Ajuste de Curvas

Andréa Maria Pedrosa Valli

Laboratório de Computação de Alto Desempenho (LCAD)
Departamento de Informática
Universidade Federal do Espírito Santo - UFES, Vitória, ES, Brasil

Ajuste de Curvas

O objetivo da regressão linear (ajuste) é encontrar uma função

$$u = a_1 g_1(x) + a_2 g_2(x) + \cdots + a_m g_m(x)$$

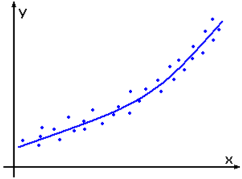
que "melhor" se ajusta a uma tabela de dados,

x_k	x_1	x_2	x_3	$\cdots$	x_n
y_k	y_1	y_2	y_3	$\cdots$	y_n

oriunda de n experimentos e podendo conter erros de medição.

- Podemos usar a regressão linear para estimar um valor $y_1 \leq y \leq y_n$ que não esteja tabelado, $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$.
- Podemos também estimar um valor y em algum ponto fora do intervalo de tabelamento (extrapolação).
- Uma aplicação, por exemplo, seria estimar a população de Vitória no ano de 2020 dada uma tabela com a população de Vitória em anos anteriores a 2015.

- Para escolher as funções do ajuste, $g_1(x), g_2(x), \cdots, g_m(x)$, podemos fazer o diagrama de dispersão ou utilizar informações *a priori sobre a relação entre as variáveis x e y do modelo.*



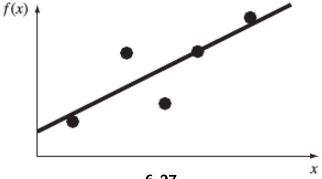
- *Caso não tenha nenhuma informação sobre as funções do ajuste, podemos fazer a regressão polinomial, ou seja, escolher as funções de interpolação como sendo polinômios de grau menor ou igual a m.*
- *Vamos ter que definir o critério para o "melhor" ajuste. Neste curso, vamos estudar o método dos quadrados mínimos.*

Ajustar uma reta a um conjunto de pares de observação: $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$. A expressão matemática do ajuste por uma reta é

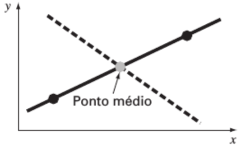
$$\begin{aligned} y &= a_0 + a_1 x + e \\ &= u + e \end{aligned}$$

onde a_0 e a_1 são coeficientes representando a intersecção com o eixo y e a inclinação, respectivamente, e é o erro ou resíduo (erro residual) entre o modelo e a observação.

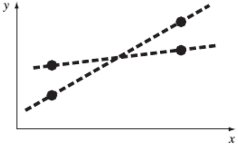
Portanto, o erro ou resíduo é a discrepância entre o valor verdadeiro de y e o valor aproximado, $u = a_0 + a_1 x$, previsto pela equação linear.



1. Minimizar a soma dos erros residuais (Min $\sum_{k=1}^{n} e_k$): qualquer reta passando pelo ponto médio resulta em um mínimo (não fornece solução única).



2. Minimizar a soma dos valores absolutos dos erros residuais (Min $\sum_{k=1}^{n} |e_k|$): qualquer reta entre as duas retas é um mínimo (não fornece solução única).



Critério dos quadrados mínimos para ajustar uma reta:

$$\text{Min } S_r(a_0, a_1) = \text{Min } \sum_{k=1}^{n} (e_k)^2 = \text{Min } \sum_{k=1}^{n} (y_k - (a_0 + a_1 x_k))^2$$

Condições para o mínimo:

$$\begin{aligned}
\frac{\partial S_r}{\partial a_0} &= \sum_{k=1}^{n} 2(y_k - (a_0 + a_1 x_k))(-1) = 0 \\
\frac{\partial S_r}{\partial a_1} &= \sum_{k=1}^{n} 2(y_k - (a_0 + a_1 x_k))(-x_k) = 0
\end{aligned}$$

Equações Normais (sistema simétrico):

$$
\begin{aligned}
\left(\sum_{k=1}^{n} 1\right)a_0 + \left(\sum_{k=1}^{n} x_k\right)a_1 &= \sum_{k=1}^{n} y_k \\
\left(\sum_{k=1}^{n} x_k\right)a_0 + \left(\sum_{k=1}^{n} x_k^2\right)a_1 &= \sum_{k=1}^{n} x_k y_k
\end{aligned}
$$

