

Núcleos de Desenvolvimento

Regressão Linear Multivariada

Regressão linear multivariada é a expressão matemática que descreve uma variável como combinação linear de outras variáveis.

A ideia é escrever uma expressão matemática para uma variável y chamada de variável dependente, ou variável explicada, ou variável endógena, em termos de outras variáveis $x_1, x_2, \dots, x_n$ chamadas de variáveis independentes, ou variáveis explicativas, ou variáveis exógenas, de forma linear, ou seja:

$$Y = a_0 + a_1.X_1 + a_2.X_2 + \dots + a_n.X_n$$

Se obtivermos os valores de $Y_1, Y_2, \dots, Y_m$ e os respectivos valores das variáveis $X_{11}, X_{12}, \dots, X_{1m}; X_{21}, X_{22}, \dots, X_{2m}; \dots; X_{n1}, X_{n2}, \dots, X_{nm}$.

Então poderemos escrever:

$$Y_i = a_0 + a_1 \cdot X_{1i} + a_2 \cdot X_{2i} + \dots + a_n \cdot X_{ni}, \text{ com } i = 1, 2, \dots, m.$$

Em formato matricial:

$$Y = A \cdot X$$

$$X = \begin{bmatrix} 1 & x_{11} & x_{21} & \dots & x_{n1} \\ 1 & x_{12} & x_{22} & \dots & x_{n2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1m} & x_{2m} & \dots & x_{nm} \end{bmatrix}$$

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix} \quad A = \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_n \end{bmatrix}$$

$$Y = A * X$$

$$\begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \dots \\ \underline{Y_m} \end{pmatrix} = \begin{pmatrix} a_0 \\ a_1 \\ a_2 \\ \dots \\ a_m \end{pmatrix} * \begin{pmatrix} 1 & X_{11} & X_{21} & \dots & X_{n1} \\ 1 & X_{12} & X_{22} & \dots & X_{n2} \\ 1 & X_{13} & X_{23} & \dots & X_{n3} \\ \dots & \dots & \dots & \dots & \dots \\ 1 & \underline{X_{1m}} & \underline{X_{2m}} & \dots & \underline{X_{nm}} \end{pmatrix}$$

$$B = X * A + \varepsilon \text{ ou } Y = A * X + \varepsilon$$

$$\begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \dots \\ Y_m \end{pmatrix} = \begin{pmatrix} a_0 \\ a_1 \\ a_2 \\ \dots \\ a_n \end{pmatrix} * \begin{pmatrix} 1 & X_{11} & X_{21} & \dots & X_{n1} \\ 1 & X_{12} & X_{22} & \dots & X_{n2} \\ 1 & X_{13} & X_{23} & \dots & X_{n3} \\ \dots & \dots & \dots & \dots & \dots \\ 1 & X_{1m} & X_{2m} & \dots & X_{nm} \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ e_3 \\ \dots \\ e_m \end{pmatrix}$$

$$Y_1 = a_0 + a_1 * X_{11} + a_2 * X_{21} + a_3 * X_{31} + \dots + a_n * X_{n1} + e_1$$

$$Y_2 = a_0 + a_1 * X_{12} + a_2 * X_{22} + a_3 * X_{32} + \dots + a_n * X_{n2} + e_2$$

$$Y_3 = a_0 + a_1 * X_{13} + a_2 * X_{23} + a_3 * X_{33} + \dots + a_n * X_{n3} + e_3$$

$$\dots \quad \dots \quad \dots \quad \dots \quad \dots \quad \dots \quad \dots \quad \dots$$

$$Y_m = a_0 + a_1 * X_{1m} + a_2 * X_{2m} + a_3 * X_{3m} + \dots + a_n * X_{nm} + e_m$$

Especificação do modelo genérico:

