LILIANE LOPES CORDEIRO

ESTIMAÇÃO EM REGRESSÃO ESPACIAL INVERSA

LAVRAS - MG
2015

LILIANE LOPES CORDEIRO

ESTIMAÇÃO EM REGRESSÃO ESPACIAL INVERSA

Tese apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Estatística e Experimentação Agropecuária, área de concentração em Estatística e Experimentação Agropecuária, para a obtenção do título de Doutor.

Orientador
Dr. João Domingos Scalon

LAVRAS - MG
2015

Ficha catalográfica elaborada pelo Sistema de Geração de Ficha Catalográfica da Biblioteca Universitária da UFLA, com dados informados pelo(a) próprio(a) autor(a).

Cordeiro, Liliane Lopes.
Estimação em regressão espacial inversa / Liliane Lopes Cordeiro. – Lavras : UFLA, 2015.
89 p. : il.

Tese(doutorado)–Universidade Federal de Lavras, 2015.
Orientadora: João Domingos Scalon.
Bibliografia.

1. Calibração Modelo SAR. 2. Estimativa Pontual. 3. Estimativa intervalar. I. Universidade Federal de Lavras. II. Título.

LILIANE LOPES CORDEIRO

ESTIMAÇÃO EM REGRESSÃO ESPACIAL INVERSA

Tese apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Estatística e Experimentação Agropecuária, área de concentração em Estatística e Experimentação Agropecuária, para a obtenção do título de Doutor.

APROVADA em 30 de março de 2015.

Dr. Renato Ribeiro de Lima	UFLA
Dr. Mário Javier Ferrua Vivanco	UFLA
Dr. José Airton Rodrigues Nunes	UFLA
Dr. Flávio Bittencourt	UNIFAL

Dr. João Domingos Scalon
Orientador

LAVRAS - MG
2015

Aos meus pais, Maria Aparecida e João,
a minha irmã, Lilian e
ao meu noivo, Heverton.

DEDICO

AGRADECIMENTOS

A Deus, pelo dom da vida e por todas as bençãos realizadas.

Aos meus pais, Maria Aparecida e João, pela oportunidade que me proporcionaram nos estudos, pelo apoio, pela compreensão, pela paciência e por serem sempre meu esteio.

A minha irmã, Lilian e ao meu cunhado, Davi, pelo incentivo, apoio e amizade.

Ao Heverton, pelo carinho, apoio, companheirismo e amizade.

A toda minha família pelos constantes incentivos: avós, tios e tias, primos e primas.

À Universidade Federal de Lavras (UFLA) e ao Departamento de Ciências Exatas (DEX) pela oportunidade de aprimoramento acadêmico.

Ao Professor Doutor João Domingos Scalon pela orientação, dedicação, atenção e pelas contribuições científicas dadas para realização deste trabalho.

Aos professores do Departamento de Ciências Exatas pelos conhecimentos transmitidos durante esta caminhada.

A todos os colegas do DEX, em especial à Adriana e à Diana, pela amizade e convivência que tivemos.

À CAPES pela bolsa de estudo, essencial para esta conquista.

A todos que de forma direta ou indireta contribuíram para realização deste novo desafio em minha vida agradeço.

RESUMO

Em alguns problemas que envolvem análise de regressão pode ser de interesse obter estimativas para um valor da variável independente dado um valor da variável dependente. Esse problema é chamado de regressão inversa ou calibração. Na literatura existem dois métodos mais comumente utilizados para realizar a estimação pontual em modelos de regressão inversa: clássico e inverso. Métodos para obter estimações intervalares para o verdadeiro valor da variável independente também estão disponíveis. O principal objetivo desta tese é apresentar o problema da calibração espacial e propor métodos para a estimação pontual e intervalar em modelos que levam em consideração a estrutura de dependência espacial entre áreas vizinhas. O problema pode ser dividido em dois casos: no primeiro caso pretende-se estimar o valor da variável independente pertencente à amostra observada, enquanto que no segundo caso, o valor da variável independente a ser estimada não pertence à amostra observada. Esta tese desenvolve estimadores pontuais e intervalares para o valor da variável independente para o modelo autorregressivo espacial (SAR). Os estimadores obtidos são aplicados em dados espaciais reais. Os resultados obtidos mostram o potencial da regressão inversa em problemas onde as informações de uma região são influenciadas diretamente pelas informações das regiões vizinhas.

Palavras-chaves: Calibração. Modelo SAR. Estimativa Pontual. Estimativa intervalar.

ABSTRACT

In some issues involving regression analysis, it can be interesting to obtain estimates for a value of the independent variable, given a value of the dependent variable. This issue is determined inverse regression or calibration. In literature, there are two more commonly used methods for performing the point estimation in reverse regression models: classic and inverse. Methods to obtain interval estimations for the true value of the independent variable are also available. The main objective of this dissertation is to present the issue of spatial calibration and propose methods for the point and interval estimation in models that consider the spatial dependence structure between neighboring areas. The issue can be divided into two cases: in the first case, we intend to estimate the value of the independent variable belonging to the observed sample, while in the second case, the value of the independent variable to be estimated does not belong to the observed sample. This dissertation develops point and interval estimators for the value of the independent variable for the spatial autoregressive model (SAR). The estimators obtained are applied to real spatial data. The results obtained show the potential of inverse regression for issues in which the information from one region are directly influenced by the information from neighboring regions.

Keywords: Calibration. SAR model. Point estimate. Interval estimate.

SUMÁRIO

1 INTRODUÇÃO

Nas aplicações envolvendo regressão, é de interesse verificar se duas ou mais variáveis estão relacionadas de alguma forma. Para expressar essa relação se estabelece um modelo matemático. Em geral, é de interesse determinar o valor da variável dependente (Y) correspondente a um dado valor da variável independente (X). Entretanto, em alguns casos, o que se deseja é o inverso, ou seja, estimar o valor da variável independente, dado o valor observado da variável dependente; trata-se da regressão inversa ou calibração. A calibração é particularmente interessante quando a variável independente é mais complicada de se mensurar e/ou, muitas vezes, a obtenção dessa variável demanda mais tempo e maiores gastos. Portanto, nessas situações o objetivo é obter estimativas do valor x_0, da variável independente, dado um valor y_0, da variável dependente. Por exemplo, suponha que uma determinada droga seja aplicada para reduzir a pressão arterial. Pode-se ajustar um modelo de regressão da pressão sanguínea (Y) em função da quantidade da droga administrada (X). O interesse do médico é controlar a dosagem do medicamento a partir da observação da pressão sanguínea do paciente, ou seja, o médico tem um problema de calibração que é estimar a quantidade, x_0, da droga a ser administrada dada a leitura do valor, y_0, da pressão sanguínea.

Em geral, a regressão inversa pode ser dividida em duas etapas. Na primeira etapa, chamada experimento de calibração, selecionam-se n observações de uma variável aleatória Y, a partir de valores prefixados de X, a fim de estimar a função f que relaciona as variáveis por meio do modelo: $Y = f(x) + \epsilon$. A segunda etapa, a calibração propriamente dita, seleciona uma amostra aleatória, Y_0, de tamanho $k\,(k \geq 1)$ da variável Y correspondente a um único valor x_0 desconhecido. Com o objetivo de obter estimadores para este valor desconhecido da variável X.

Na literatura existem dois métodos mais comuns para o problema de calibração: o método clássico e o método inverso para a estimação da variável desconhecida X. A calibração controlada, em que a variável independente X é fixa, chama-se método clássico, e uma calibração aleatória, em que a variável X é considerada aleatória correspondente ao método inverso. São apresentadas neste

trabalho revisões desses métodos, bem como exemplos de aplicação.

O problema de calibração tem sido explorado por diversos autores. A maioria desses estudos assume que a relação entre as variáveis Y e X é linear e que os erros do modelo são independentes e seguem uma distribuição normal com média zero e variância constante. Utilizando essas suposições, a regressão inversa é utilizada em diversas áreas, como por exemplo, Biologia, Química, Física, Engenharia, Medicina, entre outras. Porém, no estudo de alguns fenômenos que ocorrem em áreas tais como epidemiologia, experimentos agrícolas e geológicos, a independência entre os erros nem sempre é atendida, ou seja, nem sempre a observação realizada em uma posição independe das observações realizadas nas posições vizinhas. Portanto, nessas situações devem ser utilizados métodos da estatística espacial.

A ideia central da análise espacial é encontrar um modelo inferencial que incorpore explicitamente as relações espaciais constituintes de um fenômeno. Em geral, a modelagem é iniciada pela análise exploratória associada à visualização dos dados por meio de gráficos e mapas e, posteriormente, identificam-se padrões de dependência espacial das variáveis em estudo.

Tendo em vista a utilização da regressão inversa em diversas áreas e o crescente uso de métodos para análises de dados espacialmente distribuídos, neste trabalho defende-se que é possível desenvolver métodos para realizar a regressão inversa em dados espaciais de área em que, em geral, a suposição de independência é violada. Assim, o principal objetivo será propor um modelo de calibração que leva em consideração a estrutura de dependência espacial entre áreas vizinhas. Para obter a modelagem adequada no contexto de calibração, considera-se o modelo espacial autorregressivo (SAR), com apenas uma variável explicativa. A partir do ajuste do modelo espacial, são deduzidos os estimadores pontuais e intervalares para um valor, x_0 desconhecido, da variável independente dado o valor, y_0 conhecido, da variável dependente . Os novos métodos propostos foram aplicados em um caso real de calibração espacial.

2 REFERENCIAL TEÓRICO

2.1 Estimadores pontuais

No modelo de regressão linear simples, o método clássico consiste em considerar a regressão de uma variável aleatória Y em função de outra variável aleatória X expressa por:

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \qquad \epsilon \backsim N\left(0, \sigma^2\right). \tag{2.1}$$

Para estimar um valor x_0 da variável independente, desconhecida, em função de Y_0, tem-se:

$$\hat{x}_0 = \frac{Y_0 - \hat{\beta}_0}{\hat{\beta}_1}, \tag{2.2}$$

em que os valores estimados de β_0 e β_1 são obtidos pelo método de mínimos quadrados ou pelo método de máxima verossimilhança. Vale ressaltar que sob a suposição de erros normais com média zero e variância constante, os estimadores dos parâmetros β_0 e β_1, obtidos pelo método de mínimos quadrados e pelo método de máxima verossimilhança, são iguais.

O método inverso para calcular o valor x_0, desconhecido, da variável X é obtido considerando a regressão de x em função de Y e é dado por:

$$x_i = \gamma_0 + \gamma_1 Y_i + \epsilon_i^*, \qquad \epsilon^* \backsim N\left(0,\sigma^2\right). \tag{2.3}$$

Nesse caso, o estimador de x_0 é obtido pela equação:

$$\hat{x}_0 = \hat{\gamma}_0 + \hat{\gamma}_1 Y_0, \tag{2.4}$$

em que $\hat{\gamma}_0$ e $\hat{\gamma}_1$ são estimadores de mínimos quadrados ou de máxima verossimilhança dos parâmetros γ_0 e γ_1.

Ambos estimadores pontuais (clássico e inverso) apresentam vantagens e desvantagens. Muitos estudos têm sido feitos para compreender e comparar as propriedades desses dois métodos, podendo citar Eisenhart (1939), Krutchkoff

(1967), Shukla (1972) e Williams (1969).

Eisenhart (1939) comparou as tabelas de análise de variância dos dois estimadores pontuais (clássico e inverso). Na Tabela 1 é apresentada a tabela análise de variância do estimador clássico e na Tabela 2 é apresentada análise de variância do estimador inverso.

Tabela 1 Análise de Variância para a variável independente considerando a regressão de Y em função de x, método clássico

Fonte de Variação	G.L.	Soma de Quadrados
Modelo	1	$\text{SQR}_{eg} = \hat{\beta}_1 \sum_{i=1}^{N}(x_i - \bar{x})(Y_i - \bar{Y})$
Erro	$n-2$	$\text{SQE}= \sum_{i=1}^{N}(Y_i - \hat{Y})^2$
Total	$n-1$	$\text{SQT}= \sum_{i=1}^{N}(Y_i - \bar{Y})^2$

Na Tabela 2, pode-se observar que os valores da variável x são fixados e estes valores não dependem dos valores observados de Y. A soma de quadrados da regressão, SQR_{eg}, representa a variabilidade dos valores da variável x escolhidos e esta variabilidade resulta da forma como eles são escolhidos. A SQR_{eg} mede a dependência de x em Y, mas esta é uma dependência falsa porque x não depende de Y . Finalmente, as SQR_{eg} não podem ser interpretadas como uma medida do erro dos valores de x, porque os valores da variável x são fixados e então não tem um erro. Eisenhart (1939) concluiu que, se os valores da variável x são selecionados e os correspondentes valores de Y são observados, então o estimador clássico é o estimador apropriado.

Tabela 2 Análise de Variância para a variável independente considerando a regressão de x em função de Y, método inverso

Fonte de Variação	G.L.	Soma de Quadrados
Modelo	1	$\text{SQR}_{eg} = \hat{\alpha_1} \sum_{i=1}^{N} (x_i - \bar{x})(Y_i - \bar{Y})$
Erro	$n-2$	$\text{SQE} = \sum_{i=1}^{N} (x_i - \hat{x})^2$
Total	$n-1$	$\text{SQT} = \sum_{i=1}^{N} (x_i - \bar{x})^2$

A abordagem clássica e a abordagem inversa, para o problema de calibração linear, foram comparadas em Krutchkoff (1967) pelo critério do erro quadrático médio (EQM) utilizando o procedimento de Monte Carlo. O autor concluiu que a abordagem inversa obteve um erro quadrático médio uniformemente menor que a abordagem clássica.

Krutchkoff (1969) apresenta alguns resultados obtidos ao utilizar estes procedimentos para extrapolação. Neste caso, conclui-se que existem situações de extrapolação em que o método clássico é melhor que o método inverso.

Williams (1969) discutiu em seu trabalho os resultados que foram obtidos por Krutchkoff (1967). De acordo com Williams (1969), Krutchkoff (1967) comparou o erro quadrático médio (EQM) de dois estimadores, um que tem EQM finito (método inverso) e outro que tem EQM infinito (método clássico). Estabeleceu-se o fato de que nenhum estimador não viesado tem variância finita. Por fim, Williams (1969) concluiu que o critério EQM mínimo não é adequado na comparação do método clássico com o método inverso.

Shukla (1972) também comparou os dois métodos. Segundo o autor se o número de observações é pequeno, então o estimador inverso tem menor EQM para a interpolação. Mas quando um grande número de observações é utilizado, então o estimador clássico é mais vantajoso. Se a pessoa não pode ter mais que uma observação sobre o valor desconhecido x_0 então o estimador inverso é preferível quando x_0 fica perto da média dos pontos. No entanto, se tem uma

grande amostra e nenhuma informação prévia sobre o valor desconhecido x_0, o estimador consistente (estimador clássico) é o preferido.

Naszódi (1978) considerou uma correção de viés para o estimador clássico de x_0 e após comparar o estimador corrigido e o estimador clássico, concluiu que o estimador corrigido é mais eficiente que estimador clássico, além de também ser consistente.

Segundo Thonnard (2006) a maioria dos estatísticos prefere o método clássico, devido a sua consistência e seu EQM. Além disso, os erros devem ser minimizados na direção em que ocorrem, ou seja, na direção de Y e, assim, indica-se o uso do método clássico.

2.2 Estimadores intervalares

Com o intuito de obter informações a respeito da precisão associada à estimativa obtida, muitos pesquisadores propõem estimadores intervalares para o verdadeiro valor, x_0, da variável independente, X. O intervalo de confiança e as regiões para o estimador clássico baseado na distribuição normal são abordados por Brown (1982), Eisenhart (1939), Fieller (1954), Graybill (1976), Lieberman, Miller e Hamilton (1967), Mathew e Kasala (1994), entre outros.

Eisenhart (1939) obteve a estimativa de x_0 considerando a regressão de Y em função de x e produziu uma estimativa intervalar para a variável independente com base na distribuição t de Student com $(n-2)$ graus de liberdade.

Lieberman, Miller e Hamilton (1967) consideraram intervalos simultâneos ilimitados em problemas de calibração e utilizaram mais de uma variável desconhecida, X, para obter os intervalos. Esses intervalos simultâneos ilimitados baseiam-se na estimativa de regressão linear clássica e têm a propriedade de que, pelo menos, $100P\%$ dos intervalos conterão os verdadeiros $x's$ com confiança $1-\alpha$, para algum P. Os autores apresentaram dois métodos para obtenção desses intervalos, o método de Bonferroni e uma técnica baseada em uma ideia de Lieberman e Miller (1963).

Graybill (1976) apresentou uma técnica para obtenção de um intervalo de confiança para o estimador clássico de x_0 . O teste t de Student foi utilizado para testar a hipótese nula de que β_1 é igual a zero. Se a hipótese nula não for rejeitada,

supõe-se que β_1 é zero e que não existe um intervalo de confiança. Por outro lado, se a hipótese nula for rejeitada, conclui-se que β_1 é diferente de zero e pode-se encontrar um intervalo de confiança. Porém o autor afirma que o coeficiente de confiança deste intervalo é inferior a $100\,(1-\alpha)\,\%$.

Sobre a construção de regiões de confiança para modelos multivariados encontram-se algumas referências tais com: Brown (1982), Brown e Sundberg (1987), Davis e Hayakawa (1987), Fujikoshi e Nishii (1984), Mathew e Kasala (1994), Mathew e Zha (1996), Oman (1988) e Williams (1959) .

A calibração multivariada controlada, descrita em Brown (1982), considera a calibração dos dados $Y_i\,(q\times 1)$ para valores fixados $X_i\,(p\times 1)\,,\ i=1,\ldots,n$. O autor propôs a obtenção de regiões $100\,(1-\alpha)\,\%$ de confiança exata para a variável independente, construídas a partir da distribuição t de Student multivariada.

Mathew e Sharma (2002) discutiram o problema de construir regiões de confiança para uma nova observação x_0, considerando a calibração multivariada. O problema abordado é a construção de regiões conjuntas para vários valores desconhecidos da variável explicativa. O problema é investigado quando a matriz de variâncias e covariâncias é um múltiplo escalar da matriz identidade e também é uma matriz positiva definida completamente desconhecida. São apresentados dois casos: no primeiro, a variável resposta e as variáveis explicativas têm as mesmas dimensões e as regiões de confiança são exatas, e no segundo a variável explicativa é um escalar e as regiões de confiança conjuntas são conservadoras. Os aspectos computacionais e da aplicação prática das regiões de confiança são discutidos e ilustrados por meio de exemplos.

Jose e Isaac (2007) consideraram um modelo multivariado para o problema de calibração multivariada controlada. Os autores construíram regiões de confiança conservadoras que não são vazias e invariantes sob transformações não singulares. As simulações realizadas mostraram a proximidade de abrangência das regiões de confiança para o nível de confiança adotado.

2.3 Regressão inversa não espacial

A regressão inversa tem sido utilizada em diversas áreas do conhecimento e, portanto, diversos autores buscam modelos apropriados para descreverem a

relação entre as variáveis de interesse. Na literatura, há vários estudos sobre o modelo estatístico que descreve a relação mais simples entre duas variáveis, ou seja, a equação da reta. O problema de calibração simples é bastante conhecido e tem sido discutido por Eisenhart (1939), Fieller (1954) e Graybill (1976), entre outros autores.

Em algumas situações, a relação entre a variável dependente (Y) e a variável independente (X) não é adequadamente modelada por uma linha reta. Então pode ser necessário ajustar um modelo polinomial. Alguns autores trabalharam com modelos polinomiais de segundo grau. A incerteza padrão combinada (variância) da variável independente, assim denominada por alguns autores, é calculada a partir da expansão da série de Taylor (KIRKUP; MULHOLLAND, 2004; OLIVEIRA; AGUIAR, 2009).

Nesse tópico, apresenta-se uma revisão sobre os modelos de calibração não espaciais tais como: linear simples, polinomial quadrática. Serão apresentados os estimadores pontuais e intervalares para cada um dos modelos apresentados considerando o método de estimação pontual clássico.

2.3.1 Regressão inversa linear simples

Considere os seguintes modelos, em que Y e x são relacionados por um modelo linear simples (GRAYBILL, 1976).