Na forma vetorial

$$
\begin{bmatrix} \sum_{k=1}^{n} 1 & \sum_{k=1}^{n} x_k \\ \sum_{k=1}^{n} x_k & \sum_{k=1}^{n} x_k^2 \end{bmatrix}
\begin{bmatrix} a_0 \\ a_1 \end{bmatrix}
=
\begin{bmatrix} \sum_{k=1}^{n} y_k \\ \sum_{k=1}^{n} x_k y_k \end{bmatrix}
$$

Exemplo: Ajustar uma reta, pelo método dos quadrados mínimos, ao conjunto de dados abaixo. Utilize 4 casas decimais.

x_k	1	2	3	4	5	6	7
y_k	0.5	2.5	2.0	4.0	3.5	6.0	5.5

$$\sum_{k=1}^{7} 1 = 7 \qquad \sum_{k=1}^{7} x_k = 28$$

$$\sum_{k=1}^{7} x_k^2 = 140 \qquad \sum_{k=1}^{7} y_k = 24, \qquad \sum_{k=1}^{7} x_k y_k = 119.5$$

$$\left[\begin{array}{cc|c} 7 & 28 & 24 \\ 28 & 140 & 119.5 \end{array}\right] \Rightarrow \begin{bmatrix} x_0 \\ x_1 \end{bmatrix} = \begin{bmatrix} 0.0714 \\ 0.8393 \end{bmatrix}$$

$$\Rightarrow y = 0.0714 + 0.8393x$$

$$\Rightarrow S_r(0.0714, 0.8393) = 2.9911$$

Critério dos quadrados mínimos para ajustar um polinômio de grau m:

$$\text{Min } S_r = \text{Min } \sum_{k=1}^{n} (y_k - (a_0 + a_1 x_k + \cdots + a_m x_k^m))^2$$

Condições de mínimo: $\frac{\partial S_r}{\partial a_0} = \frac{\partial S_r}{\partial a_1} = \cdots = \frac{\partial S_r}{\partial a_m} = 0$

$$\frac{\partial S_r}{\partial a_0} = \sum_{k=1}^{n} 2(y_k - (a_0 + a_1 x_k + \cdots + a_m x_k^m))(-1) = 0$$

$$\frac{\partial S_r}{\partial a_1} = \sum_{k=1}^{n} 2(y_k - (a_0 + a_1 x_k + \cdots + a_m x_k^m))(-x_k) = 0$$

$$\cdots$$

$$\frac{\partial S_r}{\partial a_m} = \sum_{k=1}^{n} 2(y_k - (a_0 + a_1 x_k + \cdots + a_m x_k^m))(-x_k^m) = 0$$

onde $S_r = S_r(a_0, a_1, \cdots, a_m)$.

Equações Normais (sistema simétrico $(m+1)\times(m+1)$):

$$\begin{bmatrix} \sum_{k=1}^{n} 1 & \sum_{k=1}^{n} x_k & \cdots & \sum_{k=1}^{n} x_k^m \\ \sum_{k=1}^{n} x_k^m & \sum_{k=1}^{n} x_k^2 & \cdots & \sum_{k=1}^{n} x_k^{m+1} \\ & & \ddots & \\ \sum_{k=1}^{n} x_k^m & \sum_{k=1}^{n} x_k^{m+1} & \cdots & \sum_{k=1}^{n} x_k^{2m} \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_m \end{bmatrix} \begin{bmatrix} \sum_{k=1}^{n} y_k \\ \sum_{k=1}^{n} x_k y_k \\ \vdots \\ \sum_{k=1}^{n} x_k^m y_k \end{bmatrix}$$

Ajustar uma curva do tipo $u = a_1 + a_2 x + a_3 e^x$ a tabela de dados

x_k	x_1	x_2	x_3	$\cdots$	x_n
y_k	y_1	y_2	y_3	$\cdots$	y_n

$$\text{Min } S_r(a_1, a_2, a_3) = \text{Min } \sum_{k=1}^{n} (y_k - (a_1 + a_2 x_k + a_3 e^{x_k}))^2$$

