Regressão Linear Simples e Múltipla

Guilherme de Alencar Barreto

gbarreto@ufc.br

Grupo de Aprendizado de Máquinas – GRAMA
Departamento de Engenharia de Teleinformática
Universidade Federal do Ceará – UFC
http://lattes.cnpq.br/8902002461422112

Conteúdo dos Slides

1. Objetivo Geral
2. Regressão Linear Simples
3. Gráfico de Dispersão (*scatterplot*)
4. Regressão Linear por Partes
5. Regressão Linear Múltipla
6. Regressão Polinomial
7. Exemplos de Aplicação

Motivação

- Em muitas aplicações das engenharias e ciências, há duas ou mais variáveis que são intrinsicamente relacionadas, sendo necessário explorar a natureza desta relação.
- A análise de regressão abrange uma série de técnicas voltadas para a modelagem e a investigação de relações entre duas ou mais variáveis aleatórias.
- Por exemplo, sabe-se que um aerogerador é um equipamento que produz energia elétrica (P, em kW) em função da velocidade do vento (v, m/s).

Motivação (cont.-1)

- Podemos usar a análise de regressão para construir um modelo matemático que represente fidedignamente a relação determinística (i.e., de causa e efeito) entre P e v.
- Esse modelo pode ser usado, então, para predizer o valor da potência gerada para uma dada velocidade do vento.
- O modelo pode ser usado também para fins de detecção de falhas e monitoramento do equipamento.
- Porém, há diversos fatores de natureza aleatória que interferem no processo de modelagem do equipamento.

Definição do Problema

- Suponha que haja uma única variável de saída, y.
- Suponha também que a variável y está relacionada com k variáveis de entrada:

$$x_1, x_2, \ldots, x_k \tag{1}$$

- A variável y é também chamada de variável de resposta ou variável dependente.
- As variáveis x_j, $j = 1, \ldots, k$ são também chamadas de variáveis de entradas, variáveis regressoras ou ainda variáveis independentes.

Definição do Problema (cont.-1)

- Assume-se que a variável y é uma variável aleatória e que as variáveis x_j são medidas com erro desprezível.
- As variáveis x_j são freqüentemente controladas pelo experimentador (usuário).
- A relação entre y e x_j, $j = 1, \ldots, k$, é caracterizada por um modelo matemático chamado **equação de regressão**.
- A equação de regressão é ajustada a um conjunto de dados.
- Em algumas situações, o experimentador saberá a forma exata da verdadeira relação funcional $f(\cdot)$ entre y e x_j, $j = 1, \ldots, k$, representada como

$$y = f(x_1, x_2, \ldots, x_k).$$

Definição do Problema (cont.-2)

- No entanto, na maioria dos casos, a verdadeira relação funcional $f(\cdot)$ é desconhecida.
- Cabe ao experimentador escolher uma função apropriada para aproximar $f(\cdot)$.
- É comum usar um modelo polinomial como função aproximadora.
- Primeiramente, iremos tratar o caso em que há apenas uma variável de saída e uma de entrada (regressão simples).
- Em seguida, trataremos o caso em que há uma variável de saída e várias de entrada (regressão múltipla).

Parte I

Regressão Linear Simples

Objetivo

Desejamos determinar a relação entre uma única variável de entrada x e uma variável de saída y.

Suposições

- A variável x é uma variável matemática contínua, controlável pelo experimentador.
- A verdadeira relação entre x e y é definida por uma reta.
- O valor observado de y, para cada valor de x, é uma variável aleatória.
- Isto significa que y está sujeita a distorções oriundas de fenômenos aleatórios, os chamados *erros aleatórios*.

- Como supomos que y é uma variável aleatória, ela pode ser descrita pelo seguinte modelo:

$$y = \beta_0 + \beta_1 x + \varepsilon, \tag{2}$$

  em que β_0 (intercepto) e β_1 (inclinação) são constantes desconhecidas; e ε denota o ruído aleatório.
- Esta equação reflete nossas suposições sobre o processo gerador dos dados, a saber
  1. O processo tem uma componente determinística, que é linear: $f(x) = \beta_0 + \beta_1 x$.
  2. A componente determinística é contaminada aditivamente com ruído aleatório para gerar a saída observada: $y = f(x) + \varepsilon$.
  3. Assume-se, em geral, que o ruído aleatório é gaussiano com média zero e variância desconhecida σ_ε^2.

- O ruído ε é uma abstração matemática usada como modelo probabilístico das incertezas inerentes ao processo de medição.
- Como ε é uma variável aleatória de média zero, o valor esperado de y para cada valor de x é dado por

$$E[y|x] = \beta_0 + \beta_1 x. \tag{3}$$

- A Eq. (3) é determinística, já que não possui componentes estocásticas.
- Essa equação corresponde a uma curva "média", que caso conhecêssemos β_0 e β_1, seria o modelo determinístico exato do processo gerador dos dados.
- O problema é que, na prática, não conhecemos os coeficientes β_0 e β_1, nem a variância do ruído.

- Vamos supor que temos n pares de observações (medições) feitas com o equipamento adequado:

$$\{(x_1, y_1),\ (x_2, y_2),\ \ldots, (x_n, y_n)\} \tag{4}$$

- Estes dados devem obedecer à seguinte relação funcional:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, 2, \ldots, n \tag{5}$$

  em que também se assume que os valores $\{\varepsilon_i\}_{i=1}^{n}$ são descorrelacionados entre si, ou seja, $E[\varepsilon_i \varepsilon_j] = 0$, para $i \neq j$.
- Em outras palavras, o processo $\{\varepsilon_i\}_{i=1}^{n}$ não tem memória. Ou seja, o ruído em um instante de tempo não está relacionado estatisticamente com o ruído em outro instante qualquer.

