Capítulo 13

Regressão linear e polinomial

Neste capítulo, pretendemos ajustar retas ou polinômios a um conjunto de pontos experimentais.

Regressão linear

A tabela a seguir relaciona a densidade (g/cm^3) do sódio em função da temperatura (^{o}C):

Temperatura (^{o}C)	Densidade(g/cm^3)
100	0,927
200	0,904
300	0,882
400	0,859
500	0,934
600	0,809
700	0,783
800	0,757

Quando representamos estes dados em gráfico, dão a impressão de ficar numa reta que poderia ser traçada com uma régua "a olho". Porém, no caso de os pontos estarem mais dispersos, o ajustamento a olho é bastante subjetivo e inexato. (Além disso, ajustamento a olho requer que todos os pontos estejam primeiramente colocados num gráfico. No caso de, por exemplo, 100 observações, isto seria bastante tedioso.)

Nosso objetivo é ajustar uma reta y = a + bx aos pontos do diagrama de dispersão, utilizando técnicas matemáticas. O famoso método dos quadrados mínimos de Gauss responde à pergunta "o que é um bom ajustamento" com as seguintes equações para calcular os valores dos fatores a e b:

$$b = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \qquad (1)$$

$$a = \overline{y} - b\overline{x}$$

As médias de x e y são definidas por $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$; $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ (2)

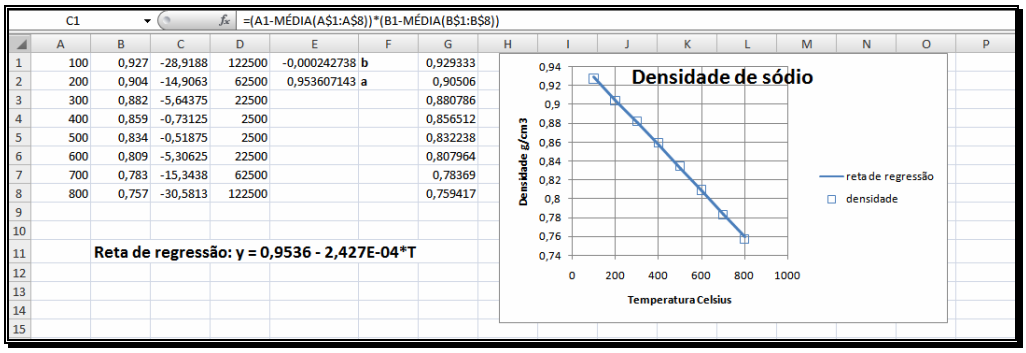
a = coeficiente linear da reta, b= coeficiente angular da reta

Apliquemos estas fórmulas ao nosso exemplo:

C1: =(A1-MÉDIA(A$1:A$8))*(B1-MÉDIA(B$1:B$8))
D1: =(A1-MÉDIA(A$1:A$8))^2, copiar as fórmulas até linha 8
E1: =SOMA(C1:C8)/SOMA(D1:D8) (=b)
E2: =MÉDIA(B1:B8)-E1*MÉDIA(A1:A8) (=a)

Na coluna G ficam os valores de y da reta de regressão
G1: =E$2+E$1*A1

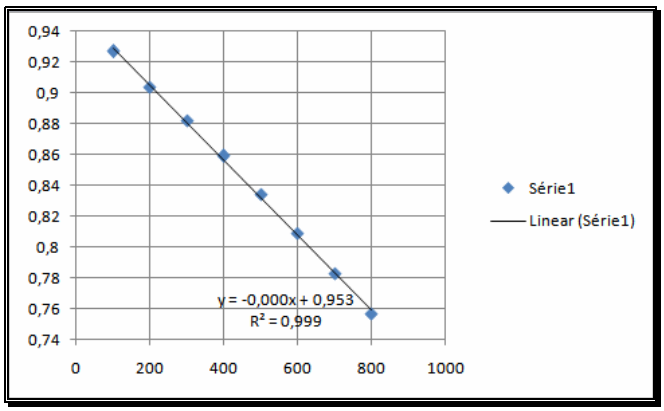
Para fazer o gráfico, deve-se levar em conta que temos de representar duas séries de dados. Veja também o capítulo 5, p. 63



