Cap. 6 Introdução aos Métodos para Dados de Painel

Artur Silva Lopes, Outubro de 2021 (ed. rev.)

ISEG–ULisboa

com base em Woodridge, J. M. (2016), *Introductory Econometrics*, 6th ed.

1 Introdução à Parte III e ao capítulo

No livro de J. Wooldridge (2016) a matéria a estudar está no cap. 13, o mais básico ou introdutório de 2 capítulos sobre o tema.

Até agora estudámos modelos para dados "puros": seccionais e de séries temporais. Os **dados de painel** têm ambas as dimensões, a seccional e a temporal ou cronológica. Antes destes, estudaremos métodos para **dados seccionais agregados ("pooled") e independentes**.

Estes dados resultam da obtenção de amostras aleatórias de uma população em pontos distintos do tempo (em geral, diferentes anos). É importante que eles sejam de observações independentes, i. e., de amostras aleatórias. Isto assegura que os erros das equações de regressão de observações distintas não estejam correlacionados.

Uma amostra seccional agregada de dados independentes difere de uma única amostra aleatória porque a amostragem da população em pontos distintos do

tempo conduz geralmente a observações que não são identicamente distribuídas. Por exemplo, as distribuições dos salários mudam ao longo do tempo num país. Isto será tratado facilmente permitindo que o termo independente mude com o tempo e, nalguns casos, também os coeficientes de declive.

Já os dados de painel, embora também com as duas dimensões, são diferentes. Para obter estes dados, também chamados longitudinais, é necessário **seguir os mesmos elementos** – indivíduos, famílias, cidades, empresas, países, etc. – **ao longo do tempo**.

Tipicamente, os elementos de um conjunto são seleccionados aleatoriamente e são depois re-entrevistados em diferentes pontos do tempo. Obviamente, não se pode assumir que as observações são independentemente distribuídas ao longo do tempo; por exemplo, a habilidade inata de uma pessoa em 2010 é a mesma em 2011.

Outro exemplo: os factores históricos e culturais que afectam a taxa de criminalidade numa cidade em 1995 também a afectam em 2000, e geralmente não são

observados. Para tratar este problema usaremos métodos especiais e, em particular, a **diferenciação** para remover atributos não observados mas (pelo menos assumidos como) constantes nos elementos da amostra.

	dad. amo.	agre. aleat.	dados de painel
independentes ?	sim		não
identic. dist.?	não		não

2 Agregação temporal de dados seccionais independentes

Há inquéritos a indivíduos, famílias, empresas, etc. que são repetidos no tempo, por vezes todos os anos. Nos EUA: os CPS, *Current Population Survey*. Se em cada período for obtida uma amostra aleatória, a agregação ou junção das amostras aleatórias dá-nos dados seccionais agregados e independentes.

Qual é a utilidade da agregação? Ao aumentar a dimensão da amostra, em princípio:

- aumenta-se a precisão dos estimadores (i.e., a sua eficiência);
- aumenta-se a potência dos testes.

Importante: há uma hipótese implícita de estabilidade na relação entre a variável dependente e as explicativas, isto é, que a relação se mantém constante no tempo.

Os problemas estatísticos são escassos: para permitir que a população possa ter distribuições diferentes em diferentes períodos de tempo, permite-se que os termos independentes sejam diferentes. Isto consegue-se facilmente com *dummies* para todos os anos excepto para o ano base, que é normalmente o mais antigo.

Exemplo: estudo da evolução da fertilidade feminina ao longo do tempo. Os dados são de inquéritos de 1972 a 1984, para os anos pares.

Variável dependente: *kids* – número de crianças nascidas de cada mulher. Questão: depois de controlados outros factores, qual é a evolução da fertilidade ao longo do tempo?

Variáveis de controlo: (anos de) educação, idade, raça, região de residência aos 16 anos e ambiente de vida aos 16 (quinta, outro rural, cidade, ou pequena cidade).

$$\widehat{kids} = \underset{(0.018)}{-0.128}\,educ + \underset{(0.138)}{0.532}\,age - \underset{(0.0016)}{0.0058}\,age^2 + \underset{(0.174)}{1.076}\,black + \underset{(0.133)}{0.217}\,east \ldots$$
$$\underset{(0.147)}{-0.053}\,farm + \ldots + \underset{(0.174)}{0.268}\,\mathbf{y74} - \underset{(0.179)}{0.097}\,\mathbf{y76} - \underset{(0.182)}{0.069}\,\mathbf{y78} - \underset{(0.180)}{0.071}\,\mathbf{y80}$$
$$\underset{(0.172)}{-0.522}\,\mathbf{y82} - \underset{(0.175)}{0.545}\,\mathbf{y84} - \underset{(3.052)}{7.742}, \quad n = 1129, \quad R^2 = 0.1295$$

O ano base é, portanto, o de 1972. Os coeficientes estimados das *dummies* mostram uma queda grande na fertilidade na primeira metade da década de 80.

