ANÁLISE DE REGRESSÃO LINEAR: ABORDAGEM TRADICIONAL E ESPACIAL EM UM ESTUDO DE CASO

Ana Julia Righetto[1], Vilma Mayumi Tachibana[2]

[1] *Programa de Pós-Graduação em Estatística e Experimentação Agronômica – ESALQ/USP*
[2] *Departamento de Matemática, Estatística e Computação – FCT/UNESP*
ajrighetto@usp.br; vilma@fct.unesp.br

1 INTRODUÇÃO

Em vários experimentos deseja-se averiguar como uma mudança ocorrida em uma ou mais variáveis ditas independentes (ou explicativas) afeta outra variável, geralmente, denominada de dependente (ou resposta). Essa relação funcional entre as variáveis pode ser obtida pela técnica denominada análise de regressão.

Recentemente esta técnica vem sendo muito utilizada na análise espacial de dados para obter um modelo que acrescente a informação de agrupamento espacial, caso não houver aleatoriedade espacial da variável resposta de interesse.

Abordando análise de regressão linear, em sua forma tradicional e em análise espacial, para a realização deste estudo trabalhou-se com duas variáveis explicativas, idade e instrução, e renda como variável resposta. São variáveis sócio-econômicas coletadas pelo IBGE no município de Presidente Epitácio – SP, no Censo 2000. Estas variáveis fazem parte dos resultados do questionário da amostra, aplicado a 10% dos domicílios, e do questionário básico aplicado a toda população; referentes aos 47 setores censitários do município.

Inicialmente, uma análise de regressão linear múltipla foi realizada para ajustar um modelo no qual a variável renda é explicada em função das variáveis idade e grau de instrução; em seguida, a dependência espacial foi incorporada ao modelo.

Os resultados obtidos também foram apresentados em forma de índices que medem a associação espacial global e local, gráfico de espalhamento e mapas, facilitando possíveis identificações de agrupamentos, áreas de transição e casos incomuns.

2 METODOLOGIA

2.1 Modelo Clássico de Regressão Linear

Seja o modelo de regressão linear múltipla (MRLM) dado por:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon} \quad , \tag{1}$$

em que **Y** é o vetor de variável resposta, **X** é a matriz de variáveis explicativas, $\boldsymbol{\beta}$ é o vetor de parâmetros do modelo (coeficientes ou pesos das variáveis explicativas) e $\boldsymbol{\varepsilon}$ é o vetor dos resíduos com distribuição Normal multivariada de dimensão n com média **0** e covariância $\sigma^2 \boldsymbol{I}$.

Os coeficientes $\hat{\boldsymbol{\beta}}$ escolhidos pelo critério de mínimos quadrados são dados por:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} \; .$$

Obtido essas estimativas, testa-se a significância das variáveis explicativas. Ou seja, a escolha entre um modelo completo com todas as variáveis explicativas ou um modelo reduzido sem um subgrupo de variáveis, respondendo a pergunta: "Vale a pena acrescentar os termos extras ao modelo?". Sempre que se acrescentam mais termos no modelo, diminui-se o erro de ajuste, restando saber se tal aumento é significativo. É preciso verificar se os termos extras contribuem para a melhor descrição da variável Y,

caso contrário, Charnet et al. (1999) recomendam, por parcimônia, optar pelo modelo reduzido.

Além disso, deve-se realizar a análise de resíduos, para comprovar que os mesmos atendem aos pressupostos do modelo: são aleatórios, têm distribuição normal e não são autocorrelacionados.

2.2 Regressão Linear Espacial

A análise espacial de dados consiste em um estudo quantitativo de fenômenos localizados em determinado espaço, sendo que a localização dos dados é muito importante em sua análise ou interpretação dos resultados.

Segundo Druck et al. (2004), compreender a distribuição espacial de dados provindos de fenômenos ocorridos no espaço constitui, nos dias atuais, um grande desafio para esclarecer questões centrais em diversas áreas do conhecimento, seja na área da saúde, em ambiente, em geologia, em agronomia, entre tantas outras áreas.

Vários tipos de dados de interesse podem caracterizar problemas de análise espacial e Assunção (2001) classifica-os de acordo com uma tipologia de quatro categorias: dados de processos pontuais; dados de interação espacial; dados de área e dados de superfície aleatória.

"Na situação de dados espaciais, quando está presente a autocorrelação espacial, as estimativas do modelo devem incorporar essa estrutura espacial, uma vez que a dependência entre as observações altera o poder explicativo do modelo. A significância dos parâmetros é usualmente superestimada, e a existência de variações em larga escala pode até mesmo induzir a presença de associações espúrias" (DRUCK et al., 2004).

