Biblioteca em C para Regressão Polinomial Utilizando o Método dos Mínimos Quadrados

Sérgio Aurélio Ferreira Soares[1], Brauliro Gonçalves Leal[1], Rodrigo Moreira Bacurau[1]

[1]Colegiado de Engenharia da Computação – Universidade Federal do Vale do São Francisco (UNIVASF) – Juazeiro – BA – Brasil

eng.sergiosoares@gmail.com, {brauliro.leal, rodrigo.bacurau}@univasf.edu.br

Abstract. *This paper presents an open source, embeddable, multiplatform, library developed in C language that aims to automate and facilitate the polynomial regression analysis using the least squares method to estimate the coefficients of the fitted curve.*

Resumo. *Este artigo apresenta uma biblioteca, open source, multiplataforma e embarcável, desenvolvida em linguagem C que objetiva automatizar e facilitar o processo de regressão polinomial utilizando o método dos mínimos quadrados para estimar os coeficientes da curva ajustada.*

1. Introdução

Investigar a relação existente entre variáveis é muito útil na ciência em geral, pois possibilita a descrição de fenômenos por meio de modelos matemáticos a partir de observações experimentais. A análise de regressão é uma técnica estatística amplamente utilizada para ajustar dados observados a funções matemáticas. A computação permitiu avanços nessa área uma vez que tornou rápida a resolução de sistemas de equações com grande número de variáveis. Pesquisadores de diversas áreas de conhecimento como, por exemplo, a meteorologia, agricultura e mercado financeiro, trabalham com grandes volumes de dados. Manipular dados nestas quantidades, pode se tornar uma tarefa trabalhosa e pouco produtiva.

Em muitas situações, uma maneira de facilitar este trabalho é desenvolver algoritmos que automatizem parcial ou totalmente estas tarefas. O grande problema desta abordagem é que a implementação de um programa que realize a regressão polinomial pode não ser trivial, já que é necessário o conhecimento de algum método computacional de solução de sistemas lineares de tamanho arbitrário. Isso desencoraja o desenvolvimento do sistema para automatização do processo e consequentemente aumenta o tempo necessário para a análise de dados.

Assim, visando automatizar e facilitar a análise de regressão polinomial, é proposta uma biblioteca escrita em linguagem C para realizar o ajuste de curvas a conjuntos de dados. A biblioteca foi construída de modo que possa ser integrada a outros *softwares* ou inserida em sistemas embarcados, devido ao seu tamanho reduzido, linguagem utilizada e simplicidade de codificação, que facilitam possíveis adaptações.

2. Revisão de literatura

A análise de regressão tem por objetivo descrever a relação existente entre duas ou mais variáveis através de um modelo matemático. Graficamente, equivale a identificar a curva matemática que melhor se ajusta aos pontos no diagrama de dispersão, Figura 1.

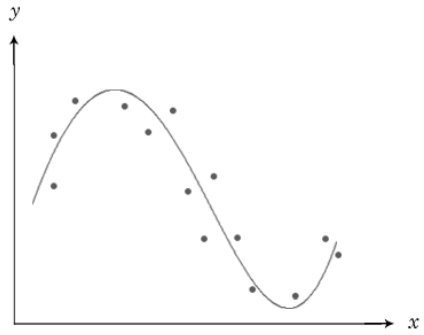

Figura 1. Diagrama de dispersão e sua respectiva curva ajustada.

Na análise de regressão obtém-se uma equação matemática cujas variáveis são denominadas dependente e independente, seus coeficientes refletem a intensidade da relação entre essas variáveis. Os modelos matemáticos de regressão fundamentam-se em três pressupostos estatísticos: a) a relação entre as variáveis dependente e independente é determinística ao invés de estocástica; b) os erros de medida são aleatórios, com distribuição normal, média zero e variância constante; e c) as variáveis explicativas não apresentam correlação entre si (Walpole *et al.*, 2007). O erro ou resíduo é a diferença entre os valores previstos pelo modelo de regressão para a variável dependente e os valores observados.