Por ex., o coeficiente estimado de $y82$, -0.522, significa que mantendo constantes ou fixos os restantes factores explicativos do modelo, se estima que em 1982 as mulheres tiveram em média menos 0.522 filhos (ou $1/2$ aproximadamente) que em 1972.

Note-se que, como a educação é um factor que está controlado, esta redução é independente da redução de fertilidade usualmente associada ao aumento da escolaridade ao longo do tempo.

Estima-se que, em média, com os restantes factores do modelo constantes, 100 mulheres com licenciatura (4 anos) terão menos 51 filhos que 100 mulheres só com o ensino secundário: $\widehat{\mathsf{E}(\Delta kids|\Delta educ} = 4) = -0.128 \times 4 = -0.512$.

Dado que as *dummies* de 82 e de 84 são individualmente muito significativas, o conjunto das *dummies* é muito significativo: a F de significância das seis é 5.84, com valor$-p = 0.000$.

Note-se que este modelo pode ser afectado por um tipo de heteroscedasticidade não usual em modelos seccionais: a de a variância variar com o tempo (através das *dummies* de ano); um remédio para esse problema consistirá em empregar os erros-padrão robustos à heteroscedasticidade.

Exemplo: não veremos em detalhe; tem *dummies* de interacção com variáveis quantitativas (ver pp. 405-6). É um modelo salarial para os anos de 1978 e de 1985 e pretendemos medir os efeitos sobre os **salários reais**, ou seja, independentemente da inflação.

$$\log(wage) = \beta_0 + \delta_0 y85 + \beta_1 educ + \delta_1 y85.educ + \beta_2 exper + \ldots + \delta_5 y85.female + u.$$

Mas não é necessário preocuparmo-nos com a deflação da variável dependente $wage$ se esta estiver logaritmizada e no caso de o modelo ter termo independente. De facto, para deflacionar faz-se

$$\log\left(\frac{wage_i}{P85}\right) = \log(wage_i) - \log(P85),$$

com P85 o deflactor para passar os salários para preços constantes de 1978: $P85 = \frac{IPC_{85}}{IPC_{78}} = \frac{107.6}{65.2} \approx 1.65$.

Ora, como P85 é igual para todos os indivíduos, não muda com a observação, é simplesmente absorvido pelo termo independente. Portanto, não precisamos fazer a passagem para preços constantes; podemos trabalhar com preços correntes.

Já não seria assim se a dependente fosse $wage$; neste caso seria necessário deflacionar para obter os salários reais.

Teste de Chow para mudança de estrutura ao longo do tempo

Já vimos que o teste de Chow serve para analisar se uma equação de regressão é igual ou diferente para 2 grupos de elementos. Ele também pode ser aplicado a 2 períodos de tempo distintos:

- a SSR_p é obtida juntando as observações dos 2 períodos de tempo;
- as SSR_1 e SSR_2 são obtidas separadamente para cada um dos períodos de tempo.

Outra forma de proceder consiste em empregar *dummies*: introduzindo *dummies* de interacção e testando a sua significância.

Pode ser calculada uma estatística de Chow para mais que 2 períodos, sejam T, permitindo que os termos independentes mudem:

$$F_{CH} = \frac{[SSR_p - (SSR_1 + SSR_2 + \ldots + SSR_T)]}{SSR_1 + SSR_2 + \ldots + SSR_T} . \frac{n - T(k+1)}{(T-1)k} \sim F_{[(T-1)k,\, n-T(k+1)]}, \text{ sob } H_0,$$

com SSR_p a SSR obtida do modelo estimado com o conjunto de todas as observações e com *dummies* para permitir termos independentes diferentes, e SSR_1 a SSR_T as $SSRs$ para cada um dos (T) períodos de tempo. k é o

número de variáveis explicativas, não incluindo o termo independente nem as *dummies*.

Análise de política com dados seccionais agregados independentes

Estes dados podem ser muito úteis para investigar o impacto de uma alteração de política ou de um acontecimento. O exemplo seguinte é do segundo tipo mas a análise da eficácia de políticas pode ser feita com estes dados.

Exemplo: que efeito teve a localização de um incinerador de lixo nos preços das habitações de North Andover, Massachusetts?

- 1978: circulação de rumores de que seria construído um incinerador;
- 1981: início da construção;
- 1985: início da operação do incinerador.
- Dados: preços de casas vendidas em 1978 (antes) e em 1981 (depois).

Hipótese: os preços das casas localizadas próximo do incinerador baixaram em relação aos das mais afastadas(?).