Condições de mínimo: $\frac{\partial S_r}{\partial a_1} = \frac{\partial S_r}{\partial a_2} = \frac{\partial S_r}{\partial a_3} = 0$

$$\begin{aligned}
\frac{\partial S_r}{\partial a_1} &= \sum_{k=1}^{n} 2(y_k - (a_1 + a_2 x_k + a_3 e^{x_k}))(-1) = 0 \\
\frac{\partial S_r}{\partial a_2} &= \sum_{k=1}^{n} 2(y_k - (a_1 + a_2 x_k + a_3 e^{x_k}))(-x_k) = 0 \\
\frac{\partial S_r}{\partial a_3} &= \sum_{k=1}^{n} 2(y_k - (a_1 + a_2 x_k + a_3 e^{x_k}))(-e^{x_k}) = 0
\end{aligned}$$

Equações Normais (sistema simétrico):

$$
\begin{bmatrix}
\sum_{k=1}^{n} 1 & \sum_{k=1}^{n} x_k & \sum_{k=1}^{n} e^{x_k} \\
\sum_{k=1}^{n} x_k & \sum_{k=1}^{n} x_k^2 & \sum_{k=1}^{n} x_k e^{x_k} \\
\sum_{k=1}^{n} e^{x_k} & \sum_{k=1}^{n} x_k e^{x_k} & \sum_{k=1}^{n} e^{2x_k}
\end{bmatrix}
\begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix}
\begin{bmatrix}
\sum_{k=1}^{n} y_k \\
\sum_{k=1}^{n} x_k y_k \\
\sum_{k=1}^{n} e^{x_k} y_k
\end{bmatrix}
$$

Problema com n experimentos e m parâmetros:

x_k	x_1	x_2	x_3	$\cdots$	x_k
y_k	y_1	y_2	y_3	$\cdots$	y_k

$$y = a_1 g_1(x) + a_2 g_2(x) + \cdots + a_m g_m(x) + e$$

$$\text{Min } S_r = \text{Min} \sum_{k=1}^{n} \left(y_k - (a_1 g_1(x_k) + a_2 g_2(x_k) + \cdots + a_m g_m(x_k)) \right)^2$$

Condições de mínimo: $\frac{\partial S_r}{\partial a_1} = \frac{\partial S_r}{\partial a_2} = \cdots = \frac{\partial S_r}{\partial a_m} = 0$

onde $S_r = S_r(a_1, a_2, \cdots, a_m)$.

$$
\begin{aligned}
\frac{\partial S_r}{\partial a_1} &= \sum_{k=1}^{n} 2(y_k - (\, a_1 g_1(x_k) + a_2 g_2(x_k) + \cdots + a_m g_m(x_k))\,)(-g_1(x_k)) = 0 \\
\frac{\partial S_r}{\partial a_2} &= \sum_{k=1}^{n} 2(y_k - (\, a_1 g_1(x_k) + a_2 g_2(x_k) + \cdots + a_m g_m(x_k))\,)(-g_2(x_k)) = 0 \\
&\vdots \\
\frac{\partial S_r}{\partial a_m} &= \sum_{k=1}^{n} 2(y_k - (\, a_1 g_1(x_k) + a_2 g_2(x_k) + \cdots + a_m g_m(x_k))\,)(-g_m(x_k)) = 0
\end{aligned}
$$

Equações Normais (sistema simétrico $m \times m$):

$$\begin{bmatrix} \sum_{k=1}^{n} g_1 g_1 & \sum_{k=1}^{n} g_1 g_2 & \cdots & \sum_{k=1}^{n} g_1 g_m \\ \sum_{k=1}^{n} g_2 g_1 & \sum_{k=1}^{n} g_2 g_2 & \cdots & \sum_{k=1}^{n} g_2 g_m \\ & & \ddots & \\ \sum_{k=1}^{n} g_m g_1 & \sum_{k=1}^{n} g_m g_2 & \cdots & \sum_{k=1}^{n} g_m g_m \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \end{bmatrix} \begin{bmatrix} \sum_{k=1}^{n} g_1 y_k \\ \sum_{k=1}^{n} g_2 y_k \\ \vdots \\ \sum_{k=1}^{n} g_m y_k \end{bmatrix}$$

onde $g_i = g_i(x_k), \;\; i = 1, 2, \cdots, m$

Se $\bar{g}_1, \bar{g}_2, \cdots, \bar{g}_m$ formam um conjunto linearmente independente

$\Rightarrow$ que o determinante do sistema normal é diferente de zero

$\Rightarrow$ que o sistema tem solução única.

