Inteligência Computacional

Regressão Simples

Slides adaptados do material disponibilizado
pelo **Prof. Dr. Guilherme de Alencar Barreto (UFC)**

Motivação

- Em muitas aplicacões da ETI há duas ou mais variáveis que são intrinsicamente relacionadas, sendo necessário explorar a natureza dessa 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

- 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, $j = 1, \ldots, k$ são também chamadas de variáveis de entradas, variáveis regressoras ou ainda variáveis independentes.

Definição do Problema

- Assume-se que a variável y é uma variável aleatória e que as variáveis x_j são medidas com erro (i.e. ruído) desprezível.
- As variáveis x_j são frequentemente 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

- 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).

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, possivelmente 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.

Regressão Linear Simples

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

$$y = \beta_0 + \beta_1 x + \varepsilon, \qquad (2)$$

  em que ε é um erro (ruído) aleatório com média zero e variância σ^2
- Daí, o valor esperado de *y* para cada valor de *x* é dado por

$$E[y|x] = \beta_0 + \beta_1 x, \qquad (3)$$

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

Regressão Linear Simples

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

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n) \qquad (4)$$

- Estes dados devem obedecer à seguinte relação funcional:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,\ i = 1, 2, \ldots, n \qquad (5)$$

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

Regressão Linear Simples

- Os dados medidos serão usados para estimar os parâmetros desconhecidos β_0 e β_1 na Eq. (2).
- 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:

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

- **Entendendo o problema:** Minimar a 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.

Regressão Linear Simples

- As estimativas de β_0 e β_1, denotadas por $\hat{\beta}_0$ e $\hat{\beta}_1$ são dadas por

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

Regressão Linear Simples

- A solução das equações normais são dadas por

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

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

em que

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

Regressão Linear Simples

- 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,

$$e_i = y_i - \hat{y}_i \tag{7}$$

chamada de *erro de estimação* ou *resíduo*.

- A soma de quadrados dos resíduos é então dada por

$$SQ_E = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(y_i - \hat{y}_i \right)^2 \tag{8}$$

Regressão Linear Simples

- Uma estimativa de σ_{ε}^{2} pode ser dada por:

$$\hat{\sigma}_{\varepsilon}^{2} = \frac{SQ_E}{n-2} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2} \tag{9}$$

- **Questão importante:** Como saber se uma equação de regressão linear é a mais adequada para modelar os dados experimentais?