- Os dados medidos serão usados para estimar os parâmetros desconhecidos β_0 e β_1 na Eq. (2), cujas estimativas são denotadas por $\hat{\beta}_0$ e $\hat{\beta}_1$.
- A equação de regressão ajustada aos dados passa a ser representada como
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \tag{6}$$
- Uma vez determinado $\hat{\beta}_0$ e $\hat{\beta}_1$ podemos usar a Eq. (6) para predizer o valor de y_i para qualquer valor de x_i.
- O erro de predição, visto que a Eq. (6) é determinística, é dado por
$$e_i = y_i - \hat{y}_i, \qquad i = 1, \ldots, n. \tag{7}$$

- A técnica de estimação a ser usada baseia na minimização da soma dos erros quadráticos, vulgarmente conhecida técnica dos mínimos quadrados ordinários (MQO).
- Assim, devemos encontrar os valores de $\hat{\beta}_0$ e $\hat{\beta}_1$ que minimizem a seguinte função objetivo:

$$J(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2. \quad (8)$$

Entendendo o Problema!

Minimizar da função-custo equivale a fazer com que a soma dos quadrados dos erros entre os valores medidos (observações) e a reta de regressão seja mínima!

- Para que as estimativas $\hat{\beta}_0$ e $\hat{\beta}_1$ minimizem $J(\hat{\beta}_0, \hat{\beta}_1)$, as seguintes condições devem ser satisfeitas:

$$\frac{\partial J(\hat{\beta}_0, \hat{\beta}_1)}{\partial \hat{\beta}_0} = -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \quad (9)$$

$$\frac{\partial J(\hat{\beta}_0, \hat{\beta}_1)}{\partial \hat{\beta}_1} = -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) x_i = 0 \quad (10)$$

- Simplificando as Eqs. (9) e (10) obtemos

$$n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \qquad (11)$$

$$\hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i x_i \qquad (12)$$

Entendendo o Problema!

As Eqs. (11) e (12) formam um sistema de equações lineares chamado de *equações normais* dos mínimos quadrados!

- A solução das equações normais (exercício proposto) é dada por

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{13}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} y_i x_i - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)\left(\sum_{i=1}^{n} x_i\right)}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2} \tag{14}$$

em que

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \quad \text{e} \quad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{15}$$

- A equação do estimador de β_1 pode ainda ser expressa como

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n x_i y_i - \frac{1}{n}\left(\sum_{i=1}^n x_i\right)\left(\sum_{i=1}^n y_i\right)}{\sum_{i=1}^n x_i^2 - \frac{1}{n}\left(\sum_{i=1}^n x_i\right)^2} \quad (16)$$

$$= \frac{\frac{\sum_{i=1}^n x_i y_i}{n} - \left(\frac{\sum_{i=1}^n x_i}{n}\right)\left(\frac{\sum_{i=1}^n y_i}{n}\right)}{\frac{\sum_{i=1}^n x_i^2}{n} - \left(\frac{\sum_{i=1}^n x_i}{n}\right)^2} \quad (17)$$

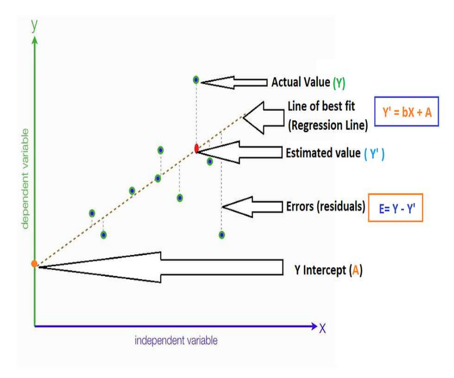
$$= \frac{\frac{\sum_{i=1}^n y_i x_i}{n} - \bar{x}\bar{y}}{\frac{\sum_{i=1}^n x_i^2}{n} - \bar{x}^2} = \frac{\hat{\sigma}_{xy}}{\hat{\sigma}_x^2} \quad (18)$$

em que $\hat{\sigma}_{xy}$ é a covariância amostral de x_i e y_i, e $\hat{\sigma}_x^2$ é a variância amostral de x_i.

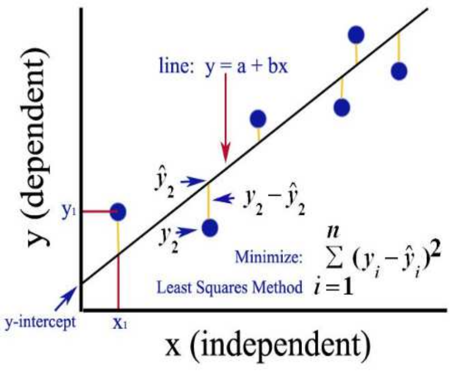
- Assim, as Eqs. (13) e (14) são os estimadores de MQ do intercepto (β_0) e da inclinação (β_1) respectivamente.

Exercício Desafio

Mostrar que $\hat{\beta}_0$ e $\hat{\beta}_1$ são estimadores não-polarizados de β_0 e β_1, respectivamente.



- Elementos que compõem o problema de regressão linear; ou seja, o ajuste de uma reta a um conjunto de pontos.



- O problema de estimação de mínimos quadrados dos parâmetros da reta pode ser entendido como um problema de posicionar uma reta tal que a soma dos valores quadráticos dos erros (segmentos verticais na figura) seja a menor possível.

Curiosidades sobre o Método dos Mínimos Quadrados

- Foi proposto em 1795 por **Carl Friedrich Gauss** (30/Abr/1777 - 23/Set/1855).

- Gauss aplicou o método no cálculo de órbitas de planetas e cometas a partir de medidas obtidas por telescópios.
- **Adrien Marie Legendre** (1752-1833) desenvolveu de forma independente o mesmo método e o publicou primeiro em 1806.