	A	B	C	D	E	F	G
1	100	0,927	-28,9188	122500	-0,000242738	b	0,929333
2	200	0,904	-14,9063	62500	0,953607143	a	0,90506
3	300	0,882	-5,64375	22500			0,880786
4	400	0,859	-0,73125	2500			0,856512
5	500	0,834	-0,51875	2500			0,832238
6	600	0,809	-5,30625	22500			0,807964
7	700	0,783	-15,3438	62500			0,78369
8	800	0,757	-30,5813	122500			0,759417

(Se tiver instalado o programa **tc²** que mencionei no último capítulo, poderia aqui, em WORD, calcular a densidade de sódio para uma temperatura dada:
T=600
d=0,9536-2,4274E-04*T = 0,808 o que corresponde bem ao o valor da tabela.)

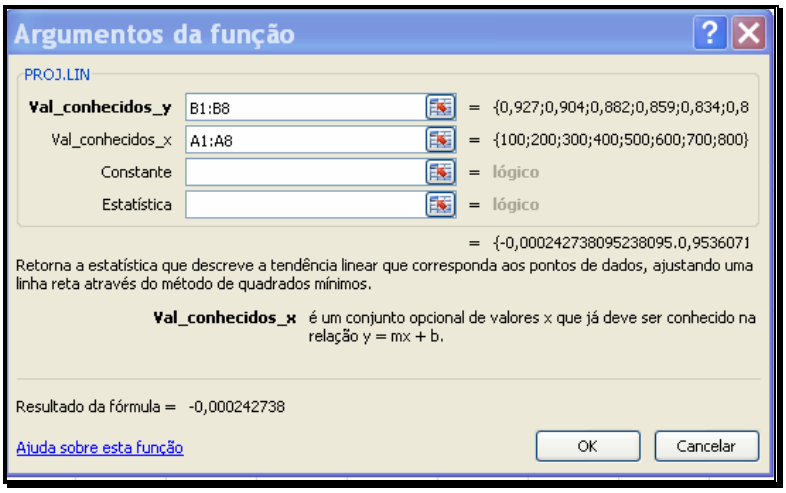
É hora de mencionar que o Excel, a partir do Excel 97, tem embutido uma ferramenta que faz tudo o que acabamos de ver, é só eleger *Layout>Linha de Tendência* com as suas opções, p. ex. a equação da linha e o valor de R^2.



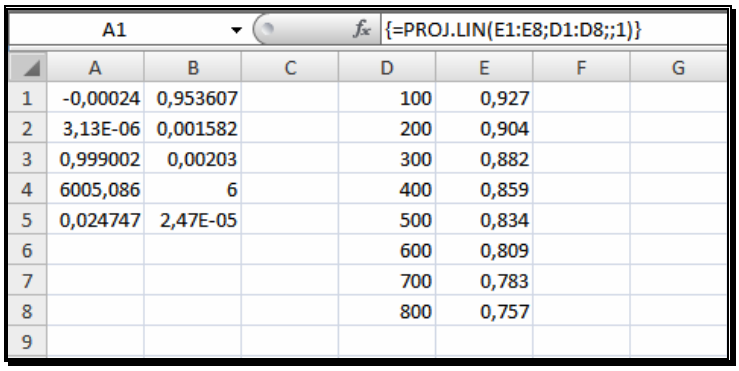
Mas, este assistente somente aparecerá depois que você selecionar um gráfico, em nosso caso *Dispersão Somente com Marcadores*. As propriedades da linha, como cor, estilo etc. podem ser variadas, é só fazer clique sobre a linha e selecionar *Formatar Linha de Tendência*.

Mas, aqui não terminam as maravilhas estadísticas do Excel. Existe a função estadística PROJ.LIN com a sintaxe PROJ.LIN(val_conhecidos_y; valconhecidos_x; constante; estatística)

Para aplicá-la, é necessário preencher as duas primeiras linhas na seguinte janela.



A janela mostra já os fatores a e b da equação da reta de regressão. Se colocarmos no último campo 1 (=VERDADEIRO), veremos a seguinte tabela.