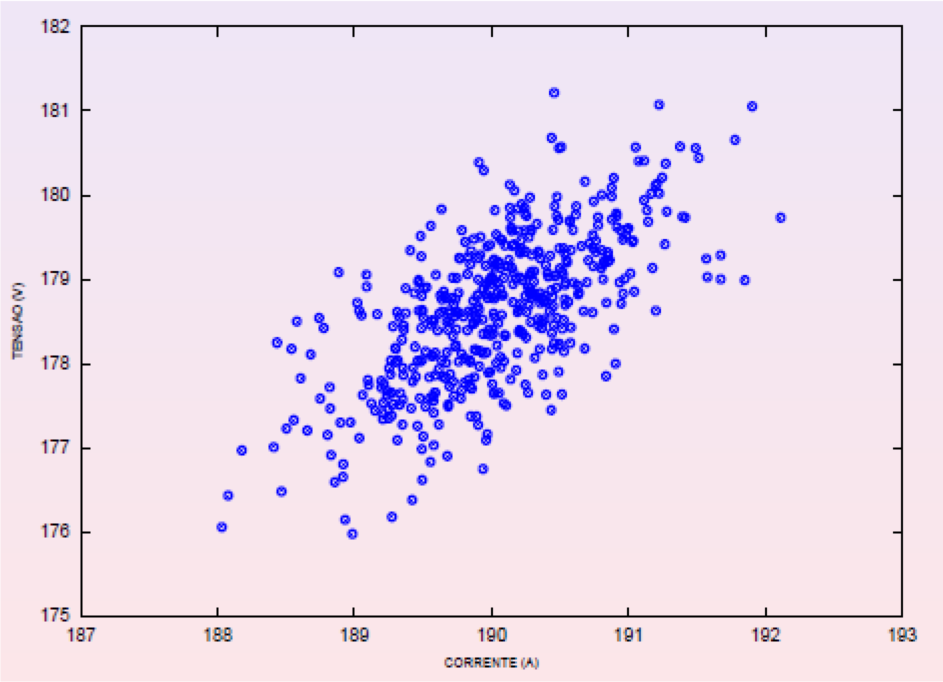
Regressão Linear Simples

- Uma primeira abordagem é puramente visual, através do gráfico de dispersão (scatterplot).
- Esse 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),'*');
```

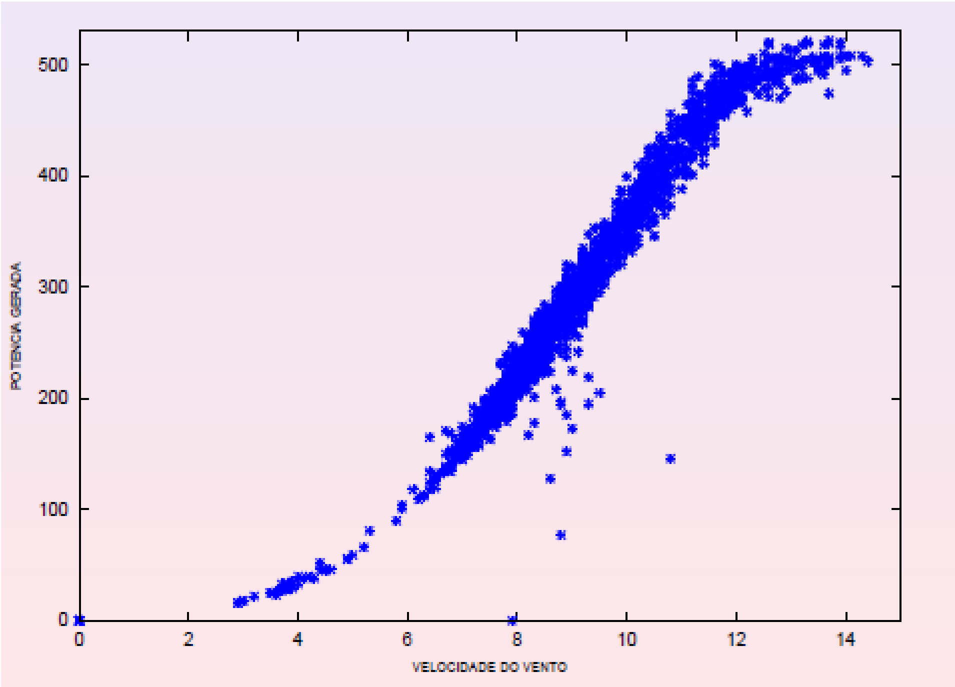
Regressão Linear Simples

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



Regressão Linear Simples

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



Regressão Linear Simples

- 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.

Regressão Linear Simples

- 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

$$\varepsilon_i \sim N\left(0, \sigma_\varepsilon^2\right)$$
$$E[\varepsilon_i, \varepsilon_j] = 0 \quad \forall i \neq j$$

Regressão Linear Simples

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 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.

Regressão Linear Simples

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.

Regressão Linear Simples

Definição - Coeficiente de Determinação

- O coeficiente de determinação é definido como

$$R^2 = 1 - \frac{SQ_E}{S_{yy}} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \tag{26}$$

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

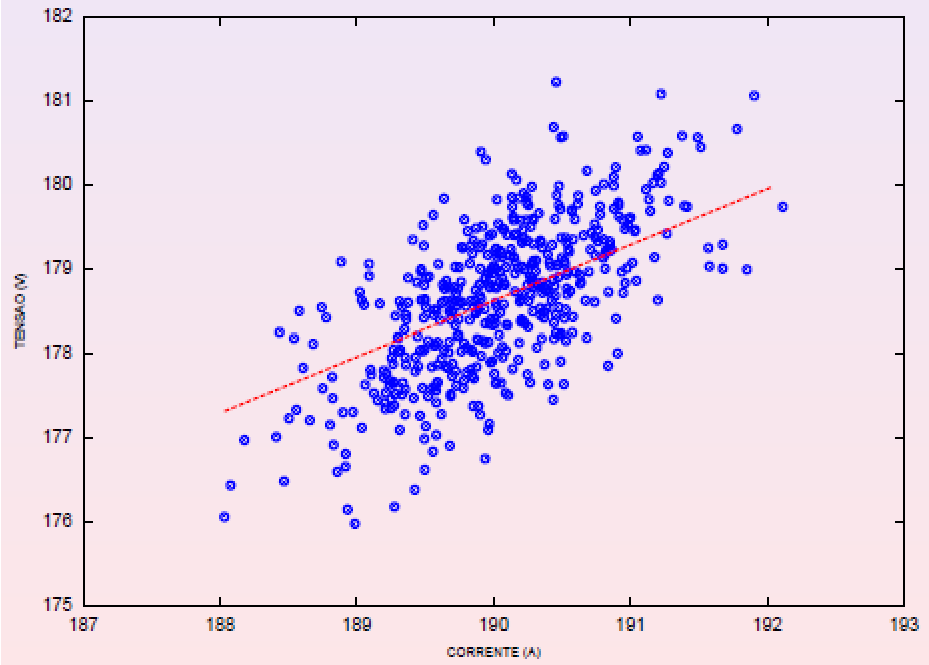
- R^2 é usada para julgar a adequação de um modelo de regressão. Em princípio, quanto mais próximo R^2 está de 1, mais adequado é o modelo de regressão.