Interpolação e Extrapolação

- Se $x_i \in [x_{min}, x_{max}]$, em que $x_{min} = \min_{\forall i}\{x_i\}$ e $x_{max} = \max_{\forall i}\{x_i\}$, dizemos que o modelo realiza uma **interpolação**.
- Caso contrário, se $x_i \notin [x_{min}, x_{max}]$ dizemos que o modelo realiza uma **extrapolação**.

Observação Importante

Normalmente, a relação linear da Eq. (6) é considerada válida apenas para $x_i \in [x_{min}, x_{max}]$. Em outras palavras, modelos de regressão linear não costumam ser válidos para fins de extrapolação.

- Usualmente em regressão linear precisamos obter uma estimativa da variância do ruído (σ_ε^2).
- Essa estimativa é feita com base na diferença entre a observação y_i e o valor predito correspondente,

$$e_i = y_i - \hat{y}_i, \tag{19}$$

chamada de *erro de estimação* ou *resíduo*.
- A soma de quadrados dos resíduos é então dada por

$$SQ_E = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2. \tag{20}$$

- Pode-se mostrar (exercício) que uma estimativa não-polarizada de σ_ε^2 é dada por:

$$\hat{\sigma}_\varepsilon^2 = \frac{SQ_E}{n-2} = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n-2}. \qquad (21)$$

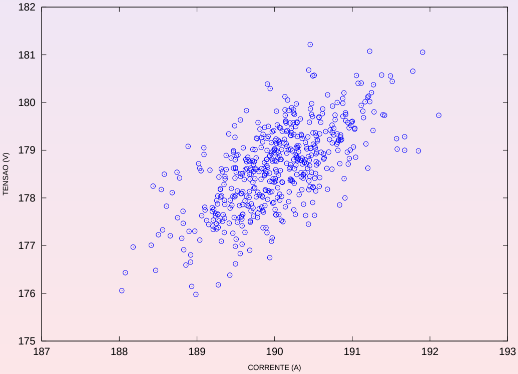
Questão Importante

Como saber se uma equação de regressão linear é a mais adequada para modelar os dados experimentais?

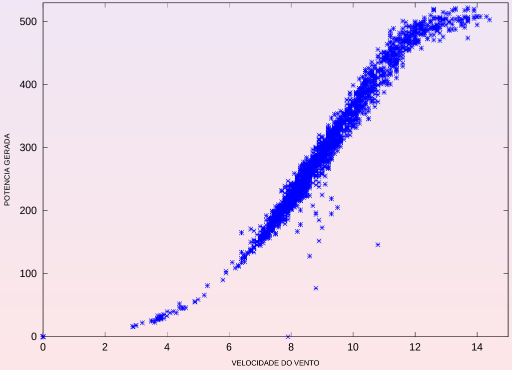
- Uma primeira abordagem é puramente visual, através do **gráfico de dispersão** (*scatterplot*).
- Este gráfico consiste em representar cada par (x_i, y_i), $i = 1, \ldots, n$, num sistema de coordenadas $x \times y$, com um ponto.
- Assumindo que os valores medidos de x e y estão dispostos, respectivamente, na primeira e segunda colunas da matriz de dados X basta usar o seguinte comando do Matlab/Octave:

```
>> plot(X(:,1), X(:,2),'*');
```

- Gráfico de dispersão para valores de x (corrente) e y (tensão) medidos em determinado equipamento elétrico ruidoso.



- Gráfico de dispersão para valores de x (velocidade do vento) e y (potência gerada) medidos em um aerogerador do parque eólico da Prainha.



- Para o primeiro gráfico de dispersão mostrado anteriormente, o modelo de regressão linear parece ser uma boa hipótese de modelagem dos dados.
- Já para o segundo gráfico de dispersão, o modelo de regressão linear não parece ser uma boa hipótese de modelagem.
- Para o segundo gráfico, um modelo polinomial de ordem maior que 1 parece ser o mais indicado.
- Mais adiante veremos como escolher um modelo mais adequado para o segundo conjunto de medidas usando regressão linear múltipla.

- Após averiguar pelo gráfico de dispersão se um modelo de regressão linear pode ser uma boa escolha, devemos estimar os parâmetros $\hat{\beta}_0$ e $\hat{\beta}_1$ da reta de regressão.
- Feito isto devemos, em seguida, calcular os resíduos $e_i = y_i - \hat{y}_i$ resultantes.
- Além de serem utilizados para estimar a variância do ruído (σ_ε^2), os resíduos são usados para validar a suposição de que os erros são gaussianos, de média zero e não-correlacionados, ou seja

$$\varepsilon_i \sim N(0, \sigma_\varepsilon^2) \quad \text{(Suposição 1)} \tag{22}$$

$$E[\varepsilon_i \varepsilon_j] = 0, \ \forall i \neq j \quad \text{(Suposição 2)} \tag{23}$$

Análise de Resíduos

(1) Construir um histograma de freqüência dos resíduos.

(2) Normalizar os resíduos, calculando-se

$$d_i = \frac{e_i}{\hat{\sigma}_\varepsilon}, \quad i = 1, \ldots, n$$

(3) Se os resíduos normalizados d_i forem $N(0,1)$, então aproximadamente 95% dos resíduos normalizados devem cair dentro do intervalo $(-2, +2)$.

(4) resíduos muito fora do intervalo $(-2, +2)$ podem indicar a presença de um *outlier*, isto, é uma observação atípica em relação ao resto dos dados.

Observações sobre Análise dos Resíduos

- O histograma dos resíduos deve ser semelhante ao esperado para dados com uma distribuição gaussiana. No Matlab, recomenda-se o uso do comando `histfit()` para facilitar a visualização da similaridade com a distribuição gaussiana.
- Alguns autores recomendam que observaçoes atípicas (*outliers*) sejam descartados.
- Outros autores acham que *outliers* fornecem informação importante sobre circunstâncias não-usuais (e.g. falhas), de interesse para o experimentador, e não devem ser descartados.

