variables* em inglês):

$$(2)\ y_{it} = \beta_0 + \beta_1 x_{1,it} + \cdots + \beta_K x_{K,it} + \gamma_2 D_{2t} + \gamma_M D_{Mt} + \varepsilon_{it}$$

onde D_{jt} é uma variável dicotómica, binária ou *dummy,* que toma o valor 1 quando $i = j$ para todo o t. Assim, $\alpha_1 = \beta_0$ e $\alpha_j = \beta_0 + \gamma_j$, para $j = 2, \ldots, M$ são as características específicas a cada entidade. Estes são os efeitos fixos (FE) específicos a cada entidade.

Apenas podemos incluir $M - 1$ variáveis *dummy* para evitar um problema de multicolinearidade (também conhecido na literatura econométrica como a 'armadilha das variáveis *dummy*'). Dito de outra forma, tipicamente exclui-se a *dummy* correspondente à entidade convencionada como sendo de referência, i.e. aquela em relação à qual se fazem comparações na interpretação das diferentes ordenadas na origem. Se cometessemos o erro de incluir todas as variáveis *dummy*, então a constante β_0 seria simplesmente proporcional à soma das variáveis *dummy,* D_{it}, …, D_{Mt}. É instrutiva a consulta ao Quadro 1, onde é imediato concluir que, a existir uma variável D_1, que tomaria o valor de 1 para a entidade 1 e 0 para as outras entidades, a soma das colunas $D_1 + D_2 + D_3$ resultaria na variável 'constante', que não é mais que uma coluna de uns.

Quadro 1: Formato adequado para a base de dados em painel

Entidade	Ano	y	x_1	x_2	D_2	D_3	P_2
1	2001				0	0	0
1	2002				0	0	1
1	2003				0	0	0
2	2001				1	0	0
2	2002				1	0	1
2	2003				1	0	0
3	2001				0	1	0
3	2002				0	1	1
3	2003				0	1	0

Fonte: elaboração própria.

Embora a especificação em FE permita a identificação de características específicas a cada entidade, no caso de um painel equilibrado[2], esta implica uma perda de graus de liberdade de $M * T - K - 1$ para $M * T - K - 1 - (M - 1) = M * T - K - M$. Com um reduzido número total de observações $M * T$, este pode ser

[2] Um painel onde existem dados para todos os anos e para todas as entidades.

um problema para a inferência. Para garantir a acuidade nas estimativas dos coeficientes, neste caso, os livros de texto em econometria como Asteriou & Hall (2015) ou Wooldridge (2013), por exemplo, sugerem que o utilizador adote uma especificação parcimoniosa, i.e., com um pequeno número total de regressores.

Contudo, se no processo que gerou os dados recolhidos tiver havido características específicas a cada entidade, e o investigador as ignorar, nesse caso este incorre noutro problema: o de enviesar as estimativas por omissão de variáveis relevantes. É por esta razão que num trabalho aplicado se sugere sempre, como boa prática, a inspiração na teoria – precisamente para nos indicar para que variáveis devemos recolher dados.

Para além dos efeitos fixos, específicos a cada entidade, podemos também, de forma alternativa ou complementar, considerar efeitos fixos relativos a cada período do tempo. Nesse caso, acrescentam-se à equação (2) *dummies* referentes a cada momento *t*, i.e.,

$$(3)\ y_{it} = \ ... + \delta_2 P_{i2} + \cdots + \delta_T P_{iT}$$

em que P_{ij} é uma variável dicotómica que toma o valor de 1 sempre que $t = j$, para todas as entidades *i*. Há $T - 1$ *dummies* deste tipo, também para evitar o já referido problema de multicolinearidade. O momento do tempo para o qual não há *dummy* é o período de referência em relação ao qual se fazem comparações na interpretação dos diferentes coeficientes δ. Novamente, é instrutiva a consulta ao Quadro 1 para comprovar que, a existirem P_1, P_2, e P_3, então a constante (que é apenas uma coluna de uns) seria obtida pela a soma das colunas , o que resultaria no problema de multicolinearidade.

Da mesma forma que os efeitos fixos específicos a cada entidade captam a heterogeneidade entre as mesmas, os efeitos fixos relativos a cada período de tempo pretendem captar o que de anómalo (se alguma coisa) se passou nesse momento.

