Estimação dos parâmetros angular e linear da equação de regressão linear simples pelo método não-paramétrico

[1]Alícia Bolfoni Dias, [2]Silvano Bolfoni Dias, [3]Luciane Flores Jacobi

[1]CEEMQ - CCNE/UFSM
e-mail:aliciabdias@mail.ufsm.br
[2]Centro de Processamento de Dados - UFSM
e-mail: silvano@cpd.ufsm.br
[3]Departamento de Estatística - CCNE/UFSM
e-mail: lfjacobi@ccne.ufsm.br

Resumo

Quando as pressuposições da análise de regressão linear simples falharem, uma alternativa para a estimação dos coeficientes da equação de regressão é o método não-paramétrico. O objetivo deste estudo foi comparar as estimativas para os coeficientes do modelo de regressão linear simples pelo método não-paramétrico. No método não-paramétrico foi considerado o estimador de Theil *apud* Daniel (1999) para o coeficiente angular (β) e os dois estimadores para o intercepto (α) propostos por Dietz *apud* Daniel (1999). Utilizou-se dados de massa corporal e estatura de crianças e adolescentes na faixa etária de 11 e 14 anos do Município de Nova Palma - RS, avaliou-se medidas para 59 meninas e 67 meninos. Obteve-se para os métodos paramétrico e não-paramétrico três equações para cada um dos gêneros, comparando-as através dos critérios Akaike Information Criteria (AIC) e Bayesian Information Criteria (BIC) e do cálculo do erro quadrado médio. Concluiu-se que as estimativas encontradas pelos métodos foram muito próximas, não havendo grandes diferenças entre AIC e BIC e os erros quadrados médios das equações.

Palavras-chave: Regressão Linear Simples, Regressão Não-Paramétrica, AIC e BIC, Erro Quadrado Médio.

Abstract

When the analysis' assumptions of simple regression are not satisfied, an alternative to estimation the coefficients of the regression equation is the nonparametric method. The objective was to compare the estimates for the coefficients of the simple linear regression model by the least squares method with the nonparametric method. We used data of corporal mass and stature of children and teenagers between 11 and 14 years old from Nova Palma (RS) town, which 59 were girls and 67 were boys. In the nonparametric method was considered the estimador of Theil apud Daniel (1999) for the slope coefficient (β) and the two estimators proposed by Dietz apud Daniel (1999) for the intercept (α). Three equations were obtained for each one of the sexes and they were compared by the Akaike Information Criteria (AIC), the Bayesian Information Criteria (BIC) and mean squared error. We concluded that the estimates we found by the two methods were very close, with small differences among AIC, BIC and the mean squared error of the three equations.
Key words: Simple Linear Regression, Nonparametric Regression, AIC and BIC, Mean Squared Error.

1. Introdução

O procedimento não-paramétrico é empregado, quando uma ou mais suposições fundamentais da análise de regressão linear simples não são válidas para estimar os coeficientes angular e linear da equação de regressão. Este método possui vantagens e desvantagens. A principal vantagem é que não são exigidas as suposições sobre a população da qual se originam os dados. Uma de suas desvantagens é de ele ser muito trabalhoso para estimar os coeficientes angular e linear da equação de regressão.

O método não-paramétrico é pouco aplicado, pois não existe um software apropriado para se estimar os coeficientes angular e linear da equação da reta, por isso o objetivo deste trabalho é implementar uma planilha do Excel à metodologia, programando-a por meio do Visual Basic for Applications (VBA).

Para exemplificar a técnica utilizada, ajustou-se a massa corporal de 126 adolescentes em função de sua estatura. Os 126 dados, sendo 59 meninas e 67

meninos, foram obtidos no Município de Nova Palma - RS.

Além disso, estimou-se os coeficientes do modelo de regressão pelo método paramétrico que foi comparado com o estimado pela metodologia proposta.

2. Metodologia

2.1. Métodos não-paramétricos para estimar os coeficientes da equação de regressão

2.1.1. O Estimador para o coeficiente angular

Para determinar a estimativa do coeficiente angular, utilizou-se a proposta de Theil *apud* Daniel (1999) determina um ponto estimado de β, usando o modelo clássico $y_i = \alpha + \beta x_i + \varepsilon_i$, i = 1,...,n. Onde x_i são constantes conhecidas, α e β são parâmetros desconhecidos, y_i um valor observado da variável contínua Y em xi, e ε_i são mutuamente independentes. Os dados consistem em n pares de observações amostrais $(x_1, y_1),..., (x_n,y_n)$.

