ANÁLISE DE REGRESSÃO

É O ESTUDO E OBTENÇÃO DE **RELAÇÕES ENTRE VARIÁVEIS.**
PRETENDE-SE SABER QUAL A
MELHOR ESTIMATIVA DOS **PARÂMETROS** DA REGRESSÃO

SIMPLES - QUANDO UMA VARIÁVEL **Y** (**DEPENDENTE**), É FUNÇÃO **APENAS** DE UMA VARIÁVEL **X** (**INDEPENDENTE**):

$$Y = f(X)$$

MÚLTIPLA - QUANDO UMA VARIÁVEL **Y** (**DEPENDENTE**), PODE SER RELACIONADA COM k **VARIÁVEIS INDEPENDENTES**:

$$Y = f(X_1, X_2, ..., X_k)$$

REGRESSÃO LINEAR SIMPLES:

$$Y = \alpha_0 + \alpha_1 X$$

REGRESSÃO LINEAR MÚLTIPLA:

$$Y = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + ... + \alpha_k X_k = \sum_{i=0}^{k} \alpha_i X_i$$

REGRESSÃO POLINOMIAL DE ORDEM k:

$$Y = \alpha_0 + \alpha_1 X + \alpha_2 X^2 + ... + \alpha_k X^k = \sum_{i=0}^{k} \alpha_i X^i$$

CÁLCULO DE PARÂMETROS

CONSISTE EM DETERMINAR OS PARÂMETROS DE UMA **FUNÇÃO QUE SE AJUSTE** A UM CONJUNTO DE PONTOS, UTILIZANDO UM MÉTODO DE OPTIMIZAÇÃO:

- **MÉTODO DA MÁXIMA VEROSIMILHANÇA**

- **MÉTODO DOS MÍNIMOS QUADRÁTICOS**

MÉTODO DOS MÍNIMOS QUADRÁTICOS

MINIMIZA-SE A FUNÇÃO CONSTITUÍDA PELO SOMATÓRIO DOS QUADRADOS DOS DESVIOS

HIPÓTESES DE APLICAÇÃO:

- **O VALOR MÉDIO DOS ERROS** É **ZERO**.

- OS **ERROS** TÊM **VARIÂNCIA COMUM**.

- OS **ERROS** SÃO **INDEPENDENTES**.

- OS VALORES $\mathbf{Y_j}$ **PARA CADA VALOR DE X** TÊM UMA **DISTRIBUIÇÃO NORMAL**.

- **OS VALORES DE X SÃO MEDIDOS SEM ERRO OU COM ERRO DESPREZÁVEL**.

O **MÉTODO DOS MÍNIMOS QUADRÁTICOS** (MMQ) BASEIA-SE NA MINIMIZAÇÃO DA FUNÇÃO:

$$S = \sum_{i=1}^{n} \left(y_i - \hat{Y}_i \right)^2$$

y_i - VALOR DE Y ENCONTRADO NA REALIZAÇÃO EXPERIMENTAL PARA O VALOR X_i.

$\hat{Y}_i$ - VALOR OBTIDO POR SUBSTITUIÇÃO DE CADA X_i NA EQUAÇÃO DA FUNÇÃO. É UMA ESTIMATIVA DO VALOR.

RECTA

$$\hat{Y}_i = \hat{\alpha}_0 + \hat{\alpha}_1 X_i$$

$\hat{\alpha}_0$ E $\hat{\alpha}_1$ - SÃO ESTIMATIVAS DOS VERDADEIROS VALORES DA **ORDENADA NA ORIGEM** E DO **COEFICIENTE ANGULAR**.

$$S = \sum_{i=1}^{n} \left[y_i - \left(\hat{\alpha}_0 - \hat{\alpha}_1 X_i \right) \right]^2$$

$$\left| \begin{array}{l} \dfrac{\partial S}{\partial \alpha_0} = 0 \\ \\ \dfrac{\partial S}{\partial \alpha_1} = 0 \end{array} \right. \quad \Rightarrow \quad \hat{\alpha}_0 \quad \text{E} \quad \hat{\alpha}_1$$