2.1. Método de estimação por mínimos quadrados

O método de estimação por mínimos quadrados consiste em obter uma função aproximadora para um conjunto de pontos a partir de seus valores (Spiegel *et al.*, 2009). Considerando um conjunto de pontos $\{ (x_1,y_1), (x_2,y_2), ..., (x_n,y_n) \}$, pode-se utilizar um polinômio de grau p, como função aproximadora para estes pontos, Equação 1. Neste caso, o erro do modelo de regressão para cada ponto do conjunto de dados é calculado pela Equação 2.

$$f(x) = a_o + a_1 x + a_2 x^2 + a_3 x^3 + \cdots + a_{p-1} x^{p-1} + a_p x^p = \sum_{k=0}^{p} a_k x^k \qquad 1$$

$$\varepsilon_i = y_i - \hat{y}_i, i = 1, 2, ..., n \qquad 2$$

em que $\hat{y}_i = f(x_i)$.

Os coeficientes a_k, k = 0, 1, ..., p, que minimizam o erro da função aproximadora f(x) podem ser obtidos solucionando o sistema de equações, Equação 4, que é obtido a partir da Equação 3.

$$\frac{\partial \xi}{\partial a_k} = 0 \qquad 3$$

em que, $\xi = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2$.

$$\begin{pmatrix} n & \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 & \dots & \sum_{i=1}^{n} x_i^p \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 & \sum_{i=1}^{n} x_i^3 & \dots & \sum_{i=1}^{n} x_i^{p+1} \\ & & \dots & & \\ \sum_{i=1}^{n} x_i^{p-1} & \sum_{i=1}^{n} x_i^p & \sum_{i=1}^{n} x_i^{p+1} & \dots & \sum_{i=1}^{n} x_i^{2p-1} \\ \sum_{i=1}^{n} x_i^p & \sum_{i=1}^{n} x_i^{p+1} & \sum_{i=1}^{n} x_i^{p+2} & \dots & \sum_{i=1}^{n} x_i^{2p} \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ \dots \\ a_{p-1} \\ a_p \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i y_i \\ \dots \\ \sum_{i=1}^{n} x_i^{p-1} y_i \\ \sum_{i=1}^{n} x_i^p y_i \end{pmatrix} \quad 4$$

2.2. Coeficiente de correlação

O estudo da correlação tem por objetivo medir e avaliar o grau de relação existente entre duas ou mais variáveis aleatórias. O coeficiente de correlação (r), é um índice que varia entre -1 e +1, indicando o quanto às diversas medidas obtidas a partir da amostra se ajustam ao modelo matemático proposto. Quanto maior o valor absoluto de *r,* maior a concordância entre os dados e a curva de regressão.

Segundo Lapponi (2005), se o coeficiente de correlação for igual a +1, os pares de valores das variáveis fazem parte de uma reta com declividade positiva. À medida que os pares de valores se afastam dessa reta, o coeficiente de correlação diminuirá de +1 em direção a -1, passando pelo valor 0. O coeficiente de determinação (R^2) indica o percentual da variância da variável dependente que pode ser explicado pela variável independente, Equação 5.

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

em que, $\overline{y} = \frac{1}{n} \sum_{i=1}^{n} x_i$.

2.3. Resolução do sistema por meio da Eliminação de Gauss com pivotação completa

Segundo Ruggiero e Lopes (1996), o método da Eliminação de Gauss consiste em transformar o sistema linear original num sistema linear equivalente com matriz de coeficientes triangular superior, pois estes são de resolução imediata.

Sendo assim, dado um sistema de equações lineares AX = B, seja M sua matriz aumentada, o método da pivotação completa consiste em escolher o elemento pivô $a_{pq} \neq 0$ de maior módulo e não pertencente à coluna dos termos independentes. Calculam-se os fatores m_i, Equação 6. Soma-se, a cada linha não pivotal, o produto da linha pivotal pelo seu fator m_i correspondente. Disso resulta uma nova matriz, cuja *q*-ésima coluna é composta de zeros, exceto o pivô. Rejeitando esta coluna e a *p*-ésima linha do pivô, tem-se uma nova matriz $M^{(1)}$, cujos índices correspondentes ao número de linhas e de colunas são diminuídos de uma unidade. Repetindo-se este processo para a nova matriz $M^{(1)}$, obtém-se $M^{(2)}$ e assim por diante. O sistema final é formado por todas as linhas

pivotais e, a partir da última linha da matriz $M^{(n-1)}$, a matriz triangular superior de M, resolve-se o sistema através de substituições retroativas Barroso *et al*. (1987).

$$m_i = -\frac{a_{ip}}{a_{pq}}, \quad \forall i \neq p \qquad 6$$

em que o elemento a_{pq} pivô e a linha *p* é a linha pivotal.

