

Ajuste de Curvas

Prof. Alexandre Salles da Cunha e Profa. Ana Paula Couto



UNIVERSIDADE FEDERAL
DE MINAS GERAIS



As duas aplicações da projeção de um vetor em um subespaço que serão abordadas neste curso são:

- A partir de uma grande quantidade de dados, definir um modelo que explique a relação entre uma variável dependente ou explicativa dado um conjunto de variáveis independentes, atributos ou preditores.
- Fatorar uma matriz A em duas matrizes: (i) matriz Q , cujas as colunas formam uma base ortonormal de $\mathcal{C}(A)$ e (ii) matriz R , matriz triangular superior.

Problema

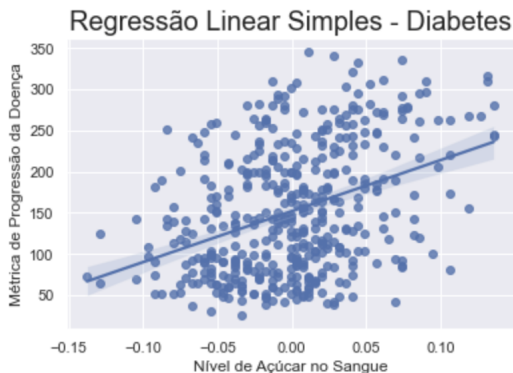
Quantificar como os atributos:

- Idade,
- Sexo,
- Índice de massa corporal,
- Pressão sanguínea,
- VLDL, LDL, HDL,
- Colesterol total,
- Triglicerídeos,
- e Nível de açúcar no sangue.

impactam no progresso da diabetes em um conjunto de 442 pacientes (https://scikit-learn.org/stable/datasets/toy_dataset.html#diabetes-dataset).

Vamos modelar o problema através de modelos de regressão linear.

Modelo 1: Atributo predictor: Nível de açúcar no sangue.

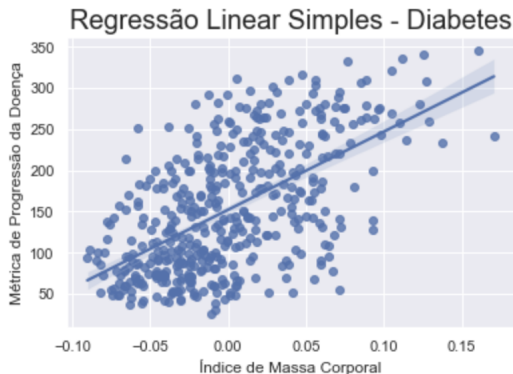


Intercepto: 152.13348416289614

Coeficientes: [619.22282068]

Coeficiente de Determinação (R2 Score): 0.14629361572293453

Modelo 2: Atributo preditor: Índice de massa corporal.



Intercepto: 152.1334841628967

Coefficientes: [949.43526038]

Coefficiente de Determinação (R2 Score): 0.3439237602253803

Modelo 3: Regressão Linear Múltipla

Intercepto: 152.1334841628965

Coefficientes: [-10.01219782 -239.81908937 519.83978679 324.39042769 -792.1841
476.74583782 101.04457032 177.06417623 751.27932109 67.62538639]

Coefficiente de Determinação (R2 Score): 0.5177494254132934



Escolha dos atributos

Atributos Selecionados: Índice de Massa Corporal

Intercepto: 152.1334841628967

Coeficientes: [949.43526038]

Coeficiente de Determinação (R2 Score): 0.3439237602253803

Atributos Selecionados: Índice de Massa Corporal e Triglicerídeos

Intercepto: 152.1334841628967

Coeficientes: [675.06977443 614.95050478]

Coeficiente de Determinação (R2 Score): 0.4594852440167805

Atributos Selecionados: Índice de Massa Corporal, Triglicerídeos, Pressão Sanguí

Intercepto: 152.1334841628966

Coeficientes: [603.07435575 543.87245014 262.27488392]

Coeficiente de Determinação (R2 Score): 0.48008281990946045

Problema

Considere um sistema linear com n incógnitas, com $m > n$ equações.