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, \ldots, n \tag{2.5}$$

$$Y_{0i} = \beta_0 + \beta_1 x_0 + \epsilon_{0i}, \quad i = n+1, \ldots, n+k, \tag{2.6}$$

em que, $\epsilon_1, \ldots, \epsilon_n$ e $\epsilon_{n+1}, \ldots, \epsilon_{n+k}$ são variáveis normais independentes e identicamente distribuídas com média zero e variância σ^2. Além disso, x_1, $\ldots$, x_n são constantes conhecidas e β_0, β_1, x_0, σ^2 são parâmetros desconhecidos.

A fim de obter maiores facilidades computacionais, pode-se considerar o modelo de regressão linear simples, centrado, para o qual a variável regressora x é redefinida como o desvio de sua própria média, $x_i - \bar{x}$, de tal modo que a

$Cov(\alpha_0, \alpha_1) = 0$. Assim o modelo se torna:

$$Y_i = \alpha_0 + \alpha_1 (x_i - \bar{x}) + \epsilon_i, \quad \epsilon_i \backsim N\left(0, \sigma^2\right), \tag{2.7}$$

em que $\alpha_1 = \beta_1$ e $\alpha_0 - \alpha_1 \bar{x} = \beta_0$.

2.3.1.1 Estimação Inversa

Para obter o estimador para x_0, segundo o método proposto por Graybill (1976), observa-se $k \geq 1$ valores de Y para um x_0 desconhecido. Sendo assim, tem-se uma amostra de tamanho $n + k$, em que x_0 é desconhecido e os valores x_i são distintos. Os k valores de Y, denotado por Y_{n+1},Y_{n+2}, . . . ,Y_{n+k}, tem distribuição normal com média $\alpha_0 + \alpha_1 (x_i - \bar{x})$ e variância σ^2.

A função de verossimilhança é dada por:

$$L\left(\alpha_0,\alpha_1, \sigma^2, x_0 : y_1,x_1, \ldots, y_n,x_n : y_{n+1}, \ldots ,y_{n+k}\right) = \left(\frac{1}{(2\pi\sigma^2)^{(n+k)/2}}\right) \times$$
$$exp\left\{-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n} (y_i - \alpha_0 - \alpha_1 (x_i - \bar{x}))^2 + \sum_{i=n+1}^{n+k} (y_i - \alpha_0 - \alpha_1 (x_0 - \bar{x}))^2\right]\right\}. \tag{2.8}$$

Para obter os estimadores de máxima verossimilhança dos parâmetros $\alpha_0, \alpha_1, \sigma^2$ e da variável independente x_0, deriva-se o *log* da função de verossimilhança, expressa em (2.8), em relação a cada um dos parâmetros e iguala-se a zero.

Portanto, os estimadores de máxima verossimilhança de α_1 e α_0, baseados nos primeiros n valores, $(x_1, Y_1), (x_2, Y_2), \ldots, (x_n, Y_n)$, são:

$$\hat{\alpha}_1 = \hat{\beta}_1 = \frac{\sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)(x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}, \tag{2.9}$$

$$\hat{\alpha}_0 = \hat{\beta}_0 + \hat{\alpha}_1 \bar{x} = \bar{Y}, \tag{2.10}$$

em que $\bar{Y} = \dfrac{\sum_{i=1}^{n} Y_i}{n}$ e $\bar{x} = \dfrac{\sum_{i=1}^{n} x_i}{n}$.

O estimador clássico de mínimos quadrados de x_0 baseado nas $n + k$ observações, é expresso por:

$$\hat{x}_0 = \bar{x} + \frac{\bar{Y}_0 - \hat{\alpha}_0}{\hat{\alpha}_1}, \tag{2.11}$$

em que $\bar{Y}_0 = \dfrac{\sum_{i=n+1}^{n+k} Y_{0i}}{n}$

Para estimar a variância σ^2, usa-se a função de verossimilhança baseada em todas $n + k$ observações. O estimador não viesado da variância é dado pela expressão:

$$\hat{\sigma}^2 = \frac{1}{n+k-3}\left(\sum_{i=1}^{n}\left(Y_i - \hat{\alpha}_0 - \hat{\alpha}_1\left(x_i - \bar{x}\right)\right)^2 + \sum_{i=n+1}^{n+k}\left(Y_i - \bar{Y}_0\right)^2\right). \tag{2.12}$$

As propriedades das distribuições dos estimadores de α_0, α_1, x_0, e σ^2 expressos nas equações (2.10), (2.9), (2.11) e (2.12), respectivamente, podem ser observadas pelo Teorema 1 (GRAYBILL, 1976).

Teorema 1: Considere o modelo de regressão linear simples, dado pela equação (2.5) e os estimadores de α_0, α_1, x_0, e σ^2 expressos pelas equações (2.10), (2.9), (2.11) e (2.12) respectivamente.

1. $\hat{\alpha}_0$ e $\hat{\alpha}_1$ são independentes;
2. $\hat{\alpha}_0 \backsim N\left(\alpha_0, \frac{\sigma^2}{n}\right)$ e $\hat{\alpha}_1 \backsim N\left(\alpha_1, \dfrac{\sigma^2}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}\right)$;
3. $U = (n + k - 3)\,\frac{\hat{\sigma}^2}{\sigma^2}$ segue a distribuição χ^2_{n+k-3};

4. $\hat{\sigma}^2$ é independente de $\hat{\alpha}_0, \hat{\alpha}_1$ e $\hat{x}_0$.

Com os resultados anteriores, pode-se discutir o intervalo de confiança para x_0 dado Y_0 em modelos de regressão linear simples, de acordo com Graybill (1976). Intuitivamente, pode-se observar que não há um intervalo de confiança útil para x_0 se α_1 é zero, porque neste caso, a linha de regressão linear simples é horizontal.

Primeiramente, precisa-se de algumas expressões, tais como:

$$\begin{aligned} E\left(\bar{Y}_0 - \hat{\alpha}_0 - \hat{\alpha}_1 (x_0 - \bar{x})\right) &= E\left(\bar{Y}_0\right) - E(\hat{\alpha}_0) - E(\hat{\alpha}_1 (x_0 - \bar{x})) \\ &= \frac{1}{k} \sum_{i=n+1}^{n+k} (\alpha_0 + \alpha_1 (x_0 - \bar{x}) - \alpha_0 - \alpha_1 (x_0 - \bar{x})) \\ &= 0, \end{aligned} \tag{2.13}$$

$$\begin{aligned} Var\left(\bar{Y}_0 - \hat{\alpha}_0 - \hat{\alpha}_1 (x_0 - \bar{x})\right) &= \sigma^2 \left(\frac{1}{k} + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right) \\ &= \sigma^2 A^2, \end{aligned} \tag{2.14}$$

em que $A^2 = \dfrac{1}{k} + \dfrac{1}{n} + \dfrac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$.

Tem-se que $\bar{Y}_0 - \hat{\alpha}_0 - \hat{\alpha}_1 (x_0 - \bar{x})$ é normalmente distribuído com média 0 e variância $\sigma^2 A^2$. Portanto, segue que:

$$\begin{aligned} Z &= \frac{\left(\bar{Y}_0 - \hat{\alpha}_0 - \hat{\alpha}_1 (x_0 - \bar{x})\right)}{\sqrt{Var\left(\bar{Y}_0 - \hat{\alpha}_0 - \hat{\alpha}_1 (x_0 - \bar{x})\right)}} \\ &= \frac{\left(\bar{Y}_0 - \hat{\alpha}_0 - \hat{\alpha}_1 (x_0 - \bar{x})\right)}{\sqrt{\sigma^2 A^2}} \backsim N(0, 1). \end{aligned} \tag{2.15}$$

Pelo Teorema 1, tem-se que:

$$\begin{aligned} U &= (n+k-3)\,\frac{\hat{\sigma}^2}{\sigma^2} \\ &= \frac{\sum_{i=1}^{n}\left(Y_i-\hat{\alpha}_0-\hat{\alpha}_1\left(x_0-\bar{x}\right)+\sum_{i=1}^{n}\left(Y_i-\bar{Y}_0\right)\right)^2}{\sigma^2} \backsim \chi^2_{(n+k-3)}. \end{aligned} \tag{2.16}$$

Pode-se mostrar que Z e U são independentes, porque $\hat{\sigma}^2$ é independente de $\hat{\alpha}_0$,$\hat{\alpha}_1$ e $\hat{x}_0$.

Assim, $z \backsim N\left(0,1\right)$, $U \backsim \chi^2_{n+k-3}$ e, uma vez que Z e U são independentes, tem-se que:

$$T = \frac{Z}{\sqrt{\frac{U}{n+k-3}}} \backsim t_{(\alpha/2;n+k-3)}. \tag{2.17}$$

Então, tem-se:

$$P\left(-t_{(\alpha/2;n+k-3)} \leq T \leq t_{(\alpha/2;n+k-3)}\right) = 1-\alpha \tag{2.18}$$

$$\Rightarrow T^2 \leq t^2_{(\alpha/2;n+k-3)} \tag{2.19}$$

$$\Rightarrow \frac{\left(\bar{Y}_0-\hat{\alpha}_0-\hat{\alpha}_1\left(x_0-\bar{x}\right)\right)^2}{\hat{\sigma}^2\left(\frac{1}{k}+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)} \leq t^2_{(\alpha/2;n+k-3)}. \tag{2.20}$$

Por fim, segue que:

$$\left(\bar{Y}_0 - \hat{\alpha}_0 - \hat{\alpha}_1 (x_0 - \bar{x})\right)^2 - \hat{\sigma}^2 \left(\frac{1}{k} + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right) t^2_{(\alpha/2;n+k-3)} \leq 0. \quad (2.21)$$

Expandindo o primeiro termo quadrático da expressão (2.21) tem-se a seguinte inequação:

$$\left(\hat{\alpha}_1^2 - \frac{\hat{\sigma}^2 t^2_{(\alpha/2;n+k-3)}}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right) (x_0 - x)^2 - 2\hat{\alpha}_1 (\bar{y}_0 - \hat{\alpha}_0)(x_0 - \bar{x}) + \\ + \left((\bar{y}_0 - \hat{\alpha}_0)^2 - \hat{\sigma}^2 t^2_{(\alpha/2;n+k-3)} \left(\frac{1}{k} + \frac{1}{n} \right) \right) \leq 0. \quad (2.22)$$

Pode-se observar que esta é uma desigualdade quadrática, a qual pode ser escrita da seguinte forma $q(x_0 - \bar{x}) = a(x_0 - \bar{x})^2 + 2b(x_0 - \bar{x}) + c \leq 0$, em que a, b e c podem ser identificados na inequação (2.22). Se os valores dos x_0, que satisfazem esta desigualdade é um intervalo, então, estes valores formam um intervalo de $100(1 - \alpha)\%$ de confiança para x_0 (GRAYBILL, 1976).

Portanto, de acordo com Graybill (1976), pode-se observar que se o discriminante $b^2 - ac$ de uma função quadrática é negativo, então a função quadrática não pode ser igual a zero (Figura 1b e Figura1d). Neste caso, há duas possíveis situações:

a) $q(x_0 - \bar{x}) < 0$ para todo x_0 (Figura 1b), então o intervalo de confiança é $-\infty < x_0 < +\infty$.

b) $q(x_0 - \bar{x}) > 0$ para todo x_0 (Figura 1d), então não há um intervalo de confiança.

Se o discriminante $b^2 - ac$ é positivo, então há também duas possibilidades (Figura 1a e Figura 1c):

a) $a > 0$, então os valores de x_0 para o qual $q\,(x_0 - \bar{x}) \leq 0$ formam um intervalo de confiança para x_0 (Figura 1a).

b) $a < 0$, os valores de x_0 em que $q\,(x_0 - \bar{x}) \leq 0$ formam dois intervalos infinitos (Figura 1c), o que não é útil.

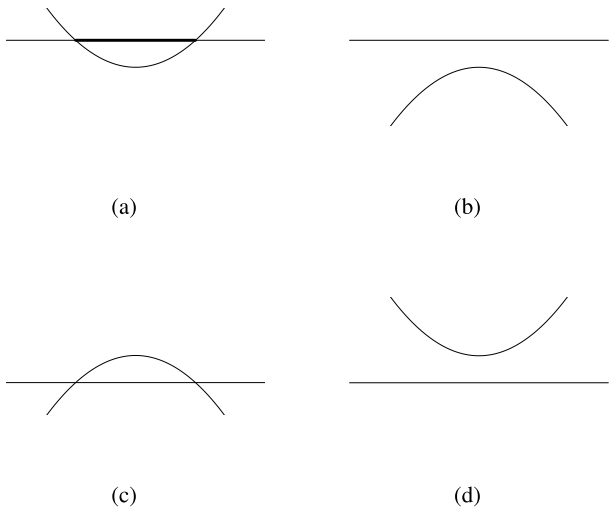


Figura 1 Quatro possibilidades de formas quadráticas

Para o caso em que $q(x_0 - \bar{x}) = 0$ se tem um único ponto. Neste caso, o intervalo de confiança para x_0 também não é útil.

Pode-se concluir que a desigualdade $q(x_0 - \bar{x}) \leq 0$ resulta em um intervalo de confiança para x_0 se, e somente se, $a > 0$ e $b^2 - ac > 0$. Expandindo $b^2 - ac$, tem-se:

$$b^2 - ac = \hat{\sigma}^2 t^2_{(\alpha/2;n+k-3)} \left(\left(\frac{1}{n} + \frac{1}{k} \right) a + \frac{\left(\bar{Y}_0 - \hat{\alpha}_0 \right)^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right). \tag{2.23}$$

Assim, se o discriminante $b^2 - ac \geq 0$ e $a \geq 0$, então (2.22) produz um

intervalo de confiança para x_0 , ou seja, se, e somente se, $\hat{\alpha}_1^2 - \frac{\hat{\sigma}^2 t^2_{\alpha/2;n+k-3}}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \geq 0$.

Simplificando, tem-se que $\frac{\hat{\alpha}_1^2 \sum_{i=1}^{n}(x_i - \bar{x})^2}{\hat{\sigma}^2} \geq t^2_{(\alpha/2;n+k-3)} = F_{(1;\alpha/2;n+k-3)}$ é um teste de tamanho α de hipóteses $H_0 : \alpha_1 = 0 \quad versus \quad H_a : \alpha_1 \neq 0$. Se α_1 está próximo de zero, a função linear é quase horizontal, portanto obtém-se um intervalo de confiança muito grande, o que não é útil.

Para obter um intervalo de confiança para x_0 usa-se o seguinte procedimento (GRAYBILL, 1976):

1. Obtém-se o estimador de x_0, o qual é dado por: $\hat{x}_0 = \bar{x} + \frac{\bar{Y}_0 - \hat{\alpha}_0}{\hat{\alpha}_1}$;

2. Realiza-se o teste: $H_0 : \alpha_1 = 0 \quad versus \quad H_a : \alpha_1 \neq 0$; rejeita H_0 se, e somente se: $\frac{\hat{\alpha}_1^2 \sum_{i=1}^{n}(x_i - \bar{x})^2}{\hat{\sigma}^2} \geq t^2_{(\alpha/2;n+k-3)} = F_{(1;\alpha/2;n+k-3)}$;

3. Se não se rejeita a hipótese H_0 , assume-se que o modelo é $y_i = \alpha_0 + \epsilon_i$. Portanto, não existe intervalo de confiança para x_0;

4. Se a hipótese H_0 é rejeitada, apresenta-se o limite inferior (LI) e o limite superior (LS) do intervalo de $100\,(1-\alpha)\,\%$ de confiança em x_0 é dado por:

$$LI = \bar{x} + \frac{\hat{\alpha}\,(\bar{y}_0 - \bar{y})}{a} - \frac{t_{(\alpha/2;n+k-3)}\hat{\sigma}}{a}\sqrt{a\left(\frac{1}{n} + \frac{1}{k}\right) + \frac{(\bar{y}_0 - \bar{y})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}, \tag{2.24}$$

$$LS = \bar{x} + \frac{\hat{\alpha}\,(\bar{y}_0 - \bar{y})}{a} + \frac{t_{(\alpha/2;n+k-3)}\hat{\sigma}}{a}\sqrt{a\left(\frac{1}{n} + \frac{1}{k}\right) + \frac{(\bar{y}_0 - \bar{y})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}, \tag{2.25}$$

em que $a = \hat{\alpha}_1^2 - \dfrac{\hat{\sigma}^2 t^2_{(\alpha/2;n+k-3)}}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \geq 0.$

Graybill (1976) afirma que este não é um intervalo de $100\,(1-\alpha)\,\%$ de confiança para x_0, mas que tem o coeficiente de confiança menor que $100\,(1-\alpha)\,\%$.

2.3.1.2 Exemplo

Para ilustrar o método de calibração na regressão linear simples, será utilizado o exemplo a seguir, descrito por Charnet et al. (1999). Toda a análise será realizada utilizando funções desenvolvidas no *software* R (R CORE TEAM, 2015). Essas funções foram disponibilizadas no anexo.

Em um estudo sobre o efeito do carbono contido em fios de aço (em porcentagem) utilizados em resistência elétrica (em μ ohms cm a 200^0C) foram obtidos os resultados apresentados na Tabela 3.

Tabela 3 Dados referentes ao efeito do carbono contido (em porcentagem) em fios de aço na resistência elétrica (μ ohms cm a 200^0C)

Carbono contido	Resistência (μ ohms cm a 200^0C)	Carbono contido	Resistência (μ ohms cm a 200^0C)
0,05	12,3	0,55	21,2
0,10	15,0	0,60	21,9
0,15	15,7	0,70	22,6
0,20	16,2	0,80	23,8
0,25	17,1	0,85	24,2
0,30	18,0	0,90	25,3
0,40	19,2	0,95	26,0
0,50	20,4		

Na Figura 2, pode-se observar que à medida que o carbono contido aumenta, a resistência aumenta aproximadamente na mesma proporção. Desta forma, supor uma relação linear entre as variáveis carbono contido e resistência é razoável.

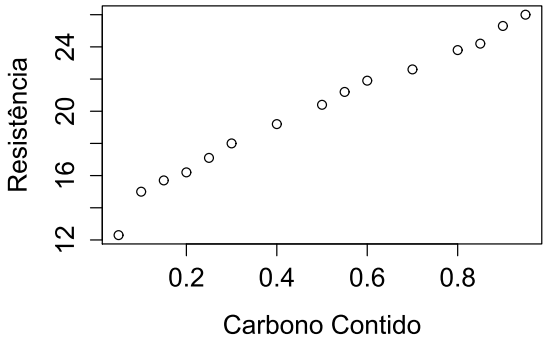


Figura 2 Gráfico de dispersão dos dados da Tabela 3

Suponha que esteja interessado em fazer estimações para um valor do carbono contido x_0, desconhecido, correspondente a um valor da resistência $Y = y_0$. Considera-se, por exemplo, um valor de resistência igual a $y_0 = 20\ \mu$ ohms cm a 200^0C. Portanto, deseja-se estimar o carbono contido x_0 dado o valor para a resistência y_0. Pode-se, então, obter o estimador pontual clássico e inverso para este exemplo.

Para obter o estimador inverso de x_0, ajusta-se o modelo de regressão linear simples de x em Y, o qual é dado por:

$$\hat{x} = -0{,}976533 + 0{,}073429Y.$$

Observa-se que para cada unidade de resistência tem-se um aumento médio de 0,073429% de carbono contido.

Logo, o estimador pontual inverso de x_0 para $y_0 = 20$ é dado por:

$$\tilde{x_0} = -0{,}976533 + 0{,}073429 \times 20 = 0{,}4920.$$

Portanto, para 20 unidades de resistência tem-se, em média, 0,4920% de carbono contido em fios de aços.

O intervalo de 95% de confiança para x_0, neste caso, é $[0{,}4665; 0{,}5176]$.

Para o cálculo do estimador clássico aplica-se o método proposto por Graybill (1976). Calcula-se o estimador pontual e intervalar. Em seguida, investiga-se a afirmação de que o intervalo proposto não é um intervalo de $100(1-\alpha)\%$ de confiança para x_0, mas tem o coeficiente de confiança menor que $100(1-\alpha)\%$. Essa afirmação é investigada por meio de uma simulação usando funções desenvolvidas no *software* R (R CORE TEAM, 2015) apresentadas no anexo.

