# Regressão e Correlação Linear

## ML205 - Estatística I

João Luis Gomes Moreira

DMAT/CCT/UFRR

03 de maio de 2021

# Conteúdo

# Introdução

Os valores que uma variável pode assumir estão associados, além dos erros experimentais, a outras variáveis cujos valores se alteram durante o experimento.

Ao se relacionar, através de um modelo matemático, a variável resposta (variável dependente) com o conjunto das variáveis explicativas (variáveis independentes), pode-se então determinar algum parâmetro, ou mesmo fazer previsão acerca do comportamento da variável resposta.

Ao se estudar a relação entre duas variáveis, deve-se inicialmente fazer um gráfico dos dados, também chamado de *diagrama de dispersão*, pois ele fornece uma idéia da forma da relação exibida por elas.

# Introdução

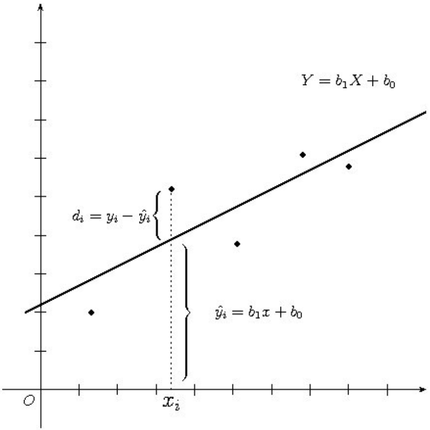
O modelo mais simples que relaciona duas variáveis $X$ e $Y$ é a equação da reta

$$Y = \beta_0 + \beta_1 . X \tag{1}$$

onde $\beta_0$ e $\beta_1$ são os parâmetros do modelo.

# Retas Possíveis

Seja uma reta arbitrária traçada sobre um dado diagrama de dispersão



# Retas Possíveis

Considerando todos os desvios de todos os $n$ pontos, temos

$$D = \sum_{i=1}^{n} {d_i}^2$$

Como medida do desvio total dos pontos observados e a reta estimada.

A magnitude de $D$ depende então obrigatoriamente da reta, ou seja, depende de $\beta_0$ e $\beta_1$.
Assim,

$$D(\beta_0, \beta_1) = \sum_{i=1}^{n} {d_i}^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Logo,

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

# Escolha da Melhor Reta

Uma maneira de estimar os coeficientes $\beta_0$ e $\beta_1$ é determinar o mínimo da função $D(\beta_0, \beta_1)$. O Método dos Mínimos Quadrados obtém o mínimo de uma função quadrática.

No processo de minimização: calculam-se as derivadas parciais de $D$ em relação a $\beta_0$ e $\beta_1$; obtem-se os valores de $\beta_0$ e $\beta_1$ igualando-se as derivadas a zero, e resolvendo o sistema de equações lineares, neste caso também chamado de sistema das equações normais.

Em nosso caso, o sistema das equações normais é dado por:

$$\left\{ \begin{array}{lcccc} (n)\beta_0 & + & (\sum x_i)\beta_1 & = & \sum y_i \\ (\sum x_i)\beta_0 & + & (\sum x_i^2)\beta_1 & = & \sum x_i y_i \end{array} \right.$$

# Escolha da Melhor Reta

Resolvendo-se o sistema encontram-se os valores de $\beta_0$ e $\beta_1$ que minimizam a função $D(\beta_0, \beta_1)$. A saber,

$$\beta_1 = \frac{n.\sum x_i y_i - \sum x_i \sum y_i}{n.\sum x_i^2 - (\sum x_i)^2}$$

$$\beta_0 = \frac{\sum y_i - (\sum x_i)\beta_1}{n}$$

A melhor reta que passa pelo diagrama de dispersão é dada, então, por :

$$Y = \beta_1 X + \beta_0 \tag{2}$$

# Coeficiente de Determinação

Mede-se a qualidade do ajuste linear simples através do coeficiente de determinação:

$$R^2 = \frac{[\sum x_i y_i - \sum x_i \sum y_i / n]^2}{[\sum x_i^2 - (\sum x_i)^2/n][\sum y_i^2 - (\sum y_i)^2/n]} \tag{3}$$

Sendo, $0 \leq R^2 \leq 1$, e quanto mais perto da unidade, melhor o ajuste.

