SCIENTIA PLENA VOL. 13, NUM. 04 2017
www.scientiaplena.org.br doi: 10.14808/sci.plena.2017.049910

O método dos mínimos quadrados no ajuste de um modelo polinomial

Evaluation of multiple least square method on adjust of an linear regression model

V. M. Silva[1*]; V. L. D. Mattos[2]

[1]*Centro de Ciências Computacionais, Universidade Federal do Rio Grande, CEP, Rio Grande-RS, Brasil*

[2]*Instituto de Matemática, Estatística e Física, Universidade Federal do Rio Grande, CEP, Rio Grande-RS, Brasil*

**vinicius.montenegro@furg.br*

(Recebido em 24 de abril de 2017; aceito em 22 de maio de 2017)

A regressão linear pode ser utilizada para determinar o padrão de variação estacional de uma série temporal, o que pode ser feito por meio do método dos mínimos quadrados. Este método consiste em minimizar a soma dos quadrados dos resíduos obtidos pela diferença entre valores observados e valores preditos, o que pode ser feito para diversos modelos lineares. Este estudo ajusta um modelo polinomial de regressão linear múltipla a uma série temporal em que existe uma quebra estrutural, avaliando seu desempenho por meio de alguns indicadores de qualidade: coeficiente de determinação, teste de hipóteses para os coeficientes do modelo, além de análise gráfica dos resíduos. O método possibilitou encontrar um modelo polinomial de sexta ordem para modelar o comportamento da série estudada: o número de professores na escola primária no Brasil, no período entre 1999 e 2014.
Palavras-chave: Séries Temporais, Regressão Linear, Método dos Mínimos Quadrados.

The linear regression can be used to determine the pattern of seasonal variation of a time series, it can be done through the least squares method. This technique consists in minimizing the sum of the squares of residues obtained by the difference between observed values and predicted values, it can be done for different linear models. This study adjusts a multiple linear regression polynomial model to a time series in which exists a structural break, evaluating its performance through of some quality indicators: coefficient of determination, hypothesis tests for the coefficients of the model, also graphical analysis of residues. The method made it possible to find a sixth order polynomial model to modeling the behavior of the studied series: the number of teacher in primary school in Brazil between 1999 and 2014.
Keywords:Time series, Linear Regression, Least Square Method.

1. INTRODUÇÃO

A análise de regressão diz respeito ao estudo da relação de uma variável dependente com outra ou outras variáveis independentes, visando determinar uma função matemática que busca descrever o comportamento desta variável dependente com base nos valores da ou das variáveis independentes. Na maioria das vezes, trata-se de uma tentativa de prever o valor médio da variável dependente em termos das demais [1, 2, 4], investigando a relação entre as variáveis envolvidas de maneira não determinística [3].

O termo regressão surgiu no século XIX, a partir dos trabalhos de Galton, que tentavam explicar certas características dos indivíduos a partir de algumas características de seus pais. O modelo matemático definido nesse estudo foi aperfeiçoado e hoje ainda é usado em diversas áreas da ciência.

Uma das aplicações da regressão trata da previsão de dados a partir de séries temporais, procedimento importante na tomada de decisão sobre fatos futuros, que se utiliza de informações retroativas. Tal prática possibilita a elaboração de um planejamento eficiente, o que pode ser um diferencial competitivo para uma organização. Dentre os métodos de previsão, existem aqueles considerados mais simples, como o método naive, assim como métodos ditos complexos, como redes neurais. Dentre estes métodos de previsão, a análise de regressão desempenha um papel importante, sendo frequentemente utilizada no processo de estimação dos parâmetros de um modelo. Como estes parâmetros são determinados a partir de amostras, conceitos de estatística

inferencial devem ser levados em consideração, por possibilitar a realização de generalizações sobre uma população com base em informações amostrais.