Depois de especificado, o modelo é estimado usando um software como o Stata que contém vários pacotes estatísticos. Isso quer dizer que, depois de carregar os dados na memória do computador, de preferência no formato apresentado no Quadro 1, são obtidas estimativas dos seguintes coeficientes: os betas, os gammas e os deltas. Muitas vezes, o objetivo do trabalho empírico é um dos seguintes: (i) saber se determinada variável é ou não significativa do ponto de vista estatístico, (ii) determinar quantitativamente a melhor estimativa de um certo coeficiente de interesse, ou ainda (iii) fazer o teste de hipóteses, que tipicamente usa uma combinação linear de alguns coeficientes. Com os coeficientes estimados, podemos também prever que valor provavelmente tomaria a variável dependente de uma determinada entidade se as respetivas variáveis explicativas tomassem outros valores. Esses exercícios são úteis e

relevantes, numa perspetiva de apoio à decisão, na medida em que permitem não só avaliar a acuidade do modelo na previsão (para tal, bastaria ter deixado uma parte dos dados de fora no momento da estimação), como também cenarizar contrafactuais, i.e. que outros valores tomaria a variável dependente se o conjunto dos regressores ou variáveis explicativas fosse diferente.

1.2. A especificação assumindo efeitos aleatórios (RE)

Esta especificação alternativa parte à mesma da equação (1), $y_{it} = \beta_1 x_{1,it} + \cdots + \beta_K x_{K,it} + (\alpha_i + \varepsilon_{it})$, com a diferença de que, agora, α_i faz parte do resíduo, u_{it}. Nesse caso, para avançar, é necessário impor a hipótese (forte) de que a correlação entre u_{it} e as variáveis explicativas é nula. Se no processo que gerou os dados tiver havido algum tipo de (cor)relação entre u_{it} e $(x_{1,it}, \ldots, x_{k,it})$, então haverá um enviesamento nas estimativas dos coeficientes, $\beta_1, \ldots, \beta_K$.

Embora seja verdade que, sob a especificação RE, podemos deixar de fora as *dummies* (D e P), sem incorrer num enviesamento por omissão de variáveis relevantes, ganhando assim nos graus de liberdade, somos obrigados a corrigir a autocorrelação nos resíduos, u_{it}, o que quer dizer que não podemos simplesmente aplicar a técnica dos mínimos quadrados (OLS), que pressupõe que estes são independentes e identicamente distribuídos (i.i.d.). Existirá autocorrelação nos erros simplesmente porque $u_{it} = \alpha_i + \varepsilon_{it}$. Assim, mesmo que ε_{it} seja i.i.d., o α_i que corresponde a uma característica específica a uma entidade, que é constante ao longo do tempo, é um efeito persistente que impede que seja i.i.d., como o OLS requer.[3]

O facto do estimador em RE apenas ser eficiente no caso em que a correlação entre e as variáveis explicativas for nula obriga o utilizador desta especificação, em particular dos modelos em dados de painel, a basear essa escolha na relação que existe na teoria e que está documentada na literatura. Dito de outra forma, quem decide estimar um modelo em RE deve sempre perguntar a si próprio se é razoável ou não assumir que o resíduo u_{it} e as variáveis explicativas não estão relacionadas entre si. Se para aquele tema que está a ser investigado, partindo da teoria económica, não houver boas razões para suspeitar de tais relações, então é seguro avançar com a especificação RE.

[3] Nesse caso, é necessário usar mínimos quadrados generalizados (*generalized least squares* ou GLS). Em muitos pacotes estatísticos, o usual é recorrer ao estimador de Swamy & Arora (1972) que é assimptoticamente eficiente com grandes.

2. Como escolher entre a especificação FE e a especificação RE?

Em termos muito pragmáticos, a escolha entre adotar a especificação FE ou a especificação RE é uma questão algo polémica. De facto, diferentes investigadores aplicados de áreas diversas têm opiniões fortes, nem sempre bem fundamentadas. O consenso entre estatísticos é que, antes de usar o Teste de Hausman[4] que permite testar qual das duas especificações usar, é preferível pensar se, para a área de estudo em causa – onde cada caso é um caso – as características específicas a cada entidade estarão ou não (cor)relacionadas com os regressores incluídos na equação a estudar. Se não houver uma boa razão para justificar uma correlação nula entre u_{it} e os *xs*, a hipótese que a especificação RE exige que se imponha, o melhor então e usar FE. Claro que, em qualquer caso, o investigador deve procurar seguir o princípio da parcimónia para garantir um elevado número de graus de liberdade. Em todo o caso, o Quadro 2 procura sistematizar um pequeno conjunto de princípios que ajudam na escolha entre as duas especificações.