Para se obter o estimador β, primeiramente forma-se todas as possíveis inclinações entre dois pontos da amostra, conforme mostra a Figura 1:

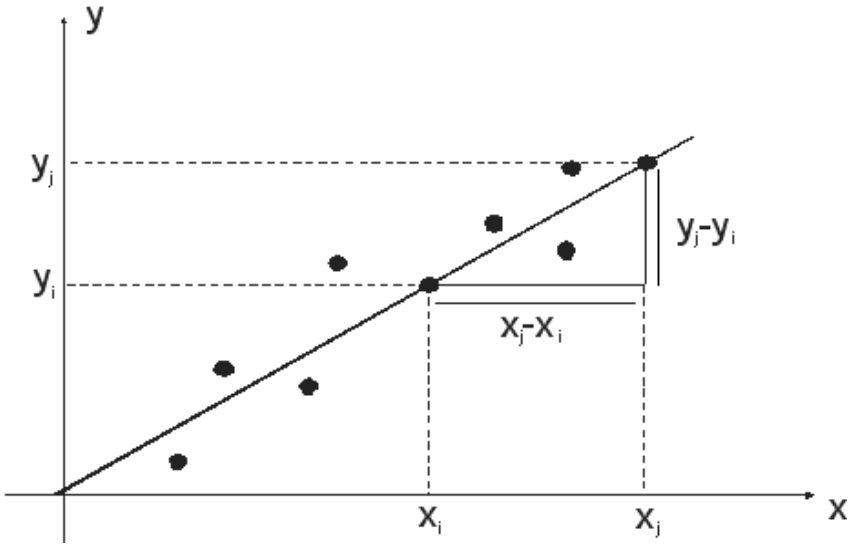

Figura 1. Interpretação geométrica dos valores de S_{ij}

Onde:

$$S_{ij} = (y_j - y_i)/(x_j - x_i) \qquad (1)$$

para $i < j$ encontra-se $N = {}_nC_2$ valores para S_{ij} que é o coeficiente angular da reta que passa por dois pontos. Assim, o cálculo de S_{ij} $\forall$ i, j, representa todas as combinações possíveis de coeficientes angulares de retas tomadas, duas a duas, em todos os pontos do diagrama de dispersão. O estimador β que se designa através de $\hat{\beta}$, é a mediana dos valores de S_{ij}. Assim $\hat{\beta}$ = mediana $\{S_{ij}\}$ (Dietz, 1989).

Deste modo para n pares de valores tem-se:

$$S_{12} = (y_2 - y_1)/(x_2 - x_1)$$
$$S_{13} = (y_3 - y_1)/(x_3 - x_1)$$
$$\vdots$$
$$S_{1n} = (y_n - y_1)/(x_n - x_1)$$
$$S_{23} = (y_3 - y_2)/(x_3 - x_2)$$
$$S_{24} = (y_4 - y_2)/(x_4 - x_2)$$
$$\vdots$$
$$S_{2n} = (y_n - y_2)/(x_n - x_2)$$
$$\vdots$$
$$S_{n-1,\,n} = (y_n - y_{n-1})/(x_n - x_{n-1})$$

Conforme Sen (1968), os S_{ij} são as inclinações das retas ligando cada par de pontos (x_i, y_i) e (x_j, y_j) onde $x_i \neq x_j$; os pares de pontos para os quais x_i é igual a x_j não são considerados.

2.1.2. Os estimadores para o coeficiente linear

Determina-se os valores do coeficiente linear da equação de regressão conforme Dietz *apud* Daniel (1999) por dois estimadores $\hat{\alpha}_1$ e $\hat{\alpha}_2$.

Para determinar $\hat{\alpha}_1$, encontra-se a partir dos n pares de valores observados, e da equação (2) todas as possíveis estimativas para α, assim tem-se:

$$Y_1 - \hat{\beta}X_1, Y_2 - \hat{\beta}X_2, \ldots, Y_i - \hat{\beta}X_i \qquad (2)$$

Após determina-se a mediana dos possíveis valores estimados para α, sendo a mesma a estimativa de $\hat{\alpha}_1$, quando se assume que os erros não são simétricos.

Para determinar $\hat{\alpha}_2$, primeiro determina-se a média para os valores da variável dependente e dos valores da variável independente consideradas em S_{ij}.