COEFICIENTE ANGULAR

$$\hat{\alpha}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x - \bar{x})^2} = \frac{n\sum_{i=1}^{n} x_i\, y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

ORDENADA NA ORIGEM

$$\hat{\alpha}_0 = \bar{y} - \hat{\alpha}_1 \bar{x} = \frac{\sum_{i=1}^{n} y_i - \hat{\alpha}_1 \sum_{i=1}^{n} x_i}{n}$$

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

VARIÂNCIA DA CORRELAÇÃO LINEAR SIMPLES

- REPRESENTA-SE POR $s^2_{y.s}$ E É UMA ESTIMATIVA DA DISPERSÃO DOS VALORES EM TORNO DA RECTA $\sigma^2_{y.s}$.

$$s^2_{y.s} = \frac{\sum \left(y_i - \hat{Y}_i\right)^2}{n-2} = \frac{\sum_{i=1}^{n} y_i^2 - \hat{\alpha}_0 \sum_{i=1}^{n} y_i - \hat{\alpha}_1 \sum_{i=1}^{n} x_i\, y_i}{n-2}$$

n - 2 É O NÚMERO DE GRAUS DE LIBERDADE.

È MENOS 2 PORQUE IMPOMOS AO SISTEMA DUAS RESTRIÇÕES: CÁLCULO DE $\hat{\alpha}_0$ E $\hat{\alpha}_1$

ERRO INERENTE AO COEFICIENTE ANGULAR

$$\alpha_1 = \hat{\alpha}_1 \pm t \frac{s_{y.x}}{\sqrt{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2}}$$

ERRO INERENTE À ORDENADA NA ORIGEM

$$\alpha_0 = \hat{\alpha}_0 \pm t\ s_{y.x} \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2}}$$

OU

$$\alpha_0 = \hat{\alpha}_0 \pm t\ s_{y.x} \sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n \sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2}}$$

t - FACTOR DE STUDENT PARA UMA **DADA PROBABILIDADE** (95 %) E **n - 2 GRAUS DE LIBERDADE**.

$$\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2 = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}$$

INTERVALOS DE INDETERMINAÇÃO DE Y

❋ QUANDO REDUZIMOS UM CONJUNTO DE PONTOS EXPERIMENTAIS (CONSTITUÍDOS POR PARES DE VALORES (**x, y**) A UMA RECTA, NORMALMENTE SERÁ PARA OBTERMOS MAIS TARDE QUER VALORES DE Y PARA DADOS VALORES DE X, QUER DE X PARA UM DADO VALOR DE Y.

EM SUMA, PARA EFECTUARMOS **PREVISÃO DE DADOS**.

- PROVA-SE QUE, PARA UM DADO VALOR DE $\mathbf{x_o}$ SE PODEM ENCONTRAR **p** VALORES DE **Y** COM UM ERRO:

$$Y_o = \hat{Y}_o \pm t\, s_{y.x} \sqrt{\frac{1}{p} + \frac{1}{n} + \frac{\left(x_o - \bar{x}\right)^2}{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2}}$$

