ECONOMETRIA

Introdução aos Métodos para Dados de Painel

Teoria e Exemplos

Esmeralda A. Ramalho

Agregação de dados seccionais de diferentes períodos de tempo

Dados seccionais: cada observação $i, i = 1, \ldots N$, representa um indivíduo, empresa, país,…

Séries temporais: cada observação $t, t = 1, \ldots T$, representa um período no tempo

Dados com ambas as dimensões: dados seccionais para diferentes períodos com $i, i = 1, \ldots N$, e $t, t = 1, \ldots T$

Formalização genérica do modelo linear

$$y_{it} = \beta_0 + \beta_1 x_{it1} + \cdots + \beta_k x_{itk} + u_{it}, i = 1, \ldots N , t = 1, \ldots T$$

Agregação de dados seccionais de diferentes períodos de tempo

Tipos de dados

- Dados seccionais agrupados / “pooled”
  - em cada momento no tempo é recolhida uma amostra aleatória, com unidades amostrais tipicamente diferentes, sendo as observações independentes
  - a distribuição idêntica pode não ocorrer, pois pode haver alteração da distribuição ao longo do tempo: dimensão temporal deve ser tida em conta

- Dados de painel / longitudinais
  - os indivíduos observados são sempre os mesmos em todos os períodos, não havendo independência ao longo do tempo
  - os momentos de medida podem não ser adjacentes
  - N*T observações

Dados seccionais agrupados

Motivação

- Aumentar a dimensão da amostra: estimadores mais precisos e inferência mais potente

- Estudar o efeito do tempo: introdução de time-dummies e interações

Nota: pode ocorrer heteroscedasticidade mesmo que a variância do erro não dependa dos regressores, basta que dependa do tempo

Dados seccionais agrupados

Introdução de time-dummies e interações

Em dados anuais, para T anos , toma-se o primeiro ano como referência e consideram-se $(T-1)$ dummies, uma para cada um dos restantes anos

- Efeito dos regressores sobre Y inalterados no tempo:
  - incluem-se essas dummies, que afectarão o intercepto do modelo, sendo que os seus coeficientes informam sobre a alteração de Y relativamente à base, controlados os factores explicativos incluidos no modelo
  - por exemplo, o coeficiente de D2017 informa quanto variou Y entre o ano base e 2017, devido a factores que não os regressores
- Efeitos parciais dos regressores alteram-se no tempo:
  - acrescentam-se também termos de interação
  - pode-se usar o teste Chow para quebra de estrutura

Dados seccionais agrupados

Exemplo: número de filhos de uma mulher nos anos de 2001, 2002 e 2003 em função da educação

Em vez do modelo

$$filhos = \beta_0 + \beta_1 educ + u_{it}, \qquad (1)$$

é mais correcto estimar

$$filhos = \beta_0 + \gamma_1 d2002 + \gamma_2 d2003 + \beta_1 educ + u_{it},$$

assumindo que ao longo deste período o número de filhos que uma mulher decide ter se alterou por motivos que não têm a ver com a educação, ou o modelo

$$filhos = \beta_0 + \gamma_1 d2002 + \gamma_2 d2003 + \beta_1 educ + \delta_1 d2002 educ + \delta_2 d2003 educ + u_{it},$$

permitindo também que o efeito da educação sobre o número de filhos se altere de ano para ano (modelo base do teste Chow, que equivale ao modelo (1) estimado separadamente para cada um dos 3 anos).

Análise de política com dados seccionais agregados

Objectivo: avaliar o impacto de uma política económica

Exemplo: impacto da construção de uma incineradora no preço das casas

- Amostra em dois momentos diferentes (antes e após a implementação) e dois grupos de indivíduos (não afectados – grupo de controle - e afectados – grupo de tratamento)
- Situação designada de "quasi-experiment", sendo uma "natural experiment" quando a alteração decorre de um factor exogeno (mudança de política, …)

Motivação para a existência de um grupo de controle: os preços das casas perto da incineradora podem-se ter alterado por outros motivos que não a sua construção

Análise de política com dados seccionais agregados

Considerar o modelo

$$preco = \beta_0 + \gamma_1 dapos + \beta_1 pertoinc + \delta_1 dapos * pertinc + \mathrm{u},$$

onde δ_1 que mede o impacto da política: se for significativo teve impacto

$\hat{\delta}_1$ é designado **estimador diferenças nas diferenças** ou **efeito médio do tratamento**