Neste caso, considera-se o modelo de regressão linear simples de Y em x:

$$Y = \beta_0 + \beta_1 x + \epsilon, \quad \epsilon \backsim N\left(0, \sigma^2\right).$$

Os parâmetros são estimados por mínimos quadrados e são apresentados na Tabela 4. Para este modelo o coeficiente de determinação é dado por $R^2 = 0{,}9792$, ou seja, 97,92% da variação da resistência são explicados pelo modelo de regressão.

Tabela 4 Valores estimados dos parâmetros β_0 e β_1

Parâmetros	Valor estimado	Erro padrão	t calculado	valor-p
β_0	13,437	0,307	43,770	$1{,}68 \times 10^{-15}$
β_1	13,335	0,539	24,740	$2{,}56 \times 10^{-12}$

Assim, o modelo de regressão linear simples de Y em x é dado por:

$$\hat{Y} = 13{,}437 + 13{,}335x.$$

Na Tabela 5, de análise de variância do modelo de regressão, observa-se que o valor p é menor que 0,0001 o que significa que o parâmetro β_1 é estatisticamente significativo. Nota-se que a estimativa da variância do erro é de 0,381.

Tabela 5 Tabela de análise de variância para modelo de regressão linear simples para o estimador clássico

Fonte	G.l.	SQ	QM	F calculado	valor p
Modelo	1	233,372	233,372	611,97	$2{,}559 \times 10^{-12}$
Resíduo	13	4,957	0,381		
Total	14	238,329			

Em um modelo de regressão linear, os resíduos têm uma relação muito forte não somente com a qualidade do ajuste feito, mas também com a confiabilidade dos testes estatísticos (CHARNET et al., 1999). Portanto, a análise de resíduos é de grande importância na verificação da qualidade de ajuste dos modelos. Uma maneira de analisar os resíduos é a partir de alguns gráficos como o apresentado na Figura 3.

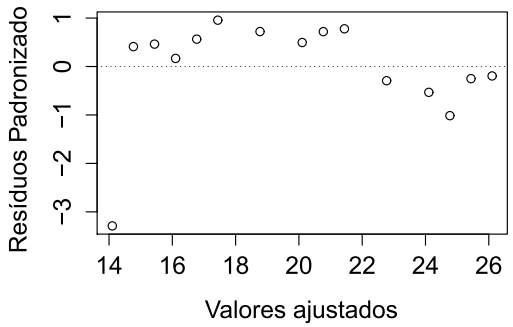


Figura 3 Gráfico dos resíduos de ajuste do modelo de regressão linear simples versus os valores ajustados de Y

Na Figura 3 é apresentado o gráfico dos resíduos em função dos valores ajustados de Y. Com disposição bem aleatória, não apresenta nenhum tipo de tendência aparente. Outra característica desse gráfico é que a faixa de variação dos resíduos ao longo dos valores ajustados de Y é constante, isso indica que provavelmente, nenhuma das suposições básicas dos modelos de regressão lineares simples esteja sendo violada. Porém, pode-se observar que o resíduo referente à primeira observação está muito afastado dos outros, portanto a primeira observação

é uma observação discrepante.

O Boxplot é um gráfico de um conjunto de dados que consiste de uma linha que se estende do valor mínimo ao valor máximo, em uma caixa com linhas verticais, traçadas no primeiro quartil (Q1), na mediana e no terceiro quartil (Q3). Os quartis, isto é, primeiro quartil, a mediana e o terceiro quartil são três valores que dividem os dados ordenados em quatro grupos com aproximadamente 25% dos valores em cada grupo. O boxplot é um gráfico utilizado para poder identificar os outlies (valores discrepantes), valores que são bastante incomuns, no sentido de estarem muito afastados da maioria dos dados.

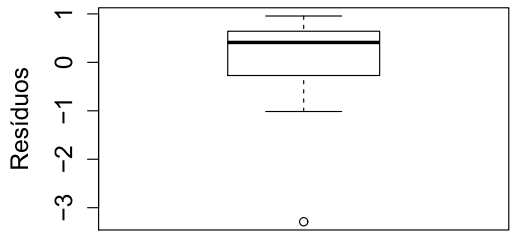


Figura 4 Boxplot dos resíduos

Segundo Charnet et al. (1999), quando uma ou mais observações discrepantes são incorporadas em um modelo de regressão, pode vir a prejudicar o ajuste do modelo. Portanto, fez-se a opção de retirar a primeira observação. Deve-se observar que a retirada de observações nem sempre é recomendada.

Após a retirada da primeira observação, ajusta-se novamente o modelo linear simples de Y em x é dado por:

$$\hat{Y} = 13{,}960 + 12{,}574x.$$

Na Tabela 6 de análise de variância do modelo de regressão, observa-se que o valor p é menor que 0,0001, o que significa que o parâmetro β_1 é estatisticamente significativo. A estimativa da variância do erro é de 0,069.

Tabela 6 Tabela de análise de variância para modelo de regressão linear simples para o estimador clássico, após retirada a primeira observação

Fonte	G.l.	SQ	QM	F calculado	valor p
Modelo	1	175,180	175,180	2534,1	$2{,}478 \times 10^{-15}$
Resíduo	12	0,830	0,069		
Total	13	176,010			

O gráfico dos resíduos não apresenta mais o valor discrepante (Figura 5a). Pelas Figuras 5a e 5b não há evidências de que alguma das suposições básicas dos modelos de regressão lineares simples esteja sendo violada. O que pode ser confirmado através de testes estatísticos de Shapiro-Wilk, Breusch-Pagan, Durbin-Watson.

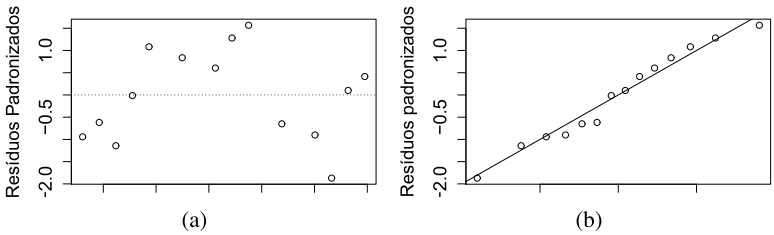


Figura 5 Gráficos de análise de resíduos para modelo de regressão linear simples

O estimador pontual clássico de x_0 dado $y_0 = 20$ é dado por:

$$\hat{x}_0 = \frac{y_0 - \hat{\beta}_0}{\hat{\beta}_1} = \frac{20 - 13{,}960}{12{,}574} = 0{,}480.$$

O intervalo de 95% de confiança para x_0, obtido de acordo com as expressões (2.24) e (2.25) é dado por [0,480; 0,528], conforme mostra a Figura 6.

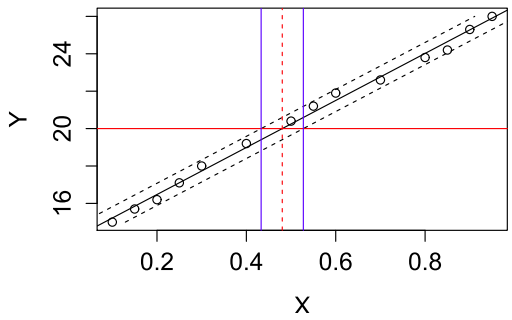


Figura 6 Intervalo de 95% de confiança para a porcentagem de carbono contido dado uma resistência de 20 μ ohms cm a 200^0C

Para os intervalos correspondentes a valores centrais de y, as amplitudes são menores e, à media que os valores se afastam da média, tem-se um aumento gradativo da amplitude. O comportamento dessas amplitudes pode ser observado nas expressões (2.24) e (2.25), visto que o termo $(\bar{y}_0 - \bar{y})^2$, que aparece dentro da raiz quadrada será igual a zero quando $\bar{y}_0 = \bar{y}$.

Segundo Graybill (1976) o intervalo de confiança proposto não é um intervalo de $100(1-\alpha)\%$ de confiança, mas tem confiança menor de $100(1-\alpha)\%$. Pode-se simular a confiança do intervalo. Essa simulação vai fornecer a confiança para o intervalo.

Considere o modelo de regressão linear centrado dado por:

$$Y = \alpha_0 + \alpha_1 (x - \bar{x}) + \epsilon, \quad \epsilon \backsim N(0,1).$$

Primeiramente, escolhe-se os n valores de Y , $Y_i, i = 1, \ldots, n$, aleatoriamente a partir de uma distribuição normal com média $Y = \alpha_0 + \alpha_1 (x_i - \bar{x})$ e desvio padrão σ. Depois, as k observações de y em x_0 devem ser escolhidos a partir de uma distribuição normal com média $Y = \alpha_0 + \alpha_1 (x_0 - \bar{x})$ e desvio padrão σ. Calculam-se as estimativas de α_0, α_1, σ^2 e x_0 . Então, calcula-se o intervalo de confiança proposto por Graybill (1976) . Há dois testes incluídos na simulação. O primeiro teste é o seguinte: $H_0 : \alpha_1 = 0$ vs $H_a : \alpha_1 \neq 0$. A hipótese H_0 é rejeitada se, e somente se $\frac{\alpha_1^2 \sum_{i=1}^{n} (x_i - \bar{x})^2}{\hat{\sigma}^2} \geq t^2_{\alpha/2;n+k-3}$. Se H_0 é rejeitada , então a simulação retorna 0 como resultado , caso contrário, retorna 1 . O outro teste verifica se

o verdadeiro valor de x_0 encontra-se no intervalo de confiança calculado. Se x_0 está no intervalo de simulação retorna 1, caso contrário 0. Este ciclo é repetido S vezes. Finalmente, a simulação tem dois resultados: o número de vezes em S que verdadeiro valor de x_0 está no intervalo de confiança e o número de vezes em S que H_0 é aceito. A confiança deste intervalo é o número de vezes que o verdadeiro valor de x_0 se encontra no intervalo de confiança menos o número de vezes que H_0 é aceito.

Repete-se 30 vezes a simulação com T = 100.000. Estes 30 valores são apresentados na Tabela 7.

Tabela 7 Confiança do intervalo apresentado pelas expressões (2.24) e (2.25)

94,87	94,77	95,27	95,41	95,43	95,15	95,02	95,31	94,99	95,21
94,71	94,65	94,88	94,85	95,13	94,78	94,85	94,58	95,03	94,86
94,81	94,96	94,87	95,10	95,19	94,64	95,24	94,82	94,87	94,95

Observa-se que a afirmação de Graybill (1976) , para este exemplo é verdadeira, pois o nível de confiança do intervalo é, em média, um pouco menor que 95%.

2.3.2 Regressão inversa quadrática

Kirkup e Mulholland (2004) discutiram o problema de calibração para dados univariados na regressão polinomial quadrática. Para os valores da variável independente $x_1, x_2, \ldots, x_n$, os valores da variável resposta são dados por $y_1, y_2, \ldots, y_n$, em que n é o número de pares de dados. Nesse caso, assume-se que a relação entre a variável independente, x, e a variável dependente, Y , é expressa por:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i, \qquad \epsilon_i \backsim N\left(0, \sigma^2\right). \tag{2.26}$$

Pode-se escrever este modelo na seguinte forma matricial:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \qquad \boldsymbol{\epsilon} \backsim NM\left(\mathbf{0}, \mathbf{I}\sigma^2\right), \tag{2.27}$$

em que:

$$\mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \qquad \mathbf{X} = \begin{bmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 \end{bmatrix}, \qquad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix},$$

e

$$\boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

sendo:
$\mathbf{Y}$ um vetor $n \times 1$ cujas componentes correspondem às n observações;
$\mathbf{X}$ uma matriz de dimensão $n \times 3$ denominada matriz de incidência;
$\boldsymbol{\beta}$ um vetor 3×1 cujos elementos são os parâmetros da regressão;
$\boldsymbol{\epsilon}$ um vetor de dimensão $n \times 1$ cujas componentes são os erros.

A equação (2.27) pode ser ajustada aos dados usando o método de mínimos quadrados ou o método de máxima verossimilhança.

Desta forma, tem-se que o estimador do vetor de parâmetros β_0, β_1 e β_2 é dado por:

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y}, \tag{2.28}$$

desde que a matriz $\left(\mathbf{X}^T\mathbf{X}\right)$ seja não singular, ou seja, tenha determinante diferente de zero e, portanto, invertível.

Pode-se obter a esperança e a variância para cada elemento de $\hat{\beta}$ simultaneamente. Assim, tem-se que:

$$\begin{aligned} E\left[\hat{\boldsymbol{\beta}}\right] &= E\left[\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y}\right] \\ &= \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T E\left[\mathbf{Y}\right] \\ &= \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} \\ &= \boldsymbol{\beta}, \end{aligned} \tag{2.29}$$

$$\begin{aligned} Var\left[\hat{\boldsymbol{\beta}}\right] &= Var\left[\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y}\right] \\ &= \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T Var\left[\mathbf{Y}\right]\left[\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\right]^T \\ &= \sigma^2\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1} \\ &= \sigma^2\left(\mathbf{X}^T\mathbf{X}\right)^{-1}. \end{aligned} \tag{2.30}$$

2.3.2.1 Estimação Inversa

Uma vez que tenha sido realizado o ajuste do modelo, com o objetivo de obter o estimador para x_0 , observa-se $k \geq 1$ valores de Y para um valor x_0 desconhecido. Dado a média dos valores observados de Y, $\bar{y}_0$, para um valor x_0, o estimador pontual clássico de mínimos quadrados de x_0 é obtido pela equação (2.31) (OLIVEIRA; AGUIAR, 2009):

$$\hat{x}_0 = \frac{\hat{\beta}_1 \pm \sqrt{\hat{\beta}_1^2 - 4\hat{\beta}_2\left(\hat{\beta}_0 - \bar{y}_0\right)}}{2\hat{\beta}_2}, \tag{2.31}$$

com a raiz sendo positiva quando a função y_i é crescente e negativa quando a função y_i é decrescente.

A variância, ou incerteza padrão combinada, de $\hat{x}_0$ pode ser calculada a partir da expansão em série de Taylor da equação (2.31) em torno do ponto $P\left(\beta_0, \beta_1, \beta_2, E\left(\bar{y}_0\right)\right)$. Considera-se a função $\hat{x}_0 = f\left(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \bar{y_0}\right)$, aplica-se a

variância, desprezam-se os termos de ordem superior e a correlação entre $\hat{y}$ e os coeficientes estimados, obtendo-se assim a equação (2.32) (OLIVEIRA; AGUIAR, 2009):

$$\begin{aligned} Var\left(\hat{x}_0\right) &= \left(\frac{\partial \hat{x}_0}{\partial \bar{y}_0}\right)^2 Var\left(\bar{y}_0\right) + \left(\frac{\partial \hat{x}_0}{\partial \hat{\beta}_0}\right)^2 Var\left(\hat{\beta}_0\right) + \\ &+ \left(\frac{\partial \hat{x}_0}{\partial \hat{\beta}_1}\right)^2 Var\left(\hat{\beta}_1\right) + \left(\frac{\partial \hat{x}_0}{\partial \hat{\beta}_2}\right)^2 Var\left(\hat{\beta}_2\right) + 2\left(\frac{\partial \hat{x}_0}{\partial \hat{\beta}_0}\frac{\partial \hat{x}_0}{\partial \hat{\beta}_1}\right)^2 Cov\left(\hat{\beta}_0, \hat{\beta}_1\right) + \\ &+ 2\left(\frac{\partial \hat{x}_0}{\partial \hat{\beta}_1}\frac{\partial \hat{x}_0}{\partial \hat{\beta}_2}\right)^2 Cov\left(\hat{\beta}_0, \hat{\beta}_2\right) + 2\left(\frac{\partial \hat{x}_0}{\partial \hat{\beta}_1}\frac{\partial \hat{x}_0}{\partial \hat{\beta}_2}\right)^2 Cov\left(\hat{\beta}_1, \hat{\beta}_2\right), \end{aligned} \tag{2.32}$$

em que as variâncias e as covariâncias dos parâmetros são obtidas pela equação (2.30) e a variância de uma resposta y_0 é estimada pela variância da regressão. Se $\bar{y}_0$ é a média de k observações dependentes, tem-se:

$$\hat{\sigma}^2_{\bar{y}_0} = \frac{\hat{\sigma}^2}{k}. \tag{2.33}$$

em que:

$$\hat{\sigma}^2 = \frac{\mathbf{Y}^T\mathbf{Y} - \hat{\beta^T}\mathbf{X}^T\mathbf{Y}}{n-3}. \tag{2.34}$$

Os termos correspondentes às derivadas parciais de $\hat{x}_0$ em relação a cada um dos parâmetros da equação (2.32) são expressos por:

$$\frac{\partial \hat{x}_0}{\partial \beta_0} = \frac{-1}{\sqrt{\beta_1^2 - 4\beta_2\left(\beta_0 - \bar{y}_0\right)}}; \tag{2.35}$$

$$\frac{\partial \hat{x}_0}{\partial \beta_1} = \frac{-1 + \beta_1 / \sqrt{\beta_1^2 - 4\beta_2\left(\beta_0 - \bar{y}_0\right)}}{2\beta_2}; \tag{2.36}$$

$$\frac{\partial \hat{x}_0}{\partial \beta_2} = \frac{\beta_1 - \sqrt{\beta_1^2 - 4\beta_2 (\beta_0 - \bar{y}_0)}}{2\beta_2^2} - \frac{(\beta_0 - \bar{y_0})}{\beta_2 \sqrt{\beta_1^2 - 4\beta_2 (\beta_0 - \bar{y}_0)}}; \quad (2.37)$$

$$\frac{\partial \hat{x}_0}{\partial \bar{y}_0} = \frac{1}{\sqrt{\beta_1^2 - 4\beta_2 (\beta_0 - \bar{y}_0)}}. \quad (2.38)$$

Para obter um intervalo de $100\,(1-\alpha)\,\%$ de confiança para uma nova observação x_0, o erro padrão é multiplicado pelo quantil $(1-\alpha/2)$ da distribuição t de Student bicaudal para os graus de liberdade da calibração $(n-3)$ (KIRKUP; MULHOLLAND, 2004).

$$LI = \hat{x}_0 - t_{(1-\alpha/2,n-3)} \sqrt{Var\,(\hat{x}_0)}, \quad (2.39)$$

$$LS = \hat{x}_0 + t_{(1-\alpha/2,n-3)} \sqrt{Var\,(\hat{x}_0)}. \quad (2.40)$$

2.3.2.2 Exemplo

Para ilustrar o método de regressão linear quadrática será utilizado o exemplo apresentado em Charnet et al. (1999). Nesse exemplo, um experimento foi realizado para estudar a relação entre o grau de corrosão de certo metal e o tempo de exposição (em semanas) desse metal à ação da acidez do solo. Foram obtidos os resultados que podem ser observados na Tabela 8. Toda a análise foi realizada utilizando funções desenvolvidas no *software* R (R CORE TEAM, 2015), apresentadas no anexo.

Tabela 8 Dados referentes à relação entre o grau de corrosão de certo metal e o tempo de exposição (em semanas) desse metal à ação da acidez do solo

Tempo de exposição (semanas)	Grau de corrosão	Tempo de exposição (semanas)	Grau de corrosão
1	0,08	6	1,30
2	0,18	7	1,95
3	0,32	8	2,80
4	0,53	9	3,90
5	0,88	10	4,60

O diagrama de dispersão das observações do grau de corrosão versus o tempo de exposição é apresentado na Figura 7.

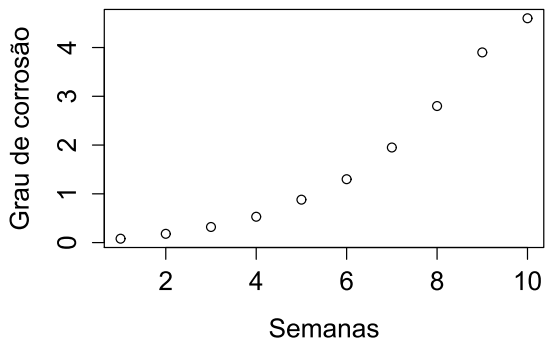


Figura 7 Gráfico de dispersão dos dados da Tabela 8

De acordo com a Figura 7, observa-se que, à medida que o tempo de exposição aumenta, o grau de corrosão também aumenta, mas não em uma mesma proporção.