Além disso, demonstra-se que que esta solução é o mínimo de $S_r(a_1, a_2, \cdots, a_m) = \sum_{k=1}^{n} (\, y_k - (a_1 g_1(x_k) + \cdots + a_m g_m(x_k))\,)^2$.

$$\bar{g}_j = \begin{bmatrix} g_j(x_1) \\ g_j(x_2) \\ \cdots \\ g_j(x_n) \end{bmatrix}, \quad j = 1, 2, \cdots, m$$

Coeficiente de determinação ($r^2 = (S_t - Sr)/S_t$): Fornece a proporção da variação total dos dados em torno da média ($\bar{y} = \frac{1}{n}\sum_{k=1}^{n} y_k$)

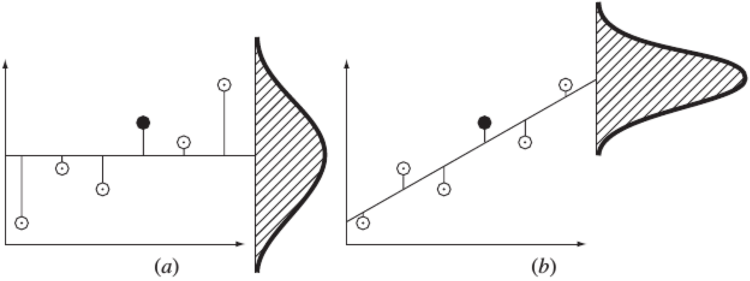
Soma total dos quadrados em torno da média ($SQTot$): o tamanho do erro residual antes do ajuste.

$$S_t = SQTot = \sum_{k=1}^{n}(y_k - \bar{y})^2$$

Soma dos quadrados dos resíduos ($SQRes$): o tamanho do erro residual que continua depois do ajuste.

$$S_r = SQRes = \sum_{k=1}^{n} e_k^2$$

Dados de regressão por uma reta mostrando (a) a dispersão dos dados em torno da média da variável dependente y e (b) a dispersão dos dados em torno da reta de melhor de ajuste.



$S_t - S_r$, quantifica a melhora ou a redução de erro quando comparamos os dados com a função do ajuste ao invés da média.

Coeficiente de determinação (r^2):

$$r^2 = \frac{S_t - S_r}{S_t} = 1 - \frac{S_r}{S_t}$$

$r^2 \approx 1 \Rightarrow$ melhor o ajuste
$r^2 = 0 \Rightarrow$ não houve nenhum melhoramento

$$S_t = \sum_{k=1}^{n} (y_k - \bar{y})^2 = \sum_{k=1}^{n} y_k^2 - 2\bar{y} \sum_{k=1}^{n} y_k + n\bar{y}^2$$

$$= \sum_{k=1}^{n} y_k^2 - 2\bar{y} n\left(\tfrac{1}{n} \sum_{k=1}^{n} y_k\right) + n\bar{y}^2 = \sum_{k=1}^{n} y_k^2 - n\left(\tfrac{1}{n} \sum_{k=1}^{n} y_k\right)^2$$

$$\Rightarrow r^2 = 1 - \frac{S_r(a_1, \cdots, a_m)}{\sum_{k=1}^{n} y_k^2 - \frac{1}{n}\left(\sum_{k=1}^{n} y_k\right)^2}$$

Ajuste por um polinômio de grau m:

$$\text{Min } S_r = \text{Min } \sum_{k=1}^{n} (y_k - (a_0 + a_1 x_k + \cdots + a_m x_k^m))^2$$

x_k	x_1	x_2	x_3	$\cdots$	x_k
y_k	y_1	y_2	y_3	$\cdots$	y_k

Variância residual: $\sigma = \frac{S_r}{n-p}$ onde n é o número de experimentos e $p = m + 1$ é o número de parâmetros.

$$\Rightarrow \sigma = \frac{S_r}{n - (m+1)}$$

A redução global da variância residual vai definir se mais parâmetros devem ou não ser incorporados ao modelo.

Exemplo: Dados históricos dos censos do Brasil [1]. Para reduzir os erros de ponto flutuante $x = \text{ano} - 1970$ e $y = \text{urbana} \times 10^{-6}$.

