# Regressão Linear Simples

O estudo da regressão aplica-se àquelas situações em que há razões para supor uma relação de causa-efeito entre duas variáveis quantitativas e se deseja expressar matematicamente essa relação.

A análise de regressão estuda o relacionamento entre uma variável $Y$, chamada de variável dependente (variável resposta) e uma variável $X$, chamada de variável independente (variável explicativa). Este relacionamento é representado por um modelo matemático, isto é, por uma equação que associa a variável dependente com a variável independente. Este modelo é designado por modelo de regressão linear simples, cuja equação é:

$$Y = \alpha + \beta X$$

sendo:

$Y$ : variável dependente

$\alpha$ : coeficiente linear ou intercepto;

$\beta$ : coeficiente angular ou inclinação da reta

$X$ : variável independente

A reta de regressão (verdadeira) seria obtida se fossem conhecidos os valores de $X$ e $Y$ para todos os indivíduos da população. No entanto, o mais comum é estudar a regressão entre $X$ e $Y$ utilizando uma amostra da população. Portanto, devemos calcular $\hat{\alpha}$ e $\hat{\beta}$, que são estimativas de $\alpha$ e $\beta$, respectivamente.

O coeficiente $\beta$ é estimado da seguinte maneira:

$$\hat{\beta} = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{n \sum x_i^2 - (\sum x_i)^2}$$

O coeficiente linear $\alpha$ é estimado por:

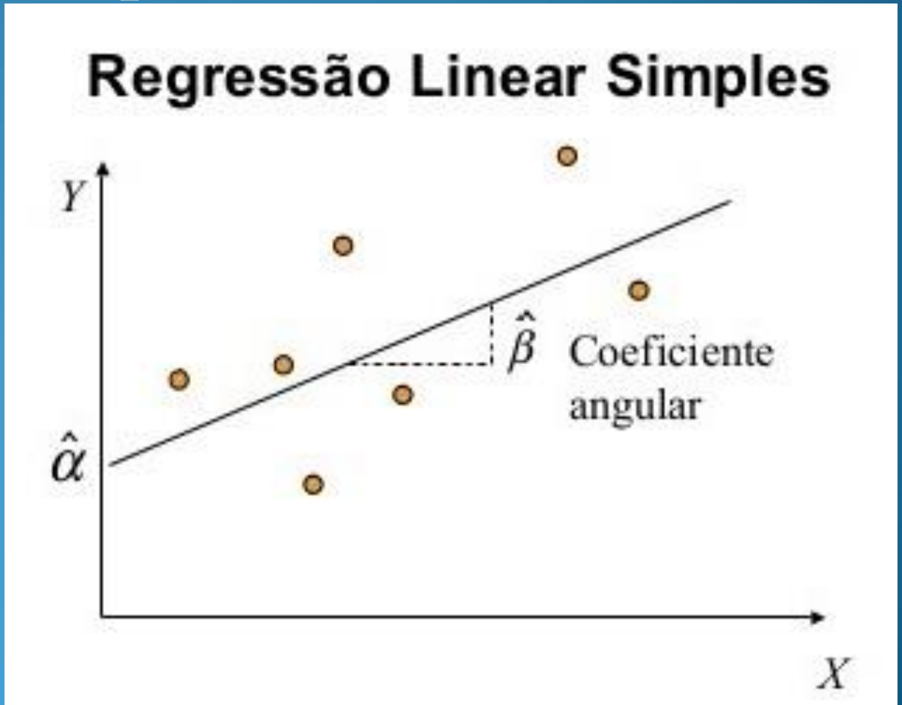
$$\hat{\alpha} = \bar{y} - \hat{\beta}\,\bar{x}$$

onde $\bar{y}$ e $\bar{x}$ são as médias amostrais de $Y$ e $X$, respectivamente.

A reta estimada de regressão linear simples (RLS) é:

$$\hat{Y} = \hat{\alpha} + \hat{\beta} X$$

Logo, $\hat{\alpha}$ é ponto onde a reta corta o eixo das ordenadas (eixo y). O coeficiente angular $\hat{\beta}$ representa o quanto varia a média de $Y$ para um aumento de uma unidade da variável $X$.



