Dados em painel

Wooldridge, capítulo 13

Painel verdadeiro x *cross sections* empilhadas

- Terminologia dados em painel é frequentemente utilizada de forma livre para quaisquer conjuntos de dados que têm tanto uma dimensão *cross-section* quanto uma dimensão temporal
- Mais precisamente corresponde somente a dados que seguem as mesmas unidades de *cross-section* ao longo do tempo
- De outra forma são somente *cross sections* empilhadas

Pooling, *cross sections* empilhadas ou agrupamento independente de cortes transversais

- Amostragem aleatória de uma população em períodos de tempo diferentes provavelmente levará a observações que não são identicamente distribuídas
- Fácil de lidar na prática: num modelo de regressão múltipla fazer com que o intercepto e, em alguns casos, a inclinação mudem ao longo do tempo
- Pode ser usado para avaliar alterações de política

Dados de painel ou dados longitudinais

- Os mesmos indivíduos, famílias, empresas, cidades, estados, países são acompanhados ao longo do tempo
- Não é possível supor que as observações sejam independentemente distribuídas ao longo do tempo
- Métodos especiais foram desenvolvidos para analisar dados em painel
- Avaliação de alterações de política

Pooling

- $y_{it} = \beta_0 + \beta_1 x_{it1} + \ldots \beta_k x_{itk} + u_{it}$
- Podemos querer empilhar *cross sections* somente para ter amostras maiores
- Podemos querer empilhar *cross sections* para investigar o efeito do tempo
- Podemos querer empilhar *cross sections* para investigar se relações mudaram ao longo do tempo

Exemplo 1: Fertilidade das mulheres ao longo do tempo

- Após termos controlado todos os outros fatores observados, o que aconteceu com a taxa de fertilidade ao longo do tempo nos EUA?
- Controles: anos de educação, idade, raça, região do país em que as mulheres residiam quando tinham 16 anos e ambiente em que viviam quando tinham essa mesma idade

TABLE 13.1 Determinants of Women's Fertility

Dependent Variable: *kids*

Independent Variables	Coefficients	Standard Errors
educ	−.128	.018
age	.532	.138
age^2	−.0058	.0016
black	1.076	.174
east	.217	.133
northcen	.363	.121
west	.198	.167
farm	−.053	.147
othrural	−.163	.175
town	.084	.124
smcity	.212	.160
y74	.268	.173
y76	−.097	.179
y78	−.069	.182
y80	−.071	.183
y82	−.522	.172
y84	−.545	.175
constant	−7.742	3.052
$n = 1{,}129$ $R^2 = .1295$ $\bar{R}^2 = .1162$		

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

$$\log(wage) = \beta_0 + \delta_0 y85 + \beta_1 educ + \delta_1 y85 \cdot educ + \beta_2 exper + \beta_3 exper^2 + \beta_4 union + \beta_5 female + \delta_5 y85 \cdot female + u,$$

$$\begin{aligned}\log(wage) = {} & .459 + .118\, y85 + .0747\, educ + .0185\, y85 \cdot educ \\ & \quad (.093) \quad (.124) \quad\quad (.0067) \quad\quad (.0094) \\ & + .0296\, exper - .00040\, exper^2 + .202\, union \\ & \quad (.0036) \quad\quad (.00008) \quad\quad (.030) \\ & - .317\, female + .085\, y85 \cdot female \\ & \quad (.037) \quad\quad (.051) \\ n = 1{,}084, & \; R^2 = .426, \; \bar{R}^2 = .422.\end{aligned}$$

Dados em painel de dois períodos

- Dados em painel podem ser usados para lidar com alguns tipos de viés de variável omitida
- Antes (capítulo 9): adição de um valor defasado de y para controlar fatores omitidos
- Variável dependente defasada fornece uma maneira simples de explicar fatores históricos que causam diferenças correntes na variável dependente que são difíceis de explicar de outras maneiras
- Equação de criminalidade: inclusão da taxa de criminalidade do ano anterior para controlar o fato de que cidades distintas têm taxas de criminalidade historicamente diferentes

- Modo alternativo de usar dados em painel: separar os fatores não observados que afetam a variável dependente em dois tipos, os que são constantes e os que variam ao longo do tempo