Entendendo Melhor

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

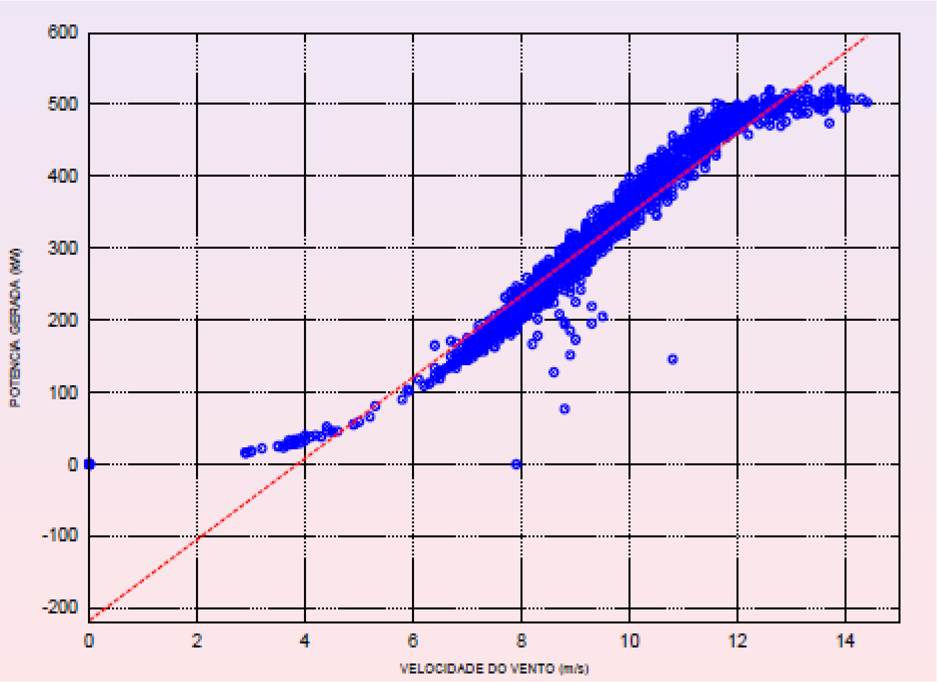
Regressão Linear Simples

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



Regressão Linear Simples

- 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.



Regressão Linear Simples

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?

Regressão Linear Simples

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.

Regressão Linear Simples

- 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

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

- Esta função pode ser linearizada por uma transformação Logarítmica

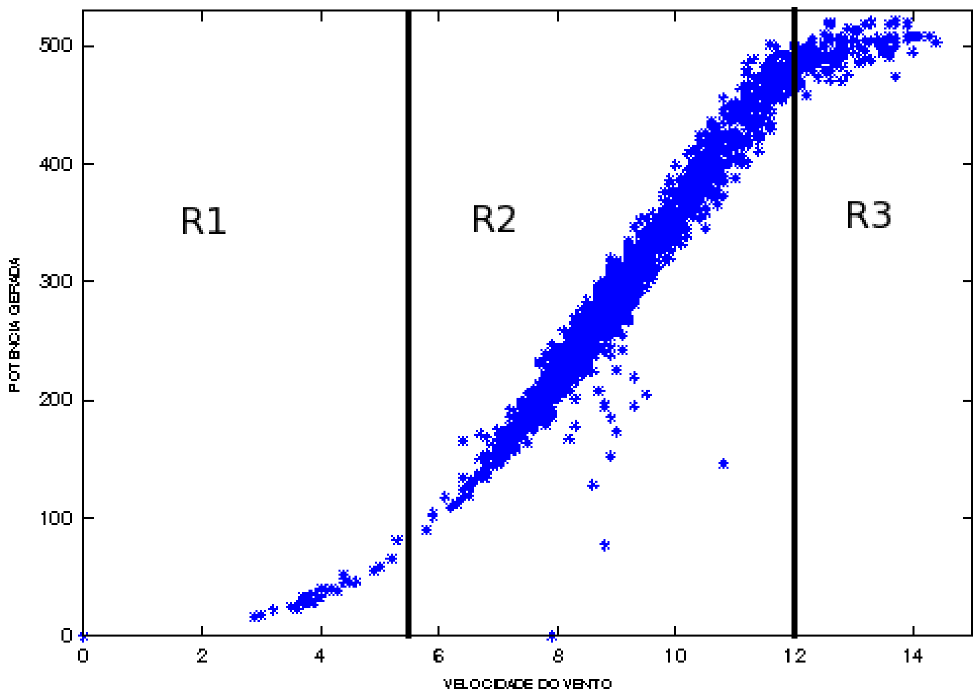
$$y^* = \ln(y) = \ln(\beta_0) + \beta_1 x + \ln(\varepsilon)$$

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

".

Regressão Linear Simples

- Uma outra opção é dividir o gráfico de dispersão em duas ou mais sub-regiões em que modelos de regressão linear sejam adequados.



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

Regressão Linear Simples

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.

Regressão Linear Simples

- Finalmente, para tratar o Caso 3, devemos lembrar que uma reta é um polinômio 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 relações não lineares e modelos polinomiais no tópico de: **regressão múltipla**.