Definição - Coeficiente de Determinação

- O coeficiente de determinação é definido como

$$R^2 = 1 - \frac{SQ_E}{S_{yy}} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \qquad (24)$$

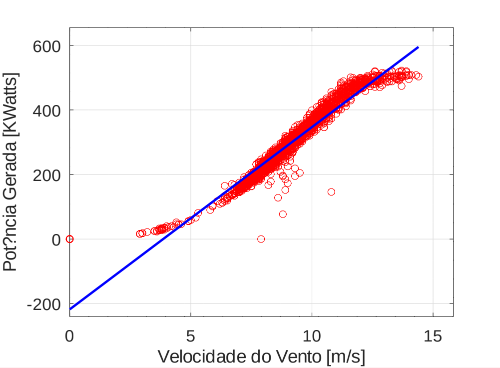
em que se nota, claramente, que $0 \leq R^2 \leq 1$.

- R^2 é usada para julgar a adequação de um modelo de regressão. Em princípio, quanto mais próximo R^2 está de 1, mais adequado é o modelo de regressão.

Entendendo Melhor

O coeficiente R^2 é entendido como a quantidade de variabilidade dos dados que o modelo de regressão é capaz de explicar.

- Qual seria reta de regressão que melhor modela os dados do aerogerador ($n = 2250$)?
- Encontramos que $\hat{\beta}_0 =$ -217,69, $\hat{\beta}_1 =$ 56,44 e $R^2 =$ 0,93.
- Apesar do alto valor de R^2, o modelo linear não é apropriado para este conjunto de dados.



Pergunta Importante

O que fazer então quando o modelo de regressão dado pela reta $y = \beta_0 + \beta_1 x + \varepsilon$ não é apropriado?

Algumas Respostas Plausíveis

- **Caso 1** - Aplicar uma transformação aos dados originais de modo a torná-los aproximadamente linear.
- **Caso 2** - Dividir o domínio original dos dados em sub-domínios, de tal modo que dentro de cada sub-domínio o modelo linear seja uma boa escolha.
- **Caso 3** - Utilizar um modelo de regressão polinomial de ordem maior que 1.

- Em algumas situações, uma função não-linear pode ser expressa através de uma reta, usando-se uma transformação adequada.
- Como exemplo, considere a função exponencial

$$y = \beta_0 e^{\beta_1 x} \varepsilon \tag{25}$$

- Esta função pode ser linearizada por uma transformação logarítmica

$$y^* = \ln y = \ln(\beta_0) + \beta_1 x + \ln(\varepsilon). \tag{26}$$

- Assume-se que os erros, $\ln(\varepsilon)$, sejam distribuídos normal e independentemente, com média 0 e variância σ_ε^2.

- Na função anterior aplicamos a transformação aos dados de saída originais y, obtendo dados transformados y^*. Fazemos então o gráfico de dispersão de $y^* \times x$.
- Outra função que pode ser linearizada por uma simples transformação (em x) é

$$y = \beta_0 + \beta_1 \left(\frac{1}{x} \right) + \varepsilon \quad (27)$$

- Usando a transformação recíproca $x^* = 1/x$, o modelo se lineariza em

$$y = \beta_0 + \beta_1 x^* + \varepsilon. \quad (28)$$

- O gráfico de dispersão $y \times z$ indicará uma relação linear.

- Algumas vezes, várias transformações podem ser empregadas conjuntamente para linearizar uma função.
- Por exemplo, considere a função

$$y = \frac{1}{\exp\{\beta_0 + \beta_1 x + \varepsilon\}} \tag{29}$$

- Fazendo $y^* = 1/y$, temos a forma linearizada da função como

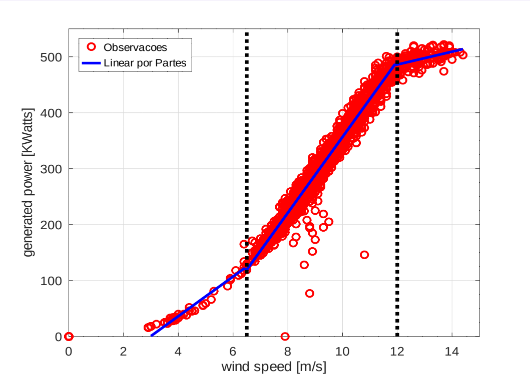
$$\ln(y^*) = \beta_0 + \beta_1 x + \varepsilon \tag{30}$$

- Considere os dados do aerogerador.

Outra Pergunta Desafio

Você consegue dividir o gráfico de dispersão em duas ou mais sub-regiões em que modelos de regressão linear sejam adequados?

- Exemplo de modelo de regressão linear por partes.



- R1: $x \in [3 \text{ a } 6,5]$, R2: $x \in [6,53 \text{ a } 12]$ e R3: $x \in [123 \text{ a } 15]$.

Exercício Proposto

Determinar a reta de regressão associada a cada uma das regiões R1, R2 e R3. Ou seja, determinar

- R1: $\hat{y} = -105,57 + 35,45x$
- R2: $\hat{y} = -321,90 + 67,86x$
- R3: $\hat{y} = 347,68 + 11,53x$

Parte II

Regressão Linear Múltipla

- Muitos problemas de regressão envolvem mais de uma variável regressora.
- Tais modelos são chamados de *modelos de regressão múltipla*.
- Em geral, a variável de saída ou resposta, y, pode ser relacionada a k variáveis de entrada.
- O modelo

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon \qquad (31)$$

é chamado de modelo de regressão linear múltipla com k variáveis de entrada.

- Os parâmetros β_j, $j = 0, 1, \ldots, k$, são chamados de coeficientes de regressão.
- O modelo da Eq. (31) descreve um hiperplano no espaço k-dimensional das variáveis de entrada $\{x_j\}$.

Conceito Importante!

O parâmetro β_j representa a mudança esperada na resposta y por unidade de mudança em x_j, quando todas as demais variáveis independentes x_i $(i \neq j)$ são mantidas constantes.