Uma das técnicas que pode ser usada na estimação dos parâmetros de um modelo é o método dos mínimos quadrados (MMQ). Neste método, para que a função calculada seja a que melhor se ajusta aos dados, deve-se minimizar a soma dos quadrados das diferenças entre os valores observados e aqueles definidos pela curva de ajuste. Entretanto, para que o resultado encontrado seja válido, é necessário que estas diferenças, denominadas de erros ou resíduos, satisfaçam determinadas condições, entre as quais, normalidade.

O estudo proposto visa analisar o desempenho deste método na estimação dos coeficientes de uma curva que permita uma previsão do número de professores atuantes no nível primário do Brasil, nos anos de 2015 e 2016, a partir do conhecimento do número de professores que atuaram em anos anteriores. Estes dados foram escolhidos em função de algumas de suas propriedades: apresentam uma queda estrutural em determinado ano, além de um dado faltante. Esta é a etapa inicial de um estudo que pretende utilizar a estratégia proposta em [6], que combina dois métodos recentemente desenvolvidos para a aplicação do MMQ: o algoritmo Christoffel de mínimos quadrados e amostragem quase-ótima que, segundo os autores, resulta em um método dos mínimos quadrados polinomial de alta precisão e estabilidade robusta em amostras pequenas. Parte-se do princípio que a variável analisada, quantidade de professores atuantes nos anos iniciais do ensino fundamental no Brasil (da primeira até a quarta série), é uma quantidade que pode variar diariamente durante o ano, o que justificaria a utilização do procedimento citado acima.

O artigo está divido como segue. A seção 2 apresenta o procedimento metodológico, onde é informado como os dados foram obtidos, além do protocolo da análise para encontrar um modelo satisfatório, enquanto a seção 3 apresenta e discute os resultados encontrados. A seção 4 apresenta as considerações finais sobre os resultados da aplicação e suas limitações, sugerindo possíveis trabalhos futuros.

2. MATERIAL E MÉTODOS

Os dados utilizados no presente estudo foram extraídos do repositório do site da *Unesco* [8] e se referem ao número de professores, de ambos os sexos, atuantes nos anos iniciais do ensino fundamental no Brasil, no período 1999 a 2014.

A análise iniciou com o preenchimento de um único valor faltante na série de dados, valor esse correspondente ao ano de 2006. Tal valor foi preenchido com o resultado da média aritmética dos valores presentes na série, correspondente a$\mu = 813712{,}4$.

Após, foi realizada uma análise exploratória da distribuição dos dados por meio de métodos analíticos e gráficos, a fim de verificar as possíveis técnicas a serem utilizadas na geração do modelo. Os métodos analíticos envolveram caracterização da tendência central, dispersão, assimetria e curtose, além de testes de hipóteses (Kolmogorov-Smirnov e Shapiro-Wilk) para verificar se existem evidências de que os dados provêm de uma distribuição normal. Os métodos gráficos envolveram a construção de histograma, *boxplot* e gráfico de probabilidade normal, que também servem para identificar lacunas e *outliers*.

Este procedimento sugeriu a utilização do Método dos Mínimos Quadrados (MMQ) para o ajuste de uma função polinomial, que é especialmente indicada para representar funções que não são rigorosamente crescentes ou decrescentes.

A forma geral de um modelo de regressão polinomial de grau *m* é apresentada na Equação 1.

$$E(y/x_i) = \widehat{y_i} = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_j x_i^j + \cdots + \beta_m x_i^m \quad (1)$$

onde:
$E(y/x_i)$ representa o valor esperado da variável dependente y na observação$i = 1, \ldots, n$;
x_i diz respeito ao valor da variável independente x na observação$i = 1, \ldots, n$;
β_0 representa o ponto em que a função corta o eixo dos y;
β_j diz respeito ao coeficiente angular associado à *j-ésima* potência da variável independente x.

O valor observado da variável dependente associada a um valor de x_i é descrito pela Equação 2:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_j x_i^j + \cdots + \beta_m x_i^m + \varepsilon_i \quad (2)$$

onde:

y_i representa o valor da variável dependente y na observação$i = 1, \ldots, n$;
ε_i corresponde ao erro na observação$i = 1, \ldots, n$, e mostra a variabilidade existente em y, não explicada por x.