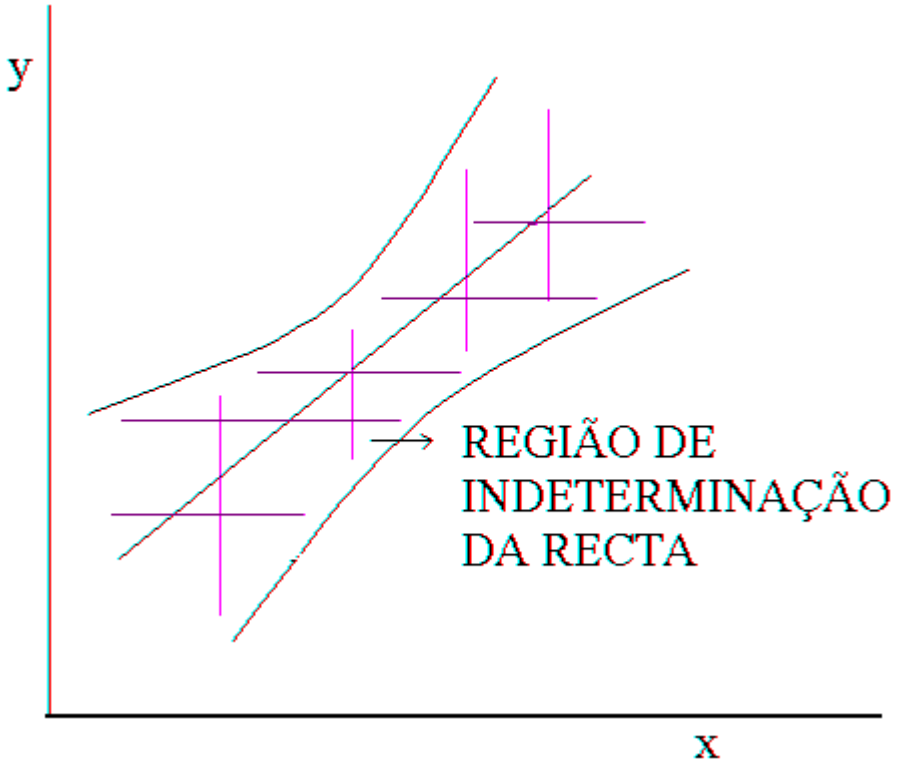
n - NÚMERO DE VALORES DE x, A PARTIR DOS QUAIS SE DEFINE A RECTA.

SE **p = 1**

$$Y_o = \hat{Y}_o \pm t\, s_{y.x} \sqrt{1 + \frac{1}{n} + \frac{\left(x_o - \bar{x}\right)^2}{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2}}$$

SE **p = ¥**

$$Y_o = \hat{Y}_o \pm t\, s_{y.x} \sqrt{\frac{1}{n} + \frac{\left(x_o - \bar{x}\right)^2}{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2}}$$



INTERPOLAÇÃO

QUANDO SE DETERMINAM VALORES DE Y, DENTRO DA GAMA DOS VALORES DE X, DETERMINADOS EXPERIMENTALMENTE.

EXTRAPOLAÇÃO

QUANDO PRETENDEMOS DETERMINAR VALORES DE X OU Y, FORA DA GAMA EXPERIMENTAL.

É PRECISO TER CUIDADO!!

PREVISÃO DE $\hat{X}$ PARA UM DADO VALOR DE Y

SERÁ $$\hat{X}_o = \frac{y_o - \hat{\alpha}_0}{\hat{\alpha}_1}$$

$$X_o = \hat{X}_o \pm \Delta \hat{X}_o$$

$\Delta \hat{X}_o$ **DEVERÁ SER OBTIDO GRAFICAMENTE, OU RESOLVENDO A EQUAÇÃO QUE NOS DÁ O INTERVALO DE INDETERMINAÇÃO DA RECTA.**

NOTAR QUE, MESMO QUE SE ADMITA QUE **X NÃO TEM ERRO** PARA SE PODER OBTER A EQUAÇÃO DA RECTA, **A PREVISÃO DE UM DADO VALOR DE X_o**, PARA UM VALOR CONHECIDO DE **Y_o**, **JÁ TEM ERRO**.

RECTA QUE PASSA PELA ORIGEM

NESTE CASO α_0 É NULA E APENAS É NECESSÁRIO DETERMINAR α_1

$$Y = \alpha_1' \; x$$

A FUNÇÃO A MINIMIZAR SERÁ:

$$S = \sum_{i=1}^{n} \left(Y_i - \alpha_1' \; x \right)^2 \qquad \Rightarrow \qquad \frac{\partial S}{\partial \alpha_1'} = 0$$

$$\alpha_1' = \frac{\sum_{i=1}^{n} x_i \; y_i}{\sum_{i=1}^{n} x_i^2}$$

$$\hat{\alpha}_1' = \alpha_1' \pm t \; s_{y.x} \; \frac{1}{\sum_{i=1}^{n} x_i^2}$$

$$s_{y.x}^2 = \frac{\sum_{i=1}^{n} \left(y_i - \alpha_1' \; x_i \right)^2}{n - 2} = \frac{\sum_{i=1}^{n} y_i^2 - \alpha_1' \sum_{i=1}^{n} x_i \; y_i}{n - 2}$$

RECTA QUE PASSA POR UM PONTO FIXO

NESTE CASO: $Y = C + \alpha_1'' \; x$

EM QUE C É UMA CONSTANTE.