Com dados espaciais é pouco provável que a hipótese das observações não correlacionadas seja verdadeira, quando a dependência espacial estiver presente. Na regressão espacial é preciso investigar os resíduos em busca de sinais de estruturas espaciais, por meio de análise gráfica, mapeamento dos resíduos ou teste de autocorrelação, como o índice de Moran.

O índice global de Moran I é uma medida de autocorrelação considerando-se o primeiro vizinho e tem a seguinte expressão:

$$I = \frac{n\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(z_i-\bar{z})(z_j-\bar{z})}{(\sum_{i=1}^{n}(z_i-\bar{z})^2)(\sum_{i\neq j}\sum w_{ij})} \qquad (2)$$

em que: n é o número de áreas (por exemplo: número de setores censitários), z_i o valor do atributo considerado na área i, $\bar{z}$ é o valor médio do atributo na região de estudo e w_{ij} os elementos da matriz normalizada de proximidade espacial, denotada por $\boldsymbol{W}$. Esta matriz é uma ferramenta básica para estimar a variabilidade espacial de dados de área, seus elementos – pelo método da rainha – são medidas de distância entre duas áreas e assumem valores 1 (para zonas com fronteiras) e 0 (quando não há fronteira). Geralmente, as linhas desta matriz são normalizadas, para que a soma dos pesos de cada linha seja igual a 1.

O valor deste índice pertence ao intervalo [-1, 1]; se o valor estiver entre 0 e +1 indica correlação direta e se estiver entre 0 e −1, correlação inversa, ou seja, quando há homogeneidade entre localidades próximas, I tende a ser positivo, enquanto se as localidades próximas forem dissimilares, o índice tende a ser negativo. A hipótese nula do teste é de independência espacial, e neste caso o valor do índice é nulo.

Para este trabalho foram utilizados os modelos de regressão permitem incorporar efeitos espaciais de forma global (como um único parâmetro). Existem duas opções para tratar a autocorrelação global num modelo de regressão: Modelo espacial auto-regressivo misto (*SAR)* ou Modelo do erro espacial (*CAR*), que são descritos a seguir, baseados em Druck et al. (2004).

2.2.1 Modelo Espacial Auto-Regressivo Misto

No modelo espacial auto-regressivo misto (*Spatial AutoRregressive – SAR* ou S*patial Lag Model*), considera-se a dependência espacial adcionando-se ao modelo de regressão um novo termo na forma de uma relação espacial para a variável resposta, ou seja, a auto a autocorrelação espacial ignorada é atribuída à variável resposta Y.

O modelo é expresso da seguinte forma:

$$Y = \rho WY + X\beta + \varepsilon \quad (3)$$

Em $\boldsymbol{W}$ que é a matriz de proximidade espacial, o produto $\boldsymbol{WY}$ expressa a dependência espacial em $\boldsymbol{W}$ e ρ é o coeficiente espacial auto-regressivo. A hipótese nula para a não existência de autocorrelação é $\rho = 0$.

Em termos de componentes individuais, esse modelo é expresso como:

$$y_i = \rho\left(\sum_j w_{ij} y_j\right) + \sum_{i=1} x_i \beta_i + \varepsilon_i \quad (4)$$

sendo w_{ij} elemento da matriz de proximidade espacial.

2.2.2 Modelo do Erro Espacial

Este modelo, também conhecido como *Spatial error model* ou ainda *Conditional AutoRegressive* – CAR, considera que os efeitos espaciais são um ruído, ou pertubação, isto é, fator que precisa ser removido. Neste caso, os efeitos de autocorrelação espacial são associados ao termo de erro $\boldsymbol{\varepsilon}$ e o modelo é expresso da seguinte forma:

$$Y = X\beta + \varepsilon\,, \qquad \varepsilon = \lambda W\varepsilon + \xi \quad (5)$$

Nas expressões dadas em (5) tem-se: $\mathbf{W\varepsilon}$ é a componente do erro com efeitos espaciais, λ é o coeficiente autoregressivo e ξ é a componente do erro com variância constante e não correlacionada.

A hipótese nula para a não-existência de autocorrelação é que $\lambda = 0$, isto é, o termo de erro não é espacialmente correlacionado.

2.2.3 Diagnóstico de Modelos com Efeitos Espaciais

Primeiramente, uma análise gráfica dos resíduos deve ser realizada para ser avaliada a qualidade do ajuste de regressão. Mapear os resíduos é importante no diagnóstico do modelo, buscando indícios de ruptura dos pressupostos de independência.