- Modelos de regressão linear múltipla são usados, em geral, como *funções aproximadoras* ou *interpoladoras*.
- Ou seja, a verdadeira relação funcional entre y e x_1, x_2, $\ldots$, x_k é desconhecida, mas dentro de certos limites das variáveis de entrada o modelo de regressão linear é uma aproximação adequada.
- Modelos mais complexos que o da Eq. (31) também podem ser analisados pelas técnicas de regressão linear múltipla.

- Por exemplo, considere o modelo de regressão linear múltipla com três variáveis de entrada:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon. \tag{32}$$

- Se fizermos $x_1 = x$, $x_2 = x^2$ e $x_3 = x^3$, então o modelo da Eq. (32) pode ser escrito como um modelo não-linear (no caso, polinomial cúbico) em uma variável de entrada:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \varepsilon. \tag{33}$$

- O método dos mínimos quadrados pode ser usado para estimar os coeficientes de regressão $\{\beta_j\}$, $j = 0, 1, \ldots, k$.
- Para isso, faremos as seguintes definições:
  1. y_i é a i-ésima observação (medida) da variavel de saída.
  2. x_{ij} é i-ésima observação da variável x_j.
- As seguintes suposições são também necessárias:
  1. Estão disponíveis $n > k$ observações (i.e., há mais equações do que incógnitas).
  2. O erro ou ruído no modelo (ε) tem média 0, variância σ_ε^2.
  3. As observações $\{\varepsilon_i\}$ são não-correlacionadas.

- Feito isto, podemos escrever o modelo da Eq. (31) em termos das observações:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i \tag{34}$$

para $i = 1, 2, \ldots, n$.

- Isto equivale a ter o seguinte sistema com n equações e $k+1$ incógnitas:

$$\begin{aligned}
y_1 &= \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \cdots + \beta_k x_{1k} + \varepsilon_1 \\
y_2 &= \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \cdots + \beta_k x_{2k} + \varepsilon_2 \\
\vdots & \quad\;\; \vdots \qquad\quad \vdots \\
y_n &= \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \cdots + \beta_k x_{nk} + \varepsilon_n
\end{aligned} \tag{35}$$

- Em forma matricial, o sistema de equações em (35) é escrito

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{36}$$

em que

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}_{n\times 1}, \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}_{n\times(k+1)},$$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}_{(k+1)\times 1} \quad \text{e} \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}_{n\times 1}.$$

- Deseja-se encontrar o vetor de estimativas dos quadrados mínimos, $\hat{\boldsymbol{\beta}}$, que minimize a seguinte função-custo:

$$J_{MQO}(\boldsymbol{\beta}) = \|\boldsymbol{\varepsilon}\|^2 = \boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon} = \sum_{i=1}^{n} \varepsilon_i^2 = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}). \tag{37}$$

- A função-custo $J(\boldsymbol{\beta})$ pode ser entendida como uma função que busca encontrar o vetor de parâmetros $\hat{\boldsymbol{\beta}}$ que produz o vetor $\boldsymbol{\varepsilon}$ de menor norma quadrática.
- A Eq. (37) pode ser decomposta em

$$\begin{aligned} J_{MQO}(\boldsymbol{\beta}) &= \mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{X}\boldsymbol{\beta} - \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{y} + \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} \qquad (38) \\ &= \mathbf{y}^T\mathbf{y} - 2\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{y} + \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} \end{aligned}$$

uma vez que $\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{y} = \mathbf{y}^T\mathbf{X}\boldsymbol{\beta}$ resulta no mesmo escalar.

- As estimativas de quadrados mínimos devem satisfazer

$$\frac{\partial J_{MQO}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -2\mathbf{X}^T\mathbf{y} + 2\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} = \mathbf{0}, \tag{39}$$

em que $\mathbf{0}$ é um vetor de zeros.

- Simplificando a Eq. (39) resulta em

$$\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} = \mathbf{X}^T\mathbf{y}. \tag{40}$$

- A Eq. (40) define as *equações normais* dos quadrados mínimos da regressão linear múltipla.

- Note que a matriz $\mathbf{X}^T\mathbf{X}$ é quadrada (dim=$(k+1) \times (k+1)$).
- Para resolver as equações normais basta multiplicar ambos os lados da Eq. (40) pela inversa de $\mathbf{X}^T\mathbf{X}$.
- Assim, a estimativa de quadrados mínimos ordinários (MQO) de $\boldsymbol{\beta}$ é dada por

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}. \tag{41}$$

- Portanto, o modelo de regressão ajustado (preditor) é definido como

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}. \tag{42}$$

- O vetor de erros de predição (resíduos) é denotado por

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}. \tag{43}$$

O Problema da Multicolinearidade

- Muitas vezes, a matriz $\mathbf{X}^T\mathbf{X}$ é é singular (ou quase!), ou seja
$$\det(\mathbf{X}^T\mathbf{X}) \approx 0$$
- Isso certamente causará problemas numéricos durante a inversão desta matriz.
- Isto ocorre geralmente quando as colunas (ou linhas) da matriz $\mathbf{X}^T\mathbf{X}$ são linearmente dependentes.
- Neste caso, dizemos que existe *multicolinearidade* em $\mathbf{X}^T\mathbf{X}$.

Entendendo a Multicolinearidade de $\mathbf{X}^T\mathbf{X}$

- Foi mencionado que a multicolinearidade em $\mathbf{X}^T\mathbf{X}$ causa a singularidade desta matriz.
- Dos livros didáticos de Álgebra Linear, a singularidade de uma matriz; ou seja, a não existência de sua inversa, pode ser verificada pelo cálculo do seu determinante.
- Se o determinante é nulo, então a matriz não possui inversa.
- Contudo, este é um procedimento muito limitado (veremos adiante o porquê) e sua interpretabilidade não é direta e útil para fins de construção do modelo de regressão.