A FUNÇAO A MINIMIZAR É:

$$S = \sum_{i=1}^{n} \left(Y_i - C - \alpha_1'' \; x \right)^2 \qquad \Rightarrow \qquad \frac{\partial S}{\partial \alpha_1''} = 0$$

Exemplo:

Na calibração de um espectrofotómetro obtiveram-se os seguintes valores:

C (g/L)	A (λ = 490 mμ)
0,204	0,04
0.306	0,06
0.408	0,08
0,510	0,11
0,612	0,13
0,714	0,15
0,816	0,18
0,918	0,20
1,020	0,23

Sabendo que o espectrofotómetro foi aferido, antes de cada leitura, para o seu valor zero, determine a recta que mais provavelmente representa estes pontos.

(Não esquecer de determinar os erros da ordenada na origem e do coeficiente angular e a região de indeterminação da recta. Neste caso deve obrigar-se a recta a passar pela origem? Interprete estatisticamente.)

PROGRAMAÇÃO DE ENSAIOS NA CORRELAÇÃO LINEAR SIMPLES

O INTERVALO DE CONFIANÇA DA RESPOSTA PREVISTA Y_o É:

$$Y_o = \hat{Y}_o \pm t\, s_{y.x} \sqrt{1 + \frac{1}{n} + \frac{(x_o - \bar{x})^2}{\sum_{i=1}^{n} (x_o - \bar{x})^2}}$$

ASSIM, O ERRO ASSOCIADO AO VALOR DA PREVISÃO DEPENDE DO VALOR x_o ESCOLHIDO.

EM PROGRAMAÇÃO CONSIDERA-SE QUE $x_o = \bar{x}$.:

$$Y_o = \hat{Y}_o \pm t\, s_{y.x} \sqrt{1 + \frac{1}{n}}$$

OU SEJA:

$$\Delta Y = t\, s_{y.x} \sqrt{\frac{n+1}{n}}$$

I - SEM ENSAIOS PRÉVIOS

SUPÕE-SE QUE OS DESVIOS $\left(y_i - \hat{Y}_i \right)$ SÃO TODOS IGUAIS (SE-LO-ÃO EM MÉDIA PARA PARA UTILIZAÇÃO DO MÉTODO DOS MÍNIMOS QUADRÁTICOS).

ENTÃO:

$$s^2_{y.s} = \frac{n\left(y_i - \hat{Y}_i \right)^2_{\overline{x}}}{n-2}$$

CONSTRUINDO UMA TABELA IDÊNTICA À QUE FOI CONSTRUÍDA PARA MEDIÇÕES INDEPENDENTES, PODEMOS DETERMINAR O NÚMERO DE PONTOS PARA DEFINIR UMA RECTA COM UM DETERMINADO RIGOR.

II - COM APOIO EM ENSAIOS PRÉVIOS

NESTE CASO JÁ TEMOS CONHECIMENTO DOS VALORES DE $s^2_{y.s}$ PARA **m** EXPERIÊNCIAS:

$$s^2_{y.s_m} = \frac{\sum_{i=1}^{m} \left(y_i - \hat{Y}_i \right)^2}{m-2}$$

DE IGUAL MODO, PARA j EXPERIÊNCIAS POSTERIORES, TEREMOS:

$$s^2_{y.s_j} = \frac{m-2}{j-2} \frac{j}{m} s^2_{y.s_m}$$

EM

CORRELAÇÕES NÃO LINEARES

COMO MÉTODO APROXIMADO, DIVIDE-SE A CURVA EM VÁRIOS TROÇOS DE TAL MODO QUE O ERRO INTRODUZIDO PELA SUBSTITUIÇÃO DE UM TROÇO CURVO POR UMA RECTA NÃO SEJA MUITO GRANDE.

