GOVERNO DO ESTADO DO CEARÁ
SECRETARIA DO PLANEJAMENTO E GESTÃO - SEPLAG
INSTITUTO DE PESQUISA E ESTRATÉGIA ECONÔMICA DO CEARÁ - IPECE

NOTA TÉCNICA
Nº 37

UMA BREVE DISCUSSÃO SOBRE OS MODELOS COM DADOS EM PAINEL

André Oliveira Ferreira Loureiro[1]
Leandro Oliveira Costa[2]

Fortaleza – CE

Março – 2009

[1] Mestre em Economia – CAEN/UFC. Analista de Políticas Públicas do IPECE.
[2] Doutorando em Economia – CAEN/UFC. Analista de Políticas Públicas do IPECE.

Notas Técnicas do Instituto de Pesquisa e Estratégia Econômica do Ceará (IPECE)

A Série Notas Técnicas do Instituto de Pesquisa e Estratégia Econômica do Ceará (IPECE) tem como objetivo a divulgação de metodologias e trabalhos elaborados pelos servidores do órgão, que possam contribuir para a discussão de diversos temas de interesse do Estado do Ceará.

Instituto de Pesquisa e Estratégia Econômica do Ceará (IPECE)
End.: Centro Administrativo do Estado Governador Virgílio Távora
Av. General Afonso Albuquerque Lima, S/N – Edifício SEPLAN – 2º andar
60830-120 – Fortaleza-CE
Telefones: (85) 3101-3521 / 3101-3496
Fax: (85) 3101-3500

www.ipece.ce.gov.br
ipece@ipece.ce.gov.br

SUMÁRIO

Apresentação

Em função de vários trabalhos do IPECE utilizarem a metodologia de dados em painel na realização de avaliações sobre diversos aspectos socioeconômicos cearenses[1], o presente trabalho busca ampliar a acessibilidade dos nossos trabalhos a essa metodologia amplamente utilizada nos artigos científicos das ciências sociais aplicadas e, principalmente, na economia. Dessa forma, a presente nota técnica apresenta um breve resumo sobre a metodologia econométrica utilizada no contexto de Dados em Painel, bem como um breve guia de como aplicá-la utilizando o software Stata[2].

Dados em Painel ou dados longitudinais são caracterizados por possuírem observações em duas dimensões que em geral são o tempo e o espaço. Este tipo de dados contém informações que possibilitam uma melhor investigação sobre a dinâmica das mudanças nas variáveis, tornando possível considerar o efeito das variáveis não-observadas. Outra vantagem é a melhoria na inferência dos parâmetros estudados, pois eles propiciam mais graus de liberdade e maior variabilidade na amostra em comparação com dados em *cross-section* ou em séries temporais, o que refina a eficiência dos estimadores econométricos. Hsiao (2006) expõe um maior detalhamento das vantagens propiciadas pela análise de Dados em Painel.

Após uma introdução que discute o modelo de dados em painel, é apresentado o conceito de heterogeneidade não-observada. São discutidos os principais modelos utilizados neste contexto: Efeitos Fixos, Primeiras Diferenças e Efeitos Aleatórios. Finalmente, é discutido o caso em que a hipótese de Exogeneidade Estrita não é valida e a utilização de variáveis instrumentais.

[1] Entre os trabalhos do IPECE que se utilizam da metodologia de dados em painel, podemos citar os artigos de Irffi, Oliveira & Barbosa (2008), Irffi et al. (2008) e Loureiro (2008).
[2] A escolha do software STATA 10.0 se deve a sua ampla utilização nas ciências sociais aplicadas.

1. Pressupostos relacionados à metodologia de Dados em Painel

Um modelo de regressão com dados em painel, com n observações em T períodos e *K* variáveis, pode ser representado da seguinte forma:

$$y_{it} = \boldsymbol{x}_{it}\boldsymbol{\beta} + \varepsilon_{it}, \quad \text{i = 1, 2, ... , n; t = 1, 2, ... , T} \quad (1)$$

onde y_{it} é a variável dependente, $\boldsymbol{x}_{it}$ é um vetor $1 \times K$ contendo as variáveis explicativas, $\boldsymbol{\beta}$ é um vetor $K \times 1$ de parâmetros a serem estimados e ε_{it} são os erros aleatórios. Os sub-índices i e t denotam a unidade observacional e o período de cada variável, respectivamente. Desta forma, em uma base de dados com dados em painel, o número total de observações corresponde a $n \times T$.