	Antes	Após	Após-antes
Cont. (longe)	β_0	$\beta_0 + \gamma_1$	$\gamma_1 = E[p_{apos} - p_{antes}\|cont] \rightarrow 1^{\underline{a}}$ dif
Trat. (perto)	$\beta_0 + \beta_1$	$\beta_0 + \gamma_1 + \beta_1 + \delta_1$	$\gamma_1 + \delta_1 = = E[p_{apos} - p_{antes}\|trat] \rightarrow 1^{\underline{a}}$ dif
Trat. – Cont.	β_1	$\beta_1 + \delta_1$	$\boldsymbol{\delta_1} \rightarrow$ dif in dif

$$\hat{\delta}_1 = (\bar{p}_{após,trat} - \bar{p}_{após,control}) - (\bar{p}_{antes,trat} - \bar{p}_{antes,control})$$

Com outras variáveis explicativas, δ_1 já não terá esta representação simples, mas o esquema de abordagem, pelo modelo e pelo quadro, será o mesmo

Análise de política com dados seccionais agregados

Considere as seguintes equações estimadas para os anos de 1978 e 1981:

$$\widehat{ln(precasa)} = 11.49 - 0.547pertinc + 0.394a81 * pertinc \quad (1)$$

$$\widehat{ln(precasa)} = 11.18 + 0.563a81 - 0.403a81 * pertinc \quad (2)$$

$$\widehat{ln(precasa)} = 11.29 + 0.457a81 - 0.340pertinc - 0.063a81 * pertinc \quad (3)$$

onde $\ln(\text{precasa})$ é o preço das habitações, *pertinc* é uma variável binária que indica se a casa está situada perto de uma incineradora e *a81* é uma variável binária anual para 1981. Compare as estimativas do coeficiente do termo de interacção nas três equações. Porque motivo são elas tão diferentes?

Análise de política com dados seccionais agregados

Eq 1: $\widehat{ln(precasa)} = 11.49 - 0.547pertinc + 0.394a81 * pertinc$

$\widehat{ln(precasa)} =$

	1978	1981	Após-antes
Cont. (longe)	11.49	11.49	**0**
Trat. (perto)	$11.49 - 0.547$	$11.49 - 0.547 + 0.394$	0.394
Trat. – Cont.	-0.547	$-0.547 + 0.394$	$\boldsymbol{\delta_1} \rightarrow 0.394$

Assume-se que, entre 1978 e 1981, o preço das casas

- longe da inceneradora não se altera (quando é natural que tenha existido alguma variação)

Análise de política com dados seccionais agregados

Eq 2: $\widehat{ln(precasa)} = 11.18 + 0.563a81 - 0.403a81 * pertinc$

$\widehat{ln(precasa)} =$

	1978	1981	Após-antes
Cont. (longe)	11.18	$11.18 + 0.563$	0.563
Trat. (perto)	11.18	$11.18 + 0.563 - 0.403$	0.563-0.403
Trat. – Cont.	**0**	-0.403	$\boldsymbol{\delta_1} \rightarrow -0.403$

Assume-se que o preço das casas perto da inceneradora comparativamente com o das casas longe

- É igual em 1978 (quando pode acontecer que a inceneradora tenha sido construida numa zona de casas mais baratas)

Análise de política com dados seccionais agregados

Eq 3: $\widehat{ln(precasa)} = 11.29 + 0.457a81 - 0.340pertinc - 0.063a81 * pertinc$

$\widehat{ln(precasa)} =$

	1978	1981	Após-antes
Cont. (longe)	11.29	$11.29 + 0.457$	0.457
Trat. (perto)	$11.29 - 0.340$	$11.29 + 0.457 - 0.340 - 0.063$	0.457-0.063
Trat. – Cont.	-0.340	$-0.340 - 0.063$	$\boldsymbol{\delta_1} \rightarrow -0.063$

- Entre 1978 e 1981 o preço das casas longe da inceneradora aumentou em $(e^{0.457} - 1)100\% = 57.9\%$ e o das casa perto da inceneradora aumentou apenas $(e^{0.457-0.063} - 1)100\% = 48.3\%$, sendo que a diferença é $(e^{-0.063} - 1)100\% = -6{,}1\ \%$
- Em 1978 o preco das casas na área onde a inceneradora veio a ser construida era inferior em $(e^{0.340} - 1)100\% = 28.8\%$ ao das casas que ficariam longe da inceneradora. Em 1981 a diferença agravou-se em $(e^{-0{,}063} - 1)100\% = 6.1\%$

Dados de painel de dois períodos de tempo

- Um conjunto de unidades amostrais são observadas em dois periodos
- Não há independencia ao longo do tempo
- Conjunto de dados rico:
  - permite analizar efeitos da passagem do tempo
  - evita problemas decorrentes de um tipo expecífico de variáveis omitidas: factores omitidos fixos no tempo durante o periodo em análise