- Supor que o modelo populacional é

$$y_{it} = \beta_0 + \delta_0 d2_t + \beta_1 x_{it1} + \ldots + \beta_k x_{itk} + a_i + u_{it} \quad (1)$$

- A variável a_i capta todos os fatores não observados, constantes no tempo
- a_i é chamado de efeito não observado, efeito fixo ou heterogeneidade não observada (heterogeneidade de indivíduo, heterogeneidade de município, etc…)

- Modelo em (1): modelo de efeitos fixos ou modelo de efeitos não observados
- u_{it} é chamado erro idiossincrático ou erro de variação temporal porque ele representa fatores não observados que mudam ao longo do tempo e afetam y_{it}

Como devemos fazer para estimar β_1 a partir de dois anos de dados de painel?

- MQO agrupado?
- Necessário assumir que a_i é não correlacionado com x_{it}
- $y_{it} = \beta_0 + \delta_0 d2_t + \beta_1 x_{it1} + \ldots + \beta_k x_{itk} + v_{it}$, onde $v_{it} = a_i + u_{it}$
- Se a_i *é* correlacionado com os x's, MQO será viesado e inconsistente porque a_i é parte do termo de erro
- Viés de heterogeneidade, mas na realidade é apenas um viés causado pela omissão de uma variável constante no tempo

Primeiras diferenças (*First-differences*)

- Podemos simplesmente subtrair um período do outro , para obter (2) $\Delta y_i = \delta_0 + \beta_1 \Delta x_{i1} + \ldots + \beta_k \Delta x_{ik} + \Delta u_i$
- Eq. (2): equação de primeiras diferenças
- Estimador MQO de β_1 de (2): estimador de primeiras diferenças
- Hipótese mais importante é que Δu_i seja não correlacionado com Δx_i
- Esta hipótese será mantida se o erro idiossincrático em cada tempo t, u_{it}, for não correlacionado com a variável explicativa em ambos os períodos de tempo

- A adição de muitas variáveis explicativas não causa dificuldades
- É possível usar o método quando se tem vários períodos

Diferenciação com vários períodos

- Simplesmente diferenciar períodos adjacentes
- Se são 3 períodos, então subtrair o período 1 do período 2, o período 2 do período 3 e ter duas observações por indivíduo
- Estimar por MQO, assumindo que os Δu_{it} são não correlacionados ao longo do tempo

Hipóteses estimador de primeiras diferenças (PD)

- Hipótese PD.1

Para cada i o modelo é $y_{it} = \beta_0 + \beta_1 x_{it1} + \ldots \beta_k x_{itk} + a_i + u_{it}$

- Hipótese PD.2 Amostragem aleatória do corte transversal
- Hipótese PD.3 $E(u_{it}|\boldsymbol{X_i}, a_i) = 0$

Implicação importante da hipótese PD.3 é que $E(\Delta u_{it}|\boldsymbol{X_i}) = 0$, t=2,…,T

- Hipótese PD.4 Cada variável explicativa muda ao longo do tempo (para pelo menos algum i) e não existem relações colineares perfeitas entre as variáveis explicativas

Sob essas 4 hipóteses o estimador de primeiras diferenças é não viesado e consistente (com um T fixo e quando $N \rightarrow \infty$

Diferenciação e variação

- $\log(salário_{it}) = \beta_0 + \delta_0 d2_t + \beta_1 educ_{it} + a_i + u_{it}$
- $\Delta\log(salário_i) = \delta_0 + \beta_1 \Delta educ_i + \Delta u_i$

- Problema: estamos interessados nos adultos que trabalham e para a maioria dos indivíduos empregados a educação não muda ao longo do tempo. Se apenas uma pequena fração da nossa amostra tiver $\Delta educ_i$ diferente de zero será difícil obter um estimador preciso de β_1, a menos que tenhamos uma amostra de tamanho bastante grande

- Hipótese PD.5 $Var(\Delta u_i|\boldsymbol{X}_i)=\sigma^2$, t=2,...,T
- Hipótese PD.6 $Cov(\Delta u_{it,}\Delta u_{is}|\boldsymbol{X_i}) = 0, t \neq s$ (as diferenças nos erros idiossincráticos são não correlacionadas)

Sob as hipóteses PD.1 a PD.6 , o estimador PD é o melhor estimador linear não viesado

