

Trabalho Computacional 4 - ALN

João Lucas Duim

15 de Maio de 2021

1 Questão 1

1.1 (a)

Como queremos uma aproximação quadrática por mínimos quadrados, ajustemos a equação $y = a + bt + ct^2$ aos dados, sendo $a = s_0$, $b = v_0$ e $c = \frac{g}{2}$ os parâmetros a serem determinados e $y = s(t)$. Temos, então, o seguinte sistema:

$$\begin{cases} a + 0,5b + 0,25c = 11 \\ a + b + c = 17 \\ a + 1,5b + 2,25c = 21 \\ a + 2b + 4c = 23 \\ a + 3b + 9c = 18 \end{cases}$$

Escrevendo na forma matricial $Ax = b$:

$$\begin{bmatrix} 1 & 0,5 & 0,25 \\ 1 & 1 & 1 \\ 1 & 1,5 & 2,25 \\ 1 & 2 & 4 \\ 1 & 3 & 9 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} 11 \\ 17 \\ 21 \\ 23 \\ 18 \end{bmatrix}$$

Vamos, então, calcular $A^T A$ e $A^T b$:

$$A^T A = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 0,5 & 1 & 1,5 & 2 & 3 \\ 0,25 & 1 & 2,25 & 4 & 9 \end{bmatrix} \begin{bmatrix} 1 & 0,5 & 0,25 \\ 1 & 1 & 1 \\ 1 & 1,5 & 2,25 \\ 1 & 2 & 4 \\ 1 & 3 & 9 \end{bmatrix} = \begin{bmatrix} 5 & 8 & 16,5 \\ 8 & 16,5 & 39,5 \\ 16,5 & 39,5 & 103,125 \end{bmatrix}$$
$$A^T b = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 0,5 & 1 & 1,5 & 2 & 3 \\ 0,25 & 1 & 2,25 & 4 & 9 \end{bmatrix} \begin{bmatrix} 11 \\ 17 \\ 21 \\ 23 \\ 18 \end{bmatrix} = \begin{bmatrix} 90 \\ 154 \\ 321 \end{bmatrix}$$

Temos o sistema de equações normais $(A^T A)\bar{x} = (A^T b)$:

$$\begin{bmatrix} 5 & 8 & 16,5 \\ 8 & 16,5 & 39,5 \\ 16,5 & 39,5 & 103,125 \end{bmatrix} \begin{bmatrix} \bar{a} \\ \bar{b} \\ \bar{c} \end{bmatrix} = \begin{bmatrix} 90 \\ 154 \\ 321 \end{bmatrix}$$

Resolvendo o sistema, encontramos:

$$\bar{x} = \begin{bmatrix} \bar{a} \\ \bar{b} \\ \bar{c} \end{bmatrix} = \begin{bmatrix} 1,9175258 \\ 20,306333 \\ -4,9720177 \end{bmatrix}$$

Portanto, os parâmetros encontrados foram $s_0 = 1,9175258$, $v_0 = 20,306333$ e $\frac{g}{2} = -4,9720177 \Rightarrow g = -9,9440353$ e, então, a equação quadrática que melhor ajusta esses dados é $y = 1,9175258 + 20,306333 \cdot t - 4,9720177 \cdot t^2$.

Veja, na figura 1, os pontos dados e a função ajustada no gráfico:

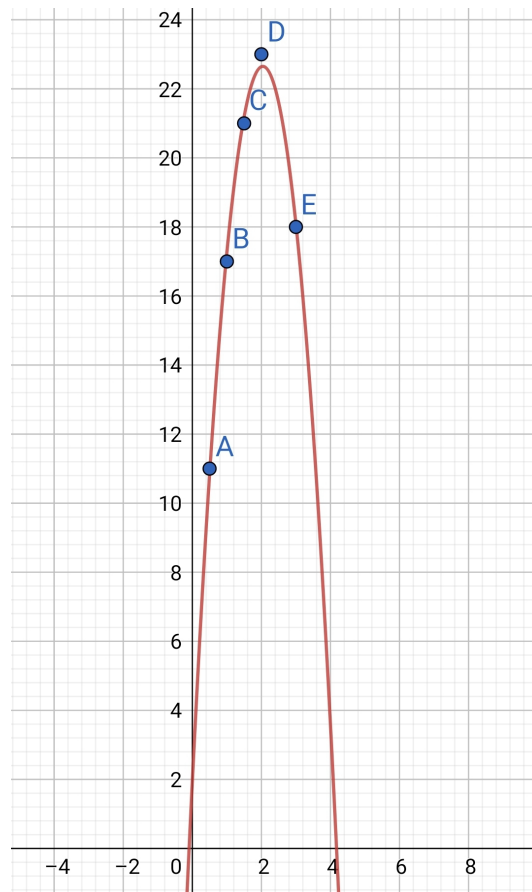


Figure 1: Gráfico da função ajustada

1.2 (b)

Conforme encontrado no item anterior (note que todos os dados trabalhados estão em conformidade com o SI), a melhor estimativa para a altura na qual o objeto foi solto é $s_0 = 1,9175258\text{m}$, para a velocidade inicial é $v_0 = 20,306333\text{m/s}$ e para a sua aceleração da gravidade é $g = -9,9440353\text{m/s}^2$.