$rprice$ – preços das casas em termos reais (em p. constantes de 1978);

$$nearinc = \begin{cases} 1, & \text{se a casa é próxima do incinerador,} \\ 0, & \text{no caso contrário} \end{cases}$$

Com dados de 1981, um economista estima

$$\widehat{rprice} = \underset{(3093.0)}{101307.5} - \underset{(5827.71)}{30688.27} nearinc$$

Logo, 101307.5 é o preço médio das casas da amostra afastadas do incinerador (que representa o grupo base ou padrão). E -30688.27 é a diferença média na amostra dos preços entre as casas próximas e as não próximas. Como $|t| = |\frac{-30688.27}{5827.71}| = 5.27$, temos uma rejeição forte da hipótese de igualdade dos preços médios das casas próximas e não próximas. Mas este resultado permite afirmar que terá sido a localização próxima do incinerador a causar a depressão dos preços? Não, obviamente.

Aliás, a regressão estimada com dados de 1978 (antes sequer dos rumores):

$$\widehat{rprice} = \underset{(2653.79)}{82517.23} - \underset{(4744.59)}{18824.37} nearinc$$

Logo, mesmo antes dos rumores, em média, as casas próximas do incinerador já eram 18824.73 USD mais baratas que as afastadas; e a diferença é estatisticamente significativa.

O incinerador terá sido construído em terrenos mais baratos. Como saber, então, se a construção do incinerador deprimiu os preços das casas próximas?

A diferença de preços médios agravou-se de 1978 para 1981: de 18824 passou para 30688. Esta diferença (de diferenças) é

$$\hat{\delta}_1 = -30688.27 - (-18824.37) = -11863.9,$$

e esta é a estimativa do efeito da construção do incinerador sobre os preços das casas próximas. Este estimador, $\hat{\delta}_1$, é usualmente chamado de estimador da **diferença das diferenças**, porque se pode escrever

$$\hat{\delta}_1 = (\overline{rprice}_{81,nr} - \overline{rprice}_{81,fr}) - (\overline{rprice}_{78,nr} - \overline{rprice}_{78,fr}),$$

com $nr =$ *near the incinerator site* e $fr =$ *farther away from the site*.

Ou seja, $\hat{\delta}_1$ é a **diferença no tempo da diferença média dos preços das casas nas duas localizações.**

Mas é estatisticamente significativo? Precisamos do se, que podemos obter da regressão

$$rprice = \beta_0 + \delta_0\, y81 + \beta_1 nearinc + \delta_1\, y81 \times nearinc + v,$$

com os dados agregados dos 2 anos. Os coeficientes representam:

$\beta_0 =$ o preço médio de uma casa afastada do incinerador em 1978,

$\delta_0 =$ a variação dos preços das casas afastadas entre 1978 e 1981,

$\beta_1 =$ o efeito de localização que não se deve à presença do incinerador (apenas à zona),

$\delta_1 =$ a variação (média) dos preços devida ao incinerador (desde que não existam outras razões que tenham justificado esta variação, que é assumido por hipótese).

Modelo estimado:

$$\widehat{rprice} = \underset{}{82517.23} + \underset{}{18790.29}\, y81 - \underset{(4875.32)}{18824.37} nearinc - \underset{(7456.65)}{11863.90}\, y81 \times nearinc,$$

como $t_{\delta_1} = \frac{-11863.90}{7456.65} = -1.59$, só é marginalmente significativo contra uma alternativa unilateral.

Mas uma análise mais aprofundada produz resultados mais fortes. Foram incluídas

mais características das casas porque:

i) os tipos de casas vendidas próximo do incinerador em 1981 podem ser sistematicamente diferentes dos das de 1978; ora, é importante controlar as características das casas;

ii) incluir mais factores explicativos faz reduzir $\hat{\sigma}_u^2$, o que pode reduzir $\widehat{\text{Var}}(\hat{\delta}_1)$.

$$\begin{aligned}\widehat{rprice} = \; & 13807.67 + \underset{(2798.75)}{13928.48}\, y81 - \underset{(4453.42)}{3780.34} nearinc - \underset{(4987.27)}{14177.93}\, y81 \times nearinc \\ & \ldots age + \ldots insts + \ldots land + \ldots area + \ldots rooms + \ldots baths.\end{aligned}$$

onde age representa a idade da casa, $insts$ a distância até à auto-estrada, $land$ a área do terreno, $area$ a área da casa, $rooms$ o número de quartos e $baths$ o número de casas-de-banho. $t_{\delta_1} = -\frac{14177.93}{4987.27} = -2.84$, e o efeito estimado é, agora, muito mais significativo.