Primeiramente, ajustou-se um modelo linear simples que não apresentou um bom ajuste aos dados, de acordo com a análise de resíduos. Em seguida, ajustou-se o modelo quadrático, obtendo um melhor ajuste aos dados. Por fim, calculou-se a estimativa pontual para variável independente desconhecida, x_0, e o respectivo intervalo de confiança.

A equação linear simples ajustada por mínimos quadrados foi:

$$\hat{y} = -1{,}1393 + 0{,}5079x.$$

Para o modelo de regressão linear simples, o coeficiente de determinação obtido é dado por $R^2 = 0{,}8919$. Na Tabela 9 apresenta-se a análise de variância do modelo de regressão linear simples, onde observa-se que o valor p é menor que 0,0001, indicando que o parâmetro β_1 é estatisticamente diferente de zero.

Tabela 9 Tabela de análise de variância para o modelo de regressão linear simples para o estimador clássico

Fonte	G.l.	SQ	QM	F calculado	p-valor
Modelo	1	21,2801	21,2801	75,271	$2{,}425 \times 10^{-05}$
Resíduo	8	2,2617	0,2827		
Total	9	23,5418			

Para verificar a qualidade de ajuste do modelo, pode-se fazer uma análise gráfica dos resíduos conforme é apresentado na Figura 8.

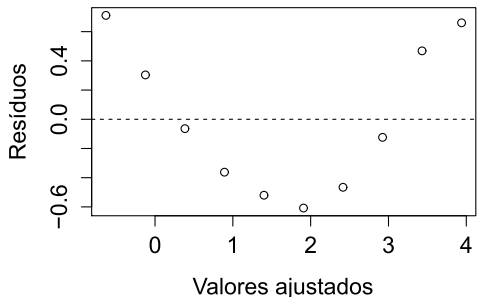


Figura 8 Gráfico de análise de resíduos para modelo de regressão linear simples

O padrão bem definido da Figura 8 indica que a relação entre o tempo de exposição e o grau de corrosão não deve ser linear simples, mas sim quadrática.

Portanto, incorporou-se a esse modelo um termo quadrático, ou seja, considerou-se o modelo de regressão linear quadrático, dado por:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i \quad \epsilon_i \backsim N(0, \sigma^2).$$

Ajustando-se esse modelo aos dados, obtém-se:

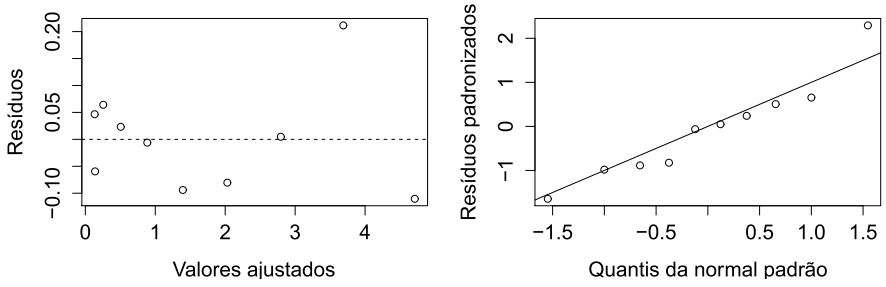
$$\hat{y} = 0{,}2740 - 0{,}1988x + 0{,}0642x^2.$$

Para o modelo de regressão linear quadrático o coeficiente de determinação foi $R^2 = 0{,}9955$. Na Tabela 10 de análise de variância do modelo de regressão, observa-se que o termo quadrático é significativo, pois o valor-p é menor que 0,0001.

Tabela 10 Tabela de análise de variância para modelo de regressão linear simples para o estimador clássico

Fonte	G.l.	SQ	QM	F calculado	valor-p
Efeito Linear	1	21,2801	21,2801	1803,06	$1{,}048 \times 10^{-09}$
Efeito quadrático	1	2,1791	2,1791	184,63	$2{,}750 \times 10^{-06}$
Resíduo	7	0,0826	0,0118		
Total	9	23,5418			

A Figura 9a e a Figura 9b sugere um melhor ajuste do modelo uma vez que o gráfico dos resíduos em função dos valores ajustados não apresenta nenhum tipo de tendência aparente. Também não há evidências contra a normalidade dos resíduos.



(a) (b)

Figura 9 Gráficos de análise de resíduos para modelo de regressão linear simples quadrática

Agora, suponha que se tem uma nova observação para o grau de corrosão do metal, dado por $y_0 = 2{,}0$ e deseja-se estimar o tempo de exposição (em semanas), x_0, desconhecido.

O estimador pontual clássico de x_0 dado $y_0 = 2{,}0$ é dado por:

$$\begin{aligned} \hat{x}_0 &= \frac{-\hat{\beta}_1 + \sqrt{\hat{\beta}_1^2 - 4\hat{\beta}_2(\hat{\beta}_0 - \bar{y}_0)}}{2\hat{\beta}_2} \\ &= \frac{0{,}1988 + \sqrt{0{,}0395 - 4 \cdot 0{,}0642\,(0{,}2740 - 2)}}{2 \cdot 0{,}0642} \\ &= 6{,}9564. \end{aligned}$$

Portanto, o tempo de exposição esperado é de, aproximadamente, 6,96 semanas dado um grau de corrosão igual a 2,0 .

Pode-se calcular o intervalo de confiança para esta nova medida obtida x_0 . Primeiramente, estima-se a variância de $\hat{x}_0$, dado por $\hat{V}(\hat{x}_0)$, expressa pela equação (2.32), a qual foi $\hat{V}(\hat{x}_0) = 0{,}02924$.

Considerando um nível de significância 5%, obtém-se o intervalo de confiança para x_0 , expresso pelas equações (2.39) e (2.40), dado por [5,5854; 8,3276], conforme mostra a Figura 10.

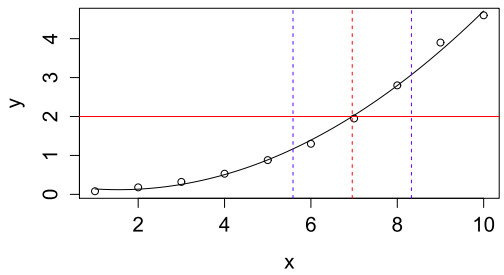


Figura 10 Intervalo de 95% de confiança para o tempo de exposição dado um grau de corrosão igual 2,0

2.4 Estatística espacial

A análise espacial pode ser definida como o estudo quantitativo de fenômenos distribuídos no espaço. Portanto, a ideia central é incorporar informações espaciais à análise que se deseja fazer. Desse modo, em análise espacial, utiliza-se de técnicas que explicitamente incorporam as localizações ou arranjos espaciais dos objetos ou fenômenos em questão (DINIZ, 2000).

Um exemplo, citado por Druck et al. (2004), em que se incorpora o espaço às análises realizadas, é quando se deseja investigar se existe alguma concentração espacial na distribuição de roubos. Deseja-se saber se os roubos que ocorrem em determinadas áreas estão correlacionados com características sócio-econômicas dessas áreas. Pode-se notar que a análise espacial está presente, uma vez que a localização relativa das características sócio-econômicas influenciam o número de roubos e estão sendo exploradas na análise.

Segundo Diniz (2000), a análise espacial produz resultados diferentes das abordagens não espaciais, uma vez que os seus resultados, em geral, são mais robustos por incorporarem a dimensão espacial.

De acordo com Cressie (1991), pode-se caracterizar a análise espacial considerando três tipos de dados:

a) Dados de eventos (ou padrões pontuais)- fenômenos em que se conhece a localização exata no espaço, denominados processos pontuais. Esses dados são representados por pontos distribuídos, normalmente, numa

superfície bidimensional. Por exemplo: localizações de crimes, localizações de casos de doenças e localizações de uma espécie vegetal.

b) Dados de superfícies contínuas (ou geoestatísticos) - são caracterizados pela continuidade espacial da variável aleatória de interesse. Em geral, esse tipo de dado é resultante de levantamento de recursos naturais que incluem mapas geológicos, topográficos, ecológicos, fitogeográficos e pedológicos.

c) Dados de áreas com contagens e taxas agregadas (ou *lattice*) - são caracterizados por constituírem valores agregados de uma determinada variável numa região ou mapa que se encontra dividido em sub-áreas. Podem ser dados associados a levantamentos populacionais, tais como o número de indivíduos doentes por município de uma região ou dados associados a *pixels* de imagens de microscopia ou satélites.

Independentemente do tipo de dado coletado, a análise espacial apresenta um conjunto de procedimentos cuja finalidade é a escolha de um modelo inferencial que considere explicitamente os relacionamentos espaciais presentes no fenômeno. Nesta tese aborda-se, exclusivamente, dados de áreas.

2.5 Análise espacial de dados de áreas

Na análise de dados de área, trabalha-se com eventos agregados cuja localização está associada a áreas delimitadas por polígonos fechados. Na prática, as divisões geográficas que resultam nas áreas são, em geral, de caráter político e geofísico, geralmente, caracterizadas por bairros, municípios, estados ou setores censitários e *pixels*.

A forma inicial de apresentação de dados de áreas é o uso de mapas coloridos do fenômeno de interesse na região. Se houver padrão espacial, espera-se encontrar cores próximas geograficamente. Esses mapas são, geralmente, úteis para uma análise exploratória. No estudo em que envolvem dados de áreas, as técnicas de análise exploratória são importantes ferramentas que auxiliam no desenvolvimento da modelagem estatística espacial. Tais técnicas contribuem para a

visualização e extração de informações.

2.5.1 Estrutura de dependência espacial

De acordo com Druck et al. (2004), a dependência espacial é um conceito importante na compreensão e análise de fenômenos espaciais. Essa afirmação se baseia na citação de Tobler (1970), sendo denominada de Primeira Lei da Geografia: todos os elementos no espaço estão relacionados, porém elementos mais próximos no espaço estão mais relacionados. Também, na afirmação de Cressie (1991) de que a dependência espacial está presente em todas as direções, mas fica mais fraca à medida que aumenta a dispersão na localização dos dados.

De um modo geral, a maior parte dos eventos apresenta uma relação que depende da distância entre si. Como exemplo, se for encontrada poluição em uma amostra de uma parte de um lago, é provável que locais próximos a essa amostra também estejam poluídos.

Segundo Waller e Gotway (2004), na análise de dados de área, o grau de dependência espacial ou similaridade é avaliado através da autocorrelação espacial que pode ser medida através do índice de Moran. A aplicação desse índice depende da definição de uma matriz de vizinhança ou matriz de proximidade espacial.

2.5.2 Matriz de proximidade espacial

A matriz de proximidade espacial é uma ferramenta básica para estimar a variabilidade espacial de dados de área. Vários são os critérios usados para definir a matriz de vizinhança. Dado um conjunto de n áreas $\{A_1, \ldots, A_n\}$, constrói-se a matriz $W(nxn)$, em que cada um dos elementos w_{ij} é um indicador de proximidade entre A_i e A_j. Esse indicador de proximidade pode ser calculado a partir de um dos seguintes critérios (WALLER; GOTWAY, 2004):

a) $w_{ij} = 1$, se o centróide de A_i está a uma determinada distância de A_j caso contrário $w_{ij} = 0$;

b) $w_{ij} = 1$, se A_i compartilha um lado comum com A_j, caso contrário $w_{ij} = 0$;

c) $w_{ij} = l_{ij}/l_i$, em que l_{ij} é o comprimento da fronteira entre A_i e A_j e l_i é o perímetro de A_i.

Alguns autores tais como Cressie (1991), Druck et al. (2004) e Waller e Gotway (2004) recomendam a normalização das linhas da matriz W dividindo cada elemento da matriz pelo total da linha e assim, obtém-se a soma dos pesos de cada linha igual a 1. Este procedimento simplifica vários cálculos de índices de autocorrelação espacial.

2.5.3 Autocorrelação espacial

Na estatística clássica o conceito de correlação diz respeito ao relacionamento entre duas variáveis. Se essas variáveis são correlacionadas usa-se como medida o coeficiente de correlação de Pearson (r), que mede o grau de correlação linear entre as variáveis. A informação que se busca através do cálculo da autocorrelação espacial é de quanto o valor de uma variável em uma área e o seu vizinho mais próximo são parecidos e quão diferentes do vizinho mais distante. A autocorrelação espacial ocorre quando observações organizadas no espaço influenciam-se mutuamente.

A autocorrelação espacial pode ser medida por meio de diversas técnicas tais como o índice global (e local) de Moran, o índice de Geary e o correlograma. Em análise de dados de áreas, o índice de Moran é o mais utilizado.

2.5.3.1 Índice de Moran

O índice global de Moran é calculado, conforme Waller e Gotway (2004), por:

$$\hat{I} = \frac{n}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\left(Y_i - \bar{Y}\right)\left(Y_j - \bar{Y}\right)}{\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2}, \tag{2.41}$$

em que:
n é o número de áreas ou de observações;
Y_i é a variável aleatória na área i;
Y_j é a variável aleatória na área j;
$\bar{Y}$ é a média amostral da variável aleatória em toda região, dada por $\bar{Y} = \frac{\sum_{i=1}^{n} Y_i}{n}$;
w_{ij} são os elementos da matriz de proximidade espacial normalizada nas linhas.

De acordo com Waller e Gotway (2004), o índice global de Moran pode assumir qualquer valor no conjunto dos reais. Porém, na maior parte dos casos este se encontra no intervalo $[-1; 1]$. Um valor do índice global de Moran próximo de zero indica ausência de autocorrelação espacial. Valores positivos indicam que valores da variável em áreas vizinhas tendem a ser similares entre si, enquanto valores negativos indicam dissimilaridade entre os valores dessa variável em áreas vizinhas.

Após calcular o índice global de Moran, é importante realizar inferências a respeito do valor encontrado, deve-se submeter tal valor a um teste de significância. Para avaliar a significância do índice, será preciso associá-lo a uma distribuição amostral. O mais comum é considerar que os dados são resultados da realização de uma distribuição normal.

Sob a suposição de normalidade, de acordo com Cliff e Ord (2004), o valor esperado da estatística de Moran na ausência de autocorrelação espacial é dado por:

$$E\left(\hat{I}\right) = -\frac{1}{n-1}, \tag{2.42}$$

em que se aproxima de zero quando n aumenta. E a variância é expressa por:

$$Var\left(\hat{I}\right) = -\frac{n^2 S_1 - nS_2 + 3S_0^2}{(n-1)(n+1)S_0^2} - \left(\frac{1}{n-1}\right)^2, \tag{2.43}$$

em que $S_0 = \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}$, $S_1 = 1/2\sum_{i=1}^{n}\sum_{j=1}^{n}\left(w_{ij} + w_{ji}\right)^2$, $S_2 = \sum_{i=1}^{n}\left(w_{i+} + w_{+j}\right)^2$,

com $w_{i+} = \sum_{j=1}^{n} w_{ij}$ e $w_{+j} = \sum_{i=1}^{n} w_{ij}$.

Com base no teste de Wald, a significância da estatística de Moran pode ser avaliada através da estatística teste expressa por:

$$z = \frac{\hat{I} - E\left(\hat{I}\right)}{\sqrt{Var\left(\hat{I}\right)}}. \tag{2.44}$$

O valor de z obtido na equação (2.44), corresponde a um quantil da distribuição normal padronizada, que corresponde a um determinado valor-p. O índice de Moran será considerado significativo se o valor-p for inferior ao valor nominal de significância previamente estabelecido. Alternativamente, pode-se obter o valor-p através de computação intensiva utilizando um teste de permutação.

2.6 Modelo espacial autorregressivo SAR

De acordo com Ywata e Albuquerque (2011), o modelo autorregressivo espacial, também conhecido como modelo SAR, é um dos modelos mais utilizados na modelagem de correlação espacial em dados de área. Assim como nos modelos AR (autorregressivos) em séries temporais, no modelo SAR incorpora-se um termo de *lag* entre os regressores da equação nos modelos lineares.

Uma das representações do modelo SAR, apresentada por Anselin (1999), é dada por:

$$Y = X\beta + \rho WY + \epsilon, \tag{2.45}$$

em que:
Y é o vetor $(n \times 1)$ de variáveis dependentes nas n áreas;
X é a matriz $(n \times (p+1))$ de variáveis independentes, com p variáveis explicativas;
β é o vetor de parâmetros$((p+1) \times 1)$;
ρ é o coeficiente espacial autorregressivo;
W é a matriz de proximidade espacial;
ϵ é um vetor $(n \times 1)$ de erros aleatórios não correlacionados que seguem uma

distribuição normal com média zero e variância constante, ou seja, $\epsilon \backsim N\left(0, I\sigma^2\right)$.

Em termos de componentes individuais, o modelo SAR pode ser expresso por:

$$y_i = \beta_0 + \sum_{i=1}^{p} x_i\beta_i + \rho\left(\sum_{j=1}^{n} w_{ij}y_j\right) + \epsilon_i. \tag{2.46}$$

2.6.1 Estimação dos parâmetros do modelo SAR

Segundo Ywata e Albuquerque (2011), a estimação dos parâmetros no modelo SAR, pelo método de mínimos quadrados ordinários, produz estimativas inconsistentes. Portanto, pode-se utilizar o método da máxima verossimilhança, a partir da hipótese de que o vetor de resíduos ϵ possui distribuição normal multivariada com média zero e covariância $\sigma^2 I$.

De acordo com Anselin (1999), outra parametrização do modelo SAR é expressa por:

$$Y = (I - \rho W)^{-1} X\beta + (I - \rho W)^{-1} \epsilon. \tag{2.47}$$

No modelo (2.47) o vetor de variáveis observadas possui distribuição (condicional a X) normal multivariada, com média condicional:

$$E(Y) = (I - \rho W)^{-1} X\beta \tag{2.48}$$

e matriz de variância condicional:

$$VAR(Y) = \sigma^2 (I - \rho W)^{-1} \left(I - \rho W^T\right)^{-1} = \sigma^2\Omega. \tag{2.49}$$

Então, a partir da distribuição de Y, obtém-se a função de log-verossimilhança $ln\ L\left(\rho, \beta, \sigma^2\right)$ expressa pela equação (2.50).

$$\begin{aligned} lnL\left(\rho,\beta,\sigma^2|Y,X\right) &= -\frac{n}{2}ln\left(2\pi\right) - \frac{n}{2}ln\left(\sigma^2\right) + ln\left(I-\rho W\right) - \\ &- \left(\frac{1}{2\sigma^2}\right)\left(Y-\rho WY-X\beta\right)^T\left(Y-\rho WY-X\beta\right). \quad (2.50) \end{aligned}$$

Maximizando-se a função de log-verossimilhança, em relação aos parâmetros do modelo, encontram-se as estimativas para os coeficientes e para a variância dos resíduos. O Jacobiano da transformação de ϵ para y é dado por:

$$J = \left|\frac{\partial\epsilon}{\partial y}\right| = \left|I-\rho W\right|. \quad (2.51)$$

De acordo com Ord (1975), a estimação por máxima verossimilhança de um modelo espacial autorregressivo consiste em explorar a decomposição do Jacobiano $|I-\rho W|$, em termos de autovalores da matriz W . Assim:

$$ln\left|I-\rho W\right| = ln\left[\prod_{i=1}^{n}\left(1-\rho\lambda_i\right)\right], \quad (2.52)$$

em que λ_i são os autovalores da matriz W. A principal dificuldade está em encontrar a estimativa do parâmetro ρ que deve ser estimado por métodos iterativos.

Para obter um estimador para os parâmetros do modelo SAR, deriva-se a função log-verossimilhança representada pela equação (2.50) em relação aos parâmetros e iguala-se a zero, resolvendo o sistema de equações resultantes.