	A	B	C	D	E	F	G
1	-0,00024	0,953607		100	0,927		
2	3,13E-06	0,001582		200	0,904		
3	0,999002	0,00203		300	0,882		
4	6005,086	6		400	0,859		
5	0,024747	2,47E-05		500	0,834		
6				600	0,809		
7				700	0,783		
8				800	0,757		
9							

(É preciso colocar nossos dados em outras células, por exemplo D1:E8, pois temos selecionado o intervalo A1:B5 para os resultados estadísticos. A fórmula =PROJ.LIN(E1:E8;D1:D8;;1) é uma fórmula matricial e deve ser inserida pressionando Ctrl+Shift+Enter.)

Os valores em A1 e B1 são, outra vez, a e b. A2 e B2 contêm os valores do *erro padrão* dos coeficientes b e a. (a e b são funções dos valores experimentais y_i. Devido à propagação dos erros, as incertezas nos y_i influenciarão também os valores de a e b. Suponhamos que as incertezas nos valores de x sejam depreciáveis.) Na célula A3 temos o valor de R^2, o coeficiente de determinação. Este valor deve ficar bem perto de 1 para que o ajustamento possa ser considerado como sendo bom. R é o coeficiente de correlação. Se R for igual a 1, existirá uma correlação perfeita na mostra – não haverá diferença entre os valores de y estimados e os valores reais. Em B3 temos o valor do erro padrão para a estimativa de y, ou o erro padrão dos resíduos. Este parâmetro calcula-se com

$$\sigma_y^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - a - bx_i)^2 \qquad (3)$$

Em nosso caso resulta $\sigma_y = \sqrt{\sigma_y^2} = \sqrt{4{,}121E-6} = 0{,}00203$

O parâmetro $\sigma_b = (\sigma^2_b)^{0,5}$ em A2 calculamos com

$$\sigma_b^2 = \frac{n\sigma_y^2}{n\sum_{i=1}^{n}x_i^2 - \left(\sum_{i=1}^{n}x_i\right)^2} \qquad (4)$$

A fórmula para Excel é =8*0,000004121/((8*SOMA(D1:D8)-SOMA(A1:A8)^2)) e dá σ_b = 3,1324E-6.

O valor para σ_a na célula B2 é determinado com

$$\sigma_a^2 = \frac{\sigma_y^2\sum_{i=1}^{n}x_i^2}{D} \qquad (5)$$

onde D significa o denominador de (4). Resultado: σ_a = 0,001582

Observe que temos também $\sigma_a = \sigma_b\sqrt{\frac{1}{n}\sum_{i=1}^{n}x_i^2}$ (6)

Em A4 aparece a estadística F, ou o valor de F observado. Com um Teste–F podemos determinar, se a relação observada entre as variáveis dependentes e independentes ocorre por acaso. Em B4 estão os graus de liberdade (número dos valores experimentais – numero de fatores, ou seja 8 – 2 = 6). A5 contém

a soma dos quadrados da regressão e B5 a soma residual dos quadrados. Para s_{reg} temos $s_{reg} = \sum_{i=1}^{n}(y_i - \bar{y})^2$ e para s_{res} temos $s_{res} = \sum_{i=1}^{n}(y_i - y_i^{'})^2$

$\bar{Y}$ significa a média dos valores experimentais, y' é um valor de y calculado, ou seja y' = a + bx. Comparando estas fórmulas com σ_y, vemos que $\sigma_y = \sqrt{(s_{res}/(n-2))}$.

No exemplo anterior, o coeficiente de determinação, R^2, é 0,990, o que indica uma forte relação entre variáveis independentes e as densidades.