- Em várias aplicações, buscamos a solução $x \in \mathbb{R}^n$, que satisfaça $Ax = b$, com $A \in \mathbb{R}^{m \times n}$ e $b \in \mathbb{R}^m$.
- Em geral, este sistema não tem solução: um vetor x que satisfaça a equação somente existe se b pertence ao $\mathcal{C}(A)$.
- Dado que b é um vetor m -dimensional e a dimensão de $\mathcal{C}(A)$ é no máximo n , o sistema linear tem solução para apenas algumas escolhas de b .
- Este sistema retangular de equações com $m > n$ é conhecido como sobredeterminado.

Problema

- Qualquer pequena variação (perturbação) nos elementos de b , pode resultar na não-solução de $Ax = b$.
- Esta situação pode ocorrer em aplicações onde dados são obtidos a partir de medições.
- Este problema também é conhecido como problema dos **mínimos quadrados** ou **regressão linear**.

Exemplo

Uma aplicação importante de mínimos quadrados é ajustar m pontos a uma reta (regressão linear simples). Vamos considerar um problema com 3 pontos: Encontre a reta mais próxima a três pontos $(0, 6)$, $(1, 0)$, $(2, 0)$.

$$\beta_0 + \beta_1(0) = 6$$

$$\beta_0 + \beta_1(1) = 0$$

$$\beta_0 + \beta_1(2) = 0$$

- O sistema linear de 3 equações e 2 variáveis não tem solução: $b = (6, 0, 0)^T$ não é uma combinação linear das colunas $(1, 1, 1)^T$ e $(0, 1, 2)^T$ de A .
- Em problemas reais, facilmente podemos ter casos com $m = 100$ pontos, que não vão ajustar exatamente em uma reta $\beta_0 + \beta_1 x$.

O que significa resolver um sistema que não tem solução?

- Como o vetor $b \notin \mathcal{C}(A)$, temos um erro (ou resíduo):

$$e = b - Ax$$

- Como o valor de e pode não ser igual à zero, uma escolha adequada de $\hat{x}$ resulta em um resíduo que é o menor possível.
- Buscamos o vetor e com menor norma Euclidiana.

Revisitando o Problema

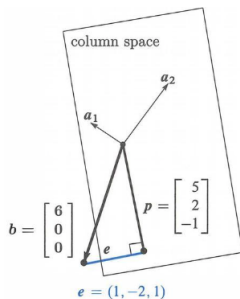
Dada a matriz $A \in \mathbb{R}^{m \times n}$, $m \geq n$ e $b \in \mathbb{R}^m$

Encontre $\hat{x} \in \mathbb{R}^n$, tal que a norma $\|b - A\hat{x}\|_2$ seja minimizada.

Como minimizar $e = b - Ax$? Através da projeção ortogonal.

- O objetivo é encontrar o vetor $\hat{x} \in \mathbb{R}^n$, tal que o vetor $A\hat{x} \in \mathbb{R}^m$ é o vetor mais próximo a b que pertence a $\mathcal{C}(A)$.
- O erro $e = b - A\hat{x}$ tem que ser ortogonal à $\mathcal{C}(A)$.

No exemplo com três pontos e ajuste a partir de uma reta:



Teorema

Seja $A \in \mathbb{R}^{m \times n}$ ($m \geq n$) e um vetor $b \in \mathbb{R}^m$ dado. Um vetor $\hat{x} \in \mathbb{R}^n$ minimiza a norma do erro $\|e\|_2 = \|b - A\hat{x}\|_2$, se e somente se $e \perp \mathcal{C}(A)$, isto é:

$$A^T e = 0_m \text{ ou } A^T A \hat{x} = A^T b \text{ ou } A \hat{x} = Pb$$

onde $P \in \mathbb{R}^{m \times m}$ é a matriz de projeção ortogonal em $\mathcal{C}(A)$. O sistema $n \times n$ de equações $A^T A \hat{x} = A^T b$, conhecido como sistema de equações normais, é não singular se e somente se A possui posto completo, i.e. $r(A) = n$. A solução $\hat{x}$ é única se e somente se A possui posto completo, ou seja, todas as colunas são LI.