Segundo Druck et al. (2004), uma elevada concentração de resíduos positivos (ou negativos) em alguma parte do mapa é um bom indicador da presença de autocorrelção espacial. Para um teste quantitativo, o mais usado é o índice I de Moran.

Os estimadores e os diagnósticos tradicionais de regressão não levam em conta os efeitos espaciais, assim, na regressão espacial as inferências, como por exemplo, as indicações de qualidade de ajuste, baseadas no coeficiente de determinação (R^2) não são confiáveis.

"O método mais usual de seleção de modelos de regressão baseia-se nos valores de máxima verossimilhança dos diferentes modelos, ponderando pela diferença no número de parâmetros estimados. Nos modelos com estrutura de dependência – espacial ou temporal – utilizam-se os critérios de informação em que a avaliação do ajuste é penalizada por uma função do número de parâmetros. Cabe observar que é necessário ainda levar em conta o número de parâmetros independentes ao se incluir funções espaciais nos modelos" (DRUCK et al., 2004).

A comparação de modelos é normalmente feita utilizando-se o logaritmo da máxima verossimilhança (LMV) e o número de coeficientes de regressão (k). Há vários critérios, sendo os mais utilizados informação de Akaike (AIC) e Bayesiano de Schwarz.

O Critério de Informação de Akaike (AIC) é expresso da seguinte forma:

$$AIC = -2*LMV + 2k \quad (6)$$

Pelo critério de Akaike, o melhor modelo é o que possuir menor valor de AIC.

O Critério Bayesiano de Schwarz (SBC) é expresso por:

$$SBC = -2*LMV + k * \ln(N), \quad (7)$$

em que N é o número de observações.

Assim como o critério de Akaike, pelo critério de Schwarz o melhor modelo é aquele que possuir o menor valor de SBC.

3 ANÁLISE

O município de Presidente Epitácio foi dividido em 47 setores censitários e neste estudo retirou-se um setor em que não havia moradores.

Os elementos w_{ij} da matriz de proximidade espacial **W** foram definidos segundo o critério de contigüidade da rainha, que de acordo com Upton e Fingleton (1985) recebem valor 1 quando o setor *i* e o setor *j* compartilham cantos e arestas e 0 em outros casos. Esses valores foram padronizados, de modo que a soma dos pesos w_{ij} em cada linha *i* era igual a 1.

Os dados sócio-econômicos dos responsáveis pelos domicílios no município foram obtidos usando-se o programa *ESTATCART*. As variáveis deste estudo são: Renda (variável resposta), Idade e Instrução (variáveis explicativas).

Para aplicação do modelo de regressão, testou-se a normalidade da variável Renda, verificando-se que esta não possuía distribuição normal. Foi realizada a transformação raiz quadrada na variável e criou-se uma nova variável denominada Renda_raiz que pelo teste de normalidade Shapiro-Wilk é normal a um nível de significância de 10%.

Prosseguindo, realizou-se uma análise exploratória de dados e cálculo dos índices de Moran local e global, em cada variável para avaliar a hipótese de dependência espacial.

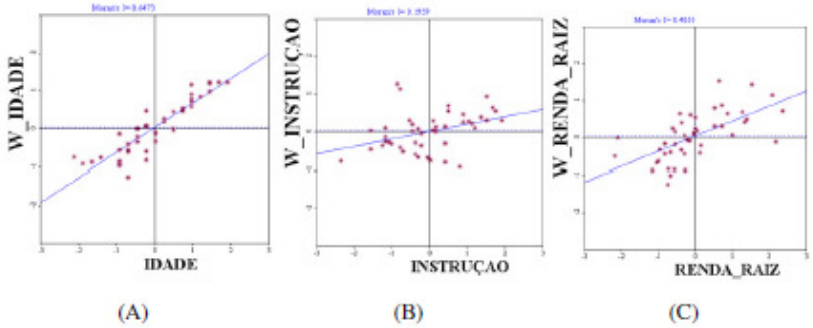
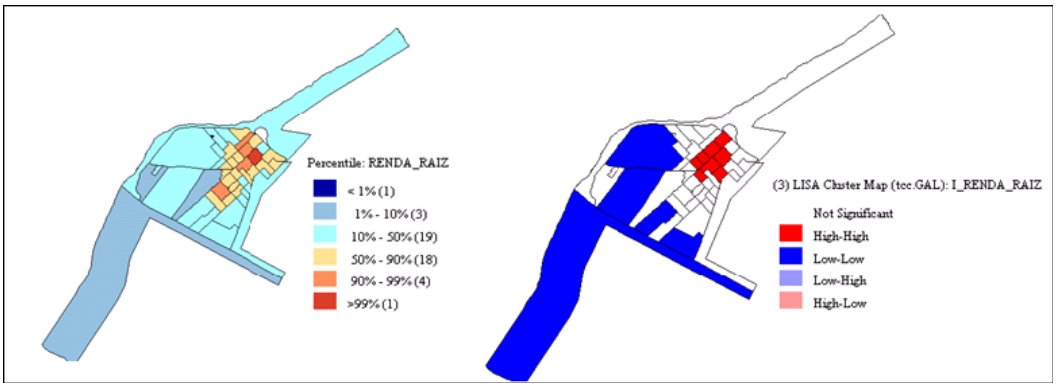