Assim, o estimador de máxima verossimilhança dos parâmetros de β, é dado por:

$$\hat{\beta} = \left(X^TX\right)^{-1}X^T\left(I-\rho W\right)Y. \quad (2.53)$$

Pode-se mostrar que o estimador de β é não viesado, pois:

$$\begin{aligned} E\left[\hat{\beta}\right] &= E\left[\left(X^TX\right)^{-1}X^T\left(I-\rho W\right)Y\right] \\ &= \left(X^TX\right)^{-1}X^T\left(I-\rho W\right)\left[\left(I-\rho W\right)^{-1}X\beta\right] \\ &= \beta. \quad (2.54) \end{aligned}$$

A variância de $\hat{\beta}$ é expressa por:

$$\begin{aligned} Var\left[\hat{\beta}\right] &= Var\left[\left(X^T X\right)^{-1} X^T (I - \rho W) Y\right] \\ &= \left[\left(X^T X\right)^{-1} X^T (I - \rho W)\right]^T \left[(I - \rho W)^{-1} X\beta\right] Var[Y] \\ &= \left(X^T X\right)^{-1} \sigma^2. \end{aligned} \tag{2.55}$$

Para estimar o parâmetro σ^2 deve-se derivar a função (2.50) em relação ao parâmetro σ^2 e igualar a zero. Então, segue que:

$$\widehat{\sigma}^2 = \frac{\left(((I - \rho W) y - x\beta)^T ((I - \rho W) y - x\beta)\right)}{n}. \tag{2.56}$$

Pode-se mostrar que o estimador de σ^2 é viesado, pois:

$$\begin{aligned} E\left[\widehat{\sigma}^2\right] &= E\left[\frac{\left(((I - \rho W) y - X\beta)^T ((I - \rho W) y - X\beta)\right)}{n}\right] \\ &= \frac{1}{n} E\left[\left(((I - \rho W) y - X\beta)^T ((I - \rho W) y - X\beta)\right)\right] \\ &= \frac{1}{n} E\left[\epsilon^T \epsilon\right] = \frac{1}{n} E[S\,QR]. \end{aligned} \tag{2.57}$$

em que $S\,QR$ é a soma de quadrados dos resíduos.

Portanto, o estimador não viesado de σ^2 é dado por:

$$\widehat{\sigma}^2 = \frac{S\,QR}{n - 2tr(S) + tr(S^T S)}, \tag{2.58}$$

em que

$$S = \left(\rho W + X\left(X^T X\right)^{-1} X^T (I - \rho W)\right). \tag{2.59}$$

3 MATERIAL E MÉTODOS

Neste capítulo e no próximo é apresentado o problema de calibração espacial e são propostos métodos para a estimação pontual e intervalar para um valor x_0, da variável independente considerando o modelo autorregressivo SAR. Os estimadores obtidos foram aplicados a um conjunto de dados descrito em Anselin (1988) que está disponível na biblioteca "spdep" do software R (R CORE TEAM, 2015).

Todos os cálculos e análises estatísticas são realizados utilizando funções desenvolvidas no software R (R CORE TEAM, 2015) que estão disponíveis no anexo.

3.1 Regressão inversa espacial

Para obter uma estrutura da modelagem da variável resposta *Y*, adequada ao contexto da calibração e levando em consideração a dependência espacial, supõem-se que a variável dependente Y é função de uma única variável independente x, (fixa e conhecida) na primeira etapa do processo de calibração. Na segunda etapa, assume-se que o valor y_0 seja função do valor x_0, desconhecido.

Para calcular os valores estimados x_0, da variável independente considerando o modelo de regressão espacial SAR, são desenvolvidas equações de estimação considerando dois casos: no primeiro caso, pretende-se estimar o valor da variável independente pertencente à amostra observada, no segundo caso, a variável a ser estimada não pertence à amostra observada.

3.1.1 Ajuste do modelo

Na primeira etapa do processo de calibração, ajusta-se o modelo espacial SAR. Inicialmente constrói-se a matriz de proximidade espacial $W(n \times n)$ segundo algum critério de acordo com a secção 2.3.3. Em seguida, estimam-se os parâmetros ρ, β_0, β_1, do modelo SAR conforme a secção 2.4.1. Após o ajuste do modelo, obtém-se os estimadores para um valor da variável independente, dado um valor y_0 da variável dependente.

3.1.2 Estimação pontual

As equações de estimação pontual são desenvolvidas com base na abordagem clássica apresentada em 2.1, ou seja, considera-se a regressão de Y em função de x, considerando o modelo SAR.

A partir dessas equações, são calculados os valores estimados pontuais da variável independente utilizando o conjunto de dados descrito por Anselin (1988).

3.1.3 Estimação intervalar

Para construir o intervalo de confiança para x_0, com base na distribuição t de Student, toma-se como base o que foi apresentado por Graybill (1976), no caso da regressão linear simples, discutido na seção 2.3.1.

A partir das expressões para o intervalo de confiança, são calculados os intervalos de confiança para os valores x_0, da variável independente utilizando o conjunto de dados descrito por Anselin (1988).

4 RESULTADOS E DISCUSSÃO

4.1 Regressão inversa espacial

O número de aplicações práticas envolvendo análises de dados espacialmente distribuídos em uma determinada área tem crescido em diversos campos do conhecimento, tais como epidemiologia, ecologia, agronomia, demografia e geologia. Diferentes modelagens podem ser determinadas conforme o tipo de problema que se deseja estudar. A modelagem abordada neste trabalho leva em consideração a divisão do espaço estudado em n regiões poligonais. Para cada região poligonal são observados os valores da variável independente x_i e os valores da variável dependente y_i. Esse tipo de abordagem é denominada de análise de dados de área, ou dados discretos no espaço ou dados de *lattice* (CRESSIE, 1991).

Para esse tipo de estudo é importante levar em consideração a estrutura de vizinhança que estabelece a relação de dependência entre as observações. Dado um conjunto de n regiões $\{A_1, \ldots, A_n\}$, é possível obter uma matriz $W\,(n \times n)$, de estrutura de vizinhança em que cada um dos elementos w_{ij} representa uma medida de proximidade entre as regiões A_i e A_j.

Uma vez definida a estrutura de vizinhança, pode-se construir modelos que refletem a dependência espacial dos dados. Dois modelos muito comumente utilizados, que incluem esse tipo de estrutura, são os modelos CAR *(conditional autoregressive)* e SAR (*simultaneous autoregressive).*

Em certas situações, assim como nos modelos de calibração apresentados na seção anterior, a variável independente pode ser mais complicada (cara ou demorada) para ser mensurada e, portanto, nessas situações é interessante obter as estimativas x_0, da variável independente dado um valor, y_0, da variável dependente, levando em consideração as suas localizações no espaço estudado.

Uma situação em que se pode ilustrar o problema de calibração para dados espaciais é, por exemplo, quando um estudo é feito em n bairros de um município com intuito de se obter uma relação entre crimes e renda. Após comprovada a existência de uma dependência espacial entre essas duas variáveis, pode-se escrever uma relação funcional entre a variável independente (renda média) e a variável dependente (número de crimes) através de um modelo espacial. Como os dados

de renda nem sempre são disponíveis e/ou confiáveis para todos os bairros e os registros de crimes são mais acurados, pode-se ter como objetivo obter informações sobre a renda média da população que é desconhecida para um determinado bairro utilizando o registros de crime desse bairro.

Com o objetivo de resolver esse tipo de questão que envolve algum grau de dependência espacial e que ainda não foi abordado na literatura, neste capítulo é proposto um modelo de calibração espacial para o modelo SAR com apenas uma variável explicativa. Primeiramente, faz-se uma breve revisão desse modelo espacial autorregressivo. Em seguida são propostos estimadores pontuais para a regressão inversa em duas situações: a primeira em que a observação a ser estimada pertence à amostra selecionada e a segunda situação em que a observação a ser estimada não pertence à amostra. Enfim, é proposto um intervalo de confiança para ambos os casos.

4.2 Modelo de calibração SAR

Em calibração espacial o objetivo é estimar o valor da variável independente x quando os valores da variável dependente y são dados, levando em consideração a dependência espacial. Para obter uma estrutura para modelagem da variável resposta Y, adequada ao contexto da calibração e levando em consideração a dependência espacial, supõe-se que a variável Y seja função de uma única variável independente x (fixa e conhecida) na primeira etapa do processo, isto é, ajusta-se o modelo espacial SAR considerando apenas uma variável independente. Além disso, na segunda etapa, assume-se que o valor y_0 seja função do valor x_0 de uma variável desconhecida.

Considera-se $n + k$ unidades, $\{A_1, \ldots, A_n, \ldots, A_{n+k}\}$. Os valores de Y e x são relacionados de acordo com o modelo SAR, da seguinte maneira:

$$y_i = \rho\left(\sum_{j=1}^{n} w_{ij}y_j\right) + \beta_0 + x_i\beta_1 + \epsilon_i, \quad i = 1, 2, \ldots, n \tag{4.1}$$

e

$$y_{0i} = \rho\left(\sum_{j=1}^{n+k} w_{0ij}y_j\right) + \beta_0 + x_{0i}\beta_1 + \epsilon_{0i}, \quad i = n+1, \ldots, n+k. \tag{4.2}$$

Em um primeiro momento, uma amostra de tamanho n é selecionada, na qual são observados os valores x_i, da variável independente e os valores y_i, da variável dependente, portanto, os valores das variáveis são conhecidos. Em um segundo momento, considera-se, k regiões não pertencentes à amostra selecionada, dos quais são observados somente os valores, Y_0, da variável dependente. Neste caso os valores, X_0, da variável independente são desconhecidos.

4.2.1 Estimação dos valores da variável independente

O objetivo deste trabalho é fazer estimações sobre o valor, x_0, da variável independente, correspondente a $Y = y_0$. Portanto, foram considerados dois tipos de situações para a estimação: a estimação de uma observação pertencente à amostra e a outra situação na qual se pretende estimar uma observação não pertencente à amostra. Os dois tipos de situações são ilustrados na Figura 11a e Figura 11b.

No problema de estimação de uma observação x_0 pertencente à amostra, tem-se n unidades espaciais para as quais são observados os valores da variável dependente, Y conhecida, Y_c, bem como os valores da variável independente, X conhecida, X_c e deseja-se estimar o valor de x_0, nos locais observados após o ajuste do modelo, embora sejam conhecidos. Portanto, a estimação de x_0 da variável independente é obtida a partir do conhecimento de X_c e Y_c e todas as n unidades são utilizadas na fase de montagem do modelo, considerando a ideia apresentada na Figura 11a.

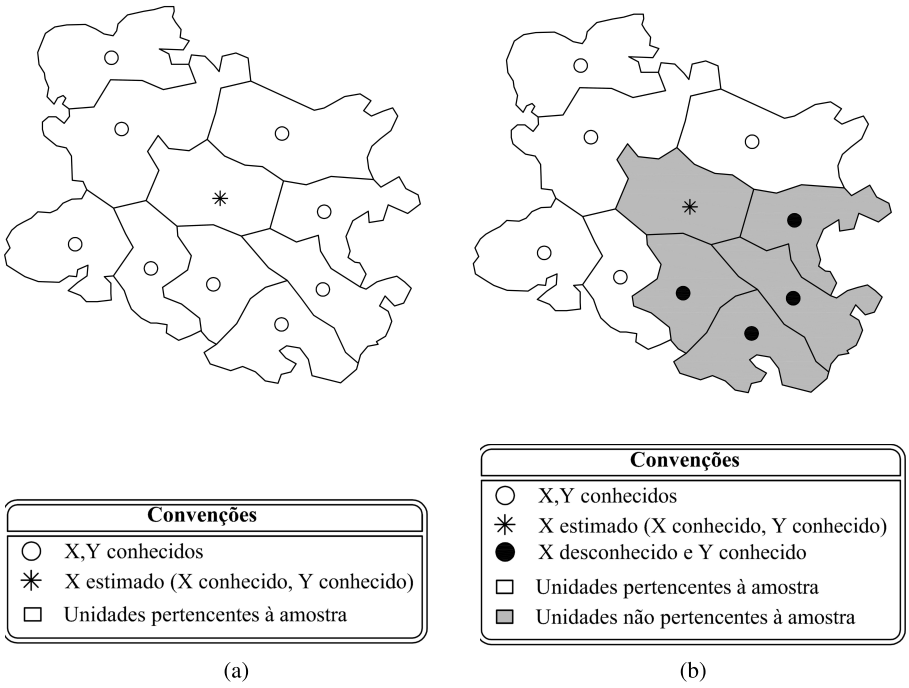


Figura 11 O problema de estimação de um valor da variável independente pertencente à amostra (a) e não pertencente à amostra (b). As áreas sombreadas são unidades de amostragem que não são utilizadas na fase de montagem do modelo

No caso em que se deseja estimar o valor x_0, de uma variável independente não pertencente à amostra selecionada, tem-se dois tipos de unidades espaciais: as unidades pertencentes à amostra, para a qual observa-se a variável dependente Y_c e as variáveis independentes X_c, e as unidades não pertencentes à amostra, para a qual observa-se somente o valor y_0 da variável dependente. Portanto, a estimação do valor x_0 da variável independente é obtida a partir do conhecimento de Y_c, X_c e Y_0. Essa situação é ilustrada na Figura 11b. As áreas sombreadas na Figura 11b são unidades de amostragem que não são utilizadas na fase de montagem do modelo, no entanto são utilizadas na fase de estimação do valor x_0 da variável independente X.

4.2.2 Estimação de x_0 pertencente à amostra

Com o interesse em estimar um valor de uma observação x_0, relativa ao valor y_0 pertencente a amostra, primeiramente, constrói-se a matriz de proximidade espacial $W(n \times n)$ entre as unidades pertencentes à amostra, em que cada um dos elementos w_{ij} é um indicador de proximidade entre as regiões A_i e A_j. Em seguida, estimam-se os parâmetros ρ, β_0, β_1, do modelo SAR expresso por (4.1) com base nas n observações da amostra, de acordo com as equações (2.52) e (2.53).

Após obter as estimativas dos parâmetros, pode-se escrever, em termos de componentes individuais, o modelo SAR da seguinte forma:

$$y_i = \hat{\beta}_0 + x_i\hat{\beta}_1 + \hat{\rho}\sum_{j=1}^{n} w_{ij}y_j. \tag{4.3}$$

Expandindo a equação (4.3), tem-se:

$$\begin{bmatrix} y_1 \\ \vdots \\ y_0 \\ \vdots \\ y_n \end{bmatrix} = \hat{\rho}\begin{bmatrix} w_{11} & \cdots & w_{10} & \cdots & w_{1n} \\ \vdots & & \vdots & & \vdots \\ w_{01} & \cdots & w_{00} & \cdots & w_{0n} \\ \vdots & & \vdots & & \vdots \\ w_{n1} & \cdots & w_{n0} & \cdots & w_{n0} \end{bmatrix}\begin{bmatrix} y_1 \\ \vdots \\ y_0 \\ \vdots \\ y_n \end{bmatrix} + \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_0 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}\begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix}.$$

Portanto, em termos de componentes individuais, o valor estimado, x_0, da variável independente, dado o valor, y_0, da variável dependente pode ser expresso pela equação:

$$\hat{x}_0 = \frac{y_0 - \hat{\rho}\left(\sum_{j=1}^{n} w_{0j}y_0\right) - \hat{\beta}_0}{\hat{\beta}_1}, \tag{4.4}$$

em que w_{0j} é o vetor proximidade da região A_0 com cada uma das regiões A_j,

$j = 1, \ldots, n$, pertencente à amostra selecionada.

4.2.3 Estimação de x_0 não pertencente à amostra

Assim como em Kato (2008) e Thomas-Agnan (2013), relacionam-se as unidades observadas pertencentes à amostra com as que não pertencem à amostra. Seja n o número de unidades pertencentes à amostra e k o número de unidades não pertencentes à amostra. Particiona-se X e Y em $X = (X_c, X_0)$ e $Y = (Y_c, Y_0)$, em que X_c com resposta Y_c são vetores correspondentes às unidades da amostra, sendo ambas variáveis conhecidas, e X_0 com resposta Y_0 são vetores correspondentes às unidades não pertencentes à amostra, com Y_0 conhecido, porém o valor de X_0 é desconhecido. Da mesma forma, particiona-se a matriz de proximidade (THOMAS-AGNAN, 2013):

$$W = \begin{pmatrix} W_{cc} & W_{c0} \\ W_{0c} & W_{00} \end{pmatrix}, \tag{4.5}$$

em que :
W_{cc} é uma matriz $n \times n$ correspondente à estrutura de vizinhança das unidades pertencentes à amostra;
W_{c0} é uma matriz $n \times k$ correspondente à estrutura de vizinhança entre as unidades pertencentes à amostra e as unidades não pertencentes à amostra;
W_{0c} é uma matriz $k \times n$ correspondente à estrutura de vizinhança entre as unidades não pertencentes à amostra e as unidades pertencentes à amostra;
W_{00} é uma matriz $k \times k$ correspondente à estrutura de vizinhança entre as unidades não pertencentes à amostra.

Considera-se, novamente, o modelo SAR expresso por (4.1) com uma única variável explicativa. Utilizando as n unidades pertencentes à amostra, obtém-se a matriz de proximidade espacial, W_{cc}. Em seguida, estimam-se os parâmetros β_0, β_1, ρ com base nas n observações pertencentes à amostra.

Suponha, que se tenha novas observações Y_0 que não pertencentes à amostra selecionada. No entanto, podem-se determinar as estruturas de proximidade

W_{c0}, W_{0c} entre as unidades não pertencentes à amostra e as unidades pertencentes à amostra e a estrutura de proximidade W_{00} entre as unidades não pertencentes à amostra. Tem-se como objetivo estimar os valores da variável independente desconhecidos, X_0 . Essa situação é ilustrada na Figura 11b em que na região sombreada são conhecidos os valores da variável dependente bem como a sua localização, no entanto, não utilizados na primeira etapa do processo de calibração, para estimar os parâmetros do modelo.

Considerando o modelo SAR com apenas uma variável explicativa e particionando as observações tem-se a seguinte expressão:

$$\begin{pmatrix} I-\rho W_{cc} & -\rho W_{c0} \\ -\rho W_{0c} & I-\rho W_{00} \end{pmatrix}\begin{pmatrix} Y_c \\ Y_0 \end{pmatrix}=\begin{pmatrix} 1 & X_c \\ 1 & X_0 \end{pmatrix}\begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}+\begin{pmatrix} \epsilon_c \\ \epsilon_0 \end{pmatrix}. \tag{4.6}$$

Após obter as estimativas dos parâmetros, ρ, β_0 e β_1, pelo método de mínimos quadrados, chega-se a:

$$\begin{pmatrix} I-\hat{\rho} W_{cc} & -\hat{\rho} W_{c0} \\ -\hat{\rho} W_{0c} & I-\hat{\rho} W_{00} \end{pmatrix}\begin{pmatrix} Y_c \\ Y_0 \end{pmatrix}=\begin{pmatrix} 1 & X_c \\ 1 & X_0 \end{pmatrix}\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix}. \tag{4.7}$$

Por meio da expressão (4.7), Obtém-se:

$$(I-\hat{\rho} W_{00})Y_0 = \hat{\rho} W_{0c} Y_c + \hat{\beta} X_0 \tag{4.8}$$

ou seja,

$$Y_0 = \hat{\rho}(W_{00} Y_0 + W_{0c} Y_c) + \hat{\beta} X_0. \tag{4.9}$$

Expandindo a expressão (4.9), observa-se:

$$\begin{bmatrix} Y_{01} \\ \vdots \\ Y_{0K} \end{bmatrix} = \hat{\rho} \begin{bmatrix} w_{0101} & \cdots & w_{010k} \\ \vdots & \vdots & \vdots \\ w_{0k01} & \cdots & w_{0k0k} \end{bmatrix} \begin{bmatrix} y_{01} \\ \vdots \\ y_{0k} \end{bmatrix} + \hat{\rho} \begin{bmatrix} w_{01c1} & \cdots & w_{01cn} \\ \vdots & \vdots & \vdots \\ w_{0kc1} & \cdots & w_{0kcn} \end{bmatrix} \begin{bmatrix} y_{c1} \\ \vdots \\ y_{cn} \end{bmatrix} + \\ + \begin{bmatrix} 1 & x_{01} \\ \vdots & \vdots \\ 1 & x_{0k} \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix}. \tag{4.10}$$

Logo, em termos de componentes individuais, pode-se escrever:

$$y_{0i} = \hat{\rho} \left(w_{0i0} Y_0 + w_{0ic} Y_c \right) + x_{0i} \hat{\beta}, \tag{4.11}$$

Portanto, o valor estimado da variável independente, desconhecido, x_0, dado o valor da variável dependente, y_0, é expresso por:

$$\hat{x}_{0i} = \frac{y_{0i} - \hat{\rho} \left[w_{0i0} Y_0 + w_{0ic} Y_c \right] - \hat{\beta}_0}{\hat{\beta}_1}. \tag{4.12}$$

4.2.4 Estimação intervalar

Apesar de a estimação pontual ser bastante útil, deve-se observar que quando obtém-se uma estimativa pontual, toda a informação presente nos dados é resumida através desse número. Portanto, é importante encontrar também um intervalo de valores possíveis para a variável. Sendo assim, a ideia é construir um intervalo em torno da estimativa pontual da variável independente, de modo que se tenha uma probabilidade conhecida de conter o verdadeiro valor x_0 que foi estimado.