Quadro 2: Orientações gerais que guiam a escolha entre especificações

Efeitos fixos (FE)	**Efeitos aleatórios (RE)**
Os α_i podem estar relacionados com os regressores (X), i.e., a correlação entre u_i e X não é nula.	Os α_i não estão relacionados com os regressores. Nesse caso, assume-se que a correlação (u_i , X) = 0.
Há interesse em determinar o impacto de regressores que mudam ao longo do tempo.	Há interesse em determinar o impacto de regressores que não mudam ao longo do tempo (como o género ou a superfície em km²).
Observa-se o mesmo conjunto de entidades (todo o universo), como por ex. todos os países da UE.	Observa-se uma amostra aleatória do universo em estudo.
Tem-se um número significativo de observações, para que a perda de graus de liberdade (que advém do uso de muitos *dummies*) não seja problemática.	Se o número de observações for pequeno, pode ser preferível optar pela especificação em RE.

Fonte: elaboração própria.

4 O Teste de Hausman (Hausman, 1978) investiga estatisticamente se a correlação entre u_{it} e *X* é nula *(hipótese nula: H_0)* ou não *(hipótese alternativa: H_a)*. Sob a hipótese nula, a utilização da especificação RE não introduz qualquer enviesamento pela omissão de variáveis relevantes. Se à estatística de teste estiver associado um *p-value* inferior a 5%, por exemplo, devemos rejeitar a hipótese nula e não se pode justificar usar a especificação RE, porque os coeficientes estimados estariam irremediavelmente enviesados.

3. Como estimar modelos de dados em painel usando o Stata

Para os alunos que estudam em instituições que conferem um grau académico, é possível adquirir uma licença temporária para a utilização do Stata sem qualquer restrição. Em março de 2019, estas licenças começavam em $45.[5] Por esta módica quantia, o aluno dispõe de pelo menos 6 meses para explorar e usar o *software*, tempo que deverá ser suficiente para correr as regressões necessárias para completar o trabalho de métodos quantitativos no âmbito de uma tese de mestrado ou de doutoramento. Licenças para trabalhar com grandes bases de dados durante um período de tempo mais alargado também estão disponíveis, a um preço mais elevado.

Esta secção sobre como estimar modelos de dados em painel em Stata é meramente introdutória, sendo o seu objetivo demonstrar como é fácil e intuitivo fazê-lo. Para além de recorrer aos extensos e sempre muito completos manuais do Stata – disponíveis eletronicamente através do comando help – o leitor interessado poderá consultar, por exemplo, Cameron & Trivedi (2010) dedicado a estimar modelos microeconométricos.

Para começar, é sempre uma boa ideia olhar com cuidado para a base de dados, com dois objetivos. Primeiro, para nos certificarmos que não existem erros ou valores que não fazem sentido. Para tal, muitas vezes, basta uma rápida vista de olhos. Segundo, devemos evitar a combinação, numa mesma base de dados, de números muito grandes com números muito pequenos. Esse problema é de fácil resolução, bastando para isso alterar as escalas nas quais estão expressas as variáveis, o que também facilita a interpretação dos coeficientes a estimar.

O passo seguinte é informar o Stata que vamos trabalhar com dados de painel. Admitindo que a base de dados está no formato do Quadro 1, usa-se o seguinte comando declarativo: xtset entidade ano seguido da tecla *Enter*.

A sequência que se segue estimará o modelo, primeiro com a especificação FE: xtreg y x1 x2, fe sendo sempre uma boa ideia guardar os resultados obtidos através do comando estimates store fixed e depois sob a especificação RE: xtreg y x1 x2, re desta vez através do GLS, seguido de estimates store random para salvar os resultados da nova regressão.

Para além dos coeficientes estimados, o *output* do Stata na sequência do comando xtreg inclui $P > |z|$ que serve para determinar se uma certa variável é estatisticamente significativa ou não, i.e. serve para testar a hipótese nula de que o coeficiente que lhe está associado é zero. Para um nível de significância do teste, convencionado em 5%, um determinado coeficiente diz-se estatisticamente significativo se o valor P (ou *P-value*) associado for inferior a 0,05. Nos quadros dos resultados de estimação que aparecem nos artigos científicos (e também nos livros de econometria) uma variável que apenas é estatisticamente significativa

[5] https://www.stata.com/order/new/edu/gradplans/student-pricing/

aos 10% é acompanhada de um só asterisco, *, aos 5%, **, e a 1% de significância, ***. Fica o exercício para o leitor replicar o raciocínio, de acordo com o qual uma variável que é estatisticamente significativa aos 5% também o será aos 10%, mas não necessariamente a 1%.