- Como não existe solução para $Ax = b$, resolvemos $A\hat{x} = p$.
- Com colunas LI, $A^T A$ é quadrada, simétrica positiva definida e inversível.

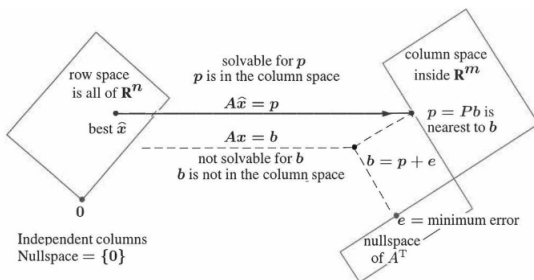


Figure 4.7: The projection $p = A\hat{x}$ is closest to b , so $\hat{x}$ minimizes $E = \|b - Ax\|^2$.

- Vamos considerar m pares de pontos $(x_1, b_1), \dots, (x_m, b_1)$.
- A reta $b = \beta_0 + \beta_1 x$ que melhor se ajusta aos pontos dados minimiza a soma das distâncias verticais $e_1, \dots, e_m$.
- A variável explicativa x está relacionada à variável resposta b por meio de um modelo linear com parâmetros β_0 e β_1 . O modelo é linear nos parâmetros.
- Temos o seguinte conjunto de equações:

$$\beta_0 + \beta_1 x_1 = b_1$$

$$\beta_0 + \beta_1 x_2 = b_2$$

$$\vdots$$

$$\beta_0 + \beta_1 x_m = b_m$$

- Em forma matricial:

$$\begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_m \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix}$$

- Como o vetor $b \notin \mathcal{C}(A)$, iremos utilizar sua projeção $p \in \mathcal{C}(A)$.
- Reduzimos o nosso problema a resolver o sistema: $A^T A \hat{x} = A^T b$, para $\hat{x} = [\beta_0 \ \beta_1]^T$.
- O erro entre o valor real b_i e o projetado $(A\hat{x})_i$ é dado por $e_i = b_i - \beta_0 - \beta_1 x_i$.

Com

$$A^T A = \begin{bmatrix} 1 & \cdots & 1 \\ x_1 & \cdots & x_m \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_m \end{bmatrix} = \begin{bmatrix} m & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix}$$

E

$$A^T b = \begin{bmatrix} 1 & \cdots & 1 \\ x_1 & \cdots & x_m \end{bmatrix} \begin{bmatrix} b_1 \\ \vdots \\ b_m \end{bmatrix} = \begin{bmatrix} \sum b_i \\ \sum x_i b_i \end{bmatrix}$$

Temos o sistema de equações lineares nos parâmetros β_0 e β_1 :

$$\begin{bmatrix} m & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \begin{bmatrix} \sum b_i \\ \sum x_i b_i \end{bmatrix}$$

- Aplicando as operações l-elementares e Eliminação de Gauss, temos o seguinte sistema linear equivalente:

$$\begin{bmatrix} m & \sum x_i \\ 0 & -\frac{1}{m}(\sum x_i)^2 + \sum x_i^2 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \begin{bmatrix} \sum b_i \\ -\frac{1}{m} \sum x_i \sum b_i + \sum x_i b_i \end{bmatrix}$$