Dados de painel de dois períodos de tempo

Modelo base – Modelo com efeitos individuais

$$y_{it} = \alpha_i + x'_{it}\beta + u_{it} \qquad (i = 1, \dots, N; t = 1, \dots, T)$$

- α_i: efeitos individuais não observáveis e invariantes no tempo / heterogeneidade individual / efeito fixo
- x_{it} - variáveis explicativas:
  - x_{it}: diferem de individuo para individuo e alteram-se no tempo
  - x_i: diferem de individuo para individuo e não se alteram no tempo
  - d_t: time dummies ou t: tendência (também pode incluir t^2, t^3, …)
  - $d_t.x_{it}$: termos de interação
- u_{it}: erro idiossincrático – difere entre indivíduos e no tempo

Invariância no tempo: muitos factores que parecem não ser totalmente invariantes (local de sede/residência, sector de actividade / profissão), não se alteram num período de dois anos, por exemplo

Dados de painel de dois períodos de tempo

O modelo pode ser re-escrito como

$$y_{it} = x'_{it}\beta + (\alpha_i + u_{it}) = x'_{it}\beta + v_{it}$$

- O erro compósito v_{it} tem duas componentes: α_i e u_{it}
- A componente de efeitos individuais α_i pode ou não ser correlacionada com x_{it}

Efeitos fixos:

- α_i e x_{it} correlacionados → x_{it} endógeno (endogeneidade respeitante apenas à parte do erro fixa no tempo)
- Estimadores "pooled" OLS enviesados devido a "heterogeneity bias"

Efeitos aleatórios:

- α_i e x_{it} não são correlacionados → x_{it} exogeno
- Estimadores "pooled" OLS válidos

Dados de painel de dois períodos de tempo

Estimador "pooled" OLS :

- Modelo:

$$y_{it} = \alpha + x'_{it}\beta + \underbrace{(\alpha_i - \alpha + u_{it})}_{v_{it}}$$

- Pressupostos: $E[x_{it}(\alpha_i + u_{it})] = 0$
  - Requer efeitos aleatórios: α_i e x_{it} não correlacionados
  - Requer exogeneidade contemporânea para u_{it} e x_{it}

- Estimação: OLS com estimador de variância de tipo cluster, para incorporar a dependência temporal do erro

Stata

regress $YX_1 \ldots X_k$, vce(cluster *clustvar*)

Dados de painel de dois períodos de tempo

Estimador das primeiras diferenças:

- Permite que x_{it} esteja correlacionado com a componente invariante no tempo, α_i, de v_{it}

- Modelo às diferenças (descarta o α_i):

$$y_{it} - y_{i,t-1} = \left(x_{it} - x_{i,t-1}\right)'\beta + \left(u_{it} - u_{i,t-1}\right) \Leftrightarrow$$
$$\Delta y_{it} = \Delta x_{it}'\beta + \Delta u_{it}$$

- Pressupostos:
  - $E(\Delta x_{it}\Delta u_{it}) = 0$
  - Requer $E(x_{it}u_{it}) = E\left(x_{it}u_{i,t-1}\right) = E\left(x_{it}u_{i,t+1}\right) = 0$ (não requer exogeneidade estrita, mas não permite regressores de tipo $y_{i,t-1}$

Dados de painel de dois períodos de tempo

- Limitação: requer x_{it} com variabilidade no tempo
  - género é invariante: é automaticamente eliminada do modelo
  - experiencia profissional varia de 1 para todos: fica constante em diferença, sendo eliminada do modelo
  - educ tem para muitos individuos variação de 0 no tempo, tendo pouca variabilidade: mantém-se no modelo mas reduz a precisão da estimação
- Se o modelo inicial incluir uma time-dummy (habitual), o modelo às diferenças terá um intercepto, que é o coeficiente da time-dummy:

$$y_{it} = \alpha + \delta d2 + x'_{it}\beta + \alpha_i + u_{it}$$

$y_{i2} = \alpha + \delta + x'_{i2}\beta + \alpha_i + u_{i2}$
$y_{i1} = \alpha + x'_{i1}\beta + \alpha_i + u_{i1}$
$\Delta y_{i2} = \delta + \Delta x'_{i2}\beta + \Delta u_{i2}$

- Estimação: OLS

Stata

regress *D.Y D.*X_1 ... *D.*X_k, vce(cluster *clustvar*)

Dados de painel de dois períodos de tempo

Considere um conjunto de dados de painel para 1980 e 1990, que incluem valores das rendas de casa e outras variáveis para cidades universitárias. Pretende-se verificar se uma presença mais forte de estudantes afecta as rendas. Propõe-se o seguinte modelo:

$$log(renda_{it}) = \beta_0 + \delta a90_t + \beta_1 log(pop_{it}) + \beta_2 log(rendim_{it}) + \beta_3 percest_{it} + \alpha_i + u_{it}$$

onde a90: observação relativa a 1990; renda: valor da renda; pop: população; rendim: rendimento per capita; percest: percentagem da população estudantil na população da cidade (durante o ano escolar).

