Regressão Linear Simples e Múltipla

Guilherme de Alencar Barreto

gbarreto@ufc.br

Grupo de Aprendizado de Máquinas – GRAMA
Programa de Pós-Graduação em Engenharia de Teleinformática
Universidade Federal do Ceará – UFC
http://www.deti.ufc.br/∼guilherme

"O processamento de sinais mudou! Não estamos mais na era na qual a informação na forma de sinais elétricos é processada por meio de tradicionais dispositivos analógicos. Nós estamos solida e, para o futuro previsível, irrevogavelmente, no âmago do processamento de sinais digitais (amostrados ou discretos no tempo) aleatórios."

Charles W. Therrien, 1992
Discrete Random Signals and Statistical Signal Processing

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 da ETI 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 $P = f(v)$, em que $f(\cdot)$ define a relação funcional 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 otimização e controle 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, $j = 1, ..., k$ 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, $j = 1, ..., k$, 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, $j = 1, ..., k$, 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 $f(\cdot)$.
- Normalmente usa-se 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.

- Daí, o valor esperado de y para cada valor de x é dado por

$$E[y|x] = \beta_0 + \beta_1 x, \tag{2}$$

em que β_0 (intercepto) e β_1 (inclinação) são constantes desconhecidas.

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

$$y = \beta_0 + \beta_1 x + \varepsilon, \tag{3}$$

em que ε é um erro (ruído) aleatório com média zero e variância σ_ε^2.

- 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:

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

em que assume-se que os valores $\{\varepsilon_i\}$ sejam variáveis aleatórias não-correlacionadas.

- Os dados medidos serão usados para estimar os parâmetros desconhecidos β_0 e β_1 na Eq. (3).
- A técnica de estimação a ser usada é a dos mínimos quadrados (MQ). Ou seja, devemos encontrar os valores de β_0 e β_1 que minimizem a seguinte função-custo:

$$J(\beta_0, \beta_1) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2. \qquad (6)$$

Entendendo o Problema!

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

- As estimativas de MQ de β_0 e β_1, denotados por $\hat{\beta}_0$ e $\hat{\beta}_1$, devem satisfazer

$$\frac{\partial J(\beta_0, \beta_1)}{\partial \beta_0} = -2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \beta_1 x_i) = 0 \quad (7)$$

$$\frac{\partial J(\beta_0, \beta_1)}{\partial \beta_1} = -2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \beta_1 x_i)x_i = 0 \quad (8)$$

- Simplificando as Eqs. (7) e (8) obtemos

$$n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \tag{9}$$

$$\hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i x_i \tag{10}$$

Entendendo o Problema!

As Eqs. (9) e (10) são chamadas *Equações Normais* dos mínimos quadrados!

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

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{11}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} y_i x_i - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)\left(\sum_{i=1}^{n} x_i\right)}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2} \tag{12}$$

em que

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \quad \text{e} \quad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{13}$$

- Assim, as Eqs. (11) e (12) são os estimadores de MQ do intercepto (β_0) e da inclinação (β_1) respectivamente.
- A equação de regressão ajustada aos dados passa a sere representada como:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \tag{14}$$

- Uma vez determinado $\hat{\beta}_0$ e $\hat{\beta}_1$ podemos usar a Eq. (14) para predizer o valor de y_i para qualquer valor de x_i.

Exercício Desafio

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

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. (14) 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,

$$e_i = y_i - \hat{y}_i, \tag{15}$$

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

$$SQ_E = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2. \tag{16}$$

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

$$\hat{\sigma}_\varepsilon^2 = \frac{SQ_E}{n-2} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}. \tag{17}$$