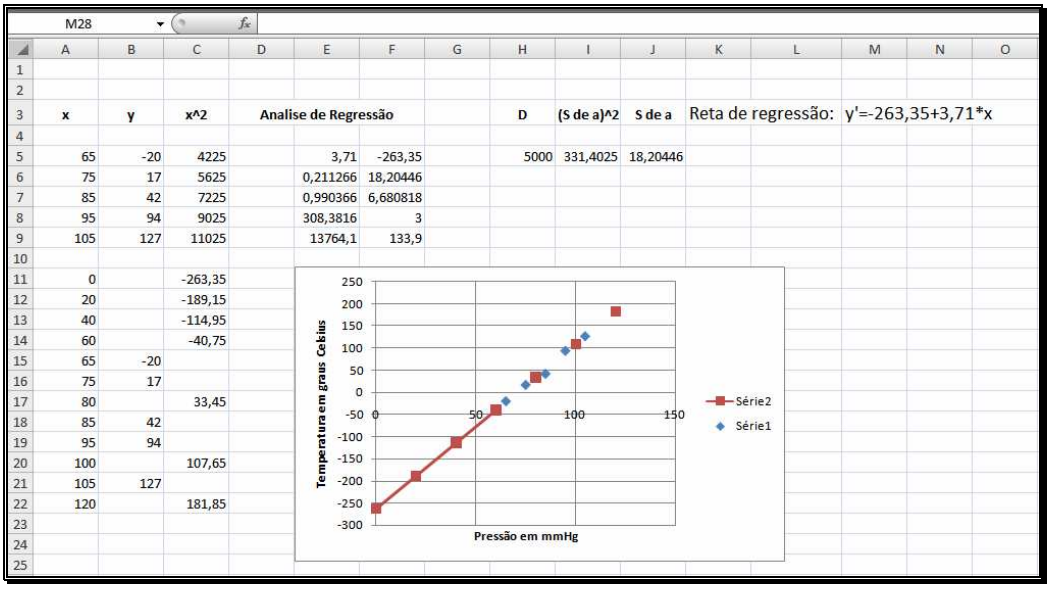
O coeficiente de determinação é definido como $R^2 = 1 - s_{res}/s_{reg}$, o que dá 1-2,426E-5/0,024747 = 0,9990.

Então, quanto maior R^2, melhor o ajuste da regressão aos dados observados.

Exemplo: Um estudante varia a temperatura de um gás quase ideal, mantendo o volume constante. Para cada valor de temperatura, ele mediu a pressão em mm Hg. O estudante obteve os seguintes valores

Pressão em mmHg	Temperatura em ^oC
65	-20
75	17
85	42
95	94
105	127

Devido à equação dos gases ideais, PV = nRT, espere-se uma relação linear entre os valores da tabela. Para confirmar esta suposição, fazemos um análise de regressão.



As entradas para a figura foram:

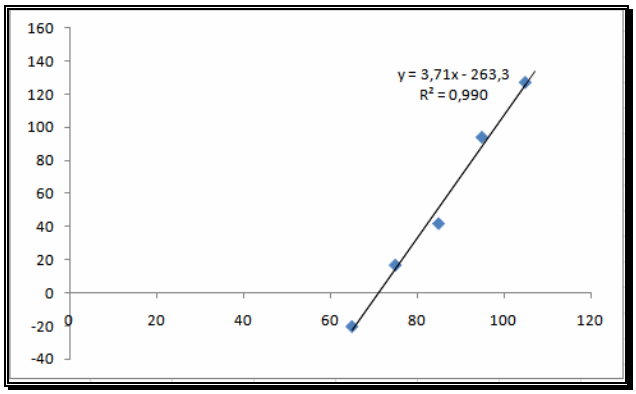
H5: =5*SOMA(C5:C9)-(SOMA(A5:A9))^2 (=D, denominador de (4))
I5: =F7^2*SOMA(C5:C9)/H5 (= σ_a^2, σ_a = desvio padrão de a, ponto de intercepção da reta com o eixo y, ou *erro padrão da intercepção*)

O bloco A11:C22 contém os dados a desenhar. Na coluna C ficam os valores y calculados com a equação de regressão. C11: =F$5+E$5*A11

O gráfico fazemos com *Inserir>Dispersão>Somente com Marcadores*.

Trata-se, neste exemplo, de um caso de extrapolação bastante duvidosa. O zero absoluto encontra-se no intervalo $a \pm \sigma_a = (-263 \pm 18)^o C$, de fato, encontra-se a –273,15ºC. Os cinco valores de temperatura (valores de y) deveriam ser marcados com barras de incerteza de ancho 2*6,7 = 13,4; 6,7 é o desvio padrão de y em F7. (Desvio padrão = standard deviation).