Se o modelo seguir todas as hipóteses clássicas de regressão[3], pode-se estimá-lo por Mínimos Quadrados Ordinários – MQO, obtendo as estimativas desejadas. As principais se referem ao erro ε, que se supõe homoscedástico e não-correlacionado no tempo e no espaço. Neste caso, ter-se-ia uma matriz de variância V da seguinte forma: $V = (\sigma^2 I_n) \otimes I_T$, onde σ^2 é a variância da regressão, $\otimes$ denota o produto de kronecker e I_n e I_T denotam matrizes identidade de ordem n e T, respectivamente. Assim, V é uma matriz de ordem $nT \times nT$. No caso de dados em painel, os problemas de heteroscedasticidade e autocorrelação podem ocorrer tanto dentro dos grupos, quanto entre os grupos, ou as duas situações simultaneamente.

O problema de heteroscedasticidade, se detectado, torna necessária a utilização do método de Mínimos Quadrados Generalizados – MQG. Segundo Greene (2003), se fosse utilizado o estimador de Mínimos Quadrados Ordinários – MQO, não levando em consideração a não-homoscedasticidade dos distúrbios, as estimativas ainda seriam não-viesadas e consistentes, mas não seriam mais eficientes. Desta forma, os testes de significância das estimativas seriam enviesados se MQO fosse utilizado. O mesmo argumento é válido na presença de autocorrelação dos erros.

[3] Para maiores detalhes dessas hipóteses, ver Greene (2003) e Davidson & MacKinnon (2004).

Se algum desses dois problemas, ou ambos, estiverem presentes no modelo, a matriz de variância do modelo deixa de ser diagonal e passa a ser da seguinte forma: $V = (\sigma^2 \Sigma) \otimes \Omega$, onde Σ e Ω representam matrizes cujos elementos podem assumir quaisquer valores.

Em função de não se conhecer a matriz de variância V do modelo, não é possível realizar estimativas dos parâmetros por MQG diretamente, sendo então necessário estimar Σ e Ω. Mas a estimação de todos os parâmetros dessas matrizes sem estabelecer qualquer padrão para as mesmas também é inviável, visto que neste caso teremos mais parâmetros a serem estimados do que observações disponíveis. Mais precisamente, em um modelo com nT observações, teremos mais nT(nT+1)/2 parâmetros na matriz de variância V para serem estimados, além dos parâmetros usuais, tornando qualquer estimativa impossível. Assim, para que se possa obter as estimativas, faz-se necessária a estimação por Mínimos Quadrados Generalizados Factíveis – MQGF, onde o padrão dessa matriz é predeterminado.[4]

Outro problema que pode surgir em dados em painel, e que inviabilizaria a utilização de MQO, é a endogeneidade. Esta ocorre quando a correlação entre alguma variável explicativa x_j e o erro é diferente de zero, isto é: $Cov(x_j, \varepsilon_{it}) \neq 0$. Wooldridge (2002) destaca as três principais fontes de endogeneidade: omissão de variáveis do modelo (heterogeneidade não-observada), erros de medição das variáveis e simultaneidade entre as variáveis.

[4] Para maiores detalhes sobre esse método, ver Greene (2003) e Wooldridge (2002).

2. Heterogeneidade Não-observada

O problema mais frequente em dados em painel é a questão da heterogeneidade não-observada. Neste caso, haveria fatores que determinam a variável dependente, mas não estão sendo considerados na equação dentro do conjunto de variáveis explicativas, por não serem diretamente observáveis ou mensuráveis. Levando em consideração a heterogeneidade não-observada, o modelo acima pode ser reescrito da seguinte forma:

$$y_{it} = \boldsymbol{x}_{it}\boldsymbol{\beta} + c_i + \varepsilon_{it}, \quad \text{i = 1, 2, ... , n; t = 1, 2, ... , T} \quad (2)$$

onde c_i representa a heterogeneidade não-observada em cada unidade observacional (no presente caso, estado) constante ao longo do tempo.

Segundo Wooldridge (2002), se c_i for correlacionado com qualquer variável em $\boldsymbol{x}_{it}$ e tentarmos aplicar MQO neste caso, as estimativas serão não só viesadas como inconsistentes.[5] As mesmas consequências ocorrem no modelo no caso em que a hipótese clássica que não haja correlação entre alguma variável explicativa x_j e o erro, $Cov(x_j, \varepsilon_{it}) = 0$, não seja válida. Assim, neste caso, somente podemos utilizar MQO se tivermos justificativas para assumir que $Cov(c_i, x_j) = 0$. Se essa hipótese for válida podemos considerar um novo termo composto, $v_{it} \equiv c_i + \varepsilon_{it}$, e estimar o modelo por MQO, visto que teríamos $Cov(v_{it}, x_j) = 0$. Esse método com dados em painel é conhecido como Mínimos Quadrados Ordinários Agrupados.