1.3 (c)

A melhor estimativa para o momento em que o objeto atingirá o chão é substituir $y = 0$ na equação ajustada e encontrar o valor de t :

$$1,9175258 + 20,306333 \cdot t - 9,9440353 \cdot t^2 = 0$$

Na forma matricial:

$$\begin{bmatrix} 1 & t & t^2 \end{bmatrix} \begin{bmatrix} 1,9175258 \\ 20,306333 \\ -4,9720177 \end{bmatrix} = 0$$

Resolvendo e descartando a solução negativa, encontramos $t = 4,1764653$. Portanto, a melhor estimativa para o momento em que o objeto atingirá o chão é 4,1764653s após o início da contagem de tempo.

Veja, na figura 2, os cálculos descritos acima feitos no Scilab, além da norma do vetor de erros:

```

--> b = [11; 17; 21; 23; 18]; abc = [0.5; 1; 1.5; 2; 3]; A = [ones(abc) abc abc^2];
--> C = A' * A
C =
5. 8. 16.5
8. 16.5 39.5
16.5 39.5 103.125
--> d = A' * b
d =
90.
154.
321.
--> xbarra = Gaussian_Elimination_4(C, d)
xbarra =
1.9175258
20.306333
-4.9720177
--> e = b - A * xbarra;
--> norm_e = norm(e)
norm_e =
0.5148745
-->

```

Figure 2: Método dos Mínimos Quadrados aplicado ao sistema em questão

2 Questão 2

2.1 (a)

O modelo exponencial fornecido é $p(t) = ce^{kt}$. Aplicando logaritmo neperiano em ambos os membros, obtemos a linearização do modelo: $\log p(t) = \log c + k \cdot t$. Ajustemos, então, a equação $y = a + bt$ aos dados, sendo $a = \log c$ e $b = k$ os parâmetros a serem determinados e $y = \log p(t)$. Temos, então, o seguinte sistema, considerando $t = 0$ para 1950 e 10 anos como unidade para t :

$$\begin{cases} a = \log 150 \\ a + b = \log 179 \\ a + 2b = \log 203 \\ a + 3b = \log 227 \\ a + 4b = \log 250 \\ a + 5b = \log 281 \end{cases}$$

Escrevendo na forma matricial $Ax = b$:

$$\begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \\ 1 & 5 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} \log 150 \\ \log 179 \\ \log 203 \\ \log 227 \\ \log 250 \\ \log 281 \end{bmatrix} = \begin{bmatrix} 5,0106353 \\ 5,1873858 \\ 5,313206 \\ 5,42495 \\ 5,5214609 \\ 5,6383547 \end{bmatrix}$$

Vamos, então, calcular $A^T A$ e $A^T b$:

$$A^T A = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 3 & 4 & 5 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \\ 1 & 5 \end{bmatrix} = \begin{bmatrix} 6 & 15 \\ 15 & 55 \end{bmatrix}$$

$$A^T b = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 3 & 4 & 5 \end{bmatrix} \begin{bmatrix} 5,0106353 \\ 5,1873858 \\ 5,313206 \\ 5,42495 \\ 5,5214609 \\ 5,6383547 \end{bmatrix} = \begin{bmatrix} 32,095993 \\ 82,366265 \end{bmatrix}$$

Temos o sistema de equações normais $(A^T A)\bar{x} = (A^T b)$:

$$\begin{bmatrix} 6 & 15 \\ 15 & 55 \end{bmatrix} \begin{bmatrix} \bar{a} \\ \bar{b} \end{bmatrix} = \begin{bmatrix} 32,095993 \\ 82,366265 \end{bmatrix}$$

Resolvendo o sistema, encontramos:

$$\bar{x} = \begin{bmatrix} \bar{a} \\ \bar{b} \end{bmatrix} = \begin{bmatrix} 5,0455774 \\ 0,1215019 \end{bmatrix}$$

Portanto, os parâmetros encontrados foram $\log c = 5,0455774 \Rightarrow c = e^{5,0455774} = 155,33396$ e $k = 0,1215019$ e, então, a equação exponencial que melhor ajusta esses dados é $y = 155,33396 \cdot e^{0,1215019 \cdot t}$.