O Excel com Linha de Tendência não faz extrapolação



Regressão parabólica

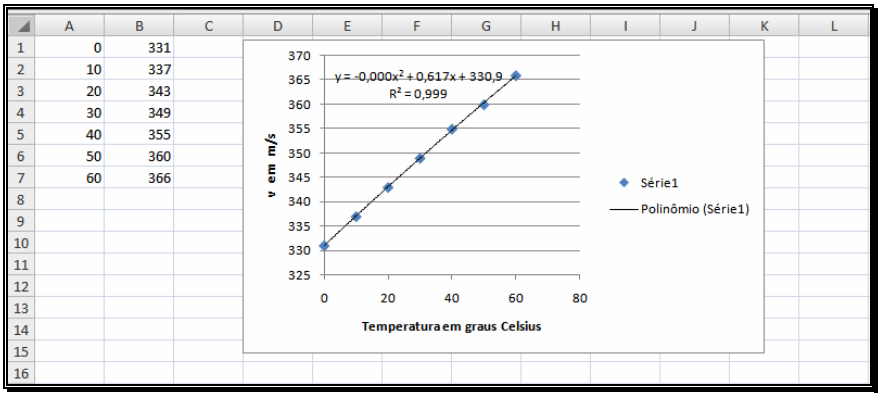
A tabela mostra os resultados experimentais correspondentes à velocidade do som em ar seca em função da temperatura.

Temperatura em ^oC	velocidade em m/s
0	331
10	337
20	343
30	349
40	355
50	360
60	366

Se busca a equação de uma **parábola** que se ajuste em forma optimal (no sentido dos mínimos quadrados) aos pontos experimentais.

A equação deve ser da forma $y = a + bx + cx^2$ onde os 3 parâmetros a,b,c devem ser determinados.

Por meio de *Linha de Tendência* obtemos o seguinte gráfico



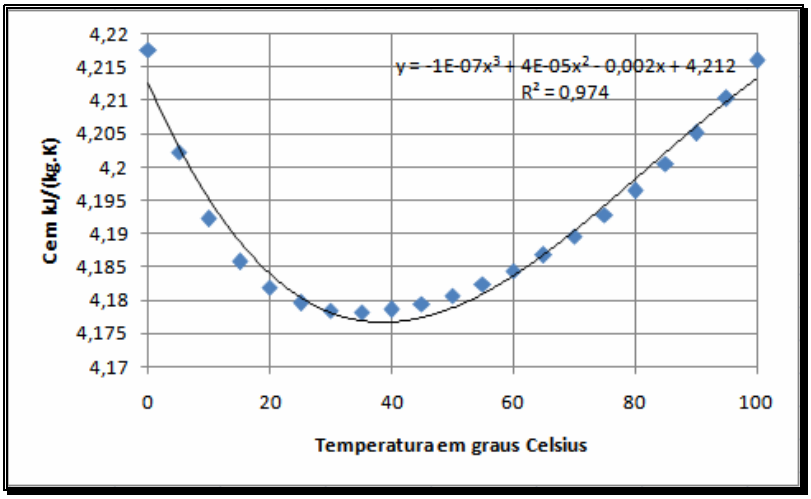
É óbvio que também houvéssemos podido utilizar um ajuste linear, mas, não é fácil predizer a curva que se esconde detrás dos dados.

No seguinte **exemplo**, capacidade térmica especifica (em kJ/(kg·K) de água em função da temperatura em graus Celsius, vamos buscar um ajuste **cúbico** da forma $y = a + bx + cx^2 + dx^3$

Temperatura ^oC	C em kJ/(kg·K)
0	4,2177
5	4,2022
10	4,1922
15	4,1858
20	4,1819
25	4,1796
30	4,1785
35	4,1782
40	4,1786
45	4,1795
50	4,1807
55	4,1824
60	4,1844
65	4,1868
70	4,1896
75	4,1928
80	4,1964

85	4,2005
90	4,2051
95	4,2103
100	4,2160

A *Linha de Tendência* produz o seguinte resultado (sem os títulos nos eixos):



Se queremos determinar os coeficientes do polinômio com mais precisão, podemos fazer um cálculo direitamente a partir das *equações normais*.