[5] Para uma discussão mais detalhada das implicações da existência da heterogeneidade não-observada nos modelos econométricos, ver Worrall & Pratt (2004).

3. Efeitos Fixos

No caso em que $Cov(c_i, x_j) \neq 0$, para que possamos estimar essa equação consistentemente, a abordagem mais usual no contexto de dados longitudinais é a de Efeitos Fixos. Neste método de estimação, mesmo permitindo que $Cov(c_i, x_j) \neq 0$, a idéia é eliminar o efeito não-observado c_i, baseado na seguinte suposição: $E(\varepsilon_{it} | \boldsymbol{x}_i, c_i) = 0$, onde $\boldsymbol{x}_i \equiv (\boldsymbol{x}_{i1}, \boldsymbol{x}_{i2}, ..., \boldsymbol{x}_{iT})$, conhecida como condição de *exogeneidade estrita*. A transformação de efeitos fixos (ou transformação *within*) é obtida em dois passos. Tirando-se a média da equação (2) no tempo obtemos:

$$\bar{y}_i = \bar{\boldsymbol{x}}_i \boldsymbol{\beta} + c_i + \bar{\varepsilon}_i \quad (3)$$

e subtraindo (3) de (2) para cada t, obtemos a equação transformada de efeitos fixos:

$$y_{it} - \bar{y}_i = (\boldsymbol{x}_{it} - \bar{\boldsymbol{x}}_i)\boldsymbol{\beta} + \varepsilon_{it} - \bar{\varepsilon}_i \quad (4)$$

ou

$$\ddot{y}_{it} = \ddot{\boldsymbol{x}}_{it} \boldsymbol{\beta} + \ddot{\varepsilon}_{it}, \quad i = 1, 2, ..., n; t = 1, 2, ..., T \quad (5)$$

removendo assim a heterogeneidade não-observada c_i.

O estimador de Efeitos Fixos é obtido ao se aplicar MQO agrupados na equação (5) e sob a hipótese de exogeneidade estrita, esse estimador é consistente. Este estimador também é conhecido como estimador *within*, por usar a variação do tempo dentro de cada unidade observacional. Outro estimador bastante utilizado a partir das transformações anteriores é o estimador *between,* que é obtido ao se aplicar MQO agrupados na equação (3), e leva em consideração somente a variação entre as unidades observacionais.

4. Efeitos Aleatórios

Outro método de estimação bastante utilizado com dados em painel é o de Efeitos Aleatórios. Assim como nos MQO agrupados, em uma análise de efeitos aleatórios, o efeito não-observado c_i é colocado junto com o termo aleatório ε_{it}. Entretanto, impõe três suposições adicionais[6]: a) $E(\varepsilon_{it}|\boldsymbol{x}_i, c_i) = 0$, b) $E(c_i|\boldsymbol{x}_i) = E(c_i) = 0$ e c) $Var(c_i^2|\boldsymbol{x}_i) = \sigma_c^2$. A primeira é a mesma do modelo de efeitos fixos, a de exogeneidade estrita. A segunda diz respeito à ortogonalidade entre c_i e cada $\boldsymbol{x}_i$ e média de c_i ser nula. A terceira se refere à homoscedasticidade de c_i.

O modelo de efeitos fixos permite a existência de correlação entre os efeitos individuais não-observados com as variáveis incluídas. Entretanto, se esses efeitos forem estritamente não-correlacionados com as variáveis explicativas, pode ser mais apropriado modelar esses efeitos como aleatoriamente distribuídos entre as unidades observacionais, utilizando o modelo de efeitos aleatórios. Em função das especificidades desse modelo, o problema de autocorrelação é uma constante, fazendo com que seja necessária a utilização de MQG factíveis.

Assim, o ponto crucial na decisão de que modelo deve ser utilizado, se efeitos fixos ou aleatórios, reside na questão se c_i e $\boldsymbol{x}_i$ são correlacionados ou não. Esse questionamento deve ser feito de acordo com os dados que se está trabalhando, examinando suas especificidades. Um teste mais formal pode ser realizado, o Teste de Hausman, baseado nas diferenças das estimativas de efeitos fixos e aleatórios. Este teste é descrito na última seção.