Para estimar um efeito em termos percentuais, usa-se um modelo com a dependente logaritmizada:

$$\widehat{\log(price)} = \ldots - 0.132\, y81 \times nearinc + \ldots$$

e o rácio$-t$ da variável relevante é agora de -2.53. Assim, controlados outros factores, estima-se que, em média, as casas construídas perto do incinerador se desvalorizaram em aproximadamente 13.2%.

Generalização

Esta metodologia é empregue frequentemente, sobretudo quando os dados resultam de uma **experiência natural** (ou quase-experiência). Uma experiência natural (EN) ocorre quando algum acontecimento exógeno – muitas vezes uma alteração de política do governo – muda o ambiente em que os elementos da população vivem.

Numa EN, há sempre:

- um **grupo de controlo**, que não é afectado pela mudança;
- um **grupo de tratamento**, que é afectado pela mudança.

Numa verdadeira experiência os 2 grupos são escolhidos explicitamente e de forma aleatória. Numa EN não: os grupos resultam da particular mudança.

Para isolar os efeitos são necessários 2 anos de dados, um antes e outro depois da mudança. Logo, na amostra, temos 4 grupos:

	antes	depois
Controlo	X	X
Tratamento	X	X

C representa o grupo de controlo e **T** o de tratamento.

$$DT_i = \begin{cases} 1, & \text{se } i \text{ pertence ao grupo de tratamento,} \\ 0, & \text{no caso contrário} \end{cases}$$

$$d_2 = \begin{cases} 1, & \text{se a observação é do segundo período,} \\ 0, & \text{no caso contrário} \end{cases}$$

ou seja, d_2 é a *dummy* para as observações do segundo período, após a mudança.

Então, a equação a estimar é

$$y = \beta_0 + \delta_0 d_2 + \beta_1 DT + \delta_1 d_2 \times DT + \underbrace{\text{outros factores}}_{erro}$$

$\hat{\delta}_1$ é o **estimador da diferença das diferenças**:

$$\hat{\delta}_1 = (\bar{y}_{2,T} - \bar{y}_{2,C}) - (\bar{y}_{1,T} - \bar{y}_{1,C}),$$

que mede o efeito estimado da (alteração de) política e onde o primeiro índice é o do ano (1 ou 2) e o segundo do grupo (C ou T). O coeficiente δ_1 é chamado por vezes o **efeito de tratamento médio**, porque mede o efeito do "tratamento" ou da política no resultado médio de y.

	antes	depois	depois−antes
controlo	β_0	$\beta_0 + \delta_0$	δ_0
tratamento	$\beta_0 + \beta_1$	$\beta_0 + \delta_0 + \beta_1 + \delta_1$	$\delta_0 + \delta_1$
tratamento−controlo	β_1	$\beta_1 + \delta_1$	δ_1

δ_1 pode ser estimado:

i) como acima, i.e., cálculo das diferenças das médias entre os 2 grupos em cada período e, em seguida, diferenciam-se os resultados no tempo, ou

ii) $$\hat{\delta}_1 = (\bar{y}_{2,T} - \bar{y}_{1,T}) - (\bar{y}_{2,C} - \bar{y}_{1,C}).$$

Atenção: quando são acrescentadas variáveis explicativas ao modelo, o estimador OLS deixa de ter estas formas simples de diferenças de médias mas a interpretação é semelhante.

3 Análise de dados de painel de 2 períodos

É o caso mais simples de dados de painel: para um conjunto seccional de elementos — indivíduos, empresas, cidades, etc. — temos 2 anos de dados, sejam $t = 1$ e $t = 2$, não necessariamente consecutivos (e.g., 1995 e 2000).

Exemplo de motivação – problema de variáveis omitidas

Apenas com dados de 1987, estimou-se o seguinte modelo seccional (para várias cidades):

$$\widehat{crmrte} = 128.38 - \underset{(3.42)}{4.16}\, unem.$$

Embora a estimativa não seja significativa ($t = \frac{-4.16}{3.42} = -1.22$), constata-se que um aumento da taxa de desemprego parece fazer diminuir a taxa de criminalidade(!). Qual será a explicação? Provavelmente um problema de "variáveis omitidas": distribuição das idades dos indivíduos, níveis educativos, forças policiais, etc.

Possível forma de resolução: incluir $crmrte_{t-1}$ como explicativa. Forma alternativa, com dados de painel: considerar os factores não observados que afectam a variável dependente como sendo de 2 tipos, os constantes e os que variam no tempo.

O **modelo de efeitos fixos** (não observados) com uma única variável explicativa:

$$y_{it} = \beta_0 + \delta_0\, d2_t + \beta_1\, x_{it} + a_i + u_{it}, \quad t = 1, 2$$

com $d2_t = \begin{cases} 1, \text{ quando } t = 2 \\ 0, \text{ quando } t = 1. \end{cases}$ Note-se que $d2_t$ não muda com i.