Trabalhando diretamente com as equações normais

Na teoria da regressão por mínimos quadrados, vemos que se obtém os parâmetros a, b, c na equação $y = a + bx + cx^2$ ou $y = a_1 + a_2x + a_3x^2$ resolvendo o seguinte sistema com respeito às incógnitas a_1, a_2, a_3

$$\begin{aligned} a_1 n + a_2 \sum x + a_3 \sum x^2 &= \sum y \\ a_1 \sum x + a_2 \sum x^2 + a_3 \sum x^3 &= \sum xy \qquad (1) \\ a_1 \sum x^2 + a_2 \sum x^3 + a_3 \sum x^4 &= \sum x^2 y \end{aligned}$$

A solução deste sistema, denominado equações normais, é fácil, pois podemos escrever (1) em forma matricial **M·A = B** com a solução **A = M^{-1}B**.
M^{-1} é a *matriz inversa* da matriz **M**. **A** é o vetor das incógnitas e **B** o vetor dos lados à direita, ou seja,

$$A = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} \text{ e } B = \begin{bmatrix} \sum y \\ \sum xy \\ \sum x^2 y \end{bmatrix} := \begin{bmatrix} Sy \\ Sxy \\ Sx^2 y \end{bmatrix} \qquad (2)$$

M é uma matriz quadrada de ordem m = 3, dada por

$$M = \begin{bmatrix} n & Sx & Sx^2 \\ Sx & Sx^2 & Sx^3 \\ Sx^2 & Sx^3 & Sx^4 \end{bmatrix} \qquad (3)$$

Determinamos a inversa da matriz, outra vez, pela função MATRIZ.INVERSO com a Matriz: A15:D17, veja a seguinte planilha que vale para o índice de refração de uma solução de açúcar em água. x = concentração, y = índice de refração n. Veja, também, capítulo 10, p. 147.

J9 =PolVal(K15:K17;J8)

	A	B	C	D	E	F	G	H	I	J	K	L
1	x	y	x^2	x^3	x^4	xy	x^2y					
2												
3	0	1,333	0	0	0	0	0		Regressão polinomial			
4	5	1,3403	25	125	625	6,7015	33,5075					
5	10	1,3479	100	1000	10000	13,479	134,79					
6	15	1,3557	225	3375	50625	20,3355	305,0325					
7	20	1,3639	400	8000	160000	27,278	545,56		Valor do polinómio em x:			
8	25	1,3723	625	15625	390625	34,3075	857,6875		x=	30		
9	30	1,3811	900	27000	810000	41,433	1242,99		y=	1,381137		
10	35	1,3902	1225	42875	1500625	48,657	1702,995					
11	40	1,3997	1600	64000	2560000	55,988	2239,52					
12	Somas:											
13	180	12,2841	5100	162000	5482500	248,1795	7062,083					
14	Equações :						Matriz invertida				Soluções	
15	9	180	5100	12,2841			0,660606	-0,06182	0,001212	a=	1,33304545	
16	180	5100	162000	248,1795			-0,06182	0,008978	-0,00021	b=	0,00141721	
17	5100	162000	5482500	7062,083			0,001212	-0,00021	5,19E-06	c=	6,1948E-06	
18												

M encontra-se no bloco A15:C17, a inversa $\mathbf{M^{-1}}$ fica em G15:I17. O vetor solução está em K15:K17. Ele foi calculado como produto matricial pela função MATRIZ.MULT: =MATRIZ.MULT(G15:I17;D15:D17), Ctrl+Shift+Enter

A15: =CONT.NÚM(A1:A11); A16: =A13; A17: =C13
B15: =A13; B16: =C13; B17: =D13
C15: =C13; C16: =D13; C17: =E13
D15: =B13; D16: =F13; D17: =G13

Finalmente, calculamos para um valor de x dado o valor do polinômio da regressão:

$$\mathbf{y = 1{,}333 + 1{,}417E\text{-}3x + 6{,}195E\text{-}6x^2}$$

No capítulo 9, pag. 121, desenvolvemos para o método de **Horner** uma *sub-rotina*. Esta vez, utilizamos uma *função*, para calcular os valores do polinômio:

```
Function PolVal(a As Variant, x As Double) As Double ' a = vetor a(i)dos coeficientes
Dim p As Double ' valor do polinômio
Dim Polgrau As Integer ' grau do polinômio
Dim i
Polgrau = a.Count - 1 ' grau do polinômio = número de coeficientes - 1
p = a(Polgrau + 1)
  For i = Polgrau To 1 Step -1
   p = p * x + a(i)
  Next i
PolVal = p ' retorna o valor do polinômio
End Function
```

Facilmente podemos inserir na planilha os dados da "velocidade do som em ar seca em função da temperatura" de acima (eliminando as linhas 10 e 11 na planilha da Regressão polinomial), para obter a mesma função que determinamos acima.