Haveria ainda a possibilidade de simplesmente não haver heterogeneidade não-observada no modelo que estamos estimando. Se isso for verdade a estimativa por MQO agrupado é eficiente e válida. A ausência de efeitos não-observados é equivalente a testar a hipótese de a variância de c_i ser nula. Um teste para verificar a existência de efeitos não-observados é o de Breusch e Pagan, baseado no multiplicador de Lagrange, que é descrito em Greene (2003) e Wooldridge (2002).

[6] Além das suposições usuais de posto e dos erros.

5. Exogeneidade Estrita e Variáveis Instrumentais

Um ponto importante a se destacar dos três modelos discutidos acima que tratam da heterogeneidade não-observada é a hipótese comum a todos eles: a de exogeneidade estrita. Embora essa suposição seja crucial para a consistência de todos esses estimadores, é também uma das mais prováveis de não ser válida. Assim, precisamos saber que procedimento deve-se utilizar se a suposição de exogeneidade estrita não for válida. Wooldridge (2002) sugere algumas soluções para esse problema, destacando a utilização de variáveis instrumentais e eliminação do efeito não-observado para que os estimadores sejam consistentes mesmo quanto à hipótese de exogeneidade estrita não for válida.

Para que possamos utilizar variáveis instrumentais, é necessária a utilização de métodos específicos para quando estas precisam ser utilizadas no modelo. O método mais utilizado nesse contexto é o método de Mínimos Quadrados em Dois Estágios – MQ2E. O objetivo principal de se utilizar esse tipo de estimação com variáveis instrumentais é resolver o problema de endogeneidade.

Uma discussão mais detalhada do método de M2QE fugiria do escopo do presente trabalho.[7] Entretanto, faz-se necessário definir o que caracteriza uma variável instrumental. Reescrevendo um modelo de regressão como o descrito na equação (1) destacando uma das variáveis contidas em $\boldsymbol{x}_{it}$ que seja endógena (isto é, $Cov(w_{it}, \varepsilon_{it}) \neq 0$), e a denotando por w_{it}, teremos:

$$y_{it} = \boldsymbol{x}_{it}\boldsymbol{\beta} + \gamma w_{it} + \varepsilon_{it}, \quad \text{i = 1, 2, ... , n; t = 1, 2, ... , T} \quad (7)$$

Sabemos que a estimação de (7) por MQO resultará em estimativas inconsistentes não só para γ, como para todos os parâmetros contidos no vetor $\boldsymbol{\beta}$. O método de variáveis instrumentais – IV possibilita uma solução geral pra o caso em que existe alguma variável endógena no modelo. Para utilizar essa abordagem, é necessária uma

[7] Para maiores detalhes sobre estimadores com variáveis instrumentais, ver Greene (2003), Davidson & MacKinnon (2004) e Wooldridge (2002).

variável observável z_{it} que sirva como instrumento (variável instrumental) e não esteja na equação (7).

Esta variável precisa satisfazer duas condições. Primeiro, z_{it} deve ser não correlacionada com o erro ε_{it}, isto é: $Cov(z_{it}, \varepsilon_{it}) = 0$. Desta forma, assim como as demais variáveis em $\boldsymbol{x}_{it}$, z_{it} é exógena na equação (7). A segunda condição diz respeito à relação entre z_{it} e a variável endógena w_{it}. Em uma projeção linear de w_{it} em todas as variáveis exógenas:

$$w_{it} = \boldsymbol{x}_{it}\boldsymbol{\delta} + \theta z_{it} + \eta_{it} \tag{8}$$

o coeficiente de z_{it} deve ser não-nulo, isto é: $\theta \neq 0$. Essa condição pode ser entendida de uma forma não tão rigorosa como: $Cov(w_{it}, z_{it}) \neq 0$. Ou seja, a variável instrumental deve ser correlacionada com a variável endógena.

Como já foi mencionado e será discutido com mais detalhes mais a frente, no presente trabalho, a variável no modelo a ser estimado que se acredita que seja endógena, é a variável de gastos em segurança pública. Assim, devemos utilizar pelo menos uma variável instrumental não somente para corrigir esse problema, como na própria determinação se a variável de gastos públicos em segurança é endógena no modelo que iremos estimar.[8]

Assim, com uma variável instrumental que satisfaça essas condições, podemos implementar o método apropriado para corrigir o problema de endogeneidade no modelo que queremos estimar, seja este problema causado pela hipótese de exogeneidade estrita não ser válida, ou haver simultaneidade entre alguma variável explicativa e a variável independente. Isto é, alguma variável explicativa, além de determinar a variável dependente, ao mesmo tempo, ser influenciada pela variável dependente.