Pelo MMQ, a minimização da soma dos quadrados dos erros para encontrar os parâmetros do modelo foi feita conforme Equação 3.

$$q = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \quad (3)$$

Para encontrar o modelo que melhor se ajustava a distribuição dos dados foram testados polinômios de diferentes graus, sendo o melhor identificado a partir de uma análise embasada em métodos analíticos e gráficos. Os métodos analíticos utilizados foram o coeficiente de determinação (Equação 4) e respectivo valor ajustado (Equação 5), além do teste para avaliar a qualidade do ajuste que utiliza a distribuição de Fischer-Snedecor (Quadro 1), que verifica a hipótese nula de que todos os coeficientes do modelo são nulos, ou seja, $H_0: \beta_i = 0 \quad para \ todo \ i$.

$$R^2 = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \quad (4)$$

$$R_a^2 = 1 - \left(\frac{n-1}{n-(k+1)}\right)(1 - R^2) \quad (5)$$

Quadro1: Resumo típico do teste F.

Fontes de Variação	**Soma dos Quadrados SQ**	**Graus Liberdade Gl**	**Quadrados Médios QM**	**F**
Regressão	$\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$	*M*	$\frac{SQ_{regressão}}{QM_{regressão}}$	$\frac{QM_{regressão}}{QM_{\text{erro}}}$
Erro (resíduos)	$\sum_{i=1}^{n} (y - \hat{y})^2$	*n-m-1*	$\frac{SQ_{erro}}{GL_{erro}}$	
Total	$\sum_{i=1}^{n} (y - \bar{y})^2$	*n-1*		

Fonte: adaptado de [9].

Obs.: Nas expressões utilizadas, n é a quantidade de observações; y_i é a i-ésima observação; $\bar{y}$ é a média das observações; $\hat{y}_i$ é o valor predito para a i-ésima observação; m é o grau do polinômio estimado.

Entre os métodos gráficos utilizados, aplicados aos resíduos, estão: histograma, *boxplot*, gráfico de probabilidade normal e diagrama de dispersão entre resíduos e valores preditos.

Finalizando, foram feitas estimações por ponto da variável em estudo para os anos de 2015 e 2016, complementadas por intervalos construídos com 95% de confiança. Este estudo será complementado com a aplicação e análise de desempenho da estratégia proposta em [6].

3. RESULTADOS E DISCUSSÃO

Entre os diferentes modelos polinomiais obtidos pela aplicação do MMQ, o identificado como apresentando melhor ajuste foi a curva de regressão polinomial de sexta ordem (*m=6*), apresentada na Equação 6:

$$\hat{y}_i = 1{,}5495x^6 - 91{,}184x^5 + 2042{,}6x^4 - 21372x^3 + 101894x^2 - 181630x + 908659 \quad (6)$$

O coeficiente de determinação deste modelo $(R^2 = 0{,}91)$ e o respectivo valor ajustado $(R_a^2 = 0{,}84)$ apresentaram evidências de que a qualidade do ajuste pode ser considerada satisfatória, pois pelo primeiro, o modelo consegue explicar *91%* das variações da variável em questão.

O Quadro2 apresenta os resultados obtidos pelo teste *F* na verificação da hipótese $H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = 0$. Como o valor-p encontrado é menor que o nível de significância adotado, rejeita-se a hipótese nula. Logo, foram encontradas evidências de que, pelo menos um dos coeficientes é diferente de zero $(\beta_i \neq 0)$, indicando que o modelo pode ser considerado satisfatório.

Quadro2: Resultados do teste F.

Fontes de Variação	**SQ**	**-gl**	**QM**	**F**	**valor p**
Regressão	2,95e^{+10}	6	4924876819	14,61	0,000349
Erro	3032775224	9	336975024,9		
Total	3,25 e^{+10}	15			

A análise dos resíduos também sugere que o modelo seja de boa qualidade, sendo encontrado para sua média um valor próximo de zero, que é o ideal, ou seja, $\mu_R = 1{,}30x10^{-9}$.