É DE SEGUIDA POSSÍVEL APLICAR O PROCEDIMENTO ANTERIOR.

Exemplo:

Realizaram-se 5 ensaios de uma grandeza Y que depende de X, segundo um lei linear, tendo obtido os seguintes resultados:

Y	X
1,02	6,04
2,08	8,02
3,07	10,18
3,98	11,94
5,02	14,09

Calcular o número de ensaios necessários para obter um erro de:

a) 1,5 %

b) 1,2 %

REGRESSÃO LINEAR MÚLTIPLA

A REGRESSÃO LINEAR MÚLTIPLA PODE SER EXPRESSA PELA RELAÇÃO:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

$$Y = \sum_{j=0}^{k} \beta_j x_j$$

OS DADOS CONSISTEMEM DUAS MATRIZES:

- ❊ - UMA (**nx1**) QUE CONTEM OS VALORES DE Y;
- ❊ - OUTRA (**nx(k+1)**) QUE CONTEM OS VALORES DAS VARIÁVEIS INDEPENDENTES

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \cdots \\ \cdots \\ \cdots \\ y_n \end{bmatrix} \qquad X = \begin{bmatrix} x_{01} & x_{11} & x_{21} & \cdots & x_{k1} \\ x_{02} & x_{12} & x_{22} & \cdots & x_{k2} \\ & & \cdots & & \\ & & \cdots & & \\ & & \cdots & & \\ x_{0n} & x_{1n} & x_{2n} & \cdots & x_{kn} \end{bmatrix}$$

EM QUE:

- ✹ A VARIÁVEL x_0 TOMA O VALOR 1.
- ✹ x_{ij} É O ENSAIO i DA VARIÁVEL x_j.

MÉTODO DOS MÍNIMOS QUADRÁTICOS

MINIMIZA-SE A FUNÇÃO DO SOMATÓRIO DOS QUADRADOS DOS DESVIOS:

$$SRM = \sum_{i=1}^{n} \left[y_i - \left(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} \right) \right]^2$$

A IGUALIZAÇÃO A ZERO DAS DERIVADAS EM ORDEM AOS PARÂMETROS:

$$\frac{\partial SRM}{\partial \beta_0} = 0$$
$$\frac{\partial SRM}{\partial \beta_1} = 0$$
$$\cdots \quad \cdots \quad \cdots$$
$$\frac{\partial SRM}{\partial \beta_k} = 0$$

CONDUZ AO SISTEMA DE (**k+1**) EQUAÇÕES LINEARES, A PARTIR DO QUAL É POSSÍVEL DETERMINAR OS PARÂMETROS:

$$\begin{cases} \beta_0 \sum x_0 x_0 + \beta_1 \sum x_0 x_1 + \cdots + \beta_k \sum x_0 x_k = \sum x_0 y \\ \beta_0 \sum x_1 x_0 + \beta_1 \sum x_1 x_1 + \cdots + \beta_k \sum x_1 x_k = \sum x_1 y \\ \qquad\qquad\qquad\qquad\qquad \cdots \\ \qquad\qquad\qquad\qquad\qquad \cdots \\ \qquad\qquad\qquad\qquad\qquad \cdots \\ \beta_0 \sum x_k x_0 + \beta_1 \sum x_k x_1 + \cdots + \beta_k \sum x_k x_k = \sum x_k y \end{cases}$$

EM QUE:

$$\sum x_j x_j = \sum_{i=1}^{n} x_{ij}^2$$
$$\sum x_j x_k = \sum_{i=1}^{n} x_{ij} x_{ik}$$
$$\sum x_j y = \sum_{i=1}^{n} x_{ji} y_i$$

EM NOTAÇÃO MATRICIAL SERÁ:

$$X^T X B = X^T Y \qquad \text{EM QUE} \qquad B = [\beta_0 \ \beta_1 \ \cdots \ \beta_k]$$