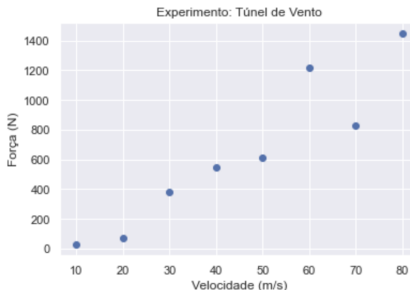
- Com solução:

$$\beta_1 = \frac{\sum x_i \sum b_i - m \sum x_i b_i}{(\sum x_i)^2 - m \sum x_i^2}$$

$$\beta_0 = \frac{\sum b_i - \beta_1 \sum x_i}{m}$$

Um indivíduo está suspenso em um túnel de vento e a força de resistência do ar é medida para vários níveis de velocidade do vento. Os resultados do experimento são descritos na seguinte tabela:

Velocidade (m/s)	10	20	30	40	50	60	70	80
Força (N)	25	70	380	550	610	1220	830	1450



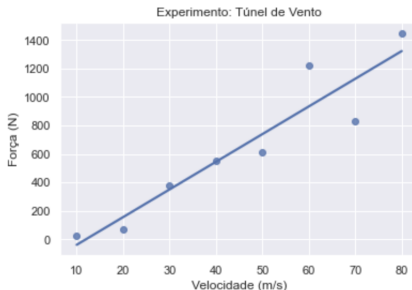
- Pontos indicam que a força aumenta à medida que a velocidade aumenta.
- Pontos apresentam dispersão significativa, particularmente em velocidades mais elevadas.
- Vamos modelar este problema utilizando uma reta.

i	x_i	b_i	x_i^2	$x_i b_i$
1	10	25	100	250
2	20	70	400	1.400
3	30	380	900	11.400
4	40	550	1.600	22.000
5	50	610	2.500	30.500
6	60	1220	3.600	73.200
7	70	830	4.900	58.100
8	80	1450	6.400	116.000
Σ	360	5.135	20.400	312.850

$$\beta_1 = \frac{\sum x_i \sum b_i - m \sum x_i b_i}{(\sum x_i)^2 - m \sum x_i^2} = 19,47$$

$$\beta_0 = \frac{\sum b_i - \beta_1 \sum x_i}{m} = -234,28$$

$$b = -234,28 + 19,47x$$



Intercepto: -234.28571428571456

Coefficientes: [19.4702381]

Coefficiente de Determinação (R2 Score): 0.8804852467812262

- Embora a reta pareça se ajustar bem aos dados, a equação prevê forças negativas fisicamente inconsistentes no caso de velocidades baixas (inferiores a 10 m/s).
- Vamos buscar um ajuste, com uma função explicativa em uma classe distinta, mas com modelo ainda linear nos parâmetros.

- Seja um modelo com p parâmetros β_i , $i = 0, 1, \dots, p-1$, que relaciona a variável dependente b com as $p-1$ variáveis independentes x_i através da equação:

$$b = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_{p-1} x_{p-1}.$$

- Este modelo é chamado de modelo de Regressão Linear Múltipla e possui o seguinte conjunto de equações normais:

$$\begin{bmatrix} m & \sum x_{1i} & \sum x_{2i} & \cdots & \sum x_{p-1,i} \\ \sum x_{1i} & \sum x_{1i}x_{1i} & \sum x_{2i}x_{1i} & \cdots & \sum x_{p-1,i}x_{1i} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum x_{p-1,i} & \sum x_{1i}x_{p-1,i} & \sum x_{2i}x_{p-1,i} & \cdots & \sum x_{p-1,i}x_{p-1,i} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix} = \begin{bmatrix} \sum b_i \\ \sum x_i b_i \\ \vdots \\ \sum x_{p-1,i} b_i \end{bmatrix}$$

Considere que a relação entre a renda familiar em SM (variável dependente), anos de estudo e idade do responsável pela família (variáveis independentes):

i	Renda (b_i)	Anos Estudo (x_1)	Idade (x_2)
1	4	1	20
2	8	4	30
3	10	6	40
4	12	7	50

seja modelada através de uma regressão linear múltipla. Defina os parâmetros $\beta_0, \beta_1, \beta_2$.