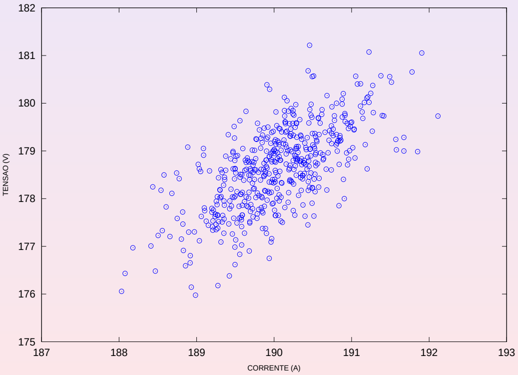
Questão Importante

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

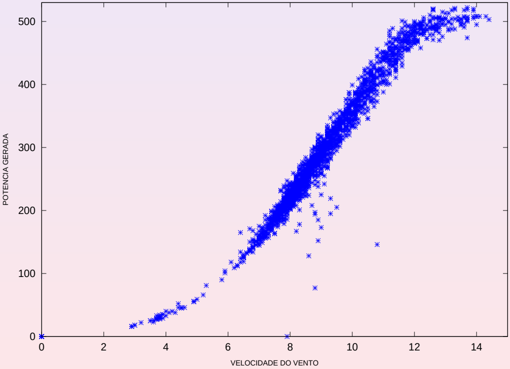
- Uma primeira abordagem é puramente visual, através do **gráfico de dispersão** (*scatterplot*).
- Este gráfico consiste em representar cada par (x_i, y_i), $i = 1, ..., n$, num sistema de coordenadas $x \times y$, com um ponto.
- Assumindo que os valores medidos de x e y estão dispostos, respectivamente, 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

$$\varepsilon_i \sim N(0, \sigma_\varepsilon^2) \quad \text{(Suposição 1)} \tag{18}$$

$$E[\varepsilon_i \varepsilon_j] = 0, \ \forall i \neq j \quad \text{(Suposição 2)} \tag{19}$$

Análise de Resíduos

(1) Construir um histograma de freqüência dos resíduos.

(2) Normalizar os resíduos, calculando-se

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

(3) Se os erros e_i forem $N(0, \sigma_\varepsilon^2)$, 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

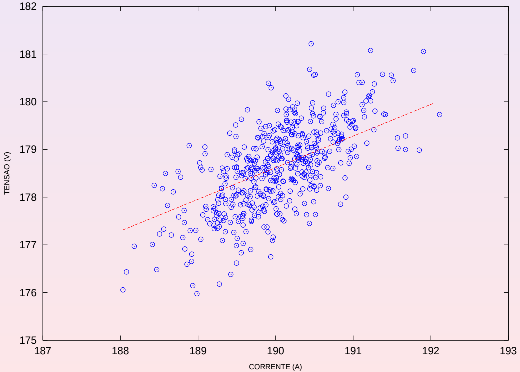
$$R^2 = 1 - \frac{SQ_E}{S_{yy}} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \tag{20}$$

  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.

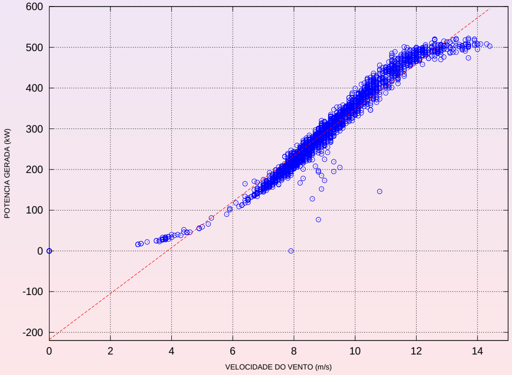
Entendendo Melhor

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

- Considere o gráfico de dispersão é mostrado abaixo ($n = 500$). Encontrar a reta de regressão correspondente.
- Encontramos que $\hat{\beta}_0 = 8{,}51$, $\hat{\beta}_1 = 0{,}90$ e $R^2 = 0{,}44$.



- 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.



Pergunta Importante

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

Possível Resposta

- Desistir de tudo e procurar outro emprego?