Assim, o termo independente é β_0 para $t = 1$ e $\beta_0 + \delta_0$ para $t = 2$. Por exemplo, num intervalo de 5 anos pode haver uma mudança significativa, comum a todos os elementos.

A variável a_i captura todos os efeitos não observados mas constantes no tempo que também afectam y_{it}. (Não tem o índice t porque se supõe não variar com o tempo). É chamado de **efeito não observado** ou **efeito fixo** (no tempo).

Por vezes a_i também é chamado de heterogeneidade não observada (ou individual, ou da empresa, ou da cidade, etc.).

u_{it} é o erro idiossincrático ou erro variável no tempo, porque representa factores não observados, individuais, que mudam com o tempo.

Exemplo: $$crmrte_{it} = \beta_0 + \delta_0\, d87_t + \beta_1 unem_{it} + a_i + u_{it},$$

com $t = 1$ em 1982 e $t = 2$ em 1987; a_i é o efeito não observado, fixo, da cidade i, e representa todos os factores que afectam a criminalidade que não mudam (ou mudam pouco) com o tempo:
- a localização geográfica;
- as características da população (idade, sexo, raça);
- factores histórico-culturais;

Estes factores não têm que ser exactamente constantes; basta que sejam aproximadamente constantes no período de 5 anos.

Como estimar β_1, o parâmetro de interesse? Possível solução: **agregar os 2 anos** e usar o OLS (como anteriormente). Tem duas deficiências: a mais importante

é que o estimador OLS deverá ser **inconsistente**. Para o OLS ser consistente é necessário que o efeito não observado, a_i, esteja não correlacionado com x_{it}. De facto, a equação acima pode escrever-se

$$y_{it} = \beta_0 + \delta_0 d2_t + \beta_1 x_{it} + v_{it}, \quad t = 1, 2$$

com $v_{it} = a_i + u_{it}$, o erro compósito. Ora, como

$$\text{Cov}(x_{it}, v_{it}) = \text{Cov}(x_{it}, a_i) + \text{Cov}(x_{it}, u_{it}),$$

mesmo que $\text{Cov}(x_{it}, u_{it}) = 0, \forall t$, esta covariância só é nula se $\text{Cov}(x_{it}, a_i) = 0$. Se a_i e x_{it} estiverem correlacionados, como é **provável, o OLS é enviesado e inconsistente**. Este enviesamento (e inconsistência) é chamado de **enviesamento de heterogeneidade**, mas trata-se de um problema resultante da omissão de uma variável (suposta constante no tempo).

Exemplo: como a amostra do exemplo tem 46 cidades, com 2 anos para cada cidade, n=92 e

$$\widehat{crmrte} = 93.42 + 7.94\, d87 + \underset{(1.188)}{0.427}\, unem$$

e como $t = \frac{0.427}{1.188} = 0.36$, o coeficiente estimado de $unem$, embora positivo, não aparece como estatisticamente significativo. Ou seja, não houve melhoria significativa por usar o "OLS agregado".

O outro problema é o de **autocorrelação** dos erros. Como a_i faz parte do erro de cada observação e deve mudar pouco ou não mudar com o tempo, uma componente importante dos erros dos 2 períodos deve estar muito (positivamente) autocorrelacionada, o que torna os erros-padrão e, em geral, os métodos de inferência inválidos. Ou seja, no exemplo anterior o teste-t não deverá ser válido (e a insignificância deverá ser ainda mais forte).

Porque se recolhem dados de painel? Principal razão: para permitir que o efeito não observado esteja correlacionado com os regressores.

No exemplo: os factores não observados da cidade podem estar correlacionados com a taxa de desemprego.

Os dados de painel permitem uma solução simples: como a_i é constante no tempo, basta diferenciar os dados usando os 2 períodos de observações. Ou seja,

$$\begin{aligned} y_{i2} = &\ (\beta_0 + \delta_0) &&+\beta_1 x_{i2} + a_i + u_{i2}\ (t = 2) \\ y_{i1} = &\ \beta_0 &&+\beta_1 x_{i1} + a_i + u_{i1}\ (t = 1) \end{aligned}$$

Logo, $y_{i2} - y_{i1} = \delta_0 + \beta_1(x_{i2} - x_{i1}) + (u_{i2} - u_{i1})$, ou seja

$$\Delta\, y_i = \delta_0 + \beta_1\, \Delta x_i + \Delta u_i,$$

e o efeito não observado a_i desaparece; foi "diferenciado para longe". Note-se que o termo independente é a variação desse termo de $t = 1$ para $t = 2$.