Entendendo a Multicolinearidade de $\mathbf{X}^T\mathbf{X}$

- Não basta que o determinante da matriz seja diferente de zero, este valor tem que ser BEM diferente de zero.
- A ênfase em ser *bem diferente de zero* está relacionada ao conceito de condicionamento de uma matriz.
- A qualidade da matriz inversa está diretamente associada ao condicionamento da matriz original.
- A saber, o número de condicionamento de uma matriz **A** depende dos seus autovalores, sendo calculado como

$$c(\mathbf{A}) = \frac{|\lambda_{max}(\mathbf{A})|}{|\lambda_{min}(\mathbf{A})|}, \tag{44}$$

em que $|\cdot|$ denota o valor absoluto.

Entendendo a Multicolinearidade de $\mathbf{X}^T\mathbf{X}$

- Idealmente, o número de condicionamento deve estar próximo de 1.
- Valores muito elevados indicam uma matriz mal-condicionada.
- Muitas vezes, é melhor usar o recíproco do número de condicionamento, $r(\mathbf{A})$.
- Neste caso, se a matriz $\mathbf{A}$ é bem-condicionada, então $r(\mathbf{A})$ será próximo de 1. Se a matriz $\mathbf{A}$ é mal-condicionada, então $r(\mathbf{A})$ será próximo de zero.
- Pra finalizar, em vez do determinante de uma matriz, recomenda-se o cálculo do *posto* e do número de condicionamento para averiguar a sua invertibilidade.

Entendendo a Multicolinearidade de $\mathbf{X}^T\mathbf{X}$

- O posto (do inglês *rank*) de uma matriz qualquer $\mathbf{A}$, de dimensões $n \times m$, é dado por

$$posto(\mathbf{A}) \leq \min(n, m). \tag{45}$$

- Para uma matriz quadrada de dimensões $n \times n$, o posto reduz-se à seguinte expressão:

$$posto(\mathbf{A}) \leq n. \tag{46}$$

- O posto tem interpretação simples e direta: É o número de linhas/colunas linearmente independentes; ou seja, que não podem ser escritas como combinação linear uma das outras.

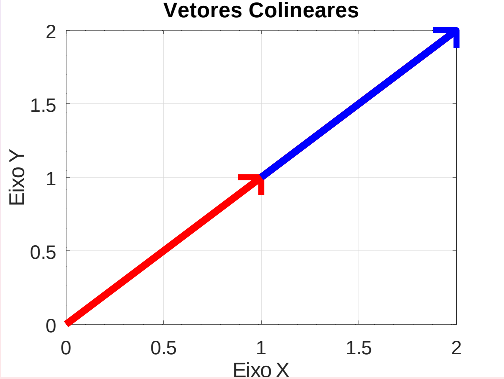
- A título de ilustração, considere a seguinte matriz 2×2:

$$\mathbf{A} = \begin{bmatrix} 1 & 2 \\ 1 & 2 \end{bmatrix} = [\ \mathbf{c}_1 \mid \mathbf{c}_2\ ] \tag{47}$$

- Por ser uma matriz simples, nota-se de cara que o determinante dessa matriz é nulo; logo, ela não possui inversa.
- Mas por que a matriz acima não possui inversa? Não é devido ao determinante nulo, mas sim uma consequência da colinearidade entre, por exemplo, as colunas da matriz $\mathbf{A}$. Pode-se tomar as linhas também.
- Percebe-se facilmente que a 2a. coluna dessa matriz é um múltiplo da 1a. coluna; e vice-versa.
- Se chamarmos a 1a. coluna de $\mathbf{c}_1$ e a 2a. coluna de $\mathbf{c}_2$, tem-se que

$$\mathbf{c}_2 = 2\mathbf{c}_1 \tag{48}$$

- Geometricamente, a colinearidade dos vetores $\mathbf{c}_1$ e $\mathbf{c}_2$ pode ser visualizada na figura abaixo.
- Os vetores em questão estão sobre a mesma reta suporte; ou seja, são múltiplos entre si e, portanto, são linearmente dependentes.



- Considere agora a seguinte matriz 2×2:

$$\mathbf{A} = \begin{bmatrix} 1 & 2 \\ 1 & 2,00000001 \end{bmatrix} = [\ \mathbf{c}_1 \mid \mathbf{c}_2\ ] \tag{49}$$

- O determinante dessa matriz é muito pequeno, mas não é nulo; logo, o software de programação vai entender que essa matriz possui inversa.
- No Octave, usando o comando **inv**, a matriz resultante é mostrada abaixo.

```
>> inv(A)
ans =

     200000002.2154942      -200000001.2154942
    -100000000.6077471       100000000.6077471
```

- A matriz inversa resultante possui com componentes com valores muito altos.
- Ou seja, um errinho numérico lá na 8a. casa decimal, transformou uma matriz singular em uma matriz invertível.

- Por isso, a análise da invertibilidade não pode se fiar apenas no valor do determinante ou mesmo do posto.
- O posto da matriz mostrada na Eq. 47 é $r(\mathbf{A}) = 1$. Já o posto da matriz da Eq. (49) é $r(\mathbf{A}) = 2$, confirmando que ela é invertível.
- Porém, a matriz da Eq. (49) é mal-condicionada, conforme pode-se ver pelas manitudes dos seus autovalores: $\lambda_{max}(\mathbf{A}) \approx 3,0$ e $\lambda_{min}(\mathbf{A}) \approx 3,33 \times 10^{-9}$.
- O número de condicionamento é $c(\mathbf{A}) \approx 1 \times 10^{9}$, indicando que é uma matriz extremamente mal-condicionada.