Não será muito difícil escrever o código VBA para realizar os passos exercidos na última planilha.

Calculamos, assim, as somas:

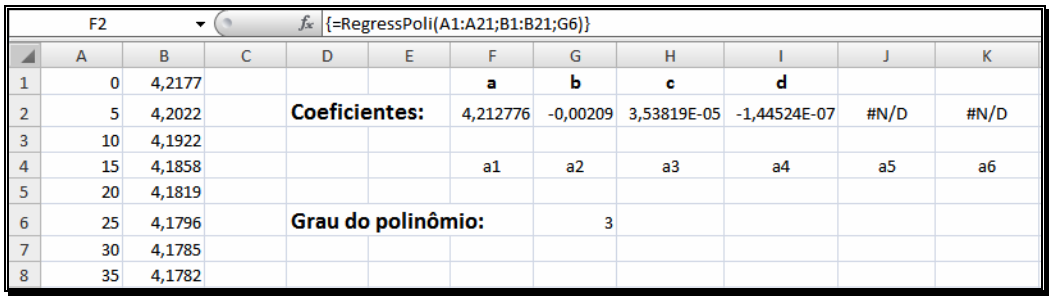
```
nx = x.Count
ReDim Sx(2 * n)
ReDim Sxy(n)
For i = 0 To 2 * n ' determinar as somas Sx
    Sx(i) = 0
    For k = 1 To nx
        Sx(i) = Sx(i) + x(k) ^ i
    Next k
Next i

For i = 0 To n 'determinar as somas Sxy
    Sxy(i) = 0
    For k = 1 To nx
        Sxy(i) = Sxy(i) + x(k) ^ i * y(k)
    Next k
Next i
```

O seguinte código cria as matrizes **M** e **B**

```
For i = 0 To n ' criar as matrizes M e B
    For j = 0 To i
        M(i + 1, j + 1) = Sx(i + j)
        M(j + 1, i + 1) = Sx(i + j)
    Next j
B(i + 1) = Sxy(i)
Next i
```

No programa completo dimensionamos, primeiro, cada matriz como matriz dinâmica (que é uma matriz que se ajusta à quantidade dos dados selecionados e que, eventualmente, podemos recortar ou ampliar).

A planilha correspondente tem o seguinte aspecto



F2 | {=RegressPoli(A1:A21;B1:B21;G6)}

	A	B	C	D	E	F	G	H	I	J	K
1	0	4,2177				**a**	**b**	**c**	**d**		
2	5	4,2022		**Coeficientes:**		4,212776	-0,00209	3,53819E-05	-1,44524E-07	#N/D	#N/D
3	10	4,1922									
4	15	4,1858				a1	a2	a3	a4	a5	a6
5	20	4,1819									
6	25	4,1796		**Grau do polinômio:**			3				
7	30	4,1785									
8	35	4,1782									

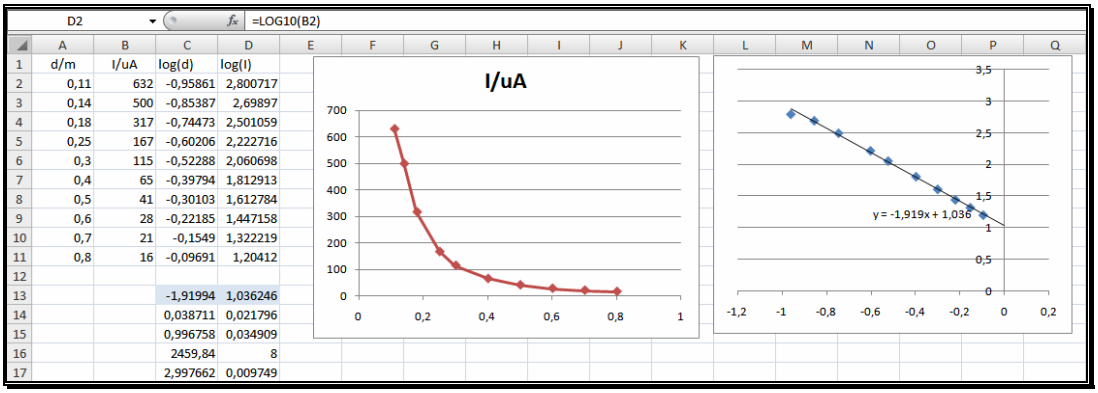
Os dados são os do exemplo da capacidade térmica específica, veja acima. A função da regressão polinomial a chamamos de "RegressPoli" e ela deve ser usada com Ctrl+Shift+Enter, pois é uma fórmula matricial. Na planilha temos previsto um polinômio até n = 5, o que é muito raro. Mas, o programa aceita polinômios de qualquer grau.