Assim como para a estimação pontual, pode-se discutir dois tipos de situações para a estimação intervalar: a estimação de uma observação pertencente à amostra e a estimação de uma observação não pertencente à amostra.

4.2.5 Intervalos de confiança para x_0 pertencente à amostra

O objetivo dessa seção é construir um intervalo de confiança para o valor x_0, da variável independente dado o valor y_0, da variável dependente considerando as n unidades das quais foram observados os valores de X e Y.

Intuitivamente, observa-se que não há um intervalo útil para x_0, quando o parâmetro β_1 é próximo de zero. Nesse caso, assume-se que o modelo é dado por $y_i = \rho\left(\sum_{j=1}^{n} w_{ij}y_j\right) + \beta_0 + \epsilon_i$, e então admite-se que x_0 não está no modelo e assim não existe um intervalo de confiança.

A distribuição de $\hat{x}_0$ é mais complicada de se obter e não é necessária para construir o intervalo de confiança para x_0. Baseado no que foi apresentado em Graybill (1976) no caso da regressão linear simples, inicialmente, demonstra-se:

$$E\left(y_0 - \hat{\rho}W_0Y - \hat{\beta}x_0\right) = E\left(y_0\right) - \hat{\rho}W_0E\left(Y\right) - E\left(\hat{\beta}x_0\right) = 0, \tag{4.13}$$

$$\begin{aligned}
Var\left(y_0 - \hat{\rho}W_0Y - \hat{\beta}x_0\right) &= Var\left(y_0\right) + \hat{\rho}W_0Var\left(Y\right)\hat{\rho}W_0^T + x_0Var\left(\hat{\beta}\right)x_0^T - \\
&- 2Cov\left(y_0, \hat{\rho}W_0Y\right) - \\
&- 2\left[Cov\left(y_0, \hat{\beta}x_0\right) - Cov\left(\hat{\rho}W_0Y, \hat{\beta}x_0\right)\right] \\
&= Var\left(y_0\right) + \hat{\rho}W_0Var\left(Y\right)\hat{\rho}W_0^T + x_0\left(X^TX\right)^{-1}\sigma^2x_0^T - \\
&- 2Cov\left(y_0, Y\right)\hat{\rho}W_0^T - \\
&- 2\left[Cov\left(y_0, \hat{\beta}\right) - \hat{\rho}W_0Cov\left(Y, \hat{\beta}\right)\right]x_0^T \\
&= \sigma^2\Omega_{00} + \hat{\rho}W_0\sigma^2\Omega\hat{\rho}W_0^T + x_0\left(X^TX\right)^{-1}\sigma^2x_0^T - \\
&- 2\sigma^2\Omega_{0j}\hat{\rho}W_0^T - 2\left(\sigma^2\Omega_{0j} - \hat{\rho}W_0\sigma^2\Omega\right)a^Tx_0^T \\
&= \sigma^2\left[x_0\left(X^TX\right)^{-1}x_0^T - 2\left(\Omega_{0j} - \hat{\rho}W_0\Omega\right)a^Tx_0^T + C\right] \\
&= \sigma^2A,
\end{aligned} \tag{4.14}$$

em que:

$$A = x_0 \left(X^T X\right)^{-1} x_0^T - 2\left(\Omega_{0j} - \hat{\rho} W_0 \Omega\right) a^T x_0^T + C; \tag{4.15}$$

$$C = \Omega_{00} + \left(\hat{\rho}^2 W_0 \Omega - 2\Omega_{0j}\hat{\rho}\right) W_0^T; \tag{4.16}$$

$$a = \left(X^T X\right)^{-1} X^T \left(I - \hat{\rho} W\right); \tag{4.17}$$

$$\Omega = (I - \rho W)^{-1} \left(I - \rho W^T\right)^{-1}. \tag{4.18}$$

Desde que, $y_0 - \hat{\rho} W_0 Y - \hat{\beta} x_0$ é normalmente distribuído com média 0 e variância $\sigma^2 A$, tem-se:

$$Z = \frac{y_0 - \hat{\rho} W_0 Y - \hat{\beta} x_0}{\sqrt{Var\left(y_0 - \hat{\rho} W_0 Y - \hat{\beta} x_0\right)}} = \frac{y_0 - \hat{\rho} W_0 Y - \hat{\beta} x_0}{\sqrt{\sigma^2 A}} \backsim N(0,1) \tag{4.19}$$

e

$$U = \left(n - 2tr(S) + tr\left(S^T S\right)\right) \frac{\hat{\sigma}^2}{\sigma^2} \backsim \chi^2_{\left(n-2tr(S)+tr\left(S^T S\right)\right)}, \tag{4.20}$$

em que S é definido pela expressão (2.59).

Assim, desde que $Z \backsim N(0,1)$ e $U \backsim \chi^2_{\left(n-2tr(S)+tr\left(S^T S\right)\right)}$, e Z e U são independentes, pode-se definir um intervalo de confiança para x_0, através da seguinte quantidade pivotal:

$$T = \frac{Z}{\sqrt{\frac{U}{n-2tr(S)+tr\left(S^T S\right)}}}, \tag{4.21}$$

a qual tem distribuição t de Student com $\left(n - 2tr(S) + tr\left(S^T S\right)\right)$ graus de liberdade

Então, tem-se:

$$P\left(-t_{(\alpha/2;n-2tr(S)+tr(S^TS))} \leq T \leq t_{(\alpha/2;n-2tr(S)+tr(S^TS))}\right) = 1-\alpha. \tag{4.22}$$

A quantidade entre parênteses em (4.22) é equivalente a:

$$T^2 \leq t^2_{(\alpha/2;n-2tr(S)+tr(S^TS))} \tag{4.23}$$

ou seja,

$$\frac{\left(y_0 - \hat{\rho}W_0Y - \hat{\beta}x_0\right)^2}{\sigma^2 A} \leq t^2_{(\alpha/2;n-2tr(S)+tr(S^TS))}. \tag{4.24}$$

Portanto, o intervalo de confiança $(1-\alpha)100\%$ de x_0, para $Y = y_0$, é definido por:

$$\left(y_0 - \hat{\rho}W_0Y - \hat{\beta}x_0\right)^2 - \sigma^2 A t^2_{(\alpha/2;n-2tr(S)+tr(S^TS))} \leq 0. \tag{4.25}$$

Na inequação dada em (4.25), somente x_0 é desconhecido. Expandindo esta inequação obtém-se uma inequação quadrática na qual pode ser escrita em função da variável desconhecida x_0, da forma $h(x_0) = ax_0^2 + bx_0 + c \leq 0$. Se $a > 0$ e o discriminante $b^2 - 4ac > 0$, então para os valores de x_0 que satisfazem a inequação (4.25), tem-se um intervalo de $100\,(1-\alpha)\,\%$ confiança.

4.2.6 Intervalos de confiança para x_0 não pertencentes à amostra

Analogamente ao caso anterior, deseja-se obter o intervalo de confiança para o valor, x_0, desconhecido da variável independente dado a um valor, y_0, conhecido da variável dependente. Nesse caso, a observação a ser estimada não pertence à amostra selecionada. Assim, como anteriormente, deve-se observar que não há um intervalo de confiança quando o parâmetro β_1 é zero.

Inicialmente, calcula-se a esperança e a variância da expressão $\epsilon_{0i} = y_{0i} - \hat{\rho}\,(w_{0i0}Y_0 + w_{0ic}Y_c) - x_{0i}\hat{\beta}$, ou seja,

$$E\left[\epsilon_{0i}\right] = E[y_{0i} - \hat{\rho}\left(w_{0i0}Y_0 + w_{0ic}Y_c\right) - x_{0i}\hat{\beta}] = 0, \tag{4.26}$$

$$\begin{aligned}
Var\left[\epsilon_{0i}\right] &= Var\left[y_{0i} - \hat{\rho}\left(w_{0i0}Y_0 + w_{0ic}Y_c\right) - x_{0i}\hat{\beta}\right] \\
&= Var\left[y_{0i}\right] + Var\left[\hat{\rho}\left(w_{0i0}Y_0 + w_{0ic}Y_c\right)\right] + Var\left[x_{0i}\hat{\beta}\right] - \\
&- 2Cov\left[y_{0i}, \hat{\rho}\left(w_{0i0}Y_0 + w_{0ic}Y_c\right)\right] - 2Cov\left[y_{0i}, x_{0i}\hat{\beta}\right] + \\
&+ 2Cov\left[\hat{\rho}\left(w_{0i0}Y_0 + w_{0ic}Y_c\right), x_{0i}\hat{\beta}\right] \\
&= Var\left[y_{0i}\right] + \hat{\rho}^2 Var\left[w_{0i0}Y_0 + w_{0ic}Y_c\right] + x_{0i}Var\left[\hat{\beta}\right]x_{0i}^T - \\
&- 2\hat{\rho}Cov\left[y_{0i}, \left(w_{0i0}Y_0 + w_{0ic}Y_c\right)\right] - 2Cov\left[y_{0i}, \hat{\beta}\right]x_{0i}^T + \\
&+ 2\hat{\rho}Cov\left[\left(w_{0i0}Y_0 + w_{0ic}Y_c\right), \hat{\beta}\right]x_{0i}^T \\
&= Var\left[y_{0i}\right] + x_{0i}Var\left[\hat{\beta}\right]x_{0i}^T + \\
&+ \hat{\rho}^2\left\{Var\left[w_{0i0}Y_0\right] + Var\left[w_{0ic}Y_c\right] + 2Cov\left[w_{0i0}Y_0, w_{0ic}Y_c\right]\right\} - \\
&- 2\hat{\rho}\left\{Cov\left[y_{0i}, w_{0i0}Y_0\right] + Cov\left[y_{0i}, w_{0ic}Y_c\right]\right\} - \\
&- 2Cov\left[y_{0i}, \hat{\beta}\right]x_{0i}^T + 2\hat{\rho}\left\{Cov\left[w_{0i0}Y_0, \hat{\beta}\right] + Cov\left[w_{0ic}Y_c, \hat{\beta}\right]\right\}x_{0i}^T \\
&= Var\left[y_{0i}\right] + x_{0i}Var\left[\hat{\beta}\right]x_{0i}^T + \\
&+ \hat{\rho}^2\left\{w_{0i0}Var\left[Y_0\right]w_{0i0}^T + w_{0ic}Var\left[Y_c\right]w_{0ic}^T + 2w_{0i0}Cov\left[Y_0, Y_c\right]w_{0ic}^T\right\} - \\
&- 2\hat{\rho}\left\{Cov\left[y_{0i}, Y_0\right]w_{0i0}^T + Cov\left[y_{0i}, Y_c\right]w_{0ic}^T\right\} - \\
&- 2Cov\left[y_{0i}, \left(X_c^T X_c\right)^{-1} X_c^T\left(I - \rho W_c\right)Y_c\right]x_{0i}^T + \\
&+ 2\hat{\rho}Cov\left[w_{0i0}Y_0, \left(X_c^T X_c\right)^{-1} X_c^T\left(I - \hat{\rho}W_c\right)Y_c\right] + \\
&+ 2\hat{\rho}Cov\left[w_{0ic}Y_c, \left(X_c^T X_c\right)^{-1} X_c^T\left(I - \hat{\rho}W_c\right)Y_c\right]x_{0i}^T.
\end{aligned} \tag{4.27}$$

Fazendo $a = \left(X_c^T X_c\right)^{-1} X_c^T\left(I - \rho W_c\right)$, segue que:

$$
\begin{aligned}
Var[\epsilon_{0i}] &= Var\left[y_{0i}\right] + x_{0i} Var\left[\beta\right] x_{0i}^T + \\
&+ \hat{\rho}^2 \left\{ w_{0i0} Var\left[Y_0\right] w_{0i0}^T + w_{0ic} Var\left[Y_c\right] w_{0ic}^T + 2 w_{0i0} Cov\left[Y_0, Y_c\right] w_{0ic}^T \right\} - \\
&- 2\hat{\rho} \left\{ Cov\left[y_{0i}, Y_0\right] w_{0i0}^T + Cov\left[y_{0i}, Y_c\right] w_{0ic}^T \right\} - \\
&- 2 Cov\left[y_{0i}, Y_c\right] a^T x_{0i}^T + \\
&+ 2\hat{\rho} \left\{ w_{0i0} Cov\left[Y_0, Y_c\right] a^T + w_{0ic} Cov\left[Y_c, Y_c\right] a^T \right\} x_{0i}^T . \qquad (4.28)
\end{aligned}
$$

Fazendo $Var[Y] = \sigma^2 \Omega$, tem-se que:

$$
\begin{aligned}
Var[\epsilon_{0i}] &= \sigma^2 \Big[\Omega_{ii} + \hat{\rho}^2 \left(w_{0i0} \Omega_{00} w_{0i0}^T + w_{0ic} \Omega_{cc} w_{0ic}^T + 2 w_{0i0} \Omega_{0c} w_{0ic}^T \right) + \\
&+ x_{0i} \left(X_c^T X_c \right)^{-1} x_{0i}^T - 2\hat{\rho} \left(\Omega_{0i0} w_{0i0}^T + \Omega_{0ic} w_{0ic}^T \right) - \\
&- 2\Omega_{0ic} a^T x_{0i}^T + 2\hat{\rho} \left(w_{0i0} \Omega_{0c} a^T + w_{0ic} \Omega_{cc} a^T \right) x_{0i}^T \Big], \qquad (4.29)
\end{aligned}
$$

fazendo,
$C = \Omega_{ii} + \hat{\rho}^2 \left(w_{0i0} \Omega_{00} w_{0i0}^T + w_{0ic} \Omega_{CC} w_{0ic}^T + 2 w_{0i0} \Omega_{0C} w_{0ic}^T \right) - 2\hat{\rho} \left(\Omega_{0i0} w_{0i0}^T + \Omega_{0ic} w_{0ic}^T \right)$
finalmente, obtém-se:

$$
\begin{aligned}
Var[\epsilon_{0i}] &= \sigma^2 \Big[x_{0i} \left(X_c^T X_c \right)^{-1} x_{0i}^T - 2 \Big(\Omega_{0ic} - \\
&- \rho \left(w_{0i0} \Omega_{0c} + w_{0ic} \Omega_{cc} \right) \Big) a^T x_{0i}^T + C \Big] \\
&= \sigma^2 A, \qquad (4.30)
\end{aligned}
$$

em que $A = x_{0i} \left(X_c^T X_c \right)^{-1} x_{0i}^T - 2 \Big(\Omega_{0ic} - \hat{\rho} \left(w_{0i0} \Omega_{0c} + w_{0ic} \Omega_{cc} \right) \Big) a^T x_{0i}^T + C$.

Desde que, $y_{0i} - \hat{\rho} \left(w_{0i0} Y_0 + w_{0ic} Y_c \right) - x_{0i} \beta$ é normalmente distribuído com média 0 e variância $\sigma^2 A$, tem-se:

$$
\begin{aligned}
Z &= \frac{y_{0i} - \hat{\rho} \left(w_{0i0} Y_0 + w_{0ic} Y_c \right) - x_{0i} \hat{\beta}}{\sqrt{Var \left(y_{0i} - \hat{\rho} \left(w_{0i0} Y_0 + w_{0ic} Y_c \right) - x_{0i} \hat{\beta} \right)}} \qquad (4.31) \\
&= \frac{y_{0i} - \hat{\rho} \left(w_{0i0} Y_0 + w_{0ic} Y_c \right) - x_{0i} \hat{\beta}}{\sqrt{\sigma^2 A}} \backsim N(0, 1)
\end{aligned}
$$

e

$$U = \left(n - 2tr\left(S\right) + tr\left(S^T S\right)\right) \frac{\hat{\sigma}^2}{\sigma^2} \backsim \chi^2_{n-2tr(S)+tr\left(S^T S\right)}. \tag{4.32}$$

Logo, desde que $Z \backsim N(0,1)$ e $U \backsim \chi^2_{\left(n-2tr(S)+tr\left(S^T S\right)\right)}$, e Z e U são independentes, pode-se escrever:

$$T = \frac{Z}{\sqrt{\frac{U}{n-2tr(S)+tr\left(S^T S\right)}}}, \tag{4.33}$$

que tem distribuição t de Student com $\left(n - 2tr\left(S\right) + tr\left(S^T S\right)\right)$ graus de liberdade.

Portanto, segue que:

$$P\left(-t_{(\alpha/2;n-2tr(S)+tr\left(S^T S\right))} \leq T \leq t_{(\alpha/2;n-2tr(S)+tr\left(S^T S\right))}\right) = 1 - \alpha \tag{4.34}$$

Assim, tem-se o intervalo de confiança $(1-\alpha)100\%$ de x_0 quando $Y = y_0$ definido por:

$$\left(y_{0i} - \hat{\rho}\left(w_{0i0}Y_0 + w_{0ic}Y_c\right) - x_{0i}\hat{\beta}\right)^2 - \sigma^2 A t^2_{(\alpha/2;n-2tr(S)+tr\left(S^T S\right))} \leq 0. \tag{4.35}$$

Expandindo a inequação (4.35) tem-se uma inequação quadrática, em função do valor desconhecido x_0, a qual pode ser escrita na forma $h\left(x_0\right) = ax_0^2 + bx_0 + c \leq 0$. Para os valores de x_0 que satisfazem à inequação tem-se um intervalo de $100\left(1-\alpha\right)\%$ confiança, desde que $a > 0$ e $b^2 - 4ac > 0$.

4.3 Aplicação

Para ilustrar o modelo de calibração proposto, utiliza-se o conjunto de dados descrito em Anselin (1988) e que está disponível na biblioteca "spdep" do *software* R (R CORE TEAM, 2015). Esse conjunto de dados inclui observações dos roubos residenciais e roubos de veículos por mil domicílios (crimes), renda média familiar (em mil dólares) e o valor médio da habitação (em mil dólares) em

cada um dos 49 bairros da cidade de Columbus, Ohio, EUA.

Na Figura 12, pode-se observar a região de estudo, a cidade de Columbus, dividida em 49 bairros (unidades).

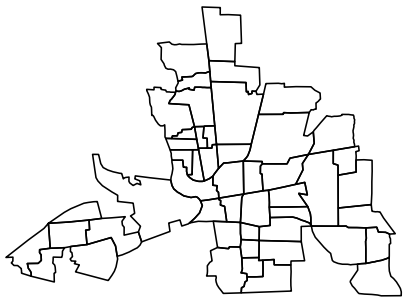


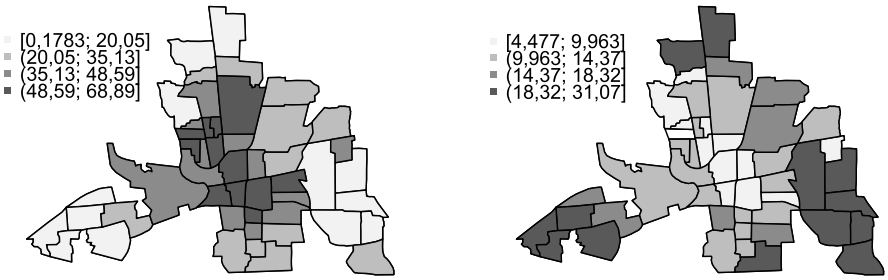
Figura 12 Divisão política da cidade de Columbus, Ohio, EUA

Como o objetivo é de aplicar o modelo de calibração proposto neste trabalho, considerou-se o número de crimes como a variável dependente de interesse Y e a renda média familiar dos bairros foi utilizada como a variável independente X.