Regularização de Thikonov

- Os efeitos nocivos da multicolinearidade podem ser minimizados reescrevendo a Eq. (41) como

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\mathbf{X}^T\mathbf{y}. \tag{50}$$

  em que
  - $0 \leq \lambda \ll 1$ é uma constante de valor pequeno.
  - $\mathbf{I}$ é uma matriz identidade de dimensão $(k+1) \times (k+1)$.
- O estimador da Eq. (50) é chamado de **mínimos quadrados regularizado** (MQR), enquanto a regressão que a utiliza é chamada de **regressão de cumeeira** (*ridge regression*).

Exercício Teórico

- Mostrar que a Eq. (50) pode ser obtida a partir da seguinte função-custo:

$$J_{MQR}(\boldsymbol{\beta}) = \|\boldsymbol{\varepsilon}\|^2 + \lambda\|\boldsymbol{\beta}\|^2. \tag{51}$$

- A função-custo da Eq. (37) é interpretada como aquela que provê estimativas dos coeficientes de regressão $\{\beta_j\}$, $j = 1, 2, \ldots, k$, que resultam na menor soma dos quadrados dos erros de estimação $\{e_i\}$, $i = 1, 2, \ldots, n$.

- Que interpretação pode ser dada à função-perda da Eq.(51)?

Regularização

- Esta função de perda leva a uma solução de compromisso (*trade-off*) ao exigir a minimização conjunta de dois termos:
  - O primeiro termo, $\|\mathbf{e}\|^2$, favorece soluções para o vetor $\boldsymbol{\beta}$ que produzem a menor norma quadrática possível do vetor de erros.
  - O segundo termo, $\|\boldsymbol{\beta}\|^2$, favorece soluções para o vetor $\boldsymbol{\beta}$ que tenham menor norma possível.
- Vetores-solução $\boldsymbol{\beta}$ de menor norma produzem coeficientes de menor magnitude, o que é interessante para evitar situações de amplificação de entradas com ruído, principalmente outliers, evitando assim saídas de valor elevado.

- Assim, a Eq. (51) busca um vetor-solução $\boldsymbol{\beta}$ que minimize $\|\mathbf{e}\|^2$ e que ao mesmo tempo tenha norma mínima. Esta função objetivo pode ser interpretada como o lagrangiano de um problema de otimização com restrições de igualdade:

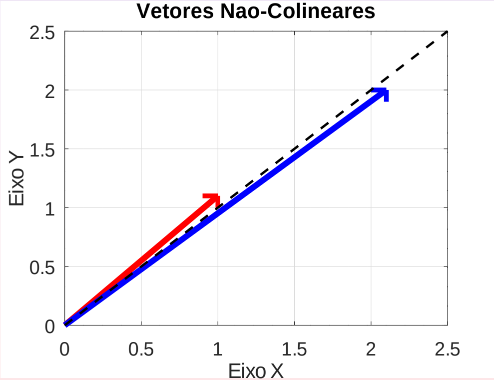
$$\begin{aligned} \text{Minimizar} \qquad & J_{MQO}(\boldsymbol{\beta}) = \frac{1}{2}\|\mathbf{e}\|^2 \\ \text{sujeito a} \qquad & \|\boldsymbol{\beta}\|^2 = c \end{aligned} \tag{52}$$

onde c é uma constante. O lagrangiano é dado por

$$J_{MQR}(\boldsymbol{\beta}) = \frac{1}{2}\|\mathbf{e}\|^2 + \lambda(\|\boldsymbol{\beta}\|^2 - c), \tag{53}$$

em que λ é justamente o multiplicador de Lagrange da restrição imposta à norma de $\boldsymbol{\beta}$.

- Geometricamente, o efeito da regularização de Tikonov sobre a invertibilidade da matriz $\mathbf{X}^T\mathbf{X}$ é de "quebrar" a colinearidade dos vetores-colunas dessa matriz, desalinhando-os; ou seja, tornando-os linearmente independentes. A figura abaixo ilustra este efeito.



- De posse da matriz $\mathbf{X}$ e do vetor $\mathbf{y}$ no Octave/Matlab, há algumas alternativas para obtenção da estimativa do vetor de parâmetros $\hat{\boldsymbol{\beta}}$.
    - Implementação direta da fórmula do estimador da Eq. (41).

      ```
      » Bhat=inv(X'*X)*X'*y
      ```

    - Utilização do comando `pinv`, que usa decomposição em valores singulares (SVD) para inverter a matriz $\mathbf{X}^T\mathbf{X}$.

      ```
      » Bhat=pinv(X)*y
      ```

    - Utilização do operador `barra invertida`, que usa diferentes métodos (e.g., decomposição de Cholesky, eliminação de Gauss com pivoteamento parcial, etc.) a depender de características da matriz $\mathbf{X}^T\mathbf{X}$.

      ```
      » Bhat=(X'*X)\(X'*y)
      ```

- Das três opções, recomenda-se apenas o uso da última por apresentar melhores características de escalabilidade e robustez numérica, sem ter que inverter explicitamente a matriz $\mathbf{X}^T\mathbf{X}$.

- As versões regularizadas para obtenção da estimativa do vetor de parâmetros $\hat{\boldsymbol{\beta}}$ são mostradas a seguir.
  - Implementação direta da fórmula do estimador MQR da Eq. (50).

```
» lamb=0.01; % parametro de regularizacao
» Bhat=inv(X'*X + lamb*eye(k+1))*X'*y
```

  - Utilizando o operador `barra invertida`.

```
» Bhat=(X'*X+ lamb*eye(k+1))\(X'*y)
```

Coeficiente de Determinação na Regressão Múltipla

- O coeficiente de determinação R^2 também é usado na regressão múltipla como medida de adequação do modelo:

$$R^2 = 1 - \frac{SQ_E}{S_{yy}} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \tag{54}$$

  em que $0 \leq R^2 \leq 1$.
- No entanto, um valor alta de R^2 não implica que o modelo seja bom!
- O acréscimo de uma variável ao modelo causará sempre, um aumento em R^2, independentemente de a variável adicional ser ou não significante (informativa).

