

**LCF280 – Métodos Quantitativos para a Gestão Ambiental  
(Estudo Dirigido Especial - Regressão)**

**Regressão Linear Simples – uma revisão**

A regressão linear é útil quando a variável de interesse (dependente) se relaciona linearmente com uma ou mais variáveis (independentes). Começamos pelo modelo que, da forma mais simples possível, representa essa relação para várias observações  $i$ :

$$Y_i = \beta_0 + \beta_1 X_i \quad (\text{a equação de uma reta}) \quad (1)$$

onde

$\beta_0$  = intercepto (isto é, valor de Y quando X é zero); e

$\beta_1$  = coeficiente angular ou de inclinação da reta (mede a variação em Y dada uma variação unitária em X).

A equação (1) expressa uma **relação determinista** entre Y e X. Isto é, não há erro na leitura de Y, basta saber o valor de X.

De fato, relações deterministas não são muito comuns, pois em geral o nível de X tende a explicar apenas “parcialmente” o valor de Y.

A utilização de um gráfico de pontos nos ajuda a inferir melhor sobre a tendência linear ou não da relação entre os dados observados para Y e X. Ao observar o gráfico, pode ser mais adequado concluir que a expressão da relação entre X e Y não segue deterministicamente a tendência de uma reta.

A versão não determinista do mesmo modelo pode ser representada da seguinte forma:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad (2)$$

onde  $\varepsilon_i$  expressa o erro aleatório da  $i$ ésima observação.

**Assimilação**

Vamos supor que um conjunto de indivíduos tenha sido reagrupado em classes de idade de acordo com a seguinte tabela:

Classe de Idade	Grupo
20 a 22	1
22 a 24	2
24 a 26	3
26 a 28	4
28 a 30	5
30 a 32	6
32 a 34	7
34 a 36	8
36 a 38	9
38 a 40	10

Alguém levanta a questão: estaria a idade relacionada com o consumo médio de combustível automotivo por quilômetro percorrido? Para testar essa hipótese, foram entrevistadas 33 pessoas. Os resultados foram tabulados e são apresentados na seguinte tabela:

Grupo	Gênero	Consumo	Grupo	Gênero	Consumo	Grupo	Gênero	Consumo
1	0	1,36	3	0	1,7	7	0	2,35
1	0	0	3	1	1,66	7	0	2,29
1	1	1,36	5	0	1,42	7	0	1,88
1	1	0,89	5	0	1,76	8	0	1,93
2	1	1,21	5	0	2,01	9	0	1,44
2	1	1,54	5	1	1,49	9	1	2,64
2	0	1,37	6	1	2,25	9	0	1,9
2	0	1,59	6	1	1,29	9	1	1,46
2	0	1,11	6	0	2,16	9	1	2,08
2	0	0,78	7	1	2,37	10	0	1,57
3	1	1,22	7	0	2,2	10	1	2,76

Gênero masculino é representado pelo algarismo 1, e o feminino pelo algarismo 0.

**LCF280 – Métodos Quantitativos para a Gestão Ambiental  
(Estudo Dirigido Especial - Regressão)**

Pressuposição básica: o valor médio de  $\varepsilon$  para um dado valor de  $X$  é zero! Assim sendo:

$$E(Y) = \beta_0 + \beta_1 X$$

$\beta_0$  e  $\beta_1$  são parâmetros desconhecidos e não se conhece o formato preciso da reta  $E(Y) = \beta_0 + \beta_1 X$ . A questão, então, é como construir boas estimativas  $b_0$  e  $b_1$  para esses parâmetros que resultem em uma boa estimativa  $y$  de  $E(Y)$ ?