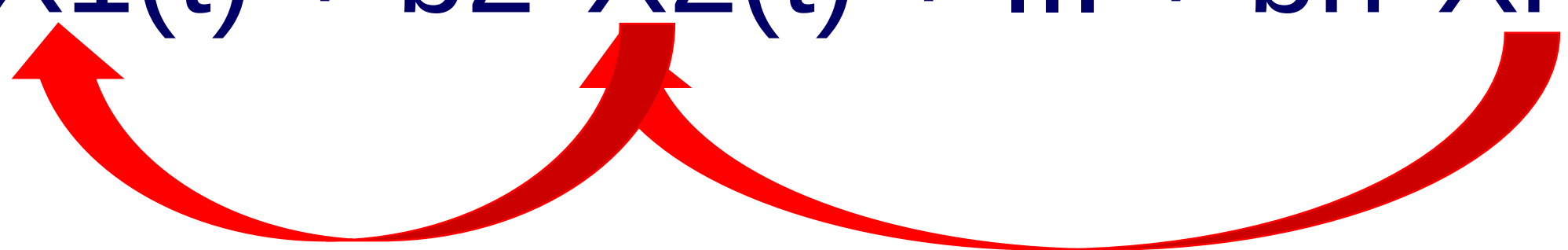
$$Y(t) = a + b1*X1(t) + b2*X2(t) + \dots + bn*Xn(t) + e(t)$$


Pressupostos para a estimação dos parâmetros (matriz A) de uma regressão linear multivariada pelo método dos mínimos quadrados:

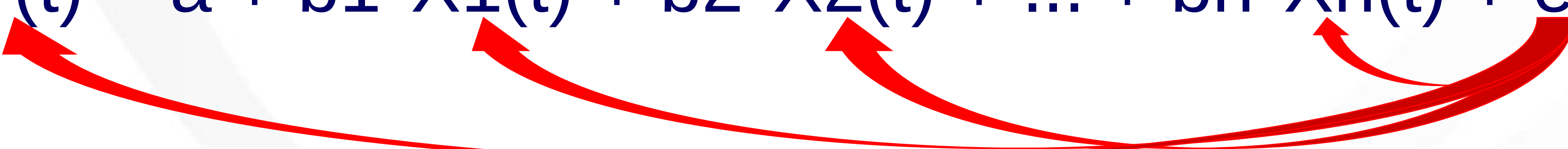
1. O tamanho da amostra deve ser grande ($n \geq 30$);
2. O modelo deve estar bem especificado, ou seja, as variáveis independentes (variáveis X's) que mais explicam a variação (ou variância) da variável dependente (Y) devem estar presentes no modelo (regressão);

3. Critério da parcimônia: o modelo deve conter o menor número possível de variáveis independentes (X's) que expliquem a maior parte da variação (ou variância) da variável dependente (Y);
4. A forma matemática do modelo (ou regressão) deve ser linear;
5. As variáveis independentes (X's) devem ser isentas de erro (desvio padrão);
6. O erro da variável dependente (Y) deve seguir distribuição de probabilidade normal com média em zero, ou seja $E(e)=0$, e variância $Var(e)$ diferente de zero ($e \sim N(0;\sigma^2_e)$);

7. Não pode haver multicolinearidade, que é a associação linear quase perfeita (em módulo, maior que 95%) entre 2 ou mais variáveis X 's;

$$Y(t) = a + b_1 * X_1(t) + b_2 * X_2(t) + \dots + b_n * X_n(t) + e(t)$$


8. Não pode haver heterocedasticidade, que é a associação entre o erro e quaisquer uma das variáveis X 's (o modelo tem de ser homocedástico);

$$Y(t) = a + b_1 * X_1(t) + b_2 * X_2(t) + \dots + b_n * X_n(t) + e(t)$$


9. Não pode haver autocorrelação serial, ou seja, $cov(e_t; e_{t-1})=0$, a correlação entre os erros no tempo deve ser nula.

Regressão Linear

Exemplo: Receita \$ x Quantidade

Admitindo que Y é a receita de uma empresa comercial em certo intervalo de tempo e que X é a quantidade vendida. Os dados que você tem são os seguintes

<u>Quantidade</u>	<u>Receita</u>
10	100
20	100
30	200
40	200
50	400



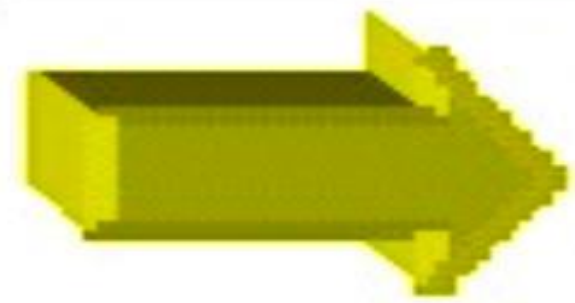
Qual a relação entre Receita e Quantidade?