Os gráficos das Figuras 1, 2, 3, 4 e 5 sugerem que a suposição da normalidade dos resíduos é verificada. No *boxplot* da Figura 1, a distribuição mostra-se com uma simetria quase perfeita, sendo possível identificar dois *outliers* (um superior e outro inferior). Estes se referem aos valores: imediatamente anterior e posterior ao valor faltante na série.

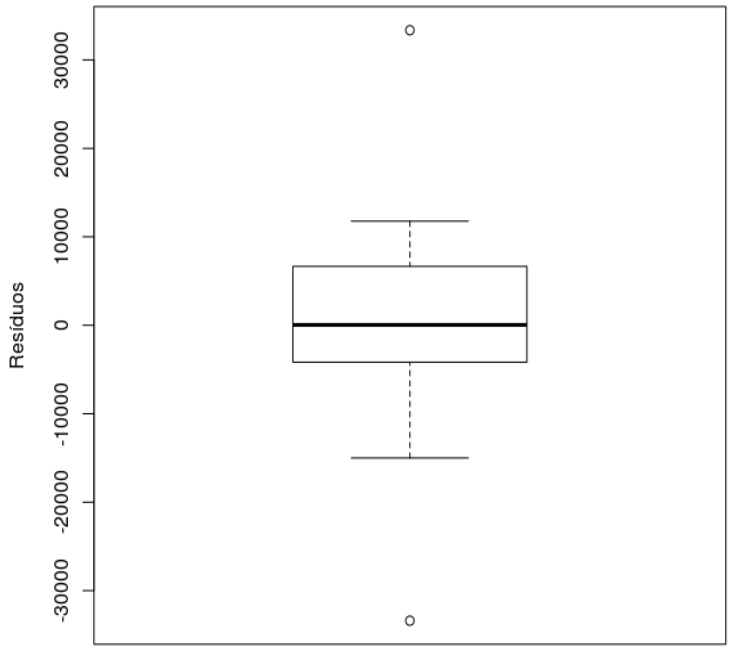

Figura 1: Boxplot de Resíduos

Na Figura 2, observa-se no gráfico de probabilidade normal dos resíduos, que os pontos não se distanciam muito da reta, o que caracteriza uma distribuição próxima à normal.

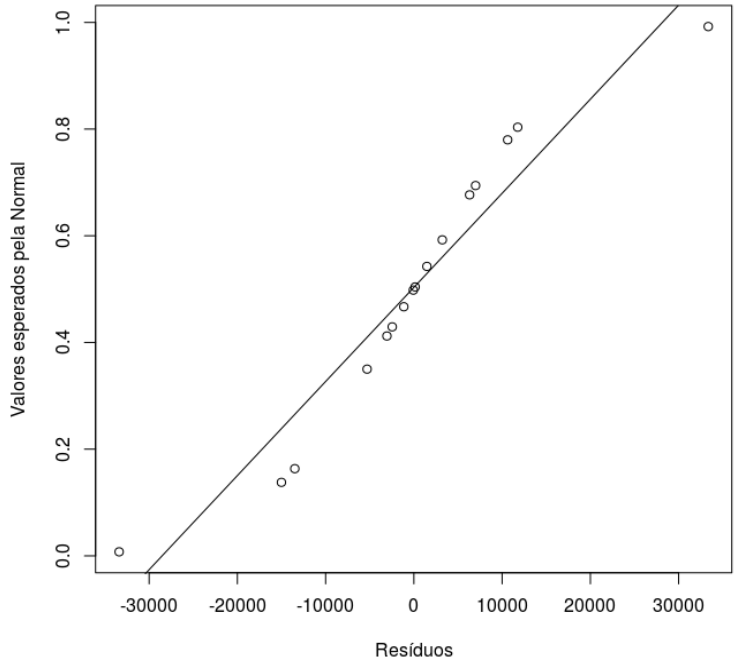

Figura 2: Gráfico de Probabilidade Normal dos Resíduos

Ainda, o teste de Shapiro-Wilk resultou em um p-valor de 0.0585, sendo esse maior que o nível de significância adotado (0,05), indicando a existência da normalidade dos resíduos. Por fim, em relação à normalidade, a Figura 3, mostra que os resíduos se distribuem igualmente abaixo e acima da linha, não sendo possível identificar um padrão.