As equações normais para duas variáveis explicativas são:

$$\begin{bmatrix} m & \sum x_{1i} & \sum x_{2i} \\ \sum x_{1i} & \sum x_{1i}^2 & \sum x_{1i}x_{2i} \\ \sum x_{2i} & \sum x_{1i}x_{2i} & \sum x_{2i}^2 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} \sum b_i \\ \sum x_{1i}b_i \\ \sum x_{2i}b_i \end{bmatrix}$$

i	x_{1i}	x_{2i}	b_i	x_{1i}^2	$x_{1i}x_{2i}$	x_{2i}^2	$x_{1i}b_i$	$x_{2i}b_i$
1	1	20	4	1	20	400	4	80
2	4	30	8	16	120	900	32	240
3	6	40	10	36	240	1.600	60	400
4	7	50	12	49	350	2.500	84	600
$\sum$	18	140	34	102	730	5.400	180	1.320

Resolvendo o sistema:
$$\begin{bmatrix} 4 & 18 & 140 \\ 18 & 102 & 730 \\ 140 & 730 & 5.400 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} 34 \\ 180 \\ 1.320 \end{bmatrix}$$

Temos:

$$\beta_0 = 1,9, \beta_1 = 1 \text{ e } \beta_2 = 0,06$$

Com o seguinte modelo:

$$b = 1,9 + 1x_1 + 0,06x_2$$

Espera-se, para cada ano adicional de estudo do responsável pela família, um aumento de 1 SM e, para cada ano de idade adicional, um aumento de 0,06 SM na renda familiar.

- Um caso particular da regressão linear múltipla é quando a variável dependente se relaciona com uma única variável independente x , usando o modelo:

$$b = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_g x^g.$$

- Este problema é modelado pelo seguinte conjunto de equações normais:

$$\begin{bmatrix} m & \sum x_i & \sum x_i^2 & \cdots & \sum x_i^g \\ \sum x_i & \sum x_i^2 & \sum x_i^3 & \cdots & \sum x_i^{g+1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum x_i^g & \sum x_i^{g+1} & \sum x_i^{g+2} & \cdots & \sum x_i^{2g} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_g \end{bmatrix} = \begin{bmatrix} \sum b_i \\ \sum x_i b_i \\ \vdots \\ \sum x_i^g b_i \end{bmatrix}$$

- Consideremos um polinômio de segundo grau com m pontos distintos.
- O sistema de equações normais a ser resolvido é:

$$\begin{bmatrix} m & \sum x_i & \sum x_i^2 \\ \sum x_i & \sum x_i^2 & \sum x_i^3 \\ \sum x_i^2 & \sum x_i^3 & \sum x_i^4 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} \sum b_i \\ \sum x_i b_i \\ \sum x_i^2 b_i \end{bmatrix}$$

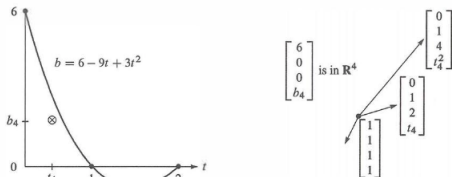


Figure 4.8: An exact fit of the parabola at $t = 0, 1, 2$ means that $p = b$ and $e = 0$. The fourth point $\otimes$ off the parabola makes $m > n$ and we need least squares: project b on $C(A)$. The figure on the right shows b —not a combination of the three columns of A .

- Nem sempre o modelo desejado é **linear nos parâmetros**.
- Podemos usar transformações para buscar um modelo equivalente ao desejado, mas que seja linear nos parâmetros.
- Alguns exemplos que podem ser linearizados:
 - **Modelo exponencial:** $b = \alpha e^{\beta x}$. Utilizado para caracterizar crescimento populacional, decaimento radioativo, entre outras aplicações.
 - **Equação de potência simples:** $b = \alpha x^{\beta}$. Utilizado para quantificar fenômenos físicos, como força de resistência do ar.
 - **Equação de taxa de crescimento da saturação:** $b = \frac{\alpha x}{\beta + x}$. Utilizada, por exemplo, para taxa de crescimento populacional sob condições limitantes.