Esta equação é a de primeira diferença. É só uma **equação seccional mas cada variável é diferenciada**. Pode ser estimada com os métodos para dados seccionais desde que as hipóteses do modelo clássico sejam satisfeitas.

A. A hipótese mais importante é a de exogeneidade: $\text{Cov}(\Delta u_i, \Delta x_i) = 0$. Ela é satisfeita se o erro idiossincrático de cada período, u_{it} não estiver correlacionado com a variável explicativa em ambos os períodos de tempo; ou seja, é a hipótese de exogeneidade estrita de ST: $\text{Cov}(u_{it}, x_{i1}) = \text{Cov}(u_{it}, x_{i2}) = 0$.

Por exemplo, x_{it} não pode ser $y_{i,t-1}$; é uma possiblidade que não pode ser admitida. Mas, ao contrário, do cap. 3, x_{it} pode estar correlacionada com variáveis não observadas desde que estas sejam constantes no tempo. O estimador OLS da equação anterior é chamado de **estimador de primeira diferença**. No exemplo, a hipótese de Cov($\Delta u_i, \Delta unem_i$) = 0 é razoável mas pode falhar.

B. Outra condição crucial é que Δx_i deve ter alguma variação "ao longo de i" (seccional). Isto falha se a variável explicativa não varia no tempo para nenhuma observação seccional ou se muda na mesma magnitude para todas as observações.

No exemplo do crime isto não constitui um problema porque a taxa de desemprego varia no tempo para quase todas as cidades. Mas se x_{it} é uma variável *dummy* para o género dos indivíduos, $\Delta x_i = 0, \forall i$, e não é possível estimar a equação de primeira diferença. Mas isto faz sentido: como permitimos que a_i esteja correlacionado com x_{it}, não podemos esperar separar o efeito de a_i em y_{it} do efeito de qualquer variável que não varia com o tempo.

C. Outra hipótese necessária para usar o OLS é a de homoscedasticidade. Por vezes é razoável assumir que a equação satisfaz todas as hipóteses do modelo

clássico. Se assim for, os estimadores OLS são centrados e a inferência é exacta.

No **exemplo** tem-se: $$\Delta \widehat{crmrte} = \underset{(4.70)}{15.40} + \underset{(0.88)}{2.22}\ \Delta unem$$

que dá uma relação positiva e estatisticamente significativa ($t = (2.22/0.88) = 2.52$) entre as taxas de crime e de desemprego. A diferenciação parece ter sido muito útil. Dado o termo independente, mesmo que $\Delta unem = 0$, prevê-se um aumento na taxa de criminalidade (crimes por 1000 pessoas) de 15.40; isto reflecte o aumento secular nas taxas de criminalidade nos EUA entre 1982 e 1987.

Em suma, a **diferenciação** de 2 anos de dados de painel é um método **eficaz, poderoso**, para resolver o problema de variáveis ou efeitos não observados. Mas tem custos:

i) os dados de painel são mais **difíceis de obter** que os seccionais. Além do inquérito inicial, é necessário acompanhar os elementos (indivíduos, empresas, etc) da amostra. Por vezes é difícil localizar algumas pessoas; algumas empresas poderão falir ou fundir-se com outras. Assim, os DP são mais fáceis de obter para escolas, cidades e países.

ii) a diferenciação (para eliminar a_i) pode **reduzir muito** a variação das explicativas. x_{it} pode variar muito mas Δx_i pode ter pouca variação, o que leva a um grande erro-padrão do estimador do seu coeficiente. Para contrariar isto, usar diferenças mais longas no tempo é, geralmente, melhor que variações ou diferenças anuais. Mas ...

Exemplo: estimação do rendimento da educação com DP. Para cada pessoa i,

$$\log(wage_{it}) = \beta_0 + \delta_0\, d2_t + \beta_1 educ_{it} + a_i + u_{it}, \quad t = 1, 2$$

onde a_i inclui a habilidade não observada que, provavelmente: $\text{Cov}(a_i, educ_{it}) \neq 0$.

Como, por definição, a habilidade inata não muda com o tempo, os métodos para DP parecem particularmente adequados para estimar o referido rendimento. A equação em primeiras diferenças é

$$\Delta \log(wage_i) = \delta_0 + \beta_1 \Delta\, educ_i + \Delta u_i,$$

Problema: os indivíduos são, geralmente, adultos trabalhadores e, para a maioria, a educação não muda com o tempo; ou seja, a precisão do estimador de β_1 tenderá

a ser baixa (se elevado). Para contrariar isto convirá obter uma amostra grande, mas isto será difícil e/ou caro. Assim, uma ideia aparentemente boa — usar uma equação diferenciada para resolver o problema de uma variável não observada omitida — pode não funcionar muito bem na prática.