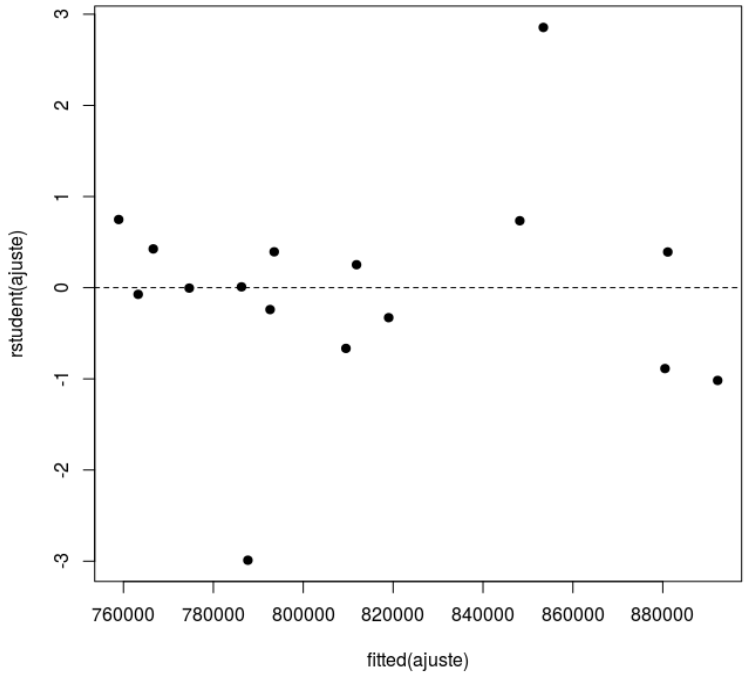

Figura 3: Distribuição Homogênea de Resíduos

No gráfico da Figura 4, observa-se que os resíduos não apresentam nenhum padrão linear de variabilidade crescente ou decrescente em função de sua distribuição no tempo, sinalizando para um modelo bem ajustado.

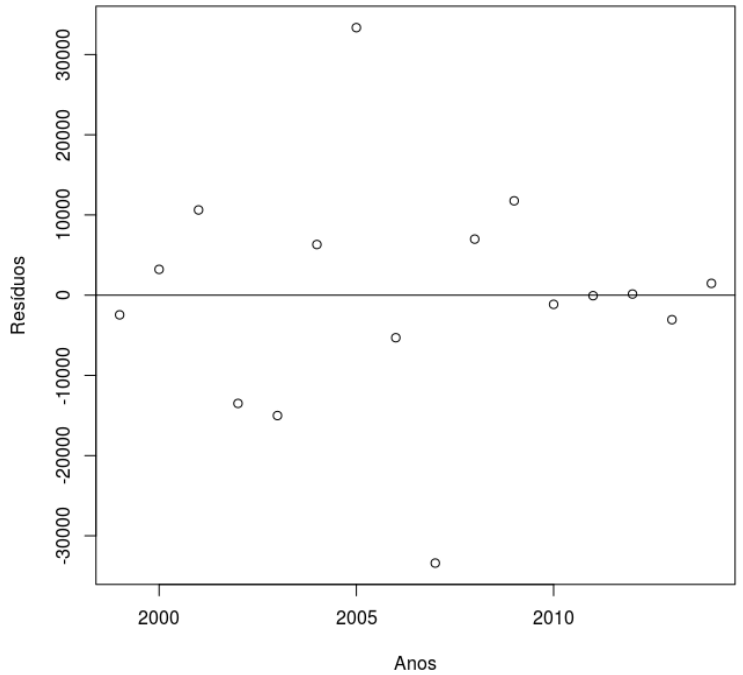

Figura 4: Distribuição dos Resíduos no Tempo.