Todos os cálculos e análises estatísticas foram realizados utilizando funções desenvolvidas no *software* R (R CORE TEAM, 2015) que estão disponíveis no anexo.

Observa-se que nessa situação pode ser mais complicado obter informações dos valores da renda familiar do que informações sobre número de crimes, devido à dificuldade de as pessoas declararem o valor real dos rendimentos familiares. Portanto, tem-se o interesse em estimar o valor da renda média familiar (X) dadas as observações dos roubos residenciais e roubos de veículos por mil domicílios (Y). Neste caso, trata-se de um problema de regressão inversa. A principal diferença da modelagem dessas observações é que se leva em consideração a distribuição espacial dos dados.

A distribuição espacial do número de roubos e da renda média familiar podem ser observados nas Figuras 13a e 13b, respectivamente. Nota-se, de uma maneira geral, que os bairros com maior número de roubos são os bairros que apresentam a menor renda familiar.



(a) (b)

Figura 13 Mapa de vizinhança dos crimes relacionados a roubos residenciais e de veículos por mil domicílios (a) e o mapa de vizinhança da renda média familiar em mil dólares (b), em Columbus, Ohio, EUA

Para descrever a estrutura global de dependência espacial das variáveis envolvidas utilizou-se o índice de Moran global (I) que mede o nível de autocorrelação espacial entre as unidades (bairros). O índice de Moran, definido pela equação (2.41), permitiu verificar que os dados são autocorrelacionados espacialmente tanto para a variável roubo como para a variável renda com valores iguais a 0,486 e 0,417, respectivamente, com valor-p <0,0001. O cálculo foi realizado com a matriz de peso espacial de acordo com o critério que se A_i compartilha um lado comum com A_j então $w_{ij} = 1$, caso contrário $w_{ij} = 0$, para a matriz de proximidade W. Pode-se observar que os níveis descritivos (valor-p) são menores que 0,05 (nível de significância), portanto as variáveis apresentam autocorrelação espacial significativa a 5% de probabilidade.

Neste trabalho são propostos dois casos de calibração. No primeiro caso, tem-se uma amostra com n unidades, das quais todas são utilizadas na primeira etapa do processo de calibração, ou seja, no ajuste do modelo. Em seguida, deseja-se estimar a renda média familiar, x_0, dado um valor do número de crimes, y_0, pertencente a esta amostra de tamanho n. No segundo caso, tem-se uma amostra com n unidades e também tem-se k unidades não pertencentes à amostra. São utilizadas apenas as n unidades pertencentes à amostra para o ajuste do modelo.

Nesse caso, deseja-se estimar a renda média familiar, x_0, de um bairro não pertencente à amostra, dado um valor do número de crimes y_0 .

4.3.1 Estimação de x_0 pertencente à amostra

Para a aplicação do primeiro caso, foi construída uma matriz de proximidade (W) para os 49 bairros. A medida de proximidade utilizada foi calculada a partir do critério que se A_i compartilha um lado comum com A_j então $w_{ij} = 1$, caso contrário $w_{ij} = 0$ (Figura 14). Em seguida, foi feita uma normalização das linhas da matriz W dividindo cada elemento da matriz pelo total da linha e assim, a soma dos pesos de cada linha foi igual a 1.

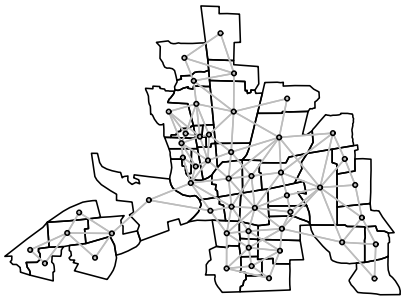


Figura 14 Contiguidades do setor censitário com, pelos menos, um ponto em comum em Columbus, EUA

Na primeira etapa do processo de calibração, estima-se pelo método de máxima verossimilhança os parâmetros (ρ, β_1, β_2) do modelo autorregressivo SAR, cuja expressão é dada na equação (2.45), utilizando todas as 49 unidades (bairros).

As estimativas dos parâmetros do modelo SAR podem ser observadas na Tabela 11.

Tabela 11 Estimativas dos parâmetros obtido para o modelo SAR

Parâmetros	Estimativas	Erro-padrão	z calc.	Valor-p
Constante	21,418	4,723	4,535	$5{,}754 \times 10^{-6}$
Renda familiar	-1,525	0,306	-4,982	$6{,}286 \times 10^{-7}$
Componente espacial (ρ)	0,393	0,128	3,056	0,0022

De acordo com a Tabela 11, verificou-se que a variável renda familiar possui um coeficiente negativo, o que significa que quanto maior a renda, menor será o número de crimes.

O coeficiente espacial autorregressivo apresentou um valor positivo sendo estatisticamente significante diferente de zero, indicando que os números de roubos apresentam uma autocorrelação espacial positiva, ou seja, os valores em bairros que fazem fronteira entre si são similares.

Na segunda etapa, suponha que se tem informações de crimes (y_0) (roubos residenciais e roubos de veículos por mil famílias) e se tem interesse em obter a estimação pontual e intervalar do valor (x_0) da variável independente (renda média familiar). O modelo ajustado na primeira etapa é utilizado para fazer as estimações do valor da variável independente (x_0).

Na Tabela 12, estão apresentados dados de crimes (por mil famílias) em função da renda familiar (em mil dólares), escolhidos de forma aleatória. Os valores pontuais estimados dos valores de x_0 dado os valores de y_0 e os respectivos intervalos de confiança são calculados conforme as expressões (4.4) e (4.25), respectivamente.

Tabela 12 Estimação pontual e intervalar da renda familiar para observações (Obs.) pertencentes à amostra

Obs.	Crimes	Renda familiar (valor real)	Renda famíliar (valor estimado)	Intervalo de 95% de confiança
3	30,627	15,956	15,918	[0,891; 30,935]
20	0,224	31,070	36,641	[23,677; 51,252]
26	40,969	8,0850	14,089	[-0,066; 30,019]
43	36,663	13,380	13,699	[-1,284; 28,776]
42	16,491	25,873	22,541	[7,613; 36,483]

Observa-se, na Tabela 12, que alguns dos intervalos apresentam limites inferiores negativos. Apesar de serem teoricamente admissíveis, do ponto de vista prático não faz sentido, pois trata-se da estimação da renda média familiar.

Na figura 15, tem-se os intervalos representados para cada uma das 49 observações. Pode-se observar, que os intervalos referentes às observações 7 e 30 não contêm o verdadeiro valor, x_0, da variável independente. Essas observações são outliers, ou seja, apresentam um comportamento diferenciado das demais observações. A observação 7 possui um valor baixo para o número de crimes por mil domicílio (0,178) e para o valor da renda média familiar (8,438 mil dólares). A observação 30 apresenta um valor elevado para o número de crimes por mil domicílio (68,892) enquanto para a renda média familiar apresenta um valor mediano (13,906 mil dólares).

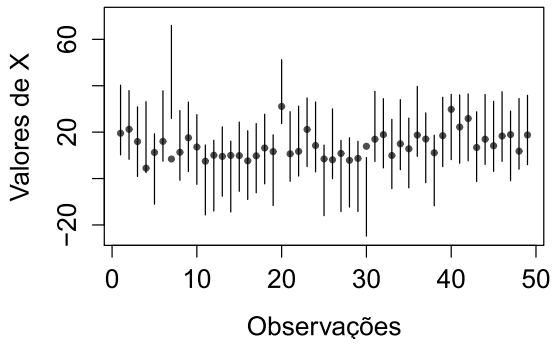


Figura 15 Intervalo de confiança para cada um valor da variável independente

Vale ressaltar que as estimações pontuais e intervalares da variável renda média, x_0, em uma determinada localidade, são influenciadas diretamente pelas regiões que compartilham um lado comum. O comportamento dessas estimações fica claro ao examinar as expressões (4.4) e (4.25), visto que o vetor de proximidade, da região em que x_0 é desconhecido em relação às demais regiões, levado em consideração.

4.3.1.1 Validação do Modelo

De acordo com Snee (1977) a validação cruzada é um método eficaz de avaliação de um modelo de regressão. A validação cruzada se baseia na habilidade de predição de um modelo construído por parte de um conjunto de dados seguido pela predição do restante. Esse tipo de validação pode ser realizado em blocos contínuos, blocos randômicos ou ainda pelo método "*leave-one-out*".

O método *leave-one-out* consiste em deixar uma amostra de fora no processo de construção do modelo e a seguir essa amostra é predita pelo modelo construído. Esse processo se repete até que todas as amostras tenham sido deixadas de fora e preditas.

O gráfico da validação cruzada que apresenta os valores preditos pelo modelo (com a retirada de uma amostra) versus os valores reais são apresentados na Figura 16.

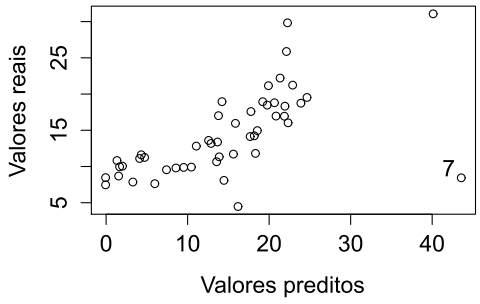


Figura 16 Gráfico de dispersão dos valores preditos e os valores reais

Como pôde observar na Figura 16, a relação entre os valores reais e os

valores preditos possui moderada relação linear. O coeficiente de correlação linear estimado com base nos dados, foi de 0,6072. Ainda na Figura 16 pode-se notar que há uma observação (Observação 7) que se destaca de todo o conjunto de observações por ter seu valor muito afastado dos outros valores. Essa observação é uma observação influente, ou seja, exerce efeito no ajuste do modelo. A observação é um valor referente a um bairro que se localiza na fronteira da cidade de Columbus. Portanto, seus valores podem ser fortemente influenciados por regiões da cidade que faz fronteira com Columbus e que não são considerados neste estudo.

4.3.2 Estimação de x_0 não pertencente à amostra

Para ilustrar o segundo caso, em que se deseja estimar uma observação não pertencente a uma amostra selecionada, particiona-se a região de estudo em dois tipos de unidades espaciais. Assim, tem-se as unidades em que as variáveis independente e dependente são conhecidas (X_c,Y_c), ou seja, estas unidades pertencem à amostra selecionada. Tem-se também as unidades em que a variável independente é desconhecida e a variável dependente é conhecida (X_0,Y_0). Para a realização da primeira etapa do processo de calibração (ajuste do modelo) são utilizadas apenas as unidades pertencentes à amostra.

Inicialmente, foram selecionados 46 bairros (veja Figura 17) dos quais todas as observações X e Y são conhecidas. Em seguida, obteve-se a matriz de proximidade normalizada para esses 46 bairros.

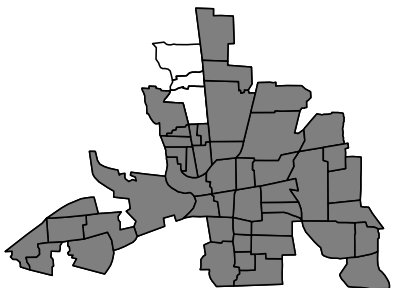


Figura 17 As áreas sombreadas são unidades de amostragem utilizadas na fase de ajuste do modelo

Após a construção da matriz de proximidade, estimou-se os parâmetros do modelo SAR pelo método de máxima verossimilhança, conforme pode-se observar na Tabela 13.

Tabela 13 Estimativas dos parâmetros obtido para o modelo SAR considerando 46 unidades espaciais

Parâmetros	Estimativas	Erro-padrão	z calc.	Valor-p
Constante	23,368	4,921	4,749	$2{,}044 \times 10^{-06}$
Renda familiar	-1,713	0.331	-5,180	2.217×10^{-07}
Componente espacial (ρ)	0,341	0,133	2,553	0,011

Para a variável renda média familiar, obteve-se um coeficiente negativo (Tabela 13), o que significa que quanto maior a renda média, menor será o número de crimes. O coeficiente espacial autorregressivo apresentou um valor positivo e é estatisticamente diferente de zero, considerando um nível de significância de 5%. Isso significa que os valores de crimes apresentam uma autocorrelação espacial positiva.

Suponha que se deseja obter a estimativa pontual e intervalar da renda familiar, x_0, desconhecida. As informações de crimes, roubos residenciais e roubos de veículos são conhecidas, assim como sua localização. Porém, não foram utilizadas para ajustar o modelo. As regiões em que o valor, x_0, da variável independente é desconhecido, são identificadas pelas regiões não sombreadas da Figura 17.

Na Tabela 14 são apresentados dados de crimes (por mil domicílio) em função da renda média familiar (em mil dólares) desconhecida, x_0. Os valores pontuais estimados dos valores de x_0 dados os valores de y_0, para as unidades não sombreadas (Figura 17) e os respectivos intervalos de confiança são calculados conforme as expressões (4.12) e (4.35), respectivamente.

Observa-se, na Tabela 14, que o intervalo referente ao valor médio estimado de 13,971 da renda familiar, apresentou limite inferior negativo, o que do ponto de vista prático não faz sentido. Porém, é teoricamente admissível.

Em um segundo momento, foram selecionadas, aleatoriamente, 40 unida-

Tabela 14 Estimação pontual para renda média familiar e os respectivos intervalos de 95% confiança , para valores não pertencentes à amostra, considerando 46 observações conhecidas

Crimes	Renda familiar (valor estimado)	Intervalo de 95% confiança
18,802	22,395	[9,442; 37,668]
32,388	16,134	[2.597; 30,076]
38,426	13,971	[-0,023; 27,353]

des (bairros), conforme ilustrado na Figura 18. Obteve-se a matriz de proximidade normalizada para esses bairros e, em seguida, estimaram-se os parâmetros do modelo SAR, conforme pode-se observar na Tabela 15.

Figura 18 As áreas sombreadas são unidades de amostragem utilizadas na fase de montagem do modelo

As estimativas dos parâmetros, considerando 40 bairros, podem ser observadas na Tabela 15. Novamente, observa-se uma relação inversa entre a renda média familiar e o número de crimes.

Tabela 15 Estimativas dos parâmetros obtido para o modelo SAR considerando 40 unidades

Parâmetros	Estimativas	Erro-padrão	z calc.	Valor-p
Constante	20,662	4,149	4,980	$6{,}352 \times 10^{-07}$
Renda familiar	-1,576	0,256	-6,1615	$7{,}204^{-10}$
Componente espacial (ρ)	0,425	0,109	3,862	$1{,}126^{-04}$

O coeficiente espacial autorregressivo apresentou um valor positivo e é estatisticamente diferente de zero. Portanto, os valores em bairros que fazem fronteira entre si são similares, considerando os 40 bairros selecionados .

Novamente, suponha que são conhecidas as informações de crimes por bairro, assim como sua localização. O objetivo é obter a estimação pontual e intervalar para uma renda média familiar (em 1000 dólares), x_0, desconhecida. As regiões não sombreadas, as quais se deseja estimar os valores da variável independente, podem ser identificadas na Figura 18.

Na Tabela 16 são apresentados dados de crimes (por mil famílias) em função da renda familiar (em mil dólares) desconhecida, x_0. Os valores pontuais estimados e os respectivos intervalos de confiança são calculados conforme as expressões (4.12) e (4.35), respectivamente.

Tabela 16 Estimação pontual para renda familiar e os respectivos intervalos com 95% de confiança, para valores não pertencentes à amostra considerando 40 observações conhecidas

Crimes	Renda famíliar (valor estimado)	Intervalo de 95% confiança
30,626	16,096	[5,173; 27,384]
0,178	40,882	[29,532; 56,548]
30,515	18,196	[7,382; 29,576]
34,001	12,855	[1,547; 23,845]
60,750	3,014	[-9,783; 13,674]
53,711	5,288	[-7,013; 15,966]
19,101	19,637	[8,887; 31,469]
27,822	14,685	[3,749; 26,062]
26,645	18,388	[7,581; 29,971]

Segundo Charnet et al. (1999), o intervalo fornece informação sobre a precisão das estimativas, no sentido de que quanto menor a amplitude do intervalo, maior a precisão. Nota-se que, para este caso, quando utilizado um número menor de observações, 40 observações, na primeira etapa do processo de calibração os intervalos tiveram menores amplitude (Tabela 16) que quando foram utilizadas 46 observações (Tabela 15). Provavelmente, isso se deve ao fato de que para o conjunto de 40 observações selecionadas houve menor valor da variância (S^2), do que quando foram selecionadas as 46 observações.

É importante observar que nas unidades em que os valores das variáveis independentes desconhecidas não foram utilizados na primeira etapa do processo de calibração, estes são consideradas na segunda etapa, para estimação pontual e intervalar dos valores da variável desconhecida, x_0, como pode ser observado nas expressões (4.12) e (4.35). Portanto, essas unidades em que o valor, x_0, da variável é desconhecida, podem influenciar nas estimativas pontuais e intervalares dos valores das variáveis independentes.

Como mencionado anteriormente, a matriz de proximidade espacial é uma ferramenta básica para estimar a variabilidade espacial de dados de área. Vários critérios são usados para definir a matriz de vizinhança. Nesta aplicação foi utilizado

o critério que se A_i compartilha um lado comum com A_j então $w_{ij} = 1$; caso contrário $w_{ij} = 0$. Portanto, se outros critérios forem utilizados os valores estimados podem ser alterados, pois as fórmulas de estimações (4.4), (4.25), (4.12) e (4.35) levam em consideração a matriz de proximidade.

5 CONSIDERAÇÕES FINAIS

Em problemas de calibração, a principal variável de interesse é o valor (ou quantidade) desconhecido (x_0) da variável independente a ser determinado. A literatura sobre modelos de calibração apresenta diversas abordagens para o tratamento do problema de calibração em modelos de regressão convencionais, com vários enfoques e aplicações nas mais diversas áreas de conhecimento. O tratamento do problema de calibração em modelos de regressão espacial, até o momento, era desconhecido.

Neste trabalho, foram apresentados os conceitos básicos de regressão inversa e os principais métodos de estimação pontual e de estimação intervalar para os modelos de regressão linear simples e polinomial quadrático. Todos os métodos foram implementados no *software* R, o que é uma contribuição da presente tese para a área. Entretanto, a principal contribuição deste trabalho é o desenvolvimento de um modelo de calibração para dados espaciais. Esse modelo difere dos estudos anteriores por propor a modelagem da variável resposta que leva em consideração a estrutura de vizinhança que estabelece a relação de dependência espacial entre as observações.

Para incorporar a dependência espacial dos dados considerou-se o modelo autorregressivo SAR com uma única variável independente. As estimativas dos parâmetros do modelo foram obtidas pelo método da máxima verossimilhança, a partir da hipótese de que o vetor de resíduos ϵ possui distribuição normal multivariada com média zero e matriz de variância e covariâncias $\sigma^2 I$.

Para a modelagem do problema de regressão inversa para dados espacialmente distribuídos foram levadas em consideração as divisões do espaço estudado em n regiões selecionadas. Para cada região, foram observados os valores das variáveis independentes x_i e os valores variáveis dependentes y_i. O problema de regressão inversa foi dividido em duas situações: a primeira em que a observação da variável independente a ser estimada pertence à amostra selecionada e a segunda situação em que a observação a ser estimada não pertence à amostra. Através das expressões de estimações desenvolvidas, pôde-se determinar valores estimados pontuais e intervalares para os valores x_0, da variável independente, em cada um

dos casos abordados. Todos os métodos da calibração espacial foram implementados no *software* R(R CORE TEAM, 2015), o que permite o uso imediato dos mesmos.

Por fim, este trabalho abre caminho para uma nova linha de pesquisa que é a calibração espacial e, portanto, com muitas questões para serem respondidas e trabalhos para serem desenvolvidos. Alguns desses trabalhos encontram-se em andamento tais como a extensão dos resultados do modelo SAR para o modelo CAR (conditional autoregressive). Uma outra linha de pesquisa seria considerar os modelos espaciais no caso multivariado, em que existe mais de uma variável dependente. A calibração de modelos geoestatísticos também pode ser considerada.