Pergunta Importante

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

Respostas Mais 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

$$y = \beta_0 e^{\beta_1 x} \varepsilon \tag{21}$$

- Esta função pode ser linearizada por uma transformação logarítmica

$$y^* = \ln y = \ln(\beta_0) + \beta_1 x + \ln(\varepsilon). \tag{22}$$

- 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

$$y = \beta_0 + \beta_1 \left(\frac{1}{x} \right) + \varepsilon \tag{23}$$

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

$$y = \beta_0 + \beta_1 x^* + \varepsilon. \tag{24}$$

- 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

$$y = \frac{1}{\exp\{\beta_0 + \beta_1 x + \varepsilon\}} \tag{25}$$

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

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

Pergunta Desafio

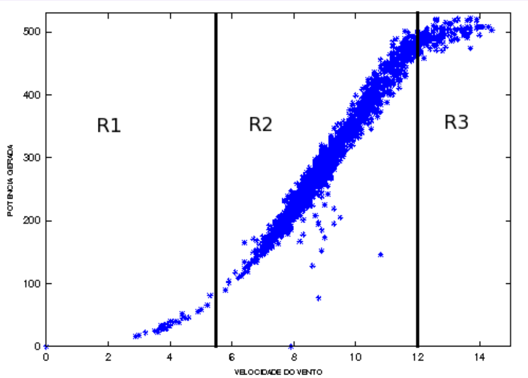
É possível "linearizar" os dados do aerogerador usando uma ou várias das transformações descritas anteriormente?

- 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?

- Que tal a divisão sugerida abaixo?



- R1: $x \in [0 - 5,5]$, R2: $x \in [5,5 - 12]$ e R3: $x \in [12 - 15]$.

Exercício Proposto

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

- R1: $\hat{y} = \hat{\beta}_0^{(1)} + \hat{\beta}_1^{(1)} x$
- R2: $\hat{y} = \hat{\beta}_0^{(2)} + \hat{\beta}_1^{(2)} x$
- R3: $\hat{y} = \hat{\beta}_0^{(3)} + \hat{\beta}_1^{(3)} x$

em que $\hat{\beta}_0^{(i)}$ e $\hat{\beta}_1^{(i)}$ definem o intercepto e a inclinação da i-ésima reta de regressão, $i = 1, 2$ e 3.

- Finalmente, para tratar o Caso 3, devemos lembrar que uma reta é um polinmio de ordem 1.
- Para tratar dados cujo gráfico de dispersão revela uma relação não-linear entre variáveis de entrada e de saída, é comum o uso de modelos polinomiais de ordem maior que 1.
- Trataremos melhor de relaçes não-lineares e modelos polinomiais no tópico a seguir: **regressão múltipla**.

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

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon \tag{27}$$

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. (27) 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, ..., 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. (27) também podem ser analisados pelas técnicas de regressão linear múltipla.

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

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon. \tag{28}$$

- Se fizermos $x_1 = x$, $x_2 = x^2$ e $x_3 = x^3$, então o modelo da Eq. (28) pode ser escrito como um modelo não-linear (no caso, polinomial cúbico) em uma variável de entrada:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \varepsilon. \tag{29}$$

- O método dos mínimos quadrados pode ser usado para estimar os coeficientes de regressão $\{\beta_j\}$, $j = 0, 1, ..., k$.
- 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. (27) em termos das observações:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i \tag{30}$$

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

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

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \cdots + \beta_k x_{1k} + \varepsilon_1 \\ y_2 &= \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \cdots + \beta_k x_{2k} + \varepsilon_2 \\ &\vdots \qquad\quad \vdots \qquad\quad \vdots \\ y_n &= \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \cdots + \beta_k x_{nk} + \varepsilon_n \end{aligned} \tag{31}$$

- Em forma matricial, o sistema de equações em (31) é escrito

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{32}$$

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:

$$J(\boldsymbol{\beta}) = \|\boldsymbol{\varepsilon}\|^2 = \boldsymbol{\varepsilon}^T \boldsymbol{\varepsilon} = \sum_{i=1}^{n} \varepsilon_i^2 = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}). \quad (33)$$