A pivotação completa busca minimizar a ampliação de erros de arredondamento, sendo recomendado especialmente na resolução de sistemas lineares de maior porte por meio de computadores Barroso *et al.* (1987).

3. Metodologia

Foi desenvolvido um conjunto de funções em linguagem C para a análise de dados por meio do ajuste de curvas polinomiais. Essas funções propiciam a obtenção dos coeficientes de regressão polinomial de um conjunto de dados arbitrário. A linguagem C foi escolhida por apresentar um código que pode ser facilmente adaptado para outras linguagens de programação, como o Java, C++ e PHP. Outra razão da escolha da linguagem C é o fato de existirem compiladores direcionados a microcontroladores como os da família PIC, da Microchip, o que faz com que a biblioteca possa ser facilmente inserida em sistemas embarcados com poucas alterações na codificação.

As funções da biblioteca estão descritas na seção 3.1. Cada função possui uma atribuição específica e implementa as definições fornecidas na seção 2. Os conceitos envolvidos se relacionam em uma sequência que se inicia pela definição do grau do polinômio do modelo de regressão, apresentado na Equação 1. Em seguida o cálculo dos coeficientes do polinômio é feito através da resolução do sistema de equações, Equação 4, utilizando o método descrito no Item 2.3. Por fim são gerados dois arquivos de saída com os resultados obtidos. Um dos arquivos possui os valores dos coeficientes do polinômio ajustado e do coeficiente de determinação R^2 e o outro uma tabela com os valores medidos, estimados e o erro relativo entre eles.

3.1. Funções da Biblioteca

A biblioteca desenvolvida é composta por sete funções. Seus protótipos e uma breve descrição são apresentados na Tabela 1. A biblioteca está disponível no SourceForge[2] acompanhada do código do programa utilizado para avaliar a biblioteca (seção 3.3).

Tabela 1 – Funções disponíveis na biblioteca

Função	Descrição
fx (double x,double a[p+1])	Calcula o valor de f(x), Equação 1
dado (double x[MAX],double y[MAX])	Carrega os dados do arquivo de entrada e armazena nos vetores
matriz (double A[p+1][p+1],double B[p+1], double x[MAX], double y[MAX])	Gera a matriz de coeficientes e o vetor de termos independentes. Equação 4
sistema (double A[p+1][p+1],double a[p+1], double B[p+1])	Soluciona o sistema de equações descrito na Equação 4

[2] http://regressao.sourceforge.net

calcula (double a[p+1], double x[MAX], double y[MAX])	Salva no arquivo de saída a tabela de valores reais e estimados.
resultado (double a[p+1],double x[MAX], double y[MAX])	Calcula o coeficiente de regressão e os salva no arquivo de resultados
pivot (int l, int m, double A[p+1][p+1], double B[p+1])	Realiza o pivoteamento completo da matriz de coeficientes. Esta função é acionada dentro da função "sistema".

3.2. Configurações da biblioteca

Para utilizar as funções contidas na biblioteca é necessário configurar o valor de cinco constantes nomeadas MAX, p, ENTRADA, SAIDA e RESULTADO. Como exemplo, podemos observar as primeiras linhas do código fonte do exemplo de uso da biblioteca disponível no sítio do projeto. As linhas estão configuradas da seguinte forma:

```
#define   MAX         501
#define   p           6
#define   ENTRADA     "c:\\regressao\\entrada.txt"
#define   SAIDA       "c:\\regressao\\saida.txt"
#define   RESULTADO   "c:\\regressao\\resultado.txt"
```

A primeira linha informa a biblioteca o número máximo de pontos (MAX) que podem ser carregados do arquivo de entrada. A segunda linha configura o grau do polinômio (p) a ser utilizado no ajuste da curva. A terceira linha informa o caminho do arquivo de entrada (ENTRADA) que deve ser do tipo texto e conter um par de números em cada linha, separados por um espaço em branco. A quarta linha informa o caminho do arquivo (SAIDA) que conterá a tabela com os dados medidos e estimados e também o erro relativo para cada valor da variável resposta. Por fim, a quinta linha define o caminho do arquivo (RESULTADO) que conterá o valor do coeficiente de determinação (R^2) e os valores dos coeficientes do polinômio estimado. Estes parâmetros devem ser configurados pelo usuário de modo a satisfazer suas preferências.