- Hipótese PD.7 Condicional a $\boldsymbol{X_i}$, os Δu_i são variáveis aleatórias normais independente e identicamente distribuídas

Com a hipótese PD.7, os estimadores PD são normalmente distribuídos e as estatísticas t e F do MQO agrupado das diferenças têm distribuições t e F exatas

Análise de decisões governamentais

Estrutura mais simples: uma amostra de indivíduos, firmas, municípios, etc... no primeiro período de tempo. Algumas unidades farão parte de um programa específico em um período de tempo posterior (tratamento); as que não farão parte estão no grupo de controle

Exemplo

- Avaliar o efeito de um programa de treinamento de pessoal sobre a produtividade dos trabalhadores de firmas manufatureiras.
- $ref_{it} = \beta_0 + \delta_0 a88_t + \beta_1 sub_{it} + a_i + u_{it}$ t=1987,1988
- Onde ref_{it} é a taxa de refugo dos produtos da firma i durante o ano t (número de itens que devem ser rejeitados devido a defeitos), $subs_{it}$ é uma variável *dummy* que assume valor igual a 1 se a firma i no ano t recebeu subsídio de treinamento de pessoal e *a88* é uma variável *dummy* igual a 1 para o ano de 1988

- Efeito fixo contém fatores como a aptidão média dos trabalhadores, capital e capacidade gerencial que são praticamente constantes ao longo de um período de dois anos
- Correlação entre o efeito fixo e programa de treinamento: os administradores do programa poderiam dar prioridade às firmas cujos trabalhadores fossem menos especializados ou poderiam conceder subsídio a empregadores com trabalhadores mais produtivos para fazer com que o programa de treinamento pareça eficiente

- $\Delta ref_i = \delta_0 + \beta_1 \Delta subs_i + \Delta u_i$
- Regredimos as mudanças na taxa de refugo nas mudanças do indicador de subsídio
- Como nenhuma firma recebeu subsídio em 1987, as mudanças do indicador de subsídio indicam se a firma recebeu subsídio em 1988

Efeitos da legislação a respeito da condução de veículos sob embriaguez sobre as fatalidades no trânsito

Exemplo 13.7

Leis de recipientes abertos: consideram ilegal os passageiros de um veículo ter em seu poder recipientes abertos de bebidas alcoólicas

Leis administrativas: autorizam a justiça suspender a carteira de habilitação do motorista preso por dirigir embriagado, mesmo antes de seu julgamento

- Uso de um único corte transversal de estados: não vai funcionar pois a adoção das leis possivelmente está relacionada com a média de fatalidades no trânsito ocorridas nos anos recentes
- Usar dados de painel de um período de tempo em que alguns estados tenham adotado novas leis

TRAFFIC1: dados de 1985 e 1990 de todos os 50 estados americanos e do Distrito de Columbia

1985: 19 estados tinham leis de recipientes abertos, enquanto 22 estados tinhas tais leis em 1990

1985: 21 estados tinham leis administrativas
1990: 29 estados tinham leis administrativas

MQO após a primeira diferenciação

$$\Delta txmort = -0{,}497 - 0{,}420\Delta abertos - 0{,}151\Delta admin$$
$$(0{,}0502) \quad (0{,}206) \qquad\qquad (0{,}117)$$

- Adoção da lei de recipientes reduziu a taxa de fatalidade no trânsito em 0,42. Efeito considerável dada que a taxa média de mortalidade em 1985 era de 2,7. Estimativa significante ao nível de 5%
- Lei administrativa tem um efeito menor e é estatisticamente insignificante
- Intercepto indica que as fatalidades no trânsito caíram substancialmente em todos os estados ao longo do período de cinco anos independente das leis

Obs: i) outras leis também afetam as fatalidades no trânsito, como as leis sobre o uso de cinto de segurança, as leis sobre o uso de capacetes por motociclistas e as leis sobre limites máximos de velocidade; ii) controlar por idade e gênero; iii) avaliar a influência que organizações como *Mothers Against Drunk Driving* têm em cada estado

- Extensão para o controle de fatores que variam ao longo do tempo pode ser feita tranquilamente
- Basta diferenciar tais variáveis e incluí-las na regressão junto com $\Delta prog$