Dados em Painel

Seção 2

Vítor Wilher, Mestre em Economia e Cientista de Dados

1 Introdução aos cortes transversais agrupados e à análise econométrica de dados em painel

Esse é o terceiro de uma sequência de três cursos em **econometria básica** ofertada pela Análise Macro. No primeiro deles, de Introdução à Econometria, vimos a estimação de modelos lineares por mínimos quadrados ordinários, com foco em dados de corte transversal (*cross-section*). No segundo, de Séries Temporais, foram vistos modelos univariados e multivariados de séries temporais, com foco particular em projetos de *forecasting*.

Nesse Curso de Dados em Painel usando o R, estaremos interessados em conjuntos de dados que apresentam características tanto de corte transversal quanto de séries temporais, classificados na literatura como *cortes transversais agrupados* e *dados em painel*. No primeiro caso, veremos *agrupamentos independentes de cortes transversais*, onde amostras coletadas de forma independente em diferentes períodos do tempo são agrupadas, com o objetivo de, por exemplo, avaliar o efeito de um determinado evento ou decisão de política. Esse é o assunto das seções 3 e 4 do nosso curso.

No segundo caso, veremos *dados em painel ou longitudinais*, onde procura-se acompanhar a mesma amostra de indivíduos, famílias, empresas, países, etc, ao longo do tempo. As seções 5, 6 e 7 cobrem os métodos simples de dados em painel, enquanto as seções seguintes dão conta de modelos mais interessantes sob o ponto de vista prático.[1]

A **econometria de dados em painel**, por suposto, é um campo em pleno desenvolvimento, a despeito do ainda escasso material em português sobre o assunto. Ademais, ainda que conjuntos de dados em painel sejam tipicamente coletados no nível microeconômico, tem se tornado bastante comum agrupar séries temporais individuais de um número de países ou indústrias de forma a analisá-las simultaneamente. A disponibilidade desse tipo de conjunto de dados, claro, tem permitido aos economistas e outros profissionais tirarem conclusões mais interessantes do que aquelas proporcionadas por uma amostra única de corte transversal ou uma única série temporal.

Uma importante vantagem de dados em painel, aliás, comparado a séries temporais e cortes transversais, é que eles permitem a identificação de certos parâmetros ou questionamentos, sem a necessidade de serem feitas premissas restritivas. Por exemplo, dados em painel tornam possível analisar mudanças *em termos individuais*. Considere uma situação onde o nível médio de consumo tenha se elevado em 2%, de um ano

[1] Essa divisão está baseada em [@wooldridge]. Sobre isso, vale notar que livros menos básicos, como [@verbeek], introduzem apenas **dados em painel** propriamente ditos. Dado o amplo uso de [@wooldridge] nas faculdades brasileiras, bem como as dificuldades metodológicas inerentes a se acompanhar o mesmo grupo ao longo do tempo, optamos aqui por fazer essa distinção.

para o outro. A análise de dados em painel pode identificar se essa mudança é resultado, por exemplo, um aumento de 2% para todos os indivíduos ou um aumento de 4% para aproximadamente metade dos indivíduos e nenhuma mudança para a outra metade - ou qualquer outra combinação possível. Em outras palavras, a análise de dados em painel não é apenas adequada para modelar ou explicar os motivos pelos quais unidades individuais se comportam de forma diferente, mas também para modelar por que uma dada unidade se comporta diferente em diferentes pontos do tempo.[2]

De forma a introduzir a notação mais básica da **econometria de dados em painel**, considere que estamos indexando todas as variáveis com i para para o indivíduo $i = 1, ..., N$ e t para o período $t = 1, ..., T$. Nesses termos, o modelo linear de regressão padrão pode ser escrito como

$$y_{it} = \beta_0 + x_{it}^{'}\beta + \varepsilon_{it} \tag{1}$$

onde x_{it} é um vetor de dimensão K contendo as variáveis explicativas, que não contém um termo de intercepto. Esse tipo de modelo impõe que o intercepto β_0 e o coeficiente de inclinação em β sejam idênticos para todos os indivíduos e períodos. O termo de erro em 1 varia ao longo dos indivíduos e tempo, capturando todos os fatores não observados que afetam y_{it}. De forma a estimar esse modelo via MQO, as premissas usuais são requeridas de forma a alcançar ausência de viés, consistência ou eficiência. Por exemplo, se $E\{\varepsilon_{it}\} = 0$ e $E\{x_{it}\varepsilon_{it}\} = 0$, o estimador de MQO será consistente para β_0 e β sob condições fracas de regularidade.[3]

Duas vantagens são bastante fortes em conjuntos de dados em painel. Dado que os mesmos são tipicamente maiores do que conjuntos de cortes transversal e séries temporais, bem como as variáveis explanatórias variam ao longo de duas dimensões (indivíduos e tempo), ao invés de apenas uma, estimadores baseados em dados em painel são frequentemente mais acurados. Mesmo com tamanhos de amostra idênticos, o uso de dados em painel, a propósito, irá frequentemente produzir estimadores mais eficientes do que uma série de cortes transversais independentes.