Coeficiente de Determinação Ajustado

- Alguns autores preferem usar o *coeficiente de determinação R^2 ajustado* (R^2_{aj}):

$$R^2_{aj} = 1 - \frac{SQ_E/(n-p)}{S_{yy}/(n-1)}, \tag{55}$$

em que $p = k + 1$.

- O valor $S_{yy}/(n-1)$ será constante, independente do número de variáveis no modelo.
- O valor $SQ_E/(n-p)$ é a média quadrática para o erro, que mudará com o acréscimo (ou retirada) de variáveis ao modelo.

Coeficiente de Determinação Ajustado

- Se forem incluídas variáveis, então p cresce e $(n-p)$ diminui.
- Se a inclusão das novas variáveis não diminuir SQ_E significativamente, tem-se que

$$\frac{SQ_E}{(n-p)} \text{ aumenta, logo } R^2_{aj} \text{ diminui.} \tag{56}$$

- Portanto, R^2_{aj} cresce apenas se a adição de um novo termo reduzir significantemente a média quadrática dos erros.

Critério de Informação de Akaike

- Outro modo comum de penalizar a adição de termos ao modelo é através do critério de informação de Akaike (AIC):

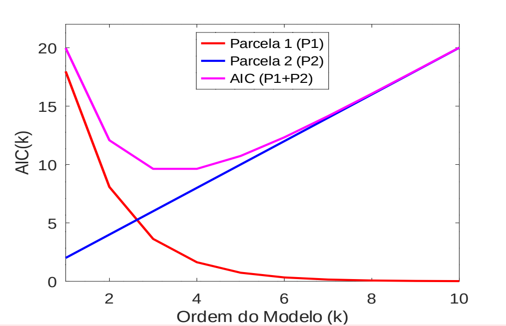
$$AIC(k) = N \ln [SQ_E(k)] + 2k, \quad k = 1, 2, \ldots \tag{57}$$

  em que k é o número de parâmetros do modelo.

- Para vários valores de k testados, escolhe-se aquele que produzir o menor $AIC(k)$:

$$k^{otimo} = \arg \min_{\forall k} \{AIC(k)\}.$$

- O primeiro termo do lado direito da Eq. (57) sempre decresce (exponencialmente) com o aumento de k.
- O segundo termo do lado direito da Eq. (57) sempre cresce (linearmente) com o aumento de k.
- A soma dos dois termos gera uma função convexa, cujo o mínimo revela o valor adequado de k.



- O modelo linear $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ é um modelo geral que pode ser usado para ajustar qualquer relação que seja *linear nos parâmetros* desconhecidos $\boldsymbol{\beta}$.
- Isso inclui a importante classe dos modelos de regressão polinomial. Por exemplo, vimos que o modelo polinomial cúbico em uma variável de entrada:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \varepsilon$$

  é um tipo de modelo de regressão múltipla se fizermos $x_1 = x$, $x_2 = x^2$ e $x_3 = x^3$.
- Modelos de regressão polinomial são amplamente usados nos casos em que a relação entre a variável de saída e de entrada é curvilínea (i.e. não-linear).

- Em regressão polinomial, a matriz $\mathbf{X}$ do modelo linear $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ passa ser definida como

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^k \\ 1 & x_2 & x_2^2 & \cdots & x_2^k \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^k \end{bmatrix}_{n\times(k+1)}$$

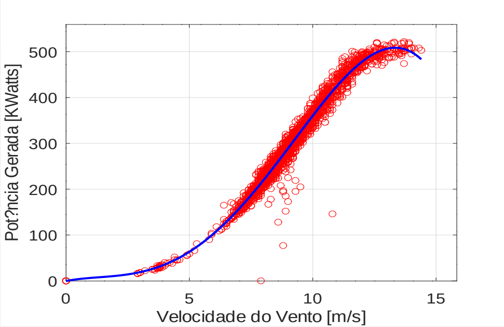
  em que x_i é a i-ésima observação da variável de entrada.

- A estimativa de quadrados mínimos $\hat{\boldsymbol{\beta}}$ é então calculada por meio da Eq. (41).
- Predições de novos valores podem ser feitas por meio da Eq. (42) e resíduos são calculados pela Eq. (50).

- Usando os dados do aerogerador ajustou-se o seguinte modelo polinomial de quarta ordem ($k = 4$):

$$\hat{y} = -0.391 + 10.37x - 5.00x^2 + 1.43x^3 - 0.068x^4$$

com $R^2 = 0.974$. A curva do modelo superposto ao gráfico de dispersão é mostrada abaixo.



Exercício Computacional - Questão 1

Q1 - Usando os dados do aerogerador, pede-se:

(1) Ajustar um modelo linear simples (reta) aos dados, determinando os valores de $\hat{\beta}_0$, $\hat{\beta}_1$ e $\hat{\sigma}_\varepsilon$.

(2) Plotar o histograma dos resíduos normalizados e verificar a porcentagem de valores que caem no intervalo [-2,+2].

Exercício Computacional - Questão 2

Q2 - Determinar a ordem mais adequada para o modelo polinomial para os dados do aerogerador usando as quantidades R^2_{aj} e $AIC(k)$.

- Mostrar resultados em uma tabela os valores de $R^2_{aj}(k)$ e $AIC(k)$ para $k = 1, 2, \ldots, 15$.
- Mostrar em um gráfico os valores de $R^2_{aj}(k) \times k$ e $AIC(k) \times k$ para $k = 1, 2, \ldots, 15$.
- Para o valor escolhido de p, pede-se:
  (1) Ajustar o modelo polinomial escolhido aos dados, determinando os valores de $\hat{\beta}_j$, $j = 0, 1, \ldots, p$.
  (2) Plotar o histograma dos resíduos normalizados e verificar a porcentagem de valores que caem no intervalo [-2,+2].