Veja, na figura 3, os pontos dados e a função ajustada no gráfico:

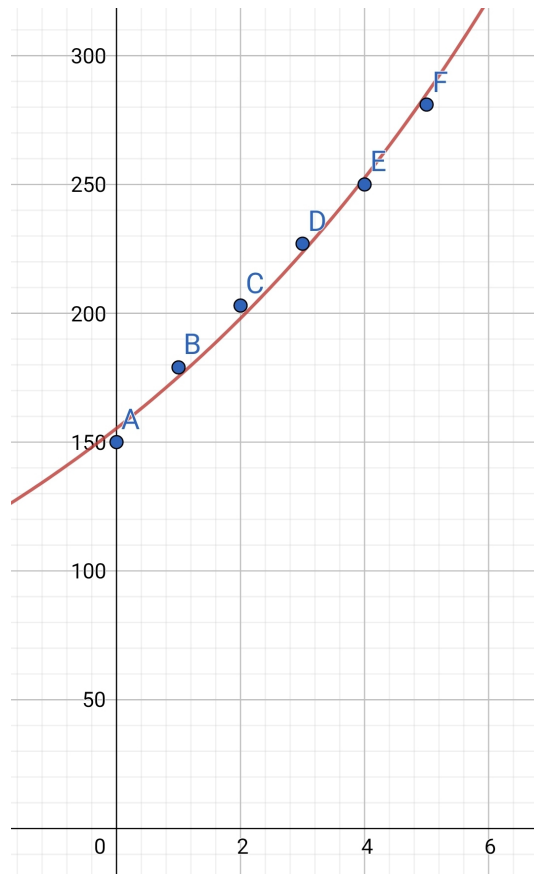


Figure 3: Gráfico da função ajustada

2.2 (b)

Para o ano 2010, temos $t = 6$. Pela equação ajustada obtida no item anterior, $p(6) = 155,33396 \cdot$

$$e^{0,1215019 \cdot 6} = e^{5,0455774} \cdot e^{0,1215019 \cdot 6} = e^{5,0455774 + 0,1215019 \cdot 6} = e^{\begin{bmatrix} 1 & 6 \end{bmatrix} \begin{bmatrix} 5,0455774 \\ 0,1215019 \end{bmatrix}} = e^{\begin{bmatrix} 1 & 6 \end{bmatrix} \cdot \bar{x}} = 322,01198.$$

A melhor estimativa para a população dos Estados Unidos em 2010 é 322,01198 milhões de habitantes. Uma rápida pesquisa na internet revela que a população real em 2010 era de 309,3 milhões de habitantes. Então, a modelagem exponencial feita superestimou a população com um erro de 4,11%. Isso ocorre porque, à medida que a população cresce, a capacidade de suporte do meio reduz (por exemplo, ocorre escassez de alimentos), o que freia o crescimento populacional, não sendo levado em conta na modelagem exponencial.

Veja, na figura 4, os cálculos descritos acima feitos no Scilab, além da norma do vetor de erros:

```

--> b = log([150, 179, 203; 227; 250; 281]); abc = [0; 1; 2; 3; 4; 5]; A = [ones(abc) abc];

--> C = A' * A
C =

    6.    15.
   15.   55.

--> d = A' * b
d =

   32.095993
   82.366265

--> xbarra = Gaussian_Elimination_4(C, d)
xbarra =

    5.0455774
    0.1215019

--> e = exp(b) - exp(A * xbarra);

--> norm_e = norm(e)
norm_e =

    10.043147

-->
-->
-->

```

Figure 4: Método dos Mínimos Quadrados aplicado ao sistema em questão

3 Questão 3

3.1 (a)

Como queremos uma aproximação quadrática por mínimos quadrados, ajustemos a equação $y = a + bt + ct^2$ aos dados, sendo a , b e c os parâmetros a serem determinados e y a média salarial em milhares de dólares no ano correspondente. Temos, então, o seguinte sistema, considerando $t = 0$ para 1970 e 5 anos como unidade para t :

$$\begin{cases} a = 29,3 \\ a + b + c = 44,7 \\ a + 2b + 4c = 143,8 \\ a + 3b + 9c = 371,6 \\ a + 4b + 16c = 597,5 \\ a + 5b + 25c = 1110,8 \\ a + 6b + 36c = 1895,6 \\ a + 7b + 49c = 2476,6 \end{cases}$$

Escrevendo na forma matricial $Ax = b$:

$$\begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 2 & 4 \\ 1 & 3 & 9 \\ 1 & 4 & 16 \\ 1 & 5 & 25 \\ 1 & 6 & 36 \\ 1 & 7 & 49 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} 29,3 \\ 44,7 \\ 143,8 \\ 371,6 \\ 597,5 \\ 1110,8 \\ 1895,6 \\ 2476,6 \end{bmatrix}$$