- Modelo exponencial:

$$b = \alpha e^{\beta x} \rightarrow \ln b = \ln \alpha + \beta x$$

- Equação de potência simples:

$$b = \alpha x^{\beta} \rightarrow \log b = \log \alpha + \beta \log x.$$

- Equação de taxa de crescimento da saturação:

$$b = \frac{\alpha x}{\beta + x} \rightarrow \frac{1}{b} = \frac{1}{\alpha} + \frac{\beta}{\alpha} x.$$

Revisitando o exemplo da medição da força em um túnel de vento, utilizando o modelo:

$$b = \alpha x^\beta \rightarrow \log b = \log \alpha + \beta \log x.$$

i	x_i	b_i	$\log x_i$	$\log b_i$	$(\log x_i)^2$	$\log x_i \log b_i$
1	10	25	1,000	1,398	1,000	1,398
2	20	70	1,301	1,845	1,693	2,401
3	30	380	1,477	2,580	2,182	3,811
4	40	550	1,602	2,740	2,567	4,390
5	50	610	1,699	2,785	2,886	4,732
6	60	1.220	1,778	3,086	3,162	5,488
7	70	830	1,845	2,919	3,404	5,386
8	80	1.450	1,903	3,161	3,622	6,016
Σ			12,606	20,515	20,516	33,622

$$\beta_1 = \frac{\sum \log x_i \sum \log b_i - m \sum \log x_i \log b_i}{(\sum \log x_i)^2 - m \sum (\log x_i)^2} = 1,9842$$

$$\beta_0 = \frac{\sum \log b_i - \beta_1 \sum \log x_i}{m} = -0.5620$$

$$\log b = -0,5620 + 1,9842 \log x$$

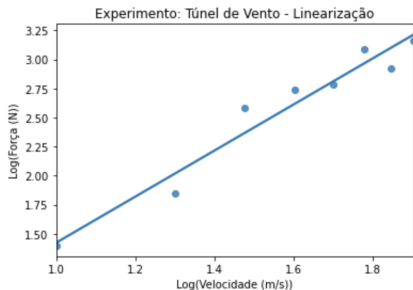
$$\log \alpha = -0,5620 \rightarrow \alpha = 10^{-0,5620} = 0,2741$$

$$\beta = 1,9842$$

Modelo final:

$$b = 0,2741x^{1,9842}$$

Reta ajustada:

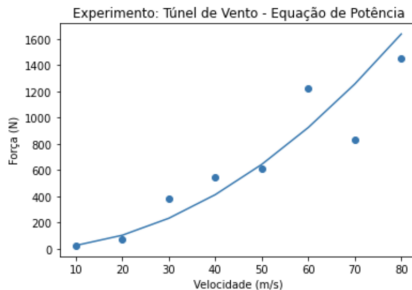


Intercepto: -0.5620318023820827

Coefficientes: $[1.98417626]$

Coefficiente de Determinação (R2 Score): 0.9480533472023586

Equação de potência parametrizada:



- $b = \alpha x^\beta \rightarrow \ln(b) = \ln(\alpha) + \beta \ln(x)$
- $b = \alpha \beta^x \rightarrow \ln(b) = \ln(\alpha) + \ln(\beta)x$
- $b = e^{\alpha + \beta x_1 + \gamma x_2} \rightarrow \ln(b) = \alpha + \beta x_1 + \gamma x_2$
- $b = \alpha x_1^\beta x_2^\gamma \rightarrow \ln(b) = \ln(\alpha) + \beta \ln(x_1) + \gamma \ln(x_2)$