Com mais variáveis explicativas, o modelo fica:

$$y_{it} = \beta_0 + \delta_0\, d2_t + \beta_1 x_{it1} + \beta_2 x_{it2} + \ldots + \beta_k x_{itk} + a_i + u_{it}, \quad t = 1, 2$$

onde, em x_{itj} o índice i é o de número de observação seccional, t é o índice de período de tempo e j o de variável.

Exemplo: *tradeoff* entre trabalhar e dormir. Num exercício de dados seccionais foram usados dados de 1975 sobre 700 pessoas. Os dados de painel de 1975 e 1981 têm dados de apenas 239 pessoas. O modelo de efeitos fixos não observados para o número de minutos de sono por semana é:

$$\begin{aligned} slpnap_{it} = \; & \beta_0 + \delta_0\, d81_t + \beta_1\, totwrk_{it} + \beta_2\, educ_{it} + \beta_3\, marr_{it} + \\ & \beta_4\, yngkid_{it} + \beta_5\, gdhlth_{it} + a_i + u_{it}, \quad t = 1, 2, \end{aligned}$$

com a_i o efeito fixo individual não observado, $marr$ uma *dummy* de estado civil, $yngkid$ uma *dummy* com valor 1 se há uma criança pequena no agregado, e $gdhlth$ uma *dummy* de *good health*.

É provável que a_i esteja correlacionado com $totwrk_{it}$: os mesmos factores (e alguns podem ser biológicos) que explicam que as pessoas durmam mais ou menos deverão estar correlacionados com o tempo a trabalhar (e.g., algumas pessoas são mais enérgicas).

Importante: **nem o género nem a raça são incluídos porque não variam com o tempo**; são parte de a_i. O principal interesse está na estimação de β_1.

Diferenciando com os 2 anos:

$$\begin{aligned}\Delta slpnap_i = \; & \delta_0 + \beta_1 \Delta\, totwrk_i + \beta_2 \Delta\, educ_i + \beta_3 \Delta\, marr_i + \beta_4 \Delta\, yngkid_i + \\ & \beta_5 \Delta\, gdhlth_i + \Delta u_i,\end{aligned}$$

e o OLS é consistente aqui desde que a variação do erro idiossincrático, Δu_i,

esteja não correlacionada com variações em todas as variáveis explicativas.

$$\widehat{\Delta slpnap}_i = \underset{}{-92.63} - \underset{(0.036)}{0.227}\,\Delta\, totwrk_i - \underset{(48.76)}{0.024}\,\Delta\, educ_i + \underset{(92.86)}{104.21}\Delta\, marr_i + \underset{(87.65)}{94.67}\,\Delta\, yngkid_i + \underset{(76.60)}{87.58}\,\Delta\, gdhlth_i.$$

$\hat{\beta}_1$ representa o *tradeoff* estimado entre trabalhar e dormir: com os outros factores fixos, a mais uma hora de trabalho estima-se que correspondem, em média, menos $0.227 \times 60 = 13.62$ minutos de sono, e é muito significativa ($t_{\beta_1} = 0.227/0.036 = 6.31$). Já as restantes variáveis são insignificantes conjuntamente: a F tem um valor-p de 0.49.

O erro-padrão (se) do coeficiente de $educ$ é muito grande. A justificação já é conhecida: das 239 pessoas, 183 (76.6%) não tem variação de $educ$ neste período de 6 anos (mas $\hat{\beta}_2$ também é muito pequeno). A mesma justificação deve servir para os coeficientes de $marr$ e $gdhlth$. E de $yngkid$?

Organização dos dados

É necessário um cuidado especial para que os dados da mesma unidade ou elemento amostral para os 2 períodos de tempo sejam **ligados**.

Suponha-se que são dados para cidades em 2 anos. O melhor é ter 2 registos para cada cidade, um para cada ano: o 1^0 registo para o 1^0 ano e o 2^0 para o 2^0 ano. E estes 2 registos devem ser adjacentes. Com 100 cidades $\times$ 2 anos $=$ 200 registos; os 2 primeiros para a 1^a cidade, etc.:

n^0 obs.	cidade	ano	homicídios	tx. desem.
1	1	1986	5	8.7
2	1	1990	8	7.2
3	2	1986	2	5.4
4	2	1990	2	5.5
⋮	⋮	⋮	⋮	⋮

Isto facilita a construção de diferenças, que são armazenadas no 2^0 registo para cada cidade e para fazer uma análise seccional agregada, que pode ser comparada com a estimação de diferenciação.(Há outra forma que não veremos.)

Assim, convém que os dados estejam empilhados 1) de acordo com as unidades seccionais, e 2) para cada uma delas.