Vamos, então, calcular $A^T A$ e $A^T b$:

$$A^T A = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ 0 & 1 & 4 & 9 & 16 & 25 & 36 & 49 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 2 & 4 \\ 1 & 3 & 9 \\ 1 & 4 & 16 \\ 1 & 5 & 25 \\ 1 & 6 & 36 \\ 1 & 7 & 49 \end{bmatrix} = \begin{bmatrix} 8 & 28 & 140 \\ 28 & 140 & 784 \\ 140 & 784 & 4676 \end{bmatrix}$$

$$A^T b = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ 0 & 1 & 4 & 9 & 16 & 25 & 36 & 49 \end{bmatrix} \begin{bmatrix} 29,3 \\ 44,7 \\ 143,8 \\ 371,6 \\ 597,5 \\ 1110,8 \\ 1895,6 \\ 2476,6 \end{bmatrix} = \begin{bmatrix} 6669,9 \\ 38100,9 \\ 230889,3 \end{bmatrix}$$

Temos o sistema de equações normais $(A^T A)\bar{x} = (A^T b)$:

$$\begin{bmatrix} 8 & 28 & 140 \\ 28 & 140 & 784 \\ 140 & 784 & 4676 \end{bmatrix} \begin{bmatrix} \bar{a} \\ \bar{b} \\ \bar{c} \end{bmatrix} = \begin{bmatrix} 6669,9 \\ 38100,9 \\ 230889,3 \end{bmatrix}$$

Resolvendo o sistema, encontramos:

$$\bar{x} = \begin{bmatrix} \bar{a} \\ \bar{b} \\ \bar{c} \end{bmatrix} = \begin{bmatrix} 57,0625 \\ -101,67321 \\ 64,716071 \end{bmatrix}$$

Portanto, os parâmetros encontrados foram $a = 57,0625$, $b = -101,67321$ e $c = 64,716071$ e, então, a equação quadrática que melhor ajusta esses dados é $y = 57,0625 - 101,67321 \cdot t + 64,716071 \cdot t^2$.

Veja, na figura 5, os cálculos descritos acima feitos no Scilab, além da norma do vetor de erros:

```

Scilab 6.1.0 Console
Arquivo Editar Controle Aplicativos ?
[Icons]
Scilab 6.1.0 Console
--> b = [29.3; 44.7; 143.8; 371.6; 597.5; 1110.8; 1895.6; 2476.6]; abc = [0; 1; 2; 3; 4; 5; 6; 7]; A = [ones(abc) abc abc^2];
--> C = A' * A
C =
8. 28. 140.
28. 140. 784.
140. 784. 4676.
--> d = A' * b
d =
6669.9
38100.9
230889.3
--> xbarra = Gaussian_Elimination_4(C, d)
xbarra =
57.0625
-101.67321
64.716071
--> e = b - A * xbarra;
--> norm_e = norm(e)
norm_e =
174.19173
-->

```

Figure 5: Método dos Mínimos Quadrados aplicado ao sistema em questão

Veja, na figura 6, os pontos dados e a função ajustada no gráfico:

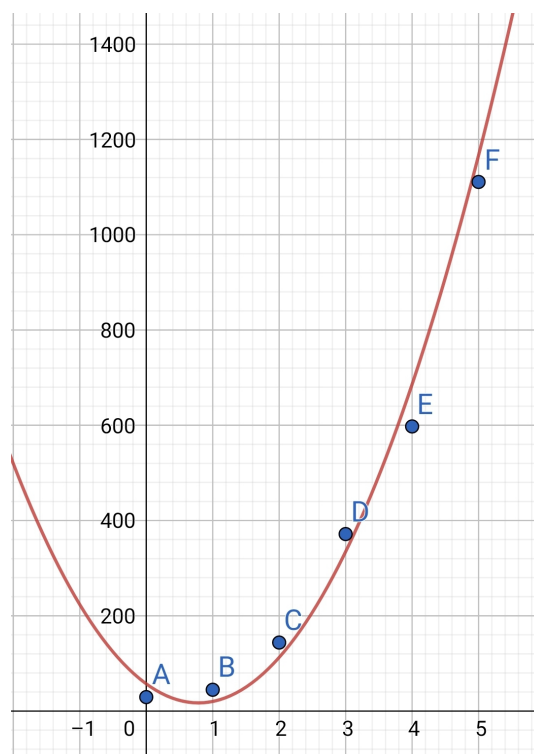


Figure 6: Gráfico da função ajustada

3.2 (b)