Figura 1 – Scatter plot do Índice de Moran das variáveis Idade (A), Instrução (B) e Renda_raiz (C)

Em relação as variáveis de interesse, nota-se que as mesmas indicam uma forte dependência espacial que pode ser confirmada pelos valores do Índice de Moran: 0,6473 para a variável Idade; 0,1959 para a variável Instrução e 0,4055 para a variável Renda_raiz. A Figura 1 apresenta os gráficos de dispersão dessas variáveis, no eixo *x* o valor do setor censitário e no eixo *y* a média dos vizinhos. As autocorrelações das variáveis deste estudo são significativas e o p-valor das variáveis Renda_raiz e Idade é 0,0010 e da variável Instrução é 0,0080. A Figura 2 mostra o mapa da distribuição da variável Renda_raiz à esquerda e ao lado direito o mapa de autocorrelação espacial local (LISA), destacando um agrupamento de setores com rendas altas (em vermelho) e outro agrupamento de setores com rendas baixas (em azul). As figuras deste trabalho foram elaboradas no GeoDa.



(A) (B)

Figura 2 – Mapa da distribuição (A) e da autocorrelação espacial local (LISA) (B) da variável Renda_raiz.

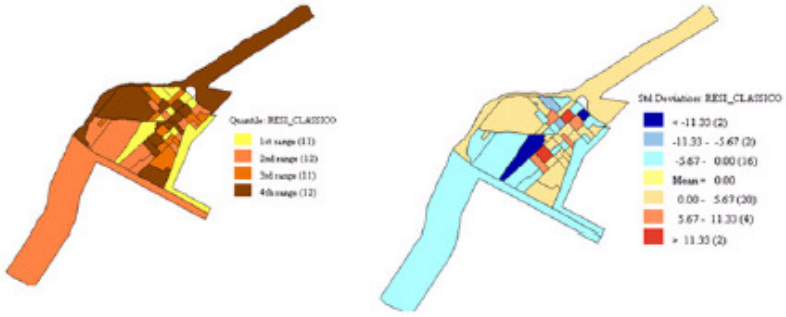
O modelo clássico de regressão linear múltipla é expresso da seguinte forma

$$Renda\_raiz = -41{,}835 + 0{,}977\ Idade + 4{,}398\ Instrução \qquad (8)$$

Neste modelo, o coeficiente de determinação é $R^2 = 0{,}653$ e o coeficiente de determinação ajustado $R^2_{aj} = 0{,}636$. Estes valores indicam que grande parte da variabilidade da variável Renda_raiz é explicada pelo modelo ajustado. As duas variáveis explicativas do modelo são significativas (p-valor <0,0001).

Outras medidas referentes ao ajuste do modelo foram obtidas: Log da Verossimilhança = -144,553; Critério Bayesiano de Schwarz (SBC) = 300,592 e Critério de Informação de Akaike = 295,106.

Os resíduos do modelo possuem distribuição normal e não estão distribuídos aleatoriamente pelo município de Presidente Epitácio-SP, como se observa na Figura 3 (A) e (B) que representa a distribuição dos resíduos em quantis e a distribuição do desvio padrão dos valores absolutos dos resíduos, respectivamente. Nota-se que há uma concentração de maiores erros na região urbana e no centro da cidade, indicando que há dependência espacial. Sendo assim, deve-se seguir a análise levando em consideração o efeito espacial do modelo.



(A) (B)

Figura 3 – Mapa da distribuição (A) e do desvio padrão (B) dos resíduos do modelo ajustado em (8).