Convém estruturar o ficheiro de dados para que a identificação de cada observação seja correcta (tanto com a unidade seccional como com o tempo). Assim, no EViews, deve-se escolher `PROC` ou clicar 2× em "`range`" e, em `Workfile structure` seleccionar `balanced panel`; em `date series` deve escolher-se a variável `year` (ano) e em `cross section series` a variável `city` (cidade), no exemplo, i.e., a variável que indexa as observações seccionalmente.

Análise de política com dados de painel de 2 períodos

Os DP são muito úteis para a análise de políticas e, em particular, para a avaliação de programas.

Num primeiro período, é obtida uma amostra de observações de elementos (indivíduos, cidades, etc.). Num período posterior, os elementos do **grupo de tratamento** participam no programa; os do **grupo de controlo** não participam.

Exemplo: avaliação do efeito de um programa de formação profissional sobre a produtividade de empresas transformadoras.

$scrap_{it}$ é a taxa de refugo, o número de produtos em 100 que são refugo devido a defeitos, e $grant_{it}$ é a *dummy* com valor 1 se a empresa i recebeu um subsídio para formação profissional (e 0 no caso contrário).

Com dados para os anos de 1987 e 1988, o modelo é

$$scrap_{it} = \beta_0 + \delta_0\, y88_t + \beta_1\, grant_{it} + a_i + u_{it}, \quad t = 1, 2$$

com $y88_t$ a *dummy* para as observações de 1988 e a_i o efeito de empresa fixo, não observado, que inclui a habilidade média dos empregados, o capital, a capacidade de gestão, etc., factores que são aproximadamente constantes num período de 2 anos.

O problema está em que a_i pode estar sistematicamente correlacionado com o facto de as empresas receberem o subsídio, ou seja, pode haver correlação entre

a produtividade das empresas e a sua obtenção de subsídio. Diferenciando para remover a_i:

$$\Delta\, scrap_i = \delta_0 + \beta_1 \Delta\, grant_i + \Delta u_i.$$

Como em 1987 nenhuma empresa recebeu subsídios, $grant_{i1} = 0, \forall i \Rightarrow \Delta grant_i = grant_{i2} - grant_{1i} = grant_{12}$. Mas é necessário diferenciar todas as variáveis, mesmo as *dummies*.

$$\widehat{\Delta scrap} = -0.564 - \underset{(0.683)}{0.739}\, \Delta\, grant,$$

isto é, estima-se que, em média, ter o subsídio terá reduzido a taxa de refugo em 0.739. Mas não é estatisticamente significativo.

Resultados mais fortes obtêm-se com $\log(scrap)$:

$$\Delta \widehat{\log scrap} = -0.057 - \underset{(0.164)}{0.317}\, \Delta\, grant,$$

com $t_{\beta_1} = -1.93$, que é marginalmente significativa, e com $[\exp(-0.317) - 1] \times 100 = -27.2\%$ é agora a redução média estimada devido aos subsídios.

Generalização

Seja y_{it} a "variável de resultado" e $prog_{it}$ a *dummy* de participação no programa (unitária se i participou no programa em t e 0 no caso contrário). O modelo mais simples de efeitos não observados é

$$y_{it} = \beta_0 + \delta_0\, d2_t + \beta_1\, prog_{it} + a_i + u_{it}.$$

Se a participação no programa só ocorreu no segundo período, o estimador OLS de β_1 da equação diferenciada é simplesmente

$$\hat{\beta}_1 = \overline{\Delta\, y_{treat}} - \overline{\Delta\, y_{control}},$$

ou seja, é a variação média de y nos 2 períodos para os grupos de tratamento e de controlo. É a versão de DP do estimador da diferença das diferenças para 2 amostras seccionais agregadas (p. 16).

Qual a vantagem dos DP? Podemos diferenciar y no tempo para as mesmas unidades seccionais. Isto é importante porque nos permite controlar os efeitos específicos dos elementos da amostra (pessoas, empresas, cidades, etc.).

Mas se existir participação no programa nos 2 períodos, $\hat{\beta}_1$ já não se pode escrever assim, embora seja interpretado da mesma forma: é a variação do valor médio de y devida à participação no programa.

Controlar factores que variam no tempo é fácil: diferenciamos essas variáveis e incluimo-las com $\Delta prog$. Ver o exemplo das leis para eliminar a condução sob efeito de alcóol nos EUA (p. 419 de W).

E com mais que 2 períodos de observações? A diferenciação continua a ser eficaz para resolver os mesmos problemas de variáveis omitidas.

Todavia, a diferenciação deixa de eliminar a perspectiva temporal (só se perde 1 observação) e os erros do modelo diferenciado tenderão a ser autocorrelacionados (a não ser em certos casos especiais, pouco interessantes). O tratamento desta autocorrelação requer procedimentos que não estudámos (ver W seccção 13.5).