O gráfico apresentado na Figura 5 mostra o cruzamento entre valores preditos e resíduos, observando-se uma leve tendência de aumento da variabilidade dos resíduos correspondentes a maiores valores preditos, o que não é satisfatório.

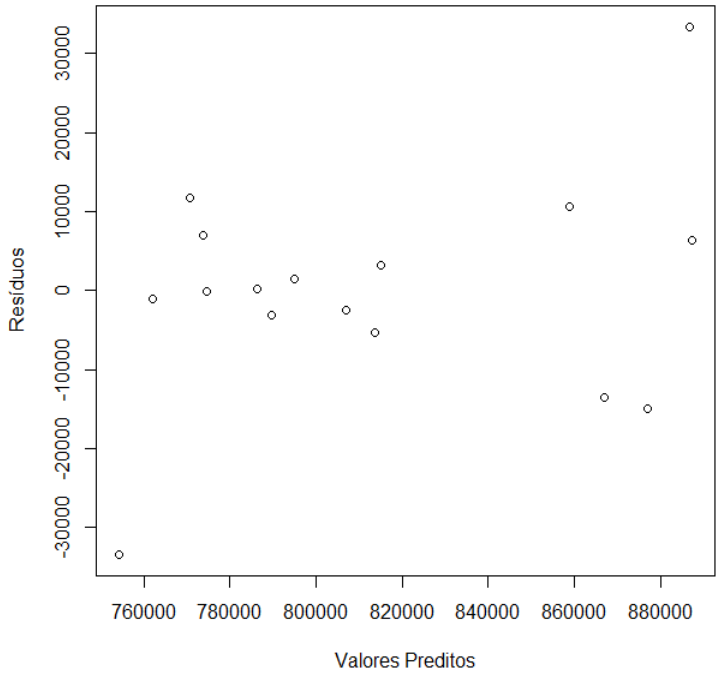

Figura 5: Gráfico de Valores Preditos x Resíduos

A Figura 6 mostra o comportamento dos valores observados e dos valores preditos no período analisado, evidenciando que o modelo encontrado pode ser considerado razoável, concordando com os resultados da análise estatística realizada.