Regressão Linear

Exemplo (Receita \$ x Quantidade) Tabela de Resultados

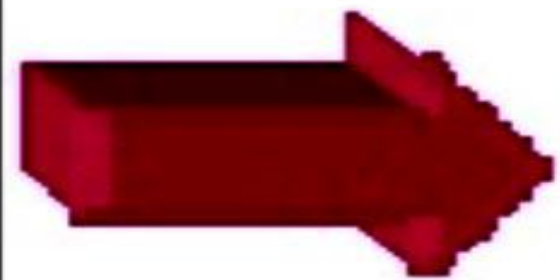
x_i	y_i	x_i^2	y_i^2	$x_i y_i$
10	100	100	10000	1000
20	100	400	10000	2000
30	200	900	40000	6000
40	200	1600	40000	8000
50	400	2500	160000	20000
150	1000	5500	260000	37000

Regressão Linear



$$S_{xx} = \sum (X_i - \bar{X})^2 = \sum X^2 - \frac{1}{n} (\sum X)^2$$

$$S_{xx} = 5500 - \frac{1}{5} (150)^2 = 1000$$



$$S_{yy} = \sum (Y_i - \bar{Y})^2 = \sum Y^2 - \frac{1}{n} (\sum Y)^2$$

$$S_{yy} = 260000 - \frac{1}{5} (1000)^2 = 60000$$



$$S_{xy} = \sum (X_i - \bar{X})(Y_i - \bar{Y}) = \sum XY - \frac{1}{n} \sum X \sum Y$$

$$S_{xy} = 37000 - \frac{1}{5} 150 \cdot 1000 = 7000$$

Regressão Linear

Exemplo (Receita \$ x Quantidade) Cálculos

$$S_{xx}=1000$$

$$S_{yy}=60000$$

$$S_{xy}=7000$$

$$B = \frac{S_{xy}}{S_{xx}} = \frac{7000}{1000} = 7$$

Declive

$$A = \bar{Y} - B\bar{X} =$$

$$A = 200 - 7.30 = -10$$

Interseção

$$\hat{y}_i = -10 + 7x_i$$

Regressão Linear

Erro padrão da Estimativa

Observação : supõe dispersão uniforme dos pontos em torno da reta de regressão



$$S_E = \sqrt{\frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{n - 2}}$$

EQUIVALENTES



$$S_E = \sqrt{\frac{S_{yy} - BS_{xy}}{n - 2}}$$

Regressão Linear

Erro Padrão do Coeficiente
Linear (Intercepto)