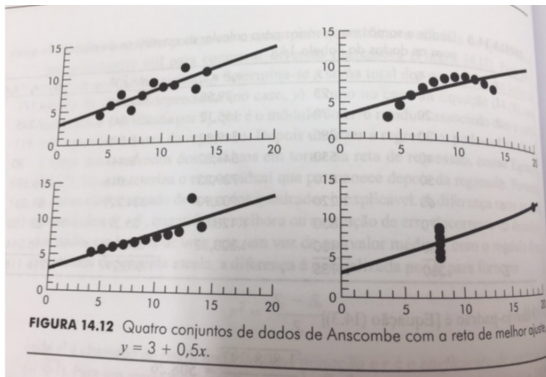
- O **Coeficiente de Determinação (R^2)** é uma métrica usada para quantificar a qualidade do ajuste obtido a partir da utilização de modelos de regressão linear.
- Para o seu cálculo, vamos considerar as seguintes medidas:
 - Média dos valores observados: $\bar{b} = \frac{1}{m} \sum_{i=1}^m b_i$.
 - Estimativas dos valores observados dadas pelo modelo: $\hat{b}_i$.
 - Soma do erro quadrático entre os valores observados e um modelo simples que usa a média dos valores observados (SQTot):
$$\sum_{i=1}^m (b_i - \bar{b})^2$$
 - Soma do resíduo quadrático entre os valores observados e o modelo de regressão (SQRes): $\sum_{i=1}^m (b_i - \hat{b}_i)^2$

O Coeficiente de Determinação (R^2) - é dado por:

$$R^2 = \frac{SQTot - SQRes}{SQTot} = 1 - \frac{SQRes}{SQTot}$$

- R^2 assume valores entre 0 e 1: bons modelos de ajuste possuem valores de R^2 próximos de 1.
- No entanto, em alguns casos, o valor de R^2 pode ser próximo de 1, mas a qualidade do ajuste pode não ser necessariamente bom.

$$R^2 = 0,67$$



Métodos Numéricos Aplicados com Matlab para Engenheiros e Cientistas - Steven C. Chapra

- Se a matriz A tem posto completo ($r(A) = n$), $A^T A$ é quadrada, simétrica positiva definida.
- O método numérico clássico para resolver este sistema de equações de dimensão n é a **fatoração de Cholesky**, com o custo computacional de $O(n^3)$.
- O sistema de equações normais se reduz a:

$$LL^T \hat{x} = A^T b,$$

onde L é uma matriz triangular inferior.

Algoritmo de Ajuste de Curvas via Equações Normais

- 1 Calcule a matriz $A^T A$ e o vetor $A^T b$.
- 2 Fatore a matriz $A^T A = LL^T$, usando a fatoração de Cholesky.
- 3 Resolva o sistema triangular inferior $Ly = A^T b$.
- 4 Resolva o sistema triangular superior $L^T x = y$.

- Ao considerarmos somente o custo computacional, resolver o problema de ajuste de curvas através de equações normais pode ser a melhor escolha.
- No entanto, na presença de erros de arredondamento, resolver o sistema de equações normais se torna um problema instável numericamente.
- Fatorações como a QR e SVD são numericamente mais estáveis.

Considere a seguinte matriz A obtida a partir de modelo de regressão polinomial de grau 3.

$$A = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 20 & 400 & 8.000 \\ 1 & 30 & 900 & 27.000 \\ 1 & 50 & 2.500 & 125.000 \end{bmatrix}$$

$$A^T A = \begin{bmatrix} 4 & 101 & 3.801 & 160.001 \\ 101 & 3.801 & 160.001 & 7.220.001 \\ 3.801 & 160.001 & 7.220.001 & 340.000.001 \\ 160.001 & 7.220.001 & 340.000.001 & 16.418.000.001 \end{bmatrix}$$

$$\kappa_2(A) = \frac{\sqrt{\lambda_{\max}(A^T A)}}{\sqrt{\lambda_{\min}(A^T A)}} = 150.412$$

$$\kappa_2(ATA) = \frac{\lambda_{\max}(A^T A)}{\lambda_{\min}(A^T A)} = 22.623.833.610$$