Aqui vem, finalmente, a função "RegressPoli":

```
Function RegressPoli(x, y, n) ' regressão polinomial
Dim nx As Integer ' número dos pontos
Dim Sx() ' matriz dinâmica para as somas x
Dim Sxy() ' matriz dinâmica para as somas xy
Dim M() As Variant
Dim Inv As Variant
Dim B() ' matriz dinâmica para o vetor B das constantes
Dim A() ' matriz dinâmica para  os coeficientes ao,..,an do polinômio
Dim i As Integer, j As Integer, k As Integer

nx = x.Count
ReDim Sx(2 * n)
ReDim Sxy(n)
For i = 0 To 2 * n ' determinar as somas Sx
    Sx(i) = 0
    For k = 1 To nx
        Sx(i) = Sx(i) + x(k) ^ i
    Next k
Next i

For i = 0 To n 'determinar as somas Sxy
    Sxy(i) = 0
    For k = 1 To nx
        Sxy(i) = Sxy(i) + x(k) ^ i * y(k)
    Next k
Next i
'------------------------------------------------ continuação
```

```
'---------------------------------------------- continuação
ReDim M(1 To n + 1, 1 To n + 1)
ReDim Inv(1 To n + 1, 1 To n + 1)
ReDim B(1 To n + 1)
ReDim A(0 To n)

For i = 0 To n ' criar as matrizes M e B
    For j = 0 To i
        M(i + 1, j + 1) = Sx(i + j)
        M(j + 1, i + 1) = Sx(i + j)
    Next j
B(i + 1) = Sxy(i)
Next i
' resolver o sistema M * A = B usando inversão da matriz M
Inv = Application.MInverse(M)

For i = 1 To n + 1 'multiplicação das matrizes: A = M-1 * B
A(i - 1) = 0
    For j = 1 To n + 1
        A(i - 1) = A(i - 1) + Inv(i, j) * B(j)
    Next j
Next i
RegressPoli = A 'retornar o vetor A
End Function
```

Regressão com logaritmos



	A	B	C	D
1	d/m	I/uA	log(d)	log(I)
2	0,11	632	-0,95861	2,800717
3	0,14	500	-0,85387	2,69897
4	0,18	317	-0,74473	2,501059
5	0,25	167	-0,60206	2,222716
6	0,3	115	-0,52288	2,060698
7	0,4	65	-0,39794	1,812913
8	0,5	41	-0,30103	1,612784
9	0,6	28	-0,22185	1,447158
10	0,7	21	-0,1549	1,322219
11	0,8	16	-0,09691	1,20412
12				
13			-1,91994	1,036246
14			0,038711	0,021796
15			0,996758	0,034909
16			2459,84	8
17			2,997662	0,009749

A planilha mostra a corrente em μA de uma fotocélula em função da distância d entre lâmpada e célula. O primeiro diagrama parece exibir uma tendência hiperbólica entre I e d. O gráfico dos logaritmos mostra uma relação linear com a equação $\log y = \log a + b \cdot \log x = 1{,}036 - 1{,}919 \cdot x$. (O diagrama à direita foi feito com *Linha de Tendência linear*.)

A retransformação dos logaritmos para as unidades originais, nos dá a equação de uma função de potência: $\mathbf{y = a \cdot x^b = 10{,}86\mu A \cdot x^{-1{,}92}} \approx 10{,}9\mu A \cdot x^{-2}$, pois $a = 10^{1{,}036} = 10{,}87$.