Ano	Urbana	Rural
1940	12.880.182	28.356.133
1950	18.782.891	33.161.506
1960	31.303.034	38.767.423
1970	52.084.984	41.054.053
1980	80.436.409	38.566.297
1991	110.990.990	35.834.485
1996	123.076.831	33.993.332
2000	137.953.959	31.845.211

g	r^2	σ^2
1	0,96602	$9{,}58219 \times 10^{1}$
2	0,99776	$7{,}59261 \times 10^{0}$
3	0,99939	$2{,}57228 \times 10^{0}$
4	0,99940	$3{,}42930 \times 10^{0}$
5	0,99980	$1{,}65615 \times 10^{0}$
6	0,99998	$4{,}17203 \times 10^{-1}$

Ajustar uma curva do tipo $y = a_0 + a_1 x_1 + a_2 x_2$ a tabela abaixo

x_1	$(x_1)_1$	$(x_1)_2$	$(x_1)_3$	$\cdots$	$(x_1)_k$
x_2	$(x_2)_1$	$(x_2)_2$	$(x_2)_3$	$\cdots$	$(x_2)_k$
y_k	y_1	y_2	y_3	$\cdots$	y_k

$$\text{Min } S_r = \text{Min } \sum_{k=1}^{n} (y_k - (a_0 + a_1 (x_1)_k + a_2 (x_2)_k))^2$$

$$\begin{aligned}
(\sum_{k=1}^{n} 1) a_0 + (\sum_{k=1}^{n} (x_1)_k) a_1 + (\sum_{k=1}^{n} (x_2)_k) a_2 &= \sum_{k=1}^{n} y_k \\
(\sum_{k=1}^{n} (x_1)_k) a_0 + (\sum_{k=1}^{n} (x_1)_k^2) a_1 + (\sum_{k=1}^{n} (x_1)_k (x_2)_k) a_2 &= \sum_{k=1}^{n} (x_1)_k y_k \\
(\sum_{k=1}^{n} (x_2)_k) a_0 + (\sum_{k=1}^{n} (x_2)_k (x_1)_k) a_1 + (\sum_{k=1}^{n} (x_2)_k^2) a_2 &= \sum_{k=1}^{n} (x_2)_k y_k
\end{aligned}$$

- $y = a\,x^b$

x_k	x_1	x_2	x_3	$\cdots$	x_k
y_k	y_1	y_2	y_3	$\cdots$	y_k

$$\Rightarrow \ln(y) = \ln(a) + b\,\ln(x) \Rightarrow z = \bar{a} + \bar{b}\,\ln(x)$$

x_k	x_1	x_2	x_3	$\cdots$	x_k
$z_k = \ln(y_k)$	z_1	z_2	z_3	$\cdots$	z_k

$$\Rightarrow z = \bar{a}\,g_1(x) + \bar{b}\,g_2(x)$$
$$\Rightarrow g_1(x) = 1$$
$$\Rightarrow g_2(x) = \ln(x)$$

Utilize o método dos quadrados mínimos na tabela de dados (x_k, z_k) e determina $\bar{a}$ e $\bar{b}$ e depois $a = e^{\bar{a}}$, $b = \bar{b}$.

- $y = \frac{1}{a+bx+cx^2}$

x_k	x_1	x_2	x_3	$\cdots$	x_k
y_k	y_1	y_2	y_3	$\cdots$	y_k

$$\Rightarrow \frac{1}{y} = a + bx + cx^2 \Rightarrow z = a + bx + cx^2$$

x_k	x_1	x_2	x_3	$\cdots$	x_k
$z_k = 1/y_k$	z_1	z_2	z_3	$\cdots$	z_k

Utilize o método dos quadrados mínimos na tabela de dados (x_k, z_k) e determina a, b e c.

Bibliografia Básica

[1] Algoritmos Numéricos, Frederico F. Campos, Filho - 2ª Ed., Rio de Janeiro, LTC, 2007.

[2] Métodos Numéricos para Engenharia, Steven C. Chapa e Raymond P. Canale, Ed. McGraw-Hill, 5ª Ed., 2008.

[3] Cálculo Numérico - Aspectos Teóricos e Computacionais, Márcia A. G. Ruggiero e Vera Lúcia da Rocha Lopes, Ed. Pearson Education, 2ª Ed., 1996.