O método dos mínimos quadrados é o procedimento mais frequentemente utilizado. Definindo *erro de predição* ou *resíduo* como  $(Y - y)$ , esse método resulta na equação de reta

$$y = b_0 + b_1 X$$

que minimiza a soma dos quadrados dos resíduos  $\sum (Y_i - y_i)^2$  para todas as observações  $i$ .

$$Q = \sum (Y_i - y_i)^2 = \sum (Y_i - b_0 + b_1 X_i)^2$$

Como encontrar  $b_0$  e  $b_1$  que minimizam  $\sum (Y_i - b_0 + b_1 X_i)^2$ ? Basta fazer:

$$\begin{cases} \frac{\partial Q}{\partial b_0} = 0 \\ \frac{\partial Q}{\partial b_1} = 0 \end{cases} \Rightarrow \begin{cases} -2 \sum_{i=1}^n (Y_i - b_0 - b_1 X_i) = 0 \\ -2 \sum_{i=1}^n X_i (Y_i - b_0 - b_1 X_i) = 0 \end{cases} \Rightarrow \begin{cases} \sum_{i=1}^n Y_i - n b_0 - b_1 \sum_{i=1}^n X_i = 0 \\ \sum_{i=1}^n X_i Y_i - b_0 \sum_{i=1}^n X_i - b_1 \sum_{i=1}^n X_i^2 = 0 \end{cases}$$

e obter as “*equações normais*”:

$$\begin{aligned} \sum_{i=1}^n Y_i &= n b_0 + b_1 \sum_{i=1}^n X_i \\ \sum_{i=1}^n X_i Y_i &= b_0 \sum_{i=1}^n X_i + b_1 \sum_{i=1}^n X_i^2 \end{aligned} \quad (3)$$

**Exercício 1:**

*Ignore, inicialmente, o gênero e apresente os dados tabulados em um gráfico. Disponha o consumo no eixo vertical e comente a relação aparente entre as variáveis.*

LCF280 – Métodos Quantitativos para a Gestão Ambiental  
(Estudo Dirigido Especial - Regressão)

Que, quando resolvida (2 incógnitas e 2 equações), produz:

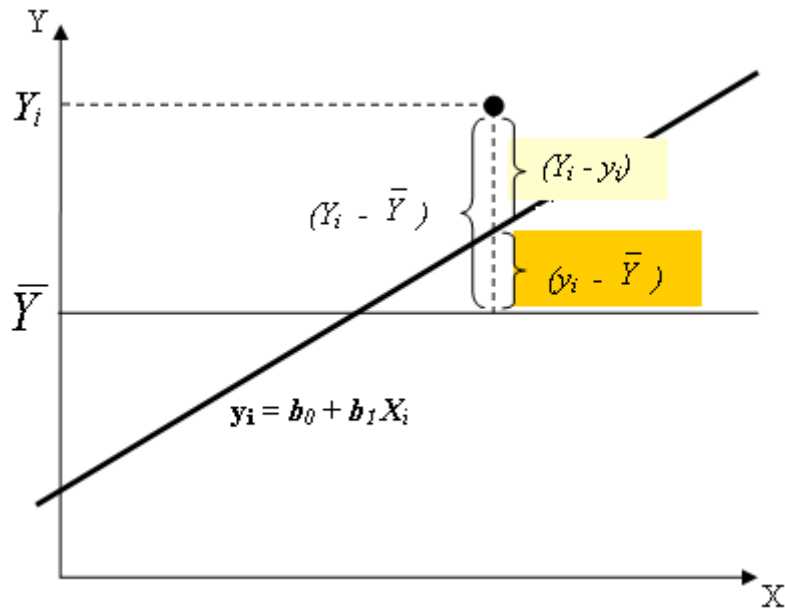
$$b_1 = S_{XY} / S_{XX} \quad e \quad b_0 = \bar{Y} - b_1 \bar{X}$$

onde  $S_{XY} = \sum_{i=1}^n X_i Y_i - \frac{\sum_{i=1}^n X_i \sum_{i=1}^n Y_i}{n}$  e  $S_{XX} = \sum_{i=1}^n X_i^2 - \frac{(\sum_{i=1}^n X_i)^2}{n}$

É interessante estudar a relação entre

$$\sum_{i=1}^n (Y_i - \bar{Y})^2 \quad e \quad \sum_{i=1}^n (Y_i - y_i)^2$$

e analisar graficamente uma forma de expressar a variação total das observações em torno da média amostral ( $\bar{Y}$ ):



**Exercício 2:**

Considere consumo uma função linear simples da classe de idade representada pelo grupo e calcule:

$$S_{XY} =$$

$$S_{XX} =$$

$$b_1 =$$

$$b_0 =$$

**Exercício 3:**

Apresente num mesmo gráfico os dados tabulados e a reta ajustada no ex. 2.