O coeficiente de determinação $R^2$ (que é o quadrado do coeficiente de correlação) é uma medida do poder explicativo do modelo. Dá a proporção da variação da variável dependente, *Y*, que é explicada em termos lineares pela variável independente, *X*.

Quanto melhor for o ajuste dos dados, maior será o valor de $R^2$ ($0 \leq R^2 \leq 1$). O coeficiente de determinação pode ser utilizado como uma medida da qualidade do ajustamento ou como medida da qualidade de confiança depositada na equação de regressão como instrumento de precisão.

A dependência de $Y$ em relação a $X$ é representada pelo coeficiente $\beta$, sendo $\beta$ estimado com base em uma amostra de dados. Para afirmar que $\hat{\beta}$ representa uma dependência real de $Y$ em relação a $X$ deve-se realizar um teste de hipóteses sobre a existência de regressão na população. O principal teste de interesse é verificar se $X$ influencia na resposta, o que é equivalente a testar:

$$H_0 : \beta = 0 \text{ (não existe RLS de } Y \text{ em } X\text{)}$$

$$H_1 : \beta \neq 0 \text{ (existe RLS de } Y \text{ em } X\text{)}$$

A estatística de teste é:

$$t = \sqrt{\frac{(n-2)R^2}{1-R^2}}$$

Rejeita-se $H_0$ quando $|t| \geq t_{\frac{\alpha}{2};\, n-2}$ . (obs: essa tabela foi dada na aula passada).

Exemplo: Em um experimento foram obtidos os resultados para teor de cálcio no solo (X) e a porcentagem de tubérculos maduros (Y). Obtenha a equação de regressão linear simples, calcule o coeficiente de determinação e conclua ao nível de 5% de significância.

| X | 0,2 | 0,3 | 0,4 | 0,5 | 0,7 | 0,8 | 1,0 | 1,1 | 1,3 |
|---|---|---|---|---|---|---|---|---|---|
| Y | 75 | 79 | 80 | 86 | 88 | 89 | 93 | 95 | 99 |

Temos que:

n = 9 $\sum x_i = 6{,}3$ $\sum y_i = 784$

$\sum x_i y_i = 572{,}7$ $\sum x_i^2 = 5{,}57$ $\sum y_i^2 = 68802$

$$\hat{\beta} = \frac{9 \text{ x } 572{,}7 - 6{,}3 \text{ x } 784}{9 \text{ x } 5{,}57 - (6{,}3)^2}$$

$$\hat{\beta} = \frac{5154{,}3 - 4939{,}2}{50{,}13 - 39{,}69} = \frac{215{,}1}{10{,}44} = 20{,}6$$

$$\hat{\alpha} = \frac{784}{9} - 20{,}6 \text{ x } \frac{6{,}3}{9} = 87{,}11 - 14{,}42 = 72{,}7$$

Assim, a reta de regressão estimada é:

$$\hat{Y} = 72{,}7 + 20{,}6\, X$$

Para calcular o coeficiente de determinação $R^2$, precisamos obter o coeficiente de correlação de Pearson, dado por:

$$r = \frac{9 \text{ x } 572{,}7 - 6{,}3 \text{ x } 784}{\sqrt{[9 \text{ x } 5{,}57 - (6{,}3)^2] \text{x} [9 \text{ x } 68802 - (784)^2]}}$$

$$r = \frac{5154{,}3 - 4939{,}2}{\sqrt{10{,}44 \text{ x } 4562}} = \frac{215{,}1}{218{,}24} = 0{,}9856$$

Portanto, $R^2 = (0{,}9856)^2 = 0{,}9714$, significando que 97,14% da variação de $Y$ é explicada pelo ajuste do modelo de regressão linear.