Assim:

$$\overline{Y}_{12} = \frac{y_2 + y_1}{2} \quad \text{e} \quad \overline{X}_{12} = \frac{x_2 + x_1}{2}$$

$$\overline{Y}_{13} = \frac{y_2 + y_3}{2} \quad \text{e} \quad \overline{X}_{13} = \frac{x_2 + x_3}{2}$$

$$\vdots$$

$$\overline{Y}_{1n} = \frac{y_n + y_1}{2} \quad \text{e} \quad \overline{X}_{1n} = \frac{x_n + x_1}{2}$$

$$\overline{Y}_{23} = \frac{y_3 + y_2}{2} \quad \text{e} \quad \overline{X}_{23} = \frac{x_3 + x_2}{2}$$

$$\overline{Y}_{24} = \frac{y_4 + y_2}{2} \quad \text{e} \quad \overline{X}_{24} = \frac{x_4 + x_2}{2}$$

$$\vdots$$

$$\overline{Y}_{2n} = \frac{y_n + y_2}{2} \quad \text{e} \quad \overline{X}_{2n} = \frac{x_n + x_2}{2}$$

$$\vdots$$

$$\overline{Y}_{n-1,n} = \frac{y_n + y_{n-1}}{2} \quad \text{e} \quad \overline{X}_{n-1,n} = \frac{x_n + x_{n-1}}{2}$$

Após aplica-se na equação (2) os $\overline{Y}_s$ e $\overline{X}_s$ encontrados. Junta-se a esses, os valores encontrados para $\hat{\alpha}_1$, determinando-se desta forma $\frac{n(n+1)}{2}$, possíveis estimativas para $\hat{\alpha}_2$ sendo que a mediana desses valores será a estimativa do intercepto, quando se considera que os erros são simétricos.

2.1.3. Critérios "Akaike Information Criteria" (AIC) e "Bayesian Information Criteria" (BIC)

As formas de comparação dos modelos comumente utilizadas são a análise dos resíduos e a avaliação da ordem do modelo cujos critérios mais usados são os AIC e BIC, que levam em conta a variância do erro, o tamanho da amostra e os valores dos coeficientes estimados. Segundo Farias, Rocha e Lima (2000):

AIC = n ln(somatório dos quadrados dos resíduos) + 2T (3)

BIC = n ln(somatório dos quadrados dos resíduos) + T ln(n) (4)

Onde: n = número de observações utilizadas;
T = número de parâmetros estimados.

2.1.4. Visual Basic for Applications (VBA)

O VBA é a linguagem de programação utilizada no programa Microsoft Excel, do tipo planilha eletrônica, onde foi implementado um algoritmo para se estimar os coeficientes angular e linear, calcular os critérios, e a construção do gráfico com o intuito de facilitar a utilização desta técnica não-paramétrica.

2.1.5. Cálculo do erro quadrado médio (EQM)

Uma outra maneira de se comparar modelos é determinar os desvios em relação aos valores observados da variável Y, ou seja, calcula-se a diferença entre o valor observado para Y e sua respectiva estimativa $\hat{y}$ determinada pelas equações ajustadas. Elevam-se esses desvios ao quadrado e divide-se a soma dos quadrados pelo número de valores amostrados, encontra-se, o erro quadrado médio determinado por:

$$EQM = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n} \qquad (5)$$

3. Resultados e discussão

Com os dados de massa corporal e estatura encontrou-se, 2211 combinações ($_{67}C_2$) possíveis de inclinações entre os pares de valores observados para os meninos, e para as meninas foram 1711 possíveis combinações ($_{59}C_2$). Para se estimar β foi usada a equação (1) dos valores de S_{ij}.

Após a determinação de todas as inclinações possíveis, determinou-se a mediana dos valores encontrados, estimando-se dessa forma o valor do coeficiente angular da equação de regressão para os meninos como sendo 75,0000 e para as meninas 90,0000.

Para estimar o parâmetro α, utilizou-se as duas metodologias propostas por Dietz *apud* Daniel (1999). Na primeira, designada por $\hat{\alpha}_1$, encontra-se a medi-

ana das condições de n termos determinados pela equação $Y_i - \hat{\beta} X_i$, onde Y_i e X_i são os valores observados para a variável dependente e independente, respectivamente e $\hat{\beta}$ é o estimador de Theil. O valor encontrado dessa forma para α foi de -71,1000 para os meninos e -93,6000 para as meninas.

Na segunda, estima-se α por $\hat{\alpha}_2$, que consiste em calcular a mediana dos $\frac{n(n+1)}{2}$ possíveis valores para α na aplicação da equação $Y_i - \hat{\beta} X_i$, onde Y_i e X_i são os valores observados para a variável dependente e independente, respectivamente, e além disso, as médias desses valores nas combinações realizadas para encontrar o estimador de β e $\hat{\beta}$ é o estimador de Theil. Os valores encontrados dessa forma para α foi de -71,0000 para os meninos e - 94,4000 para as meninas.

3.1. Comparação das equações estimadas

Nas tabelas a seguir serão mostradas as estimativas para os coeficientes angular e linear obtidos, AIC e BIC, e os erros quadrados médios das equações estimadas para as variáveis estatura e massa corporal de meninos e meninas.