REFERÊNCIAS

ANSELIN, L. **Spatial econometrics**. Dallas: University of Texas, 1999. 50 p.

ANSELIN, L. **Spatial econometrics:** methods and models. Dordrecht: Kluwer Academic, 1988. 189 p.

BROWN, P. J. Multivariate calibration. **Journal of the Royal Statistical Society Series B-Methodological**, London, v. 44, n. 3, p. 287–321, Feb. 1982.

BROWN, P. J.; SUNDBERG, R. Confidence and conflict in multivariate calibration. **Journal of the Royal Statistical Society Series B-Methodological**, London, v. 49, n. 1, p. 46–57, Mar. 1987.

CHARNET, R. et al. **Análise de modelos de regressão linear com aplicações**. Campinas: UNICAMP, 1999. 356 p.

CLIFF, A. D.; ORD, K. **Spatial processes models and applications**. London: Pion, 1981. 266 p.

CRESSIE, N. A. C. **Statistics for spatial data**. Chichester: J. Wiley, 1991. 900 p.

DAVIS, A. W.; HAYAKAWA, I. Some distribution-theory relating to confidence-regions in multivariate calibration. **Annals of the Institute of Statistical Mathematics**, Tokyo, v. 39, n. 1, p. 141–152, 1987.

DINIZ, A. **Estatística espacial:** geoprocessamento. Belo Horizonte: UFMG, 2000. 15 p. Apostila.

DRUCK, S. et al. **Análise espacial de dados geográficos**. Brasília: EMBRAPA, 2004. 209 p.

EISENHART, C. The interpretation of certain regression methods, and their use in biological and industrial research. **Annals of Mathematical Statistics**, Ann Arbor, v. 10, p. 162–186, 1939.

FIELLER, E. C. Some problems in interval estimation. **Journal of the Royal Statistical Society**, London, v. 16, p. 175–185, 1954.

FUJIKOSHI,Y.; NISHII, R. On distribution of a statistic in multivariate inverse regression analysis. **Hiroshima Mathematical Journal**, Hiroshima, v. 14, p. 215–225, 1984.

GRAYBILL, F. A. **Theory and application of linear model**. North Situate: Duxbury, 1976. 740 p.

JOSE, K. K.; ISAAC, J. Single use conservative confidence regions in multivariate controlled calibration. **Journal of Statistical Planning and Inference**, Amsterdam, v. 137, n. 4, p. 1226–1235, Apr. 2007.

KATO, T. A further exploration into the robustness of spatial autocorrelation specifications. **Journal of the Royal Statistical Society Series B-Methodological**, London, v. 48, n. 3, p. 615–638, 2008.

KIRKUP, L.; MULHOLLAND, M. Comparison of linear and non-linear equations for univariate calibration. **Journal of Chromatography A**, Amsterdam, v. 1029, n. 1/2, p. 1–11, Mar. 2004.

KRUTCHKOFF, R. G. Classical and inverse regression methods of calibration. **Technometrics**, Washington, v. 9, n. 3, p. 425–439, Aug. 1967.

KRUTCHKOFF, R. G. Classical and inverse regression methods of calibration in extrapolation. **Technometrics**, Washington, v. 11, n. 3, p. 605–608, Aug. 1969.

LIEBERMAN, G. J.; MILLER, R. G. Simultaneous tolerance intervals in regression. **Technometrics**, Washington, v. 50, n. 1/2, p. 155–156, 1963.

LIEBERMAN, G. J.; MILLER, R. G.; HAMILTON, M. A. Unlimited simultaneous discrimination intervals in regression. **Biometrika**, London, n. 54, p. 133-145, June 1967.

MATHEW, T.; KASALA, S. An exact confidence region in multivariate calibration. **Annals of Statistics**, Hayward, v. 22, n. 1, p. 94-105, Mar. 1994.

MATHEW, T.; SHARMA, M. K. Joint confidence regions in the multivariate calibration problem. **Journal of Statistical**, Hayward, v. 100, n. 2, p. 427-441, Feb. 2002.

MATHEW, T.; ZHA, W. X. Conservative confidence regions in multivariate calibration. **Annals of Statistics**, Hayward, v. 24, n. 2, p. 707-725, Apr. 1996.

NASZÓDI, L. J. Elimination of the bias in the course of calibration. **Technometrics**, Washington, v. 20, n. 2, p. 201-205, May 1978.

OLIVEIRA, E. C.; AGUIAR, P. F. Validação da metodologia da avaliação de incerteza em curvas de calibração melhor ajustadas por polinômios de segundo grau. **Química Nova**, São Paulo, v. 32, n. 6, p. 1571-1575, 2009.

OMAN, S. D. Confidence-regions in multivariate calibration. **Annals of Statistics**, Hayward, v. 16, n. 1, p. 174-187, Mar. 1988.

ORD, J. K. Estimation methods for models of spatial interaction. **Journal of the American Statistical Association**, New York, v. 70, n. 3, p. 120–126, 1975.

R CORE TEAM. **R:** a language and environment for statistical computing. Version 12.2.1. Vienna: R Foundation for Statistical Computing, 2015. Disponível em: <http://www.R-project.org/>. Acesso em: 20 jan. 2015.

SHUKLA, G. K. On the problem of calibration. **Technometrics**, Washington, v. 14, n. 3, p. 547-553, 1972.

SNEE, R. D. Validation of regression models: methods and examples. **Technometrics**, Washington, v. 19, n. 4, p. 415-428, 1977.

THOMAS-AGNAN, C. About predictions in spatial autoregressive models: optimal and almost optimal strategies. **Toulouse School of Economics Working Paper,** Saint Louis, n. 13, p. 452, 2013.

THONNARD, M. **Confidence Intervals in Inverse Regression**. 2006. 78 p. Dissertation (Master in Mathematics and Computer Science) - Technische Universiteit Eindhoven, Eindhoven, 2006.

TOBLE, W. Computer movie simulating urban growth in the Detroit region. **Economic Geography**, Worcester, v. 46, p. 234–240, June 1970.

WALLER, L. A.; GOTWAY, C. A. **Applied spatial statistics for public health data**. New York: J. Wiley, 2004. 518 p.

WILLIAMS, E. J. A note on regression methods in calibration. **Technometrics**, Washington, v. 11, n. 1, p. 189-192, Feb. 1969.

WILLIAMS, E. J. **Regression analysis**. NewYork: Wiley, 1959. 214 p.

YWATA, A. X. C.; ALBUQUERQUE, P. H. M. Métodos e modelos em econometria espacial: uma revisão. **Revista Brasileira de Biometria**, São Paulo, v. 29, n. 2, p. 273–306, 2011.

ANEXOS

ANEXO A - Código usado no *software* R para o cálculo da estimação pontual e intervalar em regressão inversa linear simples.

```
invreg<-function(x, y, y0, alfa)
{
  # Esta função retorna o intervalo de confiança de x
  # dado uma nova observação y0
  # Em que:
  #     x: vetor de variáveis independentes;
  #     y: vetor de variáveis dependentes;
  #    y0: um valor especifico da variável dependente;
  #  alfa: nivel de significância;.

  n =length(x) #tamanho de x
  xbar=mean(x) #média de x
  k=length(y0) #tamanho da nova observação y0
  ybar=mean(y) #média de y
  x1=x-xbar
  y0bar=mean(y0) #média dos valores de y0

  #Estimação do modelo centrado
  reg <- lm(y~x1)

  #valor estimado x0 dado uma nova observação y0
  x0=xbar+(y0bar-ybar)/coef(reg)[2]

  #  Valores para gerar o intervalo de confiança

  tc=qt((1-alfa/2),reg$df.residual)
  s2yx=sum(reg$residual^2)/reg$df.residual

  a1= (coef(reg)[2]^2)-((s2yx*(tc^2))/(sum((x-xbar)^2)))
  m1=xbar+(coef(reg)[2]*(y-ybar)/a)
  h1=((tc)*sqrt(s2yx)/a)*sqrt(a*((1/n)+
        (1/k))+((y-ybar)^2)/(sum((x-xbar)^2)))

  #  Generando a banda de confiança para um dado valor y0.

  li1= m1-h
  ls1= m1+h

  resultado<-cbind(y0,li,x0,ls)

  #  Plotando o intervalo de confiança

  plot(x,y,ylab="", xlab=")
       reg2=lm(y~x)
```

```
        abline(reg2)
        lines(ls1,y,lty=2)
        lines(li1,y,lty=2)
        abline(h=y0,col="red")
        abline(v=x0,col="red",lty=2)
        abline(v=ls,col="blue")
        abline(v=li,col="blue")

        #Teste para o parâmetro beta 1

        if((coef(reg)[2]^2)*sum((x-xbar)^2)/s2yx>=(tc^2))
             {return(list(resultado[1,],0))}
        else{return(c(-infinity,infinty,x0,1))}

}


Simula=function(t,x,y)
{
        # Esta função retorna a proporção de simulações em que
        # o valor real esta no intervalo;

        rejeita=0
        resultado=0

        for(h in 1:t)
{
        #estimar os parâmetros da regressão
        reg2=lm(y~x1)

        #valor da variável x0 dado y0
        x0real=xbar+ (mean(y0) - coef(reg2)[1])/coef(reg2)[2]
        delta0=coef(reg2)[1]
        delta1=coef(reg2)[2]
        sigmareal2=sum(reg2$residual^2)/reg2$df.residual

        # obter n valores de y aleatórios de uma distribuição normal
        com média delta0+delta1*(x-xbar) e desvio padrão sigma

        y1=c(1:length(x))
        for(i in 1:length(x)){
        y1[i]=rnorm(1,delta0+delta1*(x[i]-xbar),sqrt(sigmareal2))
        }

        # obter n valores de y0 aleatórios de uma distribuição normal
        com média delta0+delta1*(x0-xbar) e desvio padrão sigma

        cc=rnorm(k,delta0+delta1*(x0real-xbar),sqrt(sigmareal2))

        #obter  intervalo de confiança
        int=invreg(x,y1,cc,alfa)
        intn=cbind(int[[1]])
```

```
      int1=intn[2,1]
      int2=intn[4,1]
      int3=intn[3,1]
      int4=int[[2]]
      #plotar o gráfico do intervalo de confiança

      if(int4==1){rejeita=rejeita+1}
      if(int4==0)
{
      if(x0real<=int2 & x0real>=int1)
  {
      resultado=resultado+1
  }}}

      return(list(resultado,rejeita))
}
```

ANEXO B - Código usado no *software* R para o cálculo da estimação pontual e intervalar em regressão inversa quadrática.

```
invreg2<-function(x, y, y0, alfa)
{
  # Esta função retorna o intervalo de confiança de x
  # dado uma nova observação y0
  # Em que:
  #     x: vetor de variáveis independentes.
  #      y: vetor de variáveis dependentes.
  #    y0: um valor especifico da variável dependente.
  #  alfa: nivel de significância.

  n=length(x) #tamanho de x
  xbar=mean(x) #média de x
  k=length(y0) #tamanho da nova observação y0
  ybar=mean(y) #média de y
  y0bar=mean(y0) #média dos valores de y0

  #Estimação do modelo
  reg=lm(y~x+I(x^2));reg
  #valor estimado x0 dado uma nova observação y0
  delt=(coef(reg)[2]^2)-4*coef(reg)[3]*(coef(reg)[1]-y0bar)
  rdelt=sqrt(delt)
  #estimação da variável desconhecida x0
  x0=((-coef(reg)[2]+rdelt)/(2*coef(reg)[3]));x0

  tc=qt((1-alfa/2),reg$df.residual)
  s2yx=(sum(reg$residual^2))/reg$df.residual
  syx=sqrt(s2yx)
  dx0y0=1/rdelt
  dx0b0=-1/rdelt
  dx0b1= (-1+(coef(reg)[2])/(rdelt))/(2*coef(reg)[3])
  dx0b2=(coef(reg)[2]-rdelt)/(2*(coef(reg)[3]^2))-
```

```
      (coef(reg)[1]-y0bar)/(rdelt*coef(reg)[3]);

  #variância da variável desconhecida x0
  dx=cbind(dxb0,dxb1,dxb2)
  dx0=cbind(dx0b0,dx0b1,dx0b2)
  vary0=s2yx/k;vary0
  sigy0=sqrt(vary0)
  varb=vcov(reg)
  varx0=((dx0y0*sigy0)^2)+(dx0)%*%varb%*%t(dx0);

  #intervalo de confiança
  xL=x0-tc%*%sqrt(varx0);
  xU=x0+tc%*%sqrt(varx0);

  plot(x,y,xlim=c(min(x,xu,xL),max(x,xu,xL)),
              ylim=c(min(y,y0),max(y,y0)))
  curve((coef(reg)[1]+coef(reg)[2]*x+coef(reg)[3]*x*x),add=T)
  par(new=T)
  abline(h=y0bar,col="red")
  abline(v=x0,col="red",lty=2)
  abline(v=xu,col="blue",lty=2)
  abline(v=xL,col="blue",lty=2)

  resultado=cbind(y0,varx0,xL,x0,xU)
  return(resultado)
}
```

ANEXO C - Código usado no *software* R para o cálculo da estimação pontual e intervalar em regressão inversa espacial.

```
########Conjunto de dados biblioteca R########
require(GeoXp)
library(spdep)
data(columbus)
Y=columbus[,9]##variável dependente:CRIME####
ybar=mean(Y)
X=columbus[,8]##Variável independente: rendimento familiar##
X1=X-mean(X)
n=length(X)##tamanho da amostra##
xn=cbind(rep(1,n),X1)# vetor de variáveis centradas na média
INC1=columbus[,8]-mean(columbus[,8]) #valores de renda familiar
#centrados na média

##########################MODELO SAR################################

COL.listw=nb2listw(col.gal.nb, style="W")#lista de vizinhaça

##Ajuste do Modelo SAR##

reg= lagsarlm(CRIME~INC1, data=columbus,
              nb2listw(col.gal.nb, style="W"),type="lag",
```

```
              method="eigen", quiet=FALSE);

#parâmetros do modelo SAR##
I = diag(length(columbus$INC))
rho=reg$rho
w=listw2mat(COL.listw)
B=cbind(coef(reg)[2],coef(reg)[3])
summary(reg, correlation=TRUE, Nagelkerke=TRUE)

#####1 caso: Estimativa de x assumindo que tem-se x e Y
#####todos conhecidos

invregest<-function(x, y, i, alfa){
  # Esta função retorna o intervalo de confiança de x
  # dado uma nova observação y0

  #Estimativa da variância
  S=((rho*w)+xn%*%solve(t(xn)%*%xn)%*%t(xn)%*%(I-rho*w))
  trs=sum(diag(S))
  trss=sum(diag(t(S)%*%S))
  SSE=deviance(reg)
  s2=SSE/(n-(2*trs)+(trss))

  #componetes individuais: Estimativa da variável independente
  xci=mean(X)+((Y[i]-rho*w[i,]%*%Y-coef(reg)[2])/coef(reg)[3]);

  tc=qt((1-alfa/2),(n-(2*trs)+(trss)))
  k=s2%*%(tc^2)

  ###################Intervalo de Confiança#####################

  at=t((solve(t(xn)%*%xn))%*%t(xn)%*%(I-rho*w))
  O=(solve(I-rho*(w)))%*%(solve((I-rho*t(w))))
  s=solve(t(xn)%*%xn)
  u=-2*(O[i,]-rho*w[i,]%*%O)%*%at
  c1=s[1,1]+u[1,1]+O[i,i]+((rho^2)*w[i,]%*%O-2*rho*O[i,])%*%w[i,]
  a1=((Y[i])-rho*w[i,]%*%(Y)-coef(reg)[2])


  #########Equação  quadrática: ax2+bx+C<=0##############
  a=(((coef(reg)[3])^2)-k*s[2,2]);
  b=-(2*a1*(coef(reg)[3])+k*u[1,2]+2*k*s[1,2])
  c=((a1^2)-k*c1)
  D=(b^2)-4*a*c

  #####limites inferior e superior do intervalo de confiança#####

  xL=mean(X)+(-b-sqrt(D))/(2*a)
  xU=mean(X)+(-b+sqrt(D))/(2*a)

  ########### resultado###########
  return(xci,c(xL,xU)))
```

```
}


#####2 caso: Estimativa de x assumindo que tem-se x e Y
#####conhecidos e X desconhecidos

#######separando o conjunto de dados em duas partes######
#r são as unidades a serem consideradas na primeira etapa
#h são as unidades não consideradas

invespc=function(h,r,i,alfa){

  tcolumbus<-readOGR(dsn = system.file("etc/shapes", package="spdep"),
                     layer = "columbus")

  spols<-polygons(tcolumbus)[r]

  # ===== CRIAR LISTA DE VIZINHOS E DE PESOS ============

  # Criar lista de vizinhos a partir do objeto "Spatial Polygon"
  polgal<-poly2nb(spols)

  # Criar uma lista de pesos a partir da lista de vizinhos:
  W_polgal<-nb2listw(polgal)

  Xc1=X[r]-mean(X[r]) #valores de XC centrados na média
  nc=length(Xc1)
  xnc=cbind(rep(1,nc),Xc1)

  #####estimando o modelo com as 46 observações selecionadas ####
  regc= lagsarlm(columbus$CRIME[r]~Xc1, data=columbus,
                 nb2listw(polgal, style="W"),type="lag",
                 method="eigen", quiet=FALSE);
  ###matriz de vizinhaça das 46 observações selecionada #####
  COL.listw1=nb2listw(polgal, style="W")
  wc=listw2mat(COL.listw1)

  Ic = diag(length(columbus$INC[r]))
  rhoc=regc$rho
  Bc=cbind(coef(regc)[2],coef(regc)[3])
  ######

  Sc=(rhoc*wc+xnc%*%solve(t(xnc)%*%xnc)%*%t(xnc)%*%(Ic-rhoc*wc))
  trsc=sum(diag(Sc))
  trssc=sum(diag(t(Sc)%*%Sc))
  SSEc=deviance(regc);
  s2c=SSEc/(nc-2*trsc+trssc)

###regiões em que só se conhece os valores observados de y###
  Y0=Y[h]
###Estimativas de componentes individuais###
```

```
   x0i=mean(X[r])+(Y0[i]-rhoc*(w[h[[i]],h]%*%Y0+w[h[[i]],r]%*%Y[r])-
                    coef(regc)[2])/(coef(regc)[3]);x0i

  tcc=qt((1-alfa/2),(nc-2*trsc+trssc))
  kc=s2c%*%(tcc^2);kc

  at=t((solve(t(xnc)%*%(xnc)))%*%t(xnc)%*%(Ic-rhoc*wc))
  O=(solve(I-rhoc*(w)))%*%(solve((I-rhoc*t(w))))

  ## observção i não pertencente a amostra a ser estimada

  c=O[h[[i]],h[[i]]]+(rhoc^2)*(w[h[[i]],h]%*%O[h,h]%*%
     cbind(w[h[[i]],h])+w[h[[i]],r]%*%O[r,r]%*%w[h[[i]],r]+
     2*w[h[[i]],h]%*%O[h,r]%*%cbind(w[h[[i]],r]))-
     2*rhoc*(O[h[[i]],h]%*%cbind(w[h[[i]],h])+
     O[h[[i]],r]%*%cbind(w[h[[i]],r]))

  ####calculando o intervalo###
  l1=Y0[i]-(rhoc)*(w[h[[i]],h]%*%Y0+w[h[[i]],r]%*%Y[r]);l1
  M1=solve(t(xnc)%*%xnc)
  l2=-2*(O[h[[i]],r]-rhoc*(w[h[[i]],h]%*%O[h,r]+
                    w[h[[i]],r]%*%O[r,r]))%*%at;

  a2=Bc[,2]^2-M1[2,2]*kc;a2
  b2=-(2*(l1-Bc[,1])*Bc[,2]-kc*(2*M1[1,2]+l2[1,2]))
  c2=(l1-Bc[,1])^2-kc*(M1[1,1]+l2[1,1]+c);c2

  delta=(b2^2)-(4*a2*c2);delta
  xU=mean(X[r])+(-b2+sqrt(delta))/(2*a2)
  xL=mean(X[r])+(-b2-sqrt(delta))/(2*a2)


  list(x0i,c(xL,xU))
}
```