Uma outra vantagem típica desse tipo de conjunto de dados é que ele reduz problemas de identificação. Ainda que essa vantagem possa ser relativizada, em muitos casos ela envolve a identificação na presença de regressores endógenos ou erros de medida, robustez a omissão de variáveis e a identificação de dinâmicas individuais.

A seguir, ilustramos um problema típico que pode ser tratado através da **econometria de dados em painel**.

2 Mudanças no retorno da educação e a diferença salarial por gênero

Com base no exemplo 13.2 de (Wooldridge 2013), podemos ter uma equação *log(salário-hora)* agrupada ao longo dos anos de 1978 e 1985 tal como

$$\begin{aligned} lwage = \beta_0 + \delta_0 y_{85} + \beta_1 educ + \delta_1(y_{85} \cdot educ) + \beta_2 exper \\ + \beta_3 \frac{exper^2}{100} + \beta_4 union + \beta_5 female + \delta_5(y_{85} \cdot female) + u \end{aligned} \tag{2}$$

onde *union* é uma variável *dummy* igual a 1 se a pessoa for sindicalizada e igual a zero, caso contrário. A variável y_{85} é uma variável *dummy* igual a 1 se a observação corresponde ao ano de 1985 e zero se for de 1978. A variável $female$ também é uma variável *dummy*, igual a 1 se o indivíduo for do gênero feminimo e 0, caso contrário. Ademais, as variáveis *educ* e *exper* representam anos de escolaridade e de experiência laboral,

[2] Ver [@verbeek].

[3] Para dados em painel, propriamente dito, onde observamos de forma repetida os mesmos indivíduos, é irrealista, entretanto, assumir que os erros de diferentes períodos não são correlacionados. Por exemplo, o salário de um indivíduo será afetado por características não observadas que variam muito pouco ao longo do tempo. Como resultado, ao computar rotineiramente os erros-padrão via MQO, baseado na premissa de que os termos de erro são indepententes e identificamente distribuídos, tenderá a ser enganoso em aplicações envolvendo dados em painel. Ademais, MQO será provalvemente ineficiente em relação a um estimador que explora a correlação ao longo do tempo presente em ε_{it}. Esse tipo de problema será visto ao longo das seções mais avançadas do nosso Curso de Dados em Painel usando o R.

respectivamente. Nesses termos, o coeficiente β_1 irá medir o *retorno da educação em 1978* e δ_1 é a *diferença* do retorno da educação em 1985 relativamente a 1978. Da mesma forma, β_5 mede a diferença salarial entre homens e mulheres em 1978 e δ_5 é a mudança da diferença salarial.

Observe, por suposto, que *agrupamos* observações referentes às mesmas variáveis de dois anos distintos, fazendo uso de uma *variável dummy anual*. Feito isso, a equação 2 pode ser facilmente estimada via MQO. Abaixo, utilizamos o **R** para isso.

2.1 Usando o R para estimar *pooled cross-sections*

O arquivo `cps78_85.dta` contém todas as variáveis citadas, mais algumas outras. A extensão *dta* faz referência ao pacote STATA, de modo que precisamos carregar um pacote que lide com esse tipo de arquivo. O código abaixo importa e estima a equação 2 com a função `lm`.

```
library(foreign)
cps = read.dta("http://fmwww.bc.edu/ec-p/data/wooldridge/cps78_85.dta")
reg = lm(lwage~y85*(educ+female)+exper+I((exper^2)/100)+union,
         data=cps)
```

Podemos ver os resultados da estimação da equação 2 na tabela 1, a seguir.

Table 1:

	Dependent variable:
	lwage
y85	0.118
	(0.124)
educ	0.075^{***}
	(0.007)
female	-0.317^{***}
	(0.037)
exper	0.030^{***}
	(0.004)
I((exper^2)/100)	-0.040^{***}
	(0.008)
union	0.202^{***}
	(0.030)
y85:educ	0.018^{**}
	(0.009)
y85:female	0.085^{*}
	(0.051)
Constant	0.459^{***}
	(0.093)
Observations	1,084
R^2	0.426
Adjusted R^2	0.422
Residual Std. Error	0.413 (df = 1075)
F Statistic	99.804^{***} (df = 8; 1075)
Note:	*p<0.1; **p<0.05; ***p<0.01

Observe que identificamos um aumento no retorno da educação entre 1978 e 1985, capturado pelo coeficiente estimado $\hat{\delta}_1$, de 0.018. Ademais, houve uma queda, em termos absolutos, da diferença salarial entre homens e mulheres. Era de -0.317 em 1975, passou para -0.232 em 1985.

Referências

Wooldridge, J. M. 2013. *Introductory Econometrics: A Modern Approach*. Editora Cengage.