Tabela 1. Estimativas para os coeficientes linear (α) e angular (β), pelos métodos paramétrico e não-paramétrico; Akaike Information Criteria (AIC) e Bayesian Information Criteria (BIC) e o erro quadrado médio (EQM) para os meninos

	$\hat{\alpha}$	$\hat{\beta}$	AIC	BIC	EQM
$\hat{Y}_1$ Paramétrica	-67,8105	73,4520	565,3451	569,7545	64,9484
$\hat{Y}_2$ Não-Paramétrico ($\hat{\alpha}_1$)	-71,1000	75,0000	566,0378	570,4472	65,6233
$\hat{Y}_3$ Não-Paramétrico ($\hat{\alpha}_2$)	-71,0000	75,0000	565,8834	570,2928	65,4723

Pela análise da Tabela 1, observou-se que AIC e BIC e o erro quadrado médio pelo método paramétrico são menores quando comparados com os erros quadrados médios e AIC e BIC do método não-paramétrico.

Pelo método não-paramétrico verificou-se que AIC e BIC e o erro quadrado médio para a equação $\hat{Y}_3$ são menores que a da equação $\hat{Y}_2$, isso se deve ao fato do estimador $\hat{\alpha}_2$ estar centrado no zero, ou seja, os erros são simétricos, portanto, a estimação de α é melhor quando se usa $\hat{\alpha}_2$ para estimá-lo, embora não existindo grande diferença entre os critérios adotados para comparação.

Tabela 2. Estimativas para os coeficientes linear (α) e angular (β), pelos métodos paramétrico e não-paramétrico; Akaike Information Criteria (AIC) e Bayesian Information Criteria (BIC) e o erro quadrado médio (EQM) para as meninas

	$\hat{\alpha}$	$\hat{\beta}$	AIC	BIC	EQM
$\hat{Y}_1$ Paramétrica	-90,7390	87,8080	471,6994	475,8545	46,9729
$\hat{Y}_2$ Não-Paramétrico ($\hat{\alpha}_1$)	-93,6000	90,0000	472,1983	476,3534	47,3718
$\hat{Y}_3$ Não-Paramétrico ($\hat{\alpha}_2$)	-94,4000	90,0000	471,7780	475,9331	47,0355

Pela análise da Tabela 2, observou-se que as estimativas para os dados das meninas seguem a mesma conclusão da dos meninos, ou seja, os critérios foram menores pelo método paramétrico; e pelo método não-paramétrico a equação que obteve os menores AIC e BIC e o erro quadrado médio foi a que ao se estimar a foram considerados os erros simétricos em relação a zero.

4. Conclusão

Concluiu-se que as estimativas encontradas pelos dois métodos foram muito próximas, não havendo grandes diferenças entre os critérios utilizados para comparação das equações estimadas. Diante disso, observou-se que a metodologia não-paramétrica não traz muitas divergências em relação à paramétrica; sendo, portanto uma boa alternativa em que se tenha uma amostra de tamanho pequena, ou quando alguma suposição do método paramétrico não seja atendida.

Através da implementação do algoritmo por meio da linguagem de programação VBA, para o modelo de regressão linear simples pelo método não-paramétrico facilitou a estimação dos coeficientes angular e linear da equação, permitindo, dessa forma, sua utilização de maneira facilitada, pois esse método exige uma grande quantidade de cálculos.

Agradecimentos

Agradecimento para a Professor de Educação Física Cassiano Ricardo Rech, que forneceu os dados para a possível exemplificação deste trabalho, e para a Professora de Língua Portuguesa Ledi Bolfoni Dias pela revisão.

Referências bibliográficas

DANIEL, W.W. *Biostatistics*: A Foundation for Analysis in the Health Sciences. 7th ed. p.717 - 720. 1999.

DIETZ, E. J. Teaching regression in a Nonparametric Statistics Couse. *The American Statistican*, vol.43, n.1, p.35-40. fev.1989.

FARIAS, E.R. de; ROCHA, F.J.S; LIMA, R.C. Critérios de seleção de modelos sazonais de séries temporais: Uma aplicação usando a taxa de desemprego da região Metropolitana de Recife. III Encontro Regional de Estudos do Trabalho - ABET, 22 a 24 de novembro 2000 - Recife. Disponível em: <http:// www.race.nuca.ie.ufrj.br/abet/3seg>. Acesso em: 13 abr. 2005.

SEN, K. P. Estimates of Regression Coefficient Based on Kendall's Tau. *Journal of the American Statistical Association*, v. 63, p.1379-1389, dez 1968.