# Exemplo

Ajuste os dados abaixo a uma reta.

| $i$ | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| $x_i$ | 1,3 | 3,4 | 5,1 | 6,8 | 8,0 |
| $y_i$ | 2,0 | 5,2 | 3,8 | 6,1 | 5,8 |

Precisamos, para definir a reta, de:

$n = 5$

$\sum x_i = 24,60$

$\sum x_i^2 = 149,50$

$\sum y_i = 22,90$

$\sum x_i y_i = 127,54$

# Exemplo

Calculamos agora os parâmetros da equação da reta pretendida

$$\beta_1 = \frac{n.\sum x_i y_i - \sum x_i \sum y_i}{n.\sum x_i^2 - (\sum x_i)^2} = \frac{5.127,54 - 24,6.22,9}{5.149,5 - (24,6)^2} = 0,522$$

$$\beta_0 = \frac{\sum y_i - (\sum x_i)\beta_1}{n} = \frac{22,9 - (24,6)0,522}{5} = 2,01$$

Então, a melhor reta que passa pelos pontos dados é

$$Y = 0,522X + 2,01$$

# Introdução

Um modelo linear para relacionar uma variável resposta $Y$ com $p+1$ variáveis explicativa é

$$Y = \beta_0 + \beta_1.X_1 + \beta_2.X_2 + \cdots + \beta_p.X_p \tag{4}$$

que pode ser representado matricialmente por

$$Y = X\beta$$

# Equações Normais

O sistema de equações normais neste caso é dado por

$$\begin{bmatrix} n & \sum x_{1i} & \sum x_{2i} & \cdots & \sum x_{pi} \\ \sum x_{li} & \sum x_{1i}^2 & \sum x_{2i}x_{1i} & \cdots & \sum x_{pi}x_{1i} \\ \sum x_{2i} & \sum x_{1i}x_{2i} & \sum x_{2i}^2 & \cdots & \sum x_{pi}x_{2i} \\ \vdots & \vdots & \vdots & & \vdots \\ \sum x_{pi} & \sum x_{1i}x_{pi} & \sum x_{2i}x_{pi} & \cdots & \sum x_{pi}^2 \end{bmatrix} . \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix} = \begin{bmatrix} \sum y_i \\ \sum x_{1i}y_i \\ \sum x_{2i}y_i \\ \vdots \\ \sum x_{pi}y_i \end{bmatrix}$$

$$\text{ou } X^T X^{\beta} = X^T Y$$

Como $det(X^T X) \neq 0$:

o sistema das equações normais apresenta solução única.

# Coeficiente de Determinação

Mede-se a qualidade do ajuste linear múltiplo através do coeficiente de determinação:

$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum y_i^2 - (\sum y_i)^2/n} \tag{5}$$

onde,

$$\hat{y}_i = \beta_0 + \sum_{j=1}^{p} (x_{ji})\beta_j$$

Sendo, $0 \leq R^2 \leq 1$, e quanto mais perto da unidade, melhor o ajuste.

# Exemplo

Ajuste os dados abaixo a uma reta.

| $i$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| $x_{1i}$ | -1 | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
| $x_{2i}$ | -2 | -1 | 0 | 1 | 1 | 2 | 3 | 4 |
| $y_i$ | 13 | 11 | 9 | 4 | 11 | 9 | 1 | -1 |

O vetor $\beta$ é a solução do sistema abaixo:

$$\begin{bmatrix} n & \sum x_{1i} & \sum x_{2i} \\ \sum x_{1i} & \sum x_{1i}^2 & \sum x_{2i}x_{1i} \\ \sum x_{2i} & \sum x_{1i}x_{2i} & \sum x_{2i}^2 \end{bmatrix} \cdot \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} \sum y_i \\ \sum x_{1i}y_i \\ \sum x_{2i}y_i \end{bmatrix}$$

# Exemplo

Precisamos, para definir a reta, de:

$n = 8$ $\quad \sum x_{1i} = 22$ $\quad \sum x_{1i}^2 = 108$

$\sum x_{2i} = 8$ $\quad \sum x_{2i}^2 = 36$ $\quad \sum x_{1i}x_{2i} = 57$

$\sum x_{1i}y_i = 92$ $\quad \sum x_{2i}y_i = -5$

$\sum y_i = 57$ $\quad \sum y_i^2 = 591$

O sistema é :

$$\begin{bmatrix} 8 & 22 & 8 \\ 22 & 108 & 57 \\ 8 & 57 & 36 \end{bmatrix} \cdot \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} 57 \\ 92 \\ -5 \end{bmatrix}$$

# Exemplo

A resolução do sistema nos dá :

$$\beta_0 = 4,239 \qquad \beta_1 = 3,400 \qquad \beta_2 = -6,464$$

Então, a melhor reta que passa pelos pontos dados é

$$Y = 4,239 + 3,400X_1 - 6,464X_2$$

Com o coeficiente de determinação

$$R^2 = 0,977$$

# Introdução

Um caso especial de ajuste linear múltiplo ocorre quando $X_1 = X$, $X_2 = X^2, \cdots, X_p = X^p$. Desta forma a equação (4) torna-se

$$Y = \beta_0 + \beta_1.X + \beta_2.X^2 + \cdots + \beta_p.X^p \tag{6}$$

# Equações Normais

O sistema de equações normais neste caso é dado por

$$\begin{bmatrix} n & \sum x_i & \sum x_i^2 & \cdots & \sum x_i^p \\ \sum x_i & \sum x_i^2 & \sum x_i^3 & \cdots & \sum x_i^{p+1} \\ \sum x_i^2 & \sum x_i x^3 & \sum x_i^4 & \cdots & \sum x_i^{p+2} \\ \vdots & \vdots & \vdots & & \vdots \\ \sum x_i^p & \sum x_i x^{p+1} & \sum x_i^{p+2} & \cdots & \sum x_i^{2p} \end{bmatrix} . \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix} = \begin{bmatrix} \sum y_i \\ \sum x_i y_i \\ \sum x_i^2 y_i \\ \vdots \\ \sum x_i^p y_i \end{bmatrix}$$

O coeficiente de determinação é o mesmo dado pela equação (5).

# Exemplo

Ajuste os dados abaixo à equação $Y = \beta_0 + \beta_1 X + \beta_2 X^2$

| $i$ | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| $x_i$ | -2 | -1,5 | 0 | 1 | 2,2 | 3,1 |
| $y_i$ | -30,5 | -20,2 | -3,3 | 8,9 | 16,8 | 21,4 |

O vetor $\beta$ é a solução do sistema abaixo:

$$\begin{bmatrix} n & \sum x_i & \sum x_i^2 \\ \sum x_i & \sum x_i^2 & \sum x_i^3 \\ \sum x_i^2 & \sum x_i^3 & \sum x_i^4 \end{bmatrix} \cdot \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} \sum y_i \\ \sum x_i y_i \\ \sum x_i^2 y_i \end{bmatrix}$$

# Exemplo

Precisamos, para definir a reta, de:

$n = 6$ $\quad \sum x_i = 2,8$ $\quad \sum x_i^2 = 21,7$

$\sum x_i^3 = 30,064$ $\quad \sum x_i^4 = 137,8402$ $\quad \sum x_i y_i = 203,5$

$\sum x_i^2 y_i = 128,416$ $\quad \sum y_i = -6,9$ $\quad \sum y_i^2 = 2168,59$

O sistema é :

$$\begin{bmatrix} 6 & 2,8 & 21,7 \\ 2,8 & 21,7 & 30,064 \\ 21,7 & 30,064 & 137,8402 \end{bmatrix} \cdot \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} -6,9 \\ 203,5 \\ 128,416 \end{bmatrix}$$

# Exemplo

A resolução do sistema nos dá :

$$\beta_0 = -2,018 \qquad \beta_1 = 11,33 \qquad \beta_2 = -1,222$$

Então, a melhor reta que passa pelos pontos dados é

$$Y = -2,018 + 11,33X - 1,222X^2$$

Com o coeficiente de determinação

$$R^2 = 0,997$$

That's all, Folks !!!