[8] Somente com a variável instrumental em mãos, podemos testar se uma variável é endógena ou não em um modelo. O teste mais difundido para este fim é o teste de Hausman de endogeneidade.

Anexo: Testes frequentemente utilizados em modelos com dados em painel

A - Teste F para Heterogeneidade Não-Observada

$H_0 : c_i = c$

$$F(n-1, nT-n-K) = \frac{(R^2_{LDSV} - R^2_{MQOA})/(n-1)}{(1-R^2_{LSDV})/(nT-n-K)} \quad \text{(A.1)}$$

onde LSDV indica o estimador com variável *dummy* onde c_i é levado em consideração. Se esta estatística exceder o valor tabelado, a hipótese de heterogeneidade não-observada é válida.

B - Teste de Breusch e Pagan

$H_0 : \sigma^2_{c_i} = 0$

$$LM = \frac{nT}{2(T-1)}\left[\frac{\sum_{i=1}^{n}\left[\sum_{t=1}^{T}\hat{\varepsilon}_{it}\right]^2}{\sum_{i=1}^{n}\sum_{t=1}^{T}\hat{\varepsilon}^2_{it}} - 1\right]^2 = \frac{nT}{2(T-1)}\left[\frac{\sum_{i=1}^{n}\left(T\bar{\hat{\varepsilon}}_i\right)^2}{\sum_{i=1}^{n}\sum_{t=1}^{T}\hat{\varepsilon}^2_{it}} - 1\right]^2 \quad \text{(A.2)}$$

onde é $\hat{\varepsilon}_{it}$ resíduo da regressão de MQO agrupados e sob a hipótese nula, $LM \sim \chi^2$ com 1 grau de liberdade. Se esta estatística exceder o valor tabelado, a hipótese de heterogeneidade não-observada é válida.

C - Teste de Hausman para testar Efeitos Fixos contra Efeitos Aleatórios

Seja $\hat{\beta}_{EF}$ o vetor de estimativas de efeitos fixos e $\hat{\beta}_{EA}$ o vetor de estimativas de efeitos aleatórios, sob a hipótese nula de:

$H_0 : \hat{\beta}_{EF} - \hat{\beta}_{EA} = 0$ (i.e. efeitos aleatórios é válido), a estatística:

$$H = \left[\hat{\beta}_{EF} - \hat{\beta}_{EA}\right]' \left[V(\hat{\beta}_{EF}) - V(\hat{\beta}_{EA})\right]^{-1} \left[\hat{\beta}_{EF} - \hat{\beta}_{EA}\right] \quad \text{(A.3)}$$

possui distribuição χ^2 com K-1 graus de liberdade. Se esta estatística exceder o valor tabelado, devemos utilizar efeitos fixos.

Referências Bibliográficas

DAVIDSON, R. and MACKINNON, J. G., **Econometric Theory and Methods**, Oxford University Press, 2004.

GREENE, William H. **Econometric Analysis** 5th ed. Prentice-hall. 2003.

IRFFI, G. D.; OLIVEIRA, J.; BARBOSA, E. Análise dos Determinantes Socioeconômicos da Taxa de Mortalidade Infantil (TMI) no Ceará. **Texto para Discussão IPECE** Nº 48, 2008.

IRFFI, G. D.; TROMPIERI, N.; OLIVEIRA, J.; NOGUEIRA, C. A.; BARBOSA, M.; HOLANDA, M. Determinantes do Crescimento Econômico dos Municípios Cearenses. **Texto para Discussão IPECE** Nº 39, 2008.

HSIAO, Cheng, **Analysis of panel data: Second Edition**, Cambridge University Press, 2003.

HSIAO, Cheng, **Panel Data Analysis - Advantages and Challenges**, IEPR Working Papers, Institute of Economic Policy Research (IEPR), 2006.

LOUREIRO, A. O. F. Avaliando o Impacto do Policiamento sobre a Criminalidade no Ceará. **Texto para Discussão IPECE** Nº 53, 2008.

NERLOVE, M. **Essays in Panel Data Econometrics**. Cambridge University Press, 2002.

WOOLDRIDGE, Jeffrey M., **Econometric Analysis of Cross Section and Panel Data.** The MIT Press, Cambridge, MA, 2002.

WORRALL J. L.; PRATT T. C., On the Consequences of Ignoring Unobserved Heterogeneity when Estimating Macro-Level Models of Crime. **Social Science Research,** v. 33, p. 79-105, 2004.