SE

$$\left| \begin{array}{l} X^T X = A \\ \\ X^T Y = G \end{array} \right. \Rightarrow \quad \text{O SISTEMA SERÁ} \quad | \quad A B = G$$

RESOLVENDO VEM:

$$B = A^{-1} G$$

EM QUE A^{-1} É A MATRIZ INVERSA DA MATRIZ $A = X^T X$

DESVIO PADRÃO

$$s_{(y.x)m} = \sqrt{\frac{SRM}{n - (k+1)}}$$

INTERVALOS DE CONFIANÇA DOS PARÂMETROS

$$\beta_i = \hat{\beta}_i \pm t\, s_{(y.x)m} \frac{1}{\sqrt{\sum_{i=1}^{n} \left(x_{ij} - \bar{x}_i\right)^2}}$$

EM QUE **t** É O FACTOR DE STUDENT PARA UMA DADA PROBABILIDADE (95 %) E (**n-(k+1)**) GRAUS DE LIBERDADE.

INTERVALO DE INDETERMINAÇÃO DA CURVA

$$Y_o = \hat{Y}_o \pm t\, s_{(y.x)m} \sqrt{\frac{1}{p} + \frac{1}{n} + \sum_h \sum_j \frac{\left(x_{oh} - \bar{x}_h\right)\left(x_{oj} - \bar{x}_j\right)}{\sum_{i=1}^{n}\left(x_{ih} - \bar{x}_h\right)\left(x_{ij} - \bar{x}_j\right)}}$$

EM QUE:

p - É O NÚMERO DE DETERMINAÇÕES DE y PARA CADA CONJUNTO DE VARIÁVEIS (x_1, ..., x_k).

n - É O NÚMERO DEPONTOS USADOS PARA DEFINIR A CURVA.

k - É O NÚMERO DE VARIÁVEIS INDEPENDENTES.

2 VARIÁVEIS

SE **p**=1

$$Y_o = \hat{Y}_o \pm t\, s_{(y.x)m} \sqrt{1 + \frac{1}{n} + \frac{\left(x_{o1} - \bar{x}_1\right)\left(x_{o2} - \bar{x}_2\right)}{\sum_{i=1}^{n}\left(x_{i1} - \bar{x}_1\right)\left(x_{i2} - \bar{x}_2\right)}}$$

3 VARIÁVEIS

SE **p**=1

$$Y_o = \hat{Y}_o \pm t\, s_{(y.x)m} \sqrt{1 + \frac{1}{n} + \frac{(x_{o1} - \bar{x}_1)\ (x_{o2} - \bar{x}_2)}{\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2)} + \frac{(x_{o1} - \bar{x}_1)\ (x_{o3} - \bar{x}_3)}{\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)(x_{i3} - \bar{x}_3)} + \frac{(x_{o2} - \bar{x}_2)\ (x_{o3} - \bar{x}_3)}{\sum_{i=1}^{n}(x_{i2} - \bar{x}_2)(x_{i3} - \bar{x}_3)}}$$

REGRESSÃO POLINOMIAL

A REGRESSÃO POLINOMIAL REPRESENTA-SE:

$$Y = \alpha_0 + \alpha_1 X + \alpha_2 X^2 + \cdots + \alpha_k X^k$$

$$Y = \sum_{j=0}^{k} \alpha_j X^j$$

E PODE TRANSFORMAR-SE EM:

$$Y = \sum_{j=0}^{k} \alpha_j X_j \qquad \text{EM QUE} \qquad X^i \equiv X_i$$

ASSIM O TRATAMENTO FICA EM TUDO IDÊNTICO AO DA REGRESSÃO LINEAR MÚLTIPLA.

Exemplo:

Uma variável Y depende de duas variáveis (X_1 e X_2), segundo a equação:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

com $\beta_0 = 0$.

Determinar o valor dos parâmetros, a partir dos seguintes dados:

X_1	X_2	Y
5,6	77	17,9
5,2	76	16,3
4,8	75	15,5
3,9	73	17,3
4,6	68	18,1
4,9	66	18,5
5,4	63	19,5