Dados de painel de dois períodos de tempo

```
. gen lrenda=log(renda)
. gen a90=(ano==90)
. gen lpop=log(pop)
. gen lrendim=log(rendim)
. gen percest= est/pop*100

. regress lrenda a90 lpop lrendim percest
  Source |       SS       df       MS                  Number of obs =     128
---------+------------------------------                F(  4,   123) =  190.92
   Model |  12.1080112     4  3.02700281                Prob > F      =  0.0000
Residual |   1.9501234   123  .015854662                R-squared     =  0.8613
---------+------------------------------                Adj R-squared =  0.8568
   Total |  14.0581346   127  .110693974                Root MSE      =  .12592

------------------------------------------------------------------------------
  lrenda |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
     a90 |   .2622267   .0347632      7.543   0.000       .1934151    .3310384
    lpop |   .0406863   .0225154      1.807   0.073      -.0038815    .0852541
 lrendim |   .5714461   .0530981     10.762   0.000       .4663417    .6765504
 percest |   .0050436   .0010192      4.949   0.000       .0030262     .007061
   _cons |  -.5688069   .5348808     -1.063   0.290      -1.627571    .4899568
------------------------------------------------------------------------------
```

Dados de painel de dois períodos de tempo

```
. regress lrenda a90 lpop lrendim percest, cluster(cidade)

Linear regression                               Number of obs =     128
                                                F(  4,    63) =  629.70
                                                Prob > F      =  0.0000
                                                R-squared     =  0.8613
                                                Root MSE      =  .12592

                                  (Std. Err. adjusted for 64 clusters in cidade)
------------------------------------------------------------------------------
             |               Robust
      lrenda |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         a90 |   .2622267   .0642476     4.08   0.000     .1338382    .3906152
        lpop |   .0406863   .0288226     1.41   0.163     -.016911    .0982836
     lrendim |   .5714461   .1229035     4.65   0.000      .325843    .8170492
     percest |   .0050436   .0015268     3.30   0.002     .0019924    .0080947
       _cons |  -.5688069   1.039856    -0.55   0.586    -2.646793    1.509179
------------------------------------------------------------------------------
```

Dados de painel de dois períodos de tempo

Comentários:

- Significância individual: a 5% as conclusões são iguais com os desvios padrão habituais e os robustos – apenas lpop não é significativo para explicar as rendas

- as rendas aumentaram em média $(e^{0.262} - 1)100\% = 30.0\%$ entre 1980 e 1990 devido a factores que não os controlados pelo modelo

- Admitindo tudo o resto constante, a interpretação dos parâmetros associados a regressores significativos é a seguinte:
  - por um aumento de 1% no rendimento, as rendas aumentam 0.571%
  - por um aumento da percentagem de estudantes num ponto percentual, as rendas aumentam 0.005*100=0.5%

Dados de painel de dois períodos de tempo

```
. gen td=1
. replace td=1 if ano==80
. replace td=2 if ano==90
. xtset cidade td
       panel variable:  cidade (strongly balanced)
        time variable:  td, 1 to 2
                delta:  1 unit
. regress D.lrenda a90 D.lpop D.lrendim D.percest, noconst
      Source |       SS           df       MS      Number of obs   =        64
-------------+----------------------------------   F(4, 60)        =    624.15
       Model |   20.279024         4    5.069756   Prob > F        =    0.0000
    Residual |  .487362198        60  .008122703   R-squared       =    0.9765
-------------+----------------------------------   Adj R-squared   =    0.9750
       Total |  20.7663862        64  .324474784   Root MSE        =    .09013
------------------------------------------------------------------------------
    D.lrenda |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         a90 |   .3855214   .0368245    10.47   0.000     .3118615    .4591813
        lpop |
          D1.|   .0722456   .0883426     0.82   0.417     -.104466    .2489571
     lrendim |
          D1.|   .3099605   .0664771     4.66   0.000     .1769865    .4429346
     percest |
          D1.|   .0112033   .0041319     2.71   0.009     .0029382    .0194684
------------------------------------------------------------------------------
```