Este alinhamento permite ainda, com o comando hausman fixed random implementar o Teste de Hausman e assim obter a evidência de natureza estatística que nos ajudará a escolher entre as duas especificações alternativas.

A hipótese nula favorece a especificação RE, pelo que se o *P-value* associado à estatística de teste ($Prob > \chi 2\ ou\ Prob > Chi2$) for inferior ao nível de significância (5% ou 0,05), então, nesse caso, deve-se rejeitar a especificação RE a favor da especificação alternativa FE. De forma similar, se o *P-value* for superior ao nível de significância, optamos pela especificação RE.

4. Para quem quiser saber mais

Saber o que são modelos de dados em painel e como se estimam usando o Stata foi o objectivo desta breve nota. Com a disponibilidade de bases de dados cada vez mais ricas (o chamado *big data*), quer no setor privado, quer na Administração Pública, é uma classe de modelos que é cada vez mais procurada e, por isso, a sua existência já não pode ser ignorada pelo aluno interessado em métodos quantitativos atuais. Por outro lado, sendo o Stata o pacote estatístico comercial de eleição entre os profissionais da microeconometria (e não só, como prova a sua popularidade nos domínios da saúde, da ecologia e da biologia, só para dar três exemplos), é o *software* de referência nas aplicações estatísticas/econométricas ao mais alto nível.

Esta nota, que admitidamente é apenas uma primeira introdução a um vasto tema, termina com a indicação de algumas referências bibliográficas sobre modelos de dados em painel, para o leitor que quiser saber mais. Sendo um *Handbook* da Oxford, Baltagi (2014) é enciclopédico e a sua consulta revela o quão fulcral esta classe de modelos é na investigação aplicada dos dias que correm. Todo o género de variantes está lá, devidamente documentado. Para o leitor matematicamente mais sofisticado – mas sem perder o interesse pela aplicabilidade imediata – sugerem-se, por ordem cronológica, Cameron & Trivedi (2005), Wooldridge (2010), Pesaran (2015) e ainda Biørn (2016). Para quem ainda não ficou convencido que o Stata é a plataforma por excelência para a estimação desta classe de modelos, num ambiente que é fácil de usar, a referência incontornável é Cameron e Trivedi (2010).

Nota: De natureza marcadamente pedagógica, este texto teve por base dois vídeos sobre modelos de dados em painel, da autoria do Professor Mark Burkey,

da *A&T State University* da Carolina do Norte, EUA, que o leitor poderá consultar em: http://www.burkeyacademy.com/home/statistics-econometrics. Deverá ser útil, não apenas ao aluno que apresenta dificuldades com a língua inglesa, mas também porque, ao material de base, se acrescenta intuição, interpretação e aplicação em Stata.

Bibliografia

Asteriou, D. & S.G. Hall (2015) *Applied Econometrics*. Red Globe Press, 3rd ed.

Baltagi, B.H. (2014) *The Oxford Handbook of Panel Data (Oxford Handbooks)*. Oxford University Press.

Biørn, E. (2016) *Econometrics of Panel Data: Methods and Applications.* Oxford University Press.

Cameron, A.C., & P.K. Trivedi (2010) *Microeconometrics Using Stata: Revised Edition*. Stata Press.

Cameron, A.C., & P.K. Trivedi (2005) *Microeconometrics: Methods and Applications.* Cambridge University Press.

Durlauf, S., & L. Blume (2009) *Microeconometrics (The New Palgrave Economics Collection)*. Palgrave MacMillan.

Hausman, J. (1978) "Specification tests in econometrics". *Econometrica* 46: 1251-1271.

Pesaran, M.H. (2015) *Time Series and Panel Data Econometrics*. Oxford University Press.

Stock, J.H. & M.W. Watson (2018) *Introduction to Econometrics*. Pearson Series in Economics, 4th ed.

Studenmund, A.H. & B.K. Johnson (2017) *Using Econometrics: A Practical Guide*. Pearsson.

Swamy, P. & S. Arora (1972) "The exact finite sample properties of the estimators of coefficients in the error component regressions models". *Econometrica* 40: 261–275.

Wooldridge, J.M. (2013) *Introductory Econometrics: A Modern Approach*. South-Western Cengage Learning.

Wooldridge, J.M. (2010) *Econometric Analysis of Cross Section and Panel Data*. MIT Press.