3.3. Avaliação da Biblioteca

Para avaliar a biblioteca foi gerado um conjunto com 501 pares de dados (x,y) obtidos pela Equação 7, fazendo x variar de 0 a 500, e a partir dele foram calculados os valores de a_0, a_1,..., a_6, usando a biblioteca desenvolvida.

$$f(x) = 1{,}0 + 0{,}1x + 0{,}01x^2 + 0{,}001x^3 + 0{,}0001x^4 + 0{,}00001x^5 + 0{,}000001x^6 \qquad 7$$

A Tabela 2 apresenta os coeficientes reais e estimados calculados pelo modelo proposto, bem como o erro relativo percentual entre o real e o estimado, Equação 8. O maior valor de ERP foi aproximadamente igual a 0,222 %, enquanto que o menor foi 0,0 %. Pode-se observar que os valores dos erros são muito pequenos, o que demonstra a confiabilidade do modelo implementado. Como esperado, o valor do coeficiente de determinação (R^2) foi igual a 1,00000. Estes resultados indicam que a biblioteca funcionou e se comportou de modo satisfatório.

Tabela 2 – Coeficientes reais, estimados e erro relativo percentual (ERP)

Coeficiente	Real	Estimado	ERP(%)
a_0	1	0,9977817	0,222
a_1	0,1	0,1001854	-0,185
a_2	0,01	0,00999638	0,036
a_3	0,001	0,001000028	-0,003
a_4	0,0001	0,0000999999	0,000
a_5	0,00001	0,00001	0,000
a_6	0,000001	0,000001	0,000

$$ERP = 100\left(\frac{Real - Estimado}{Real}\right) \quad (8)$$

4. Trabalhos Relacionados

Existem várias bibliotecas estatísticas ou numéricas robustas que, dentre outras operações, realizam a regressão polinomial como, por exemplo, a GSL – GNU Scientific Library[3] ou a NAG C Library[4]. O grande problema dessas bibliotecas é a complexidade de implementação, utilização e a inviabilidade de inserção da solução em sistemas embarcados. Assim, a biblioteca proposta neste artigo se destaca e consegue suprir essas deficiências de forma satisfatória.

5. Conclusão

A análise de regressão polinomial apresenta-se como uma ferramenta poderosa e muito utilizada por pesquisadores de distintas áreas de conhecimento. A possibilidade de extrair um modelo a partir de observações experimentais é um recurso que pode ser bastante útil em diversas aplicações.

Neste trabalho foi proposta uma biblioteca que visa facilitar e automatizar o processo de análise de regressão polinomial. A biblioteca caracteriza-se por ser *open source,* multiplataforma, e poder ser inserida em sistemas embarcados. O modelo implementado foi avaliado e apresentou erro relativo percentual máximo da estimação dos coeficientes igual a 0,222 %. Esse indicador permite avaliar de forma positiva o modelo e sugerir o uso da biblioteca desenvolvida.

6. Referências

Lapponi, Juan Carlos. Estatística usando Excel. Rio de Janeiro: Elsevier, 2005.

Ruggiero, Márcia A. Gomes; Lopes, Vera Lúcia da Rocha. Cálculo Numérico: aspectos teóricos e computacionais. São Paulo: Makron Books, 1996.

Spiegel, Murray R.;Schiller, John J.; Srinivasan, R. Alu. Probability and Statistics. 3td ed. New York: McGraw-Hill, 2009.

Walpole, Ronald E.; Myers, R. H.; Myers, S. L.; Ye, Keying. Probability & Statistics for Engineers & Scientists. 8th ed. New York: PEARSON Prentice Hall, 2007.

[3] http://www.gnu.org/s/gsl/

[4] http://www.nag.com/numeric/CL/CLdescription.asp