Como queremos uma aproximação quadrática por mínimos quadrados, modelamos os dados com $s(t) = ce^{kt}$, considerando $t = 0$ para 1970 e 5 anos como unidade para t e sendo $s(t)$ a média

salarial em milhares de dólares no ano correspondente. Aplicando logaritmo neperiano em ambos os membros, obtemos a linearização do modelo: $\log s(t) = \log c + k \cdot t$. Ajustemos, então, a equação $y = a + bt$ aos dados, sendo $a = \log c$ e $b = k$ os parâmetros a serem determinados e $y = \log s(t)$. Temos, então, o seguinte sistema:

$$\begin{cases} a = \log 29,3 \\ a + b = \log 44,7 \\ a + 2b = \log 143,8 \\ a + 3b = \log 371,6 \\ a + 4b = \log 597,5 \\ a + 5b = \log 1110,8 \\ a + 6b = \log 1985,6 \\ a + 7b = \log 2476,6 \end{cases}$$

Escrevendo na forma matricial $Ax = b$:

$$\begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \\ 1 & 5 \\ 1 & 6 \\ 1 & 7 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} \log 29,3 \\ \log 44,7 \\ \log 143,8 \\ \log 371,6 \\ \log 597,5 \\ \log 1110,8 \\ \log 1985,6 \\ \log 2476,6 \end{bmatrix} = \begin{bmatrix} 3.3775875 \\ 3.7999735 \\ 4.9684234 \\ 5.917818 \\ 6.3927543 \\ 7.0128358 \\ 7.5472907 \\ 7.8146419 \end{bmatrix}$$

Vamos, então, calcular $A^T A$ e $A^T b$:

$$A^T A = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \\ 1 & 5 \\ 1 & 6 \\ 1 & 7 \end{bmatrix} = \begin{bmatrix} 8 & 28 \\ 28 & 140 \end{bmatrix}$$

$$A^T b = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 \end{bmatrix} \begin{bmatrix} 3.3775875 \\ 3.7999735 \\ 4.9684234 \\ 5.917818 \\ 6.3927543 \\ 7.0128358 \\ 7.5472907 \\ 7.8146419 \end{bmatrix} = \begin{bmatrix} 46.831325 \\ 192.11171 \end{bmatrix}$$

Temos o sistema de equações normais $(A^T A)\bar{x} = (A^T b)$:

$$\begin{bmatrix} 8 & 28 \\ 28 & 140 \end{bmatrix} \begin{bmatrix} \bar{a} \\ \bar{b} \end{bmatrix} = \begin{bmatrix} 46.831325 \\ 192.11171 \end{bmatrix}$$

Resolvendo o sistema, encontramos:

$$\bar{x} = \begin{bmatrix} \bar{a} \\ \bar{b} \end{bmatrix} = \begin{bmatrix} 3.5037431 \\ 0.6714779 \end{bmatrix}$$

Portanto, os parâmetros encontrados foram $\log c = 3.5037431 \Rightarrow c = e^{3.5037431} = 33.239640$ e $k = 0.6714779$ e, então, a equação exponencial que melhor ajusta esses dados é $y = 33.239640 \cdot e^{0.6714779 \cdot t}$.

Veja, na figura 7, os cálculos descritos acima feitos no Scilab, além da norma do vetor de erros:

```
Scilab 6.1.0 Console
Arquivo Editar Controle Aplicativos ?
[Icons]
Scilab 6.1.0 Console
--> b = log([29.3; 44.7; 143.8; 371.6; 597.5; 1110.8; 1895.6; 2476.6]); abc = [0; 1; 2; 3; 4; 5; 6; 7]; A = [ones(abc) abc];
--> C = A' * A
C =
8. 28.
28. 140.
--> d = A' * b
d =
46.831325
192.11171
--> xbarra = Gaussian_Elimination_4(C, d)
xbarra =
3.5037431
0.6714779
--> e = exp(b) - exp(A * xbarra);
--> norm_e = norm(e)
norm_e =
1201.5096
-->
-->
-->|
```

Figure 7: Método dos Mínimos Quadrados aplicado ao sistema em questão

Veja, na figura 8, os pontos dados e a função ajustada no gráfico:

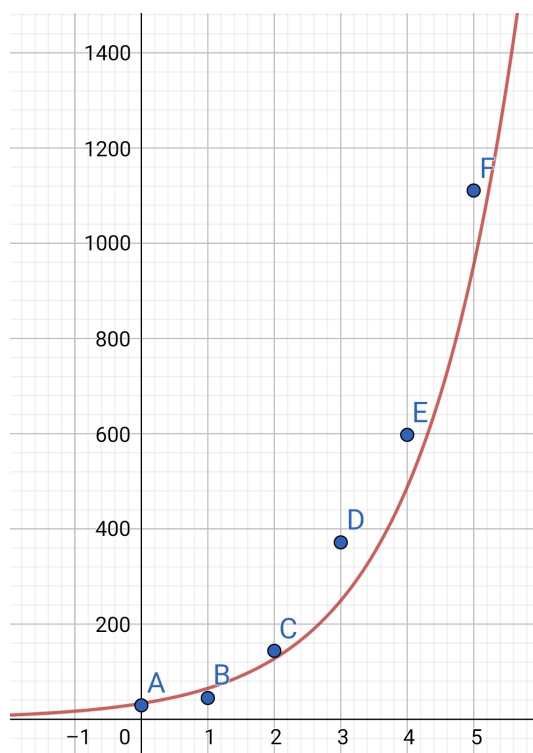


Figure 8: Gráfico da função ajustada

3.3 (c)

Uma rápida conferência nas figuras 3 e 4 mostra que a norma do vetor de erros obtida no item (a) é muito menor que a norma do vetor de erros obtida no item (b), ou seja, o modelo quadrático do item (a) se ajustou melhor aos dados. Portanto, a equação $y = 57,0625 - 101,67321 \cdot t + 64,716071 \cdot t^2$ dá uma melhor aproximação para a média salarial da liga adulta de beisebol para os anos de 1970 a 2000.

3.4 (d)

Dada a resposta do item anterior, vamos utilizar a equação $y = 57,0625 - 101,67321 \cdot t + 64,716071 \cdot t^2$ para estimar a média salarial da liga adulta de beisebol em 2010 ($t = 8$) e em 2015 ($t = 9$):

$$y(8) = 57,0625 - 101,67321 \cdot 8 + 64,716071 \cdot 8^2 = \begin{bmatrix} 1 & 8 & 64 \end{bmatrix} \begin{bmatrix} 57,0625 \\ -101,67321 \\ 64,716071 \end{bmatrix} = 3385,5054$$

$$y(9) = 57,0625 - 101,67321 \cdot 9 + 64,716071 \cdot 9^2 = \begin{bmatrix} 1 & 9 & 81 \end{bmatrix} \begin{bmatrix} 57,0625 \\ -101,67321 \\ 64,716071 \end{bmatrix} = 4384,0054$$

Portanto, a melhor estimativa para a média salarial da liga adulta de beisebol em 2010 é de 3385,51 milhares de dólares e em 2015 é de 4384,01 milhares de dólares.

4 Questão 4

Feita a leitura dos arquivos de treinamento e de teste armazenados, respectivamente, nas variáveis A_{tr} e A_{tt} , podemos proceder com o ajuste da função $y = h(x) = \alpha_0 + \sum_{i=1}^{10} \alpha_i \cdot X_i$, sendo X_i a i -ésima entrada do vetor x . Temos, então, o seguinte sistema:

$$\begin{cases} y^{(1)} = \alpha_0 + \sum_{i=1}^{10} \alpha_i \cdot X_i^{(1)} \\ y^{(2)} = \alpha_0 + \sum_{i=1}^{10} \alpha_i \cdot X_i^{(2)} \\ y^{(3)} = \alpha_0 + \sum_{i=1}^{10} \alpha_i \cdot X_i^{(3)} \\ \vdots \\ y^{(300)} = \alpha_0 + \sum_{i=1}^{10} \alpha_i \cdot X_i^{(300)} \end{cases}$$

Escrevendo na forma matricial $X\alpha = y$:

$$\begin{bmatrix} 1 & X_1^{(1)} & X_2^{(1)} & \dots & X_{10}^{(1)} \\ 1 & X_1^{(2)} & X_2^{(2)} & \dots & X_{10}^{(2)} \\ 1 & X_1^{(3)} & X_2^{(3)} & \dots & X_{10}^{(3)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_1^{(300)} & X_2^{(300)} & \dots & X_{10}^{(300)} \end{bmatrix} \begin{bmatrix} \alpha_0 \\ \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_{10} \end{bmatrix} = \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ y^{(3)} \\ \vdots \\ y^{(10)} \end{bmatrix}$$

Resolvendo o sistema $X^T X \alpha_{tr} = X^T y$ de equações normais, encontramos $\alpha_{tr} =$

$$\begin{bmatrix} -6,7579731 \\ 29,311052 \\ 2,0765803 \\ -18,730222 \\ -7,3665161 \\ 1,2222756 \\ 0,2283419 \\ 0,0503253 \\ 2,2385058 \\ 0,0249405 \\ 0,7704282 \end{bmatrix}$$