$H_0 : \beta = 0$ (não existe a regressão linear simples)

$H_1 : \beta \neq 0$ (existe a regressão linear simples)

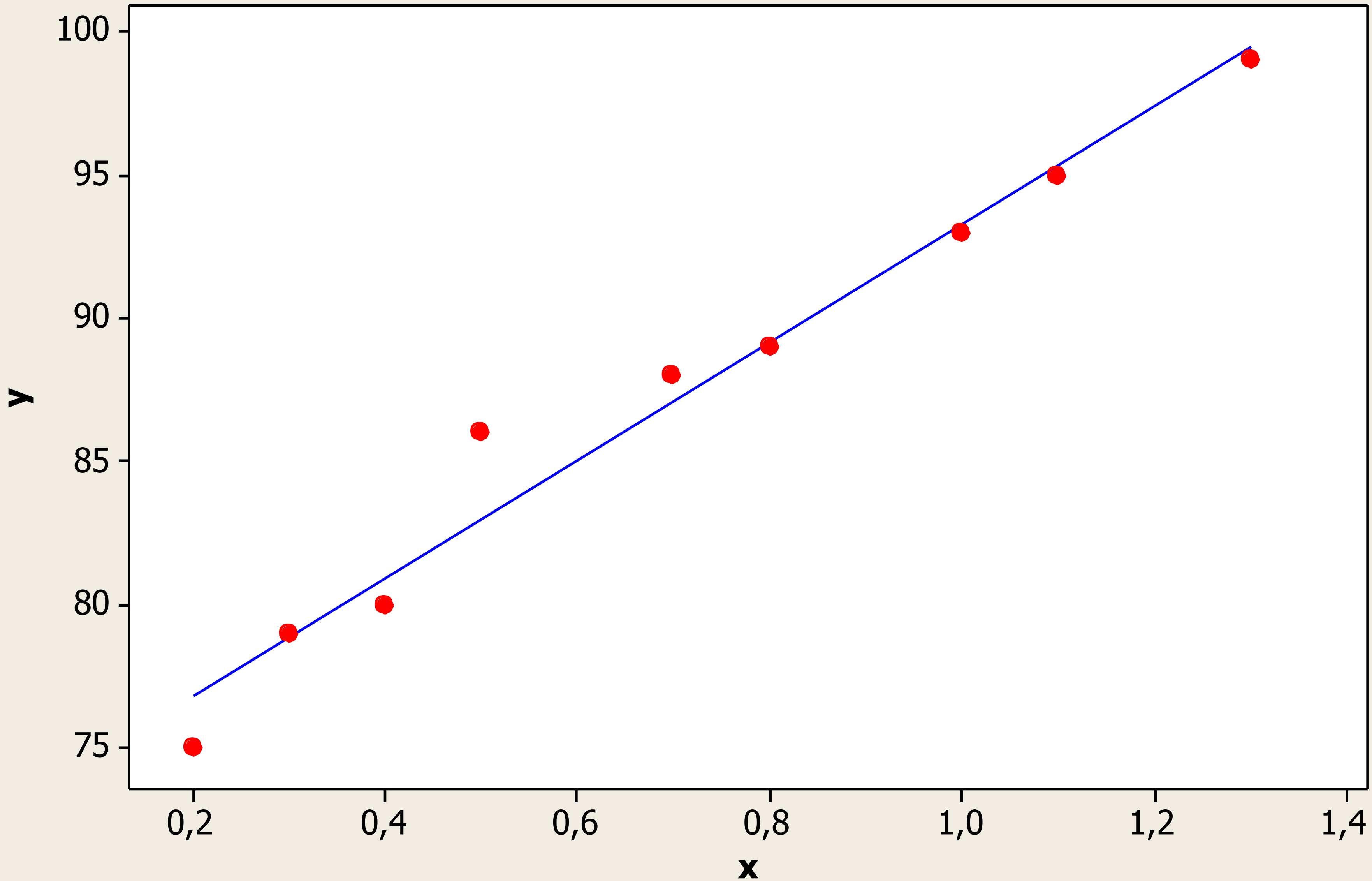
A estatística de teste é: $t = \sqrt{\dfrac{(9-2) \text{ x } 0{,}9714}{1-0{,}9714}} = 15{,}42$

Na tabela *t de Student* encontramos $t_{2,5\%;\,7} = 2{,}3646$.

Como $t = 15{,}42 > t_{2,5\%;\,7} = 2{,}3646$, rejeitamos $H_0$.

Portanto existe a regressão linear simples de *Y* em *X*.

A seguir mostramos o diagrama de dispersão e a reta de regressão estimada.


100
95
90
85
80
75
y
0,2
0,4
0,6
0,8
1,0
1,2
1,4
x