A inclusão de a90 requer a ausência de constante: equivale ao modelo sem a90 com constante

Dados de painel de dois períodos de tempo

```
. regress D.lrenda a90 D.lpop D.lrendim D.percest, noconst vce(cluster cidade)
Linear regression                               Number of obs     =         64
                                                F(4, 63)          =     691.38
                                                Prob > F          =     0.0000
                                                R-squared         =     0.9765
                                                Root MSE          =     .09013

                                   (Std. Err. adjusted for 64 clusters in cidade)
------------------------------------------------------------------------------
             |               Robust
    D.lrenda |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         a90 |   .3855214   .0487186     7.91   0.000      .288165    .4828778

        lpop |
          D1.|   .0722456   .0696796     1.04   0.304     -.066998    .2114891
             |
     lrendim |
          D1.|   .3099605   .0893099     3.47   0.001     .1314889    .4884322
             |
     percest |
          D1.|   .0112033    .002936     3.82   0.000     .0053362    .0170704
------------------------------------------------------------------------------

.
```

Análise de política com dados de painel de dois períodos de tempo

Comando alternativo:

```
. regress D.(lrenda a90 lpop lrendim percest), noconst
      Source |       SS           df       MS      Number of obs   =        64
-------------+----------------------------------   F(4, 60)        =    624.15
       Model |   20.279024         4    5.069756   Prob > F        =    0.0000
    Residual |  .487362198        60  .008122703   R-squared       =    0.9765
-------------+----------------------------------   Adj R-squared   =    0.9750
       Total |  20.7663862        64  .324474784   Root MSE        =    .09013
------------------------------------------------------------------------------
    D.lrenda |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         a90 |
          D1.|   .3855214   .0368245    10.47   0.000     .3118615    .4591813
             |
        lpop |
          D1.|   .0722456   .0883426     0.82   0.417     -.104466    .2489571
             |
     lrendim |
          D1.|   .3099605   .0664771     4.66   0.000     .1769865    .4429346
             |
     percest |
          D1.|   .0112033   .0041319     2.71   0.009     .0029382    .0194684
------------------------------------------------------------------------------
```

Análise de política com dados de painel de dois períodos de tempo

Considera-se uma amostra em condições similares da "quasi experiment": os indivíduos são observados duas vezes, uma antes do programa de intervenção e outra depois, e devem ser de dois tipos, tratados e controls

Modelo geral:

$$y_{it} = \alpha + \delta d2 + \beta prog_{it} + \alpha_i + u_{it}$$

onde $prog = 1$ se recebeu formação

Modelo às diferenças:

$$\Delta y_{it} = \delta + \beta prog_{it} + \Delta u_{it}$$

Efeito de interesse: $\beta = \overline{\Delta y}_{trat} - \overline{\Delta y}_{cont}$

Análise de política com dados de painel de dois períodos de tempo

Exemplo: Wooldridge

Pretende-se apurar se a taxa de produtos defeituosos (% de produtos defeituosos inutilizados na produção total da empresa), *scrap*, se altera em função da participação num programa de formação, (*Grant=1* se participa), ocorrido em 1988. Os dados são de painel, para 1987 e 1988 e contêm indivíduos que receberam e indivíduos que não receberam formação profissional

Modelo estimado

$$\Delta ln(\widehat{scrap}) = \underset{}{-0.057} \underset{(0.097)}{} - \underset{(0.164)}{0.317}\ \Delta grant, n = 54, R^2 = 0.067$$

- A formação reduziu a taxa de rejeição em $(e^{0.317} - 1)100\% = 27.2\%$
- A taxa de rejeição reduziu-se em $(e^{0.057} - 1)100\% = 5.9\%$ por factores que não a formação profissional

Diferenciação com mais do que dois períodos de tempo

Extensão do método das primeiras diferenças para mais períodos

A diferença faz-se entre períodos adjacentes: por exemplo, para T=3, obtêm-se duas equações, uma com as diferenças de t=2 para t=1 e outra com as diferenças de t=3 para t=2:

$$\Delta y_{it} = \Delta x'_{it}\beta + \Delta u_{it}, t = 2, 3,$$

Sendo que o número total de observações será N(T-1)

Neste caso, havendo inicialmente time-dummies, o que se faz é não as diferenciar e acrescenta-las directamente no modelo às diferenças, assim como um intercepto

$$\Delta y_{it} = \alpha + \delta_1 d2 + \delta_2 d3 + \Delta x'_{it}\beta + \Delta u_{it}, t = 2, 3$$

- Estimação: OLS

Stata

regress *D.(Y* X_1 ... X_k*)*, vce(cluster *clustvar*)