- 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. (33) pode ser decomposta em

$$\begin{aligned} J(\boldsymbol{\beta}) &= \mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{X}\boldsymbol{\beta} - \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{y} + \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} \quad (34) \\ &= \mathbf{y}^T\mathbf{y} - 2\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{y} + \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} \end{aligned}$$

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

$$\frac{\partial J(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -2\mathbf{X}^T\mathbf{y} + 2\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} = \mathbf{0}, \qquad (35)$$

  em que $\mathbf{0}$ é um vetor de zeros.
- Simplificando a Eq. (35) resulta em

$$\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} = \mathbf{X}^T\mathbf{y}. \qquad (36)$$

- A Eq. (36) 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. (36) pela inversa de $\mathbf{X}^T\mathbf{X}$.
- Assim, a estimativa de quadrados mínimos de $\boldsymbol{\beta}$ é dada por

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}. \tag{37}$$

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

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}. \tag{38}$$

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

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}. \tag{39}$$

Regularização de Thikonov

- Muitas vezes a matriz $\mathbf{X}^T\mathbf{X}$ é quase singular, 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 variáveis de entrada são intercorrelacionadas.
- Quando essa intercorrelação é grande, dizemos que existe *multicolinearidade*, ou seja, as linhas da matriz $\mathbf{X}^T\mathbf{X}$ não são linearmente independentes.

Regularização de Thikonov

- Os efeitos nocivos da multicolinearidade podem ser minimizados reescrevendo a Eq. (37) como

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\mathbf{X}^T\mathbf{y}. \tag{40}$$

  em que
  - $0 \leq \lambda \ll 1$ é uma constante de valor pequeno.
  - $\mathbf{I}$ é uma matriz identidade de dimensão $(k+1) \times (k+1)$.
- A técnica da Eq. (40) é chamada de **regularização de Tikhonov**, enquanto a regressão que a utiliza é chamada de **regressão de cumeeira** (*ridge regression*).

Exercício Teórico

- Mostrar que a Eq. (40) pode ser obtida a partir da seguinte função-custo:
$$J(\boldsymbol{\beta}) = \|\boldsymbol{\varepsilon}\|^2 + \lambda \|\boldsymbol{\beta}\|^2. \qquad (41)$$
- A função-custo da Eq. (33) é interpretada como aquela que provê estimativas dos coeficientes de regressão $\{\beta_j\}$, $j = 1, 2, ..., k$, que resultam na menor soma dos quadrados dos erros de estimação $\{e_i\}$, $i = 1, 2, ..., n$.
- Que interpretação pode ser dada à função-custo da Eq.(41)?

Coeficiente de Determinação na Regressão Múltipla

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

$$R^2 = 1 - \frac{SQ_E}{S_{yy}} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \tag{42}$$

  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}):

$$R^2_{aj} = 1 - \frac{SQ_E/(n-p)}{S_{yy}/(n-1)}, \tag{43}$$

  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.
- Portanto, R^2_{aj} crescer'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):

$$AIC(k) = N \ln\left[SQ_E(k)\right] + 2k, \quad k = 1, 2, \ldots \tag{44}$$

em que k é o número de parâmetros do modelo.

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

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

Critério de Informação de Akaike (cont.-1)

- O primeiro termo do lado direito da Eq. (44) sempre decresce (exponencialmente) com o aumento de k.
- O segundo termo do lado direito da Eq. (44) 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

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

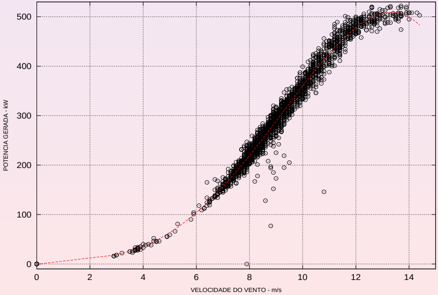
em que x_i é a i-ésima observação da variável de entrada.

- A estimativa de quadrados mínimos $\hat{\boldsymbol{\beta}}$ é então calculada por meio da Eq. (37).
- Predições de novos valores podem ser feitas por meio da Eq. (38) e resíduos são calculados pela Eq. (40).

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