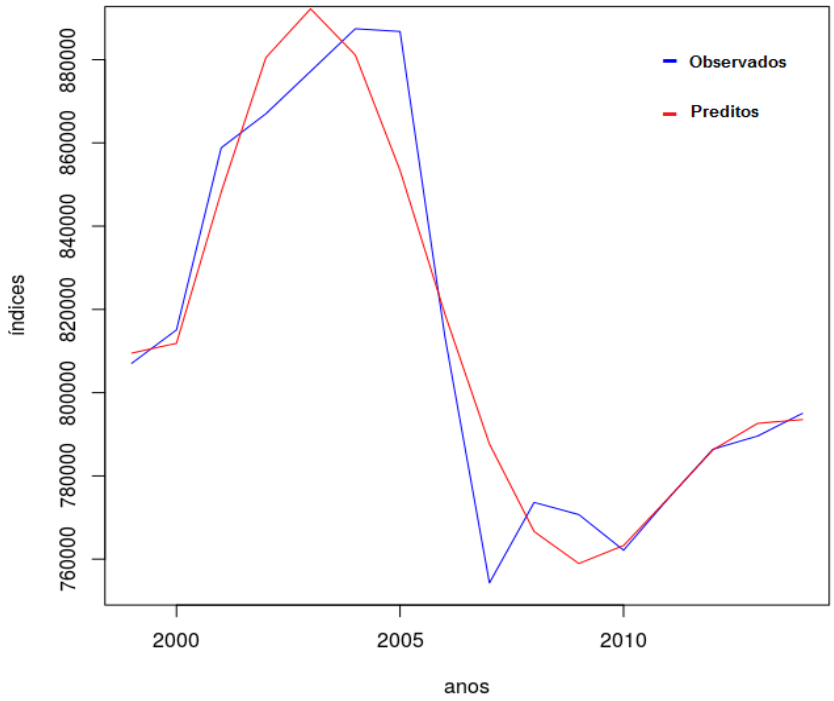

Figura 6: Gráfico de Valores Observados x Valores Preditos

Finalizando, o modelo encontrado foi utilizado para realizar uma estimação por ponto da quantidade esperada de professores nos anos iniciais do ensino fundamental no Brasil para os anos de 2015 e 2016, encontrando respectivamente, 800596,1 e 839021,2. Observa-se, entretanto, que esses valores estão bastante próximos dos valores encontrados pelos métodos mais simples de previsão (média e naive). Mais detalhes sobre estes métodos podem ser encontrados em [5].

4. CONCLUSÃO

Este estudo apresenta uma aplicação do MMQ na definição de um modelo para estimar a quantidade de professores atuantes nos anos iniciais do ensino fundamental no Brasil. A série de dados analisada se caracterizava por ser anual, apresentando algumas peculiaridades: inicialmente era ascendente, nos anos centrais apresentava uma queda brusca, além de um dado faltante, caracterizando uma quebra estrutural e, posteriormente, tornava-se novamente ascendente.

O modelo encontrado apresentou evidências de que se adapta aos dados observados como curva de ajuste. Entretanto, a análise foi desenvolvida apenas considerando modelos de regressão polinomial. Em sequência, espera-se melhorá-lo aplicando a estratégia proposta por [6] que combina o algoritmo Christoffel de mínimos quadrados e amostragem quase-ótima. Outras técnicas também poderiam ser aplicadas na busca por modelos de melhor qualidade, como os modelos lineares pertencentes à família dos modelos ARIMA, modelos auto-regressivos integrados e de médias móveis.

Salienta-se que para o contexto analisado, não foram encontradas justificativas plausíveis que explicassem a queda brusca entre os anos de 2005 e 2007, além do dado faltante referente a 2006, visto que a população brasileira aumentou neste período, o que deveria levar a um aumento no número de professores nas escolas primárias (variável pesquisada).

5. AGRADECIMENTOS

Agradecemos a CAPES pelo auxílio financeiro em forma de Bolsa de Mestrado – Demanda Social, à Universidade Federal do Rio Grande e ao Programa de Pós-Graduação em Modelagem Computacional.

6. REFERÊNCIAS BIBLIOGRÁFICAS

1. Cunha JV, Coelho ACA. Regressão linear múltipla. In: Corrar, LJ; Paulo, E; Dias filho, JM. Análise multivariada. São Paulo, Brasil: Atlas, 2009. 541 p.
2. Brooks C. Introductory Econometrics for Finance. Cambridge, Reino Unido: Cambridge University Press, 2013. 648 p.
3. Devore JL. Probabilidade e estatística: para engenharia e ciências. São Paulo, Brasil: Pioneira Thomson Learning, 2006. 692 p.
4. Gujarati, DN. Econometria Básica. 3 ed., São Paulo, Brasil: Pearson Education do Brasil, 2000.
5. Hyndman RJ, Athanasopoulos G. Forecasting: principles and practice: Melbourne, Australia: OTexts, 2015. Disponível em http://otexts.org/fpp/. Acesso em: jul de 2016.
6. Shin Y, Xiu D. On a near optimal sampling strategy for least squares polynomial regression. Journal of Computational Physics. 2016; 326:931-946, doi: 10.1016/j.jcp.2016.09.032.
7. Silva VM, Mattos VLD. Avaliação do Método dos Mínimos Quadrados no Ajuste de um Modelo de Regressão Linear. In: 7ª Conferência Sul de Modelagem Computacional. Anais da 7º MCSul, 2016, nov 16-18, Rio Grande/RS, Brasil. p. 558-565.
8. UNESCO. United Nations Educational, Scientific and Cultural Organization, Institute for Statistics. Disponível em: http://uis.unesco.org/. Acesso em: jul de 2016.
9. Zar J. Biostatistical Analysis. New Jersey, USA: Prentice Hall, 1999. 663 p.