Veja, na figura 9, os cálculos descritos acima feitos no Scilab, além do número de acertos do modelo sobre os dados do arquivo de treinamento:

```

Scilab 6.1.0 Console
Arquivo  Editar  Controle  Aplicativos ?

--> Y_tr = A_tr(:,11);
--> Y_tt = A_tt(:,11);
--> X_tr = [ones(Y_tr) A_tr(:, 1:10)];
--> X_tt = [ones(Y_tt) A_tt(:, 1:10)];
--> alfa_tr = Gaussian_Elimination_4(X_tr' * X_tr, X_tr' * Y_tr)
alfa_tr =

-6.7579731
29.311052
2.0765803
-18.730222
-7.3665161
1.2222756
0.2283419
0.0503253
2.2385058
0.0249405
0.7704282

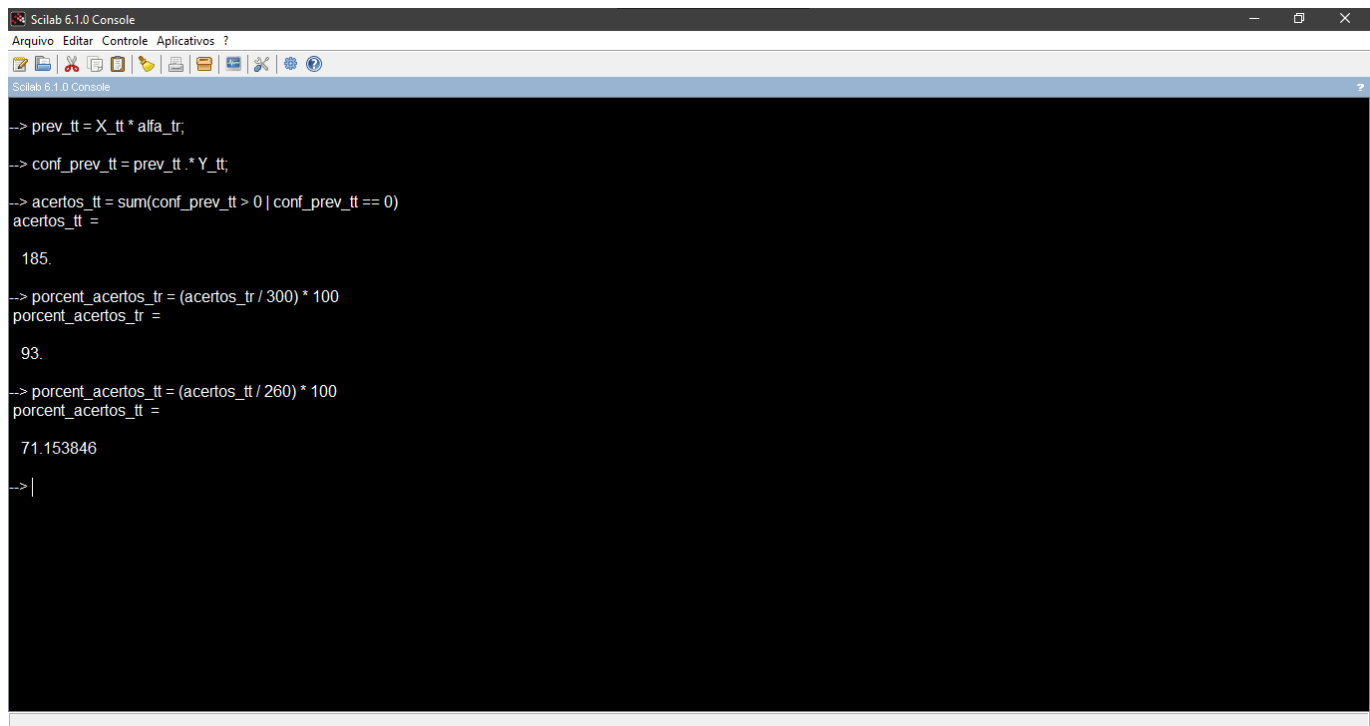
--> prev_tr = X_tr * alfa_tr;
--> conf_prev_tr = prev_tr .* Y_tr;
--> acertos_tr = sum(conf_prev_tr > 0 | conf_prev_tr == 0)
acertos_tr =

279

```

Figure 9: Ajustes do modelo aos dados do arquivo de treinamento

Veja, na figura 10, o número de acertos do modelo sobre os dados do arquivo de teste, além do cálculo da porcentagem de acertos em ambos os conjuntos de dados:

A screenshot of the Scilab 6.1.0 Console window. The window has a menu bar with 'Arquivo', 'Editar', 'Controlé', and 'Aplicativos ?'. Below the menu is a toolbar with various icons. The main area is a black console with white text. The code entered is:

```
--> prev_tt = X_tt * alfa_tr;  
--> conf_prev_tt = prev_tt .* Y_tt;  
--> acertos_tt = sum(conf_prev_tt > 0 | conf_prev_tt == 0)  
acertos_tt =  
  
185.  
  
--> percent_acertos_tr = (acertos_tr / 300) * 100  
percent_acertos_tr =  
  
93.  
  
--> percent_acertos_tt = (acertos_tt / 260) * 100  
percent_acertos_tt =  
  
71.153846  
--> |
```

Figure 10: Coleta de quantidades e porcentagens de acertos

Logo, o modelo ajustado obteve 93% de acerto sobre os dados de treinamento e obteve 71,153846% de acerto sobre os dados de teste. Vamos, então, construir a matriz de confusão:

TP é o número de casos positivos previstos como positivos; TN é o número de casos negativos previstos como negativos; FP é o número de casos negativos previstos como positivos; FN é o número de casos positivos previstos como negativos;

Veja, na figura 11, a quantidade de casos reais positivos, reais negativos, previstos positivos e previstos negativos:

```

Scilab 6.1.0 Console
Arquivo Editar Controle Aplicativos ?
qtd_positivos_reais = sum(Y_tt == 1)
qtd_positivos_reais =
60.
qtd_negativos_reais = sum(Y_tt == -1)
qtd_negativos_reais =
200.
qtd_positivos_previstos = sum(prev_tt > 0 | prev_tt == 0)
qtd_positivos_previstos =
135.
qtd_negativos_previstos = sum(prev_tt < 0)
qtd_negativos_previstos =
125.
-->|

```

Figure 11: Dados coletados a respeito da quantidade de casos positivos e negativos, reais e previstos

$$\begin{cases} TP + FN = 60 \\ TN + FP = 200 \\ TP + FP = 135 \\ TN + FN = 125 \end{cases}$$

Resolvendo, encontramos:

$$\begin{cases} TP = 60 \\ TN = 125 \\ FP = 75 \\ FN = 0 \end{cases}$$

A matriz de confusão é a seguinte, onde a coluna diz a classe prevista e a linha diz a classe real:

	P	N
P	60	0
N	75	125

Table 1: Matriz de confusão

Vamos agora calcular as outras medidas decorrentes:

Acurácia: $\frac{TP+TN}{TP+TN+FP+FN} = \frac{60+125}{60+125+75+0} = \frac{185}{260} = 0,7115385$

Precisão: $\frac{TP}{TP+FP} = \frac{60}{60+75} = \frac{60}{135} = 0,4444444$

Recall: $\frac{TP}{TP+FN} = \frac{60}{60+0} = \frac{60}{60} = 1$

Probabilidade de falso alarme: $\frac{FP}{FP+TN} = \frac{75}{75+125} = \frac{75}{200} = 0,375$

Probabilidade de falsa omissão de alarme: $\frac{FN}{FN+TN} = \frac{0}{0+125} = 0$

Comentários gerais: É esperado que o modelo tenha uma alta porcentagem de acerto sobre os dados de treinamento, porque ele foi treinado exatamente para isso, enquanto a porcentagem de acerto sobre os dados de teste é razoável. A acurácia é a fração entre o número de acertos e a quantidade total de dados, a precisão é a razão entre o número de casos positivos corretamente classificados e a quantidade total de casos positivos previstos, o recall é a razão entre o número de casos positivos corretamente classificados e a quantidade total de casos positivos reais, a probabilidade de falso alarme é a razão entre o número de casos negativos previstos como positivos e a quantidade total de casos negativos reais e a probabilidade de falsa omissão de alarme é a razão entre o número de casos positivos previstos como negativos e a quantidade total de casos negativos reais. Podemos ver que o modelo acertou na previsão de 100% dos casos positivos reais, ou seja, nenhum paciente que tem realmente câncer foi diagnosticado erroneamente. No entanto, 37,5% dos pacientes que não têm câncer foram diagnosticados com câncer. No geral, a acurácia do modelo foi de 71,15%, aproximadamente, o que é relativamente bom, bem melhor do que um chute às cegas, além de a não ocorrência de falsas omissões de alarme ser também um ótimo ponto a favor do modelo, já que poupará vidas. Por outro lado, uma porcentagem expressiva de pacientes que não portam a doença foram diagnosticados com a doença, o que causa um enorme impacto emocional. Logo, o modelo treinado pode representar um forte aliado no diagnóstico de câncer de mama, porém são necessários estudos mais precisos sobre como a doença afeta os locais do corpo de onde foram obtidos os dados, a fim de obter um modelo mais fidedigno à realidade, sendo capaz de diminuir a probabilidade de falso alarme, de aumentar a acurácia e a precisão e de manter o recall muito próximo de 1 e a probabilidade de falsa omissão de alarme muito próxima de 0.