Pelo diagnóstico de autocorrelação espacial sugerido por Anselin (2005), o modelo escolhido foi o do erro espacial (CAR ou *Spatial error mode)* dado por:

$$Renda\_raiz = -44{,}895 + 1{,}013\ Idade + 4{,}653 Instrução - 0{,}434 W\_Resíduo \qquad (9)$$

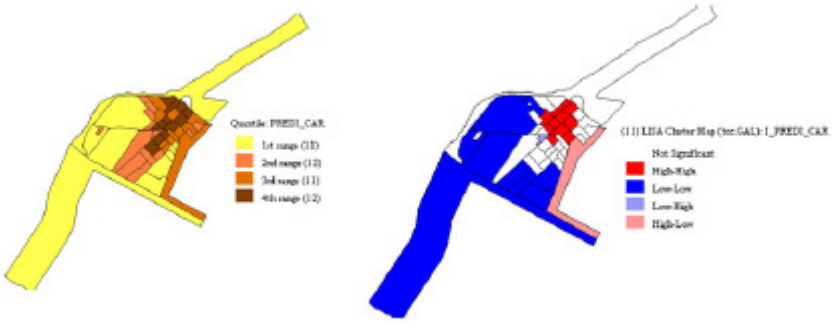
O pseudo-coeficiente de determinação é $R^2 = 0{,}676$; e todas as variáveis explicativas são importantes no modelo.

Outras medidas referentes ao modelo foram calculadas e comparadas com as medidas do modelo de regressão clássica dada em (8): o Log de verossimilhança para este modelo é -143,576 que é um valor maior do que o Log dado no modelo (8), indicando que o ajuste foi melhor; a medida dada pelo Critério Bayesiano de Schwarz

(SBC) é 298,638 e do Critério de Informação de Akaike (AIC) é 293,152; ambos os valores são menores do que os valores do modelo clássico, indicando também um melhor ajuste.

Para este modelo, a estatística I de Moran dos resíduos é 0,0972 que pode ser considerado igual a zero ao nível de significância de 10%, indicando que a inclusão da variável W_resíduo no modelo, eliminou a autocorrelação espacial.

Os valores de renda_raiz ajustados pelo modelo dado em (9) são apresentados na Figura 4 (A), na qual observa-se que as rendas mais baixas (em amarelo) estão na região rural do município e as mais altas estão na região central de Presidente Epitácio - SP. Pela Figura 4 (B), nota-se que na zona rural estão os setores com menor renda cercados por setores com menor renda (cor azul) e que mais ao centro estão os setores com maiores renda cercados de setores com rendas também maiores (cor vermelha).



(A) (B)

Figura 4 – Mapa da distribuição (A) e da autocorrelação espacial local (LISA) (B) para Renda_raiz.

4 CONSIDERAÇÕES FINAIS

Dias et al. (2002) apresentam alguns problemas que podem surgir ao trabalhar com dados agregados por área, como setores censitários, pois as áreas são pequenas e normalmente são agregadas em um mesmo setor grupos sociais distintos como favela e áreas nobres. Neste trabalho não houve problemas em utilizar setores censitários definidos pelo IBGE, pois em Presidente Epitácio não existiu esses problemas.

O modelo clássico de regressão linear gerou resíduos que estão correlacionados no espaço, necessitando-se da construção de um novo modelo que eliminasse esse aspecto, além de incorporar esses efeitos espaciais.

O modelo de regressão de erro espacial CAR obtido apresentou melhores resultados do que o modelo de regressão clássico, além de possibilitar apresentar os resultados da estimativa em forma de mapas.

5 REFERÊNCIAS BIBLIOGRÁFICAS

[1]ANSELIN, L. **Exploring spatial data with GeoDaTM: a Workbook.** University of Illinois, Urbana-Champaign, 2005.

[2] ASSUNÇÃO, R. M. **Estatística espacial com aplicações em epidemiologia, economia, sociologia.** In 7ª ESCOLA DE MODELOS DE REGRESSÃO, São Carlos: Universidade Federal de São Carlos, 2001. 130p.

[3] CHARNET, R. et al. **Análise de modelos de regressão linear com aplicações.** São Paulo: Unicamp, 1999. 354p.

[4] DIAS, T.L., OLIVEIRA, M.P.G., CÂMARA, G., CARVALHO, M.S. Problemas de escala e a relação área-indivíduo em análise espacial de dados censitários. **Informática Pública**. Belo Horizonte, V. 1, no. 4, p. 89-104, 2002.

[5]DRUCK, S. et al. **Análise espacial de dados geográficos**. Brasília, EMBRAPA, 2004 (ISBN: 85-7383-260-6).

[6] UPTON, G.; FINGLETON, B. ***Spatial Data Analysis by Example*, volume I – Point pattern and quantitative data**, Chischester: John Wiley & Sons, 1985.