**LCF280 – Métodos Quantitativos para a Gestão Ambiental  
(Estudo Dirigido Especial - Regressão)**

$(y_i - \bar{Y})$  representa a porção da variação em Y devida à variável independente X.

$(Y_i - y_i)$  representa a porção da variação em Y não devida à variável independente X.

A partir dessa análise podemos expressar a variabilidade total das observações em torno da média  $\sum_{i=1}^n (Y_i - \bar{Y})^2$ , também chamada Soma dos Quadrados Médios, como a soma do quadrado dos desvios devidos à regressão  $\sum_{i=1}^n (y_i - \bar{Y})^2$  e a soma do quadrado dos resíduos  $\sum_{i=1}^n (Y_i - y_i)^2$ . Ou seja:

<b>Soma de Quadrados em torno da média ou Total</b>	<b>=</b>	<b>Soma de Quadrados devido à Regressão</b>	<b>+</b>	<b>Soma de Quadrados devido ao Erro</b>
---	----------	---	----------	---

$$\sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n (y_i - \bar{Y})^2 + \sum_{i=1}^n (Y_i - y_i)^2$$

SQT	=	SQReg	+	SQE
-----	---	-------	---	-----

**Quanto maior a variabilidade explicada pelo modelo de regressão com relação à variabilidade não explicada, mais o modelo “ajusta” os dados!**

**Exercício 4:**

Calcule para os dados ajustados no exercício 2,

**A Soma de Quadrados Total (SQT)**

**A Soma de Quadrados da Regressão (SQReg)**

**A Soma de Quadrados do Erro (SQE)**

LCF280 – Métodos Quantitativos para a Gestão Ambiental  
(Estudo Dirigido Especial - Regressão)

Fórmulas úteis:

$$SQT = S_{YY}$$

$$SQE = S_{YY} - b_1 S_{XY}$$

$$SQReg = b_1 S_{XY} = b_1^2 S_{XX}$$

**Coeficiente de correlação**

Um coeficiente de correlação mede o “vigor” da relação entre duas variáveis. O coeficiente de correlação amostral  $r$  (ou coeficiente de correlação de *Pearson*) é um exemplo.

$$r = \frac{S_{XY}}{\sqrt{S_{XX} S_{YY}}} = b_1 \sqrt{\frac{S_{XX}}{S_{YY}}} = b_1 \sqrt{\frac{S_{XX}/(n-1)}{S_{YY}/(n-1)}} = b_1 \frac{s_X}{s_Y}$$

$r$  varia entre -1 e +1:

> 0 → correlação linear positiva

< 0 → correlação linear negativa

~ 0 → correlação nula

**Coeficiente de determinação**

É representado por  $r^2$  e mostra a proporção da variação total em Y “*explicada*” pela variável independente X.

$$r^2 = b_1^2 \frac{S_{XX}}{S_{YY}} = SQReg / SQT$$

**Exercício 5:**

Calcule o coeficiente de determinação do ajuste feito no exercício 2:

Aqueles que quiserem saber mais, procurem uma cópia do livro *Análise de Regressão – uma introdução à econometria* (Hoffman / Vieira) e:

- (i) leiam o capítulo 2 (Regressão Linear Simples); e
- (ii) estudem o capítulo 4 (Regressão Linear Múltipla) para ajustar *consumo* em função de *idade* (grupo de idades) e *gênero*.