$$S_A = S_E \sqrt{\frac{1}{n} + \frac{\bar{X}^2}{S_{XX}}}$$

$$S_A = 60,55 \sqrt{\frac{1}{5} + \frac{30^2}{1000}} = 63,51$$

Regressão Linear

Exemplo (Receita \$-Quantidade) Teste do Coeficiente de Linear

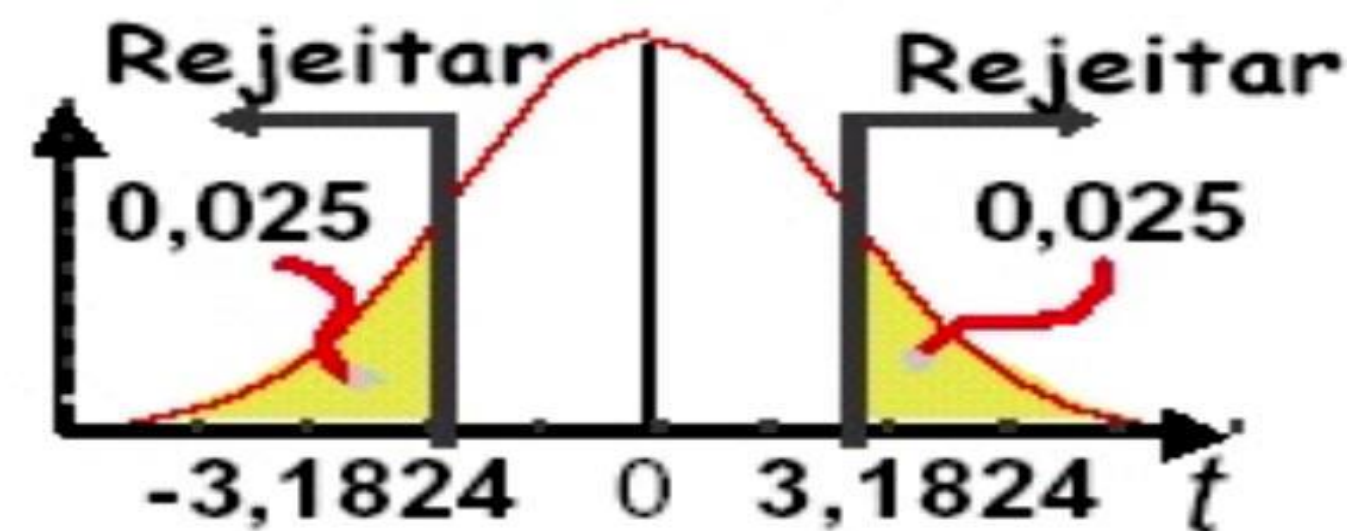
$$H_0: \alpha = 0$$

$$H_1: \alpha \neq 0$$

$$\alpha = 0,05$$

$$g.l. = 5 - 2 = 3$$

Valores Críticos:



Estatística Teste:

$$T = \frac{A - \alpha}{S_A} = \frac{-10 - 0}{63,51}$$

$$T = -0,16$$

Decisão:

Rejeitar se $T > 3,1824$

Conclusão:

Não devemos rejeitar a hipótese H_0

Regressão Linear

Intervalos de Confiança para o Intercepto

$$P(A - tS_A \leq \alpha \leq A + tS_A) = \gamma$$

Onde $t = t_{n-2, \alpha}$ tabela t-Student

Determine o intervalo de Confiança
de 95% para o Declive

$$\alpha : -10 \pm 3,18 . 63,51$$

$$P(-212,11 \leq \alpha \leq 192,11) = 0,95$$

Regressão Linear

Erro Padrão do Coeficiente
Angular (Declive)

$$S_B = \frac{S_E}{\sqrt{S_{xx}}}$$

$$S_B = \frac{60,55}{\sqrt{1000}} = 1,92$$

Regressão Linear

Exemplo (Receita \$ - Quantidade) Teste do Coeficiente de Declive

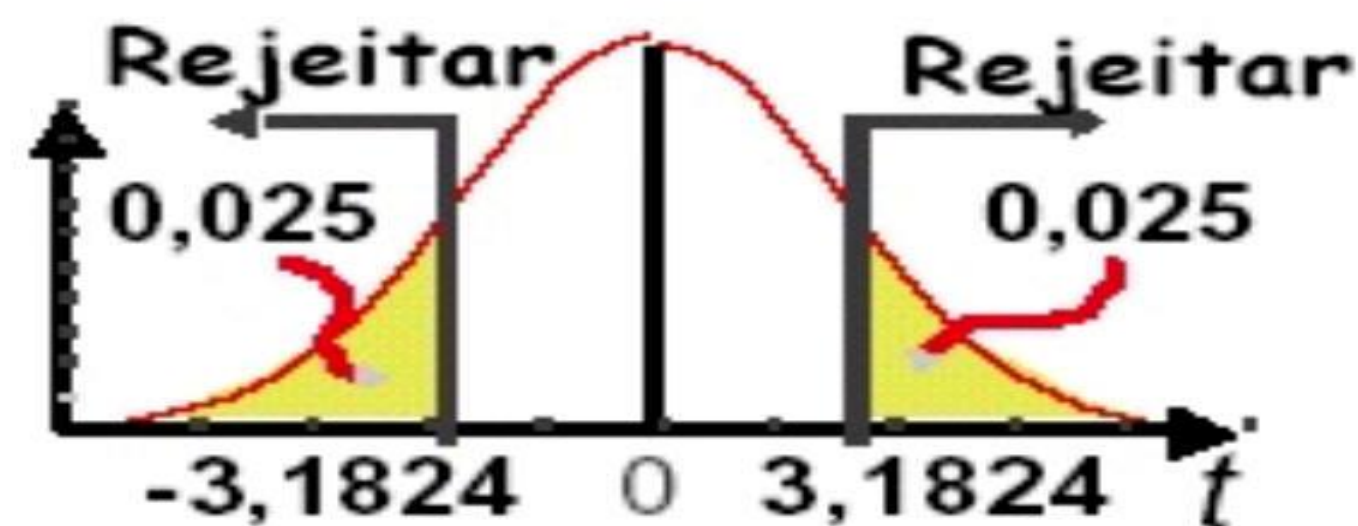
$$H_0: \beta = 0$$

$$H_1: \beta \neq 0$$

$$\alpha = 0,05$$

$$g.l. = 5 - 2 = 3$$

Valores Críticos:



Estatística Teste:

$$T = \frac{B - \beta}{S_B} = \frac{7 - 0}{1,92} = 3,66$$

Decisão:

Rejeitar se $T > 3,1824$

Conclusão:

Devemos rejeitar a hipótese H_0

Regressão Linear

Intervalos de Confiança para o Declive

$$P(B - tS_B \leq \beta \leq B + tS_B) = \gamma$$

Onde $t = t_{n-2, \alpha}$ tabela t- Student

Determine o intervalo de Confiança de 95%
para o Declive

$$\beta : 7 \pm 3,18. 1,92$$

$$P(0,91 \leq \beta \leq 13,09) = 0,95$$

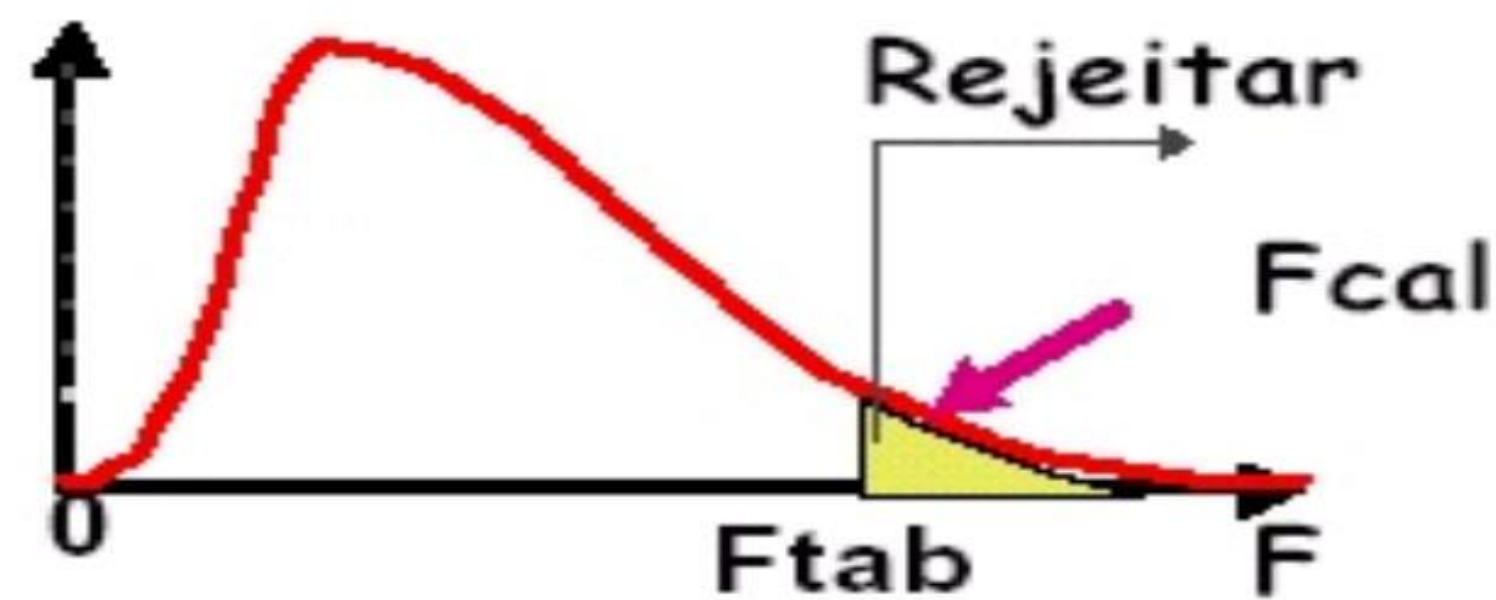
Regressão Linear

Hipóteses

$$\begin{cases} H_0 : \beta = 0 \rightarrow \text{não existe regressão} \\ H_1 : \beta \neq 0 \rightarrow \text{existe regressão} \end{cases}$$

Quadro de Análise de Variância

F.V.	G.L.	S.Q.	Q.M.	F
Regressão	1	BS_{xy}	BS_{xy}	$F_{col} = \frac{BS_{xy}}{S_E^2}$
Resíduo	$n-2$	$S_{yy} - BS_{xy}$	$S_E^2 = \frac{S_{yy} - BS_{xy}}{n-2}$	
Total	$n-1$	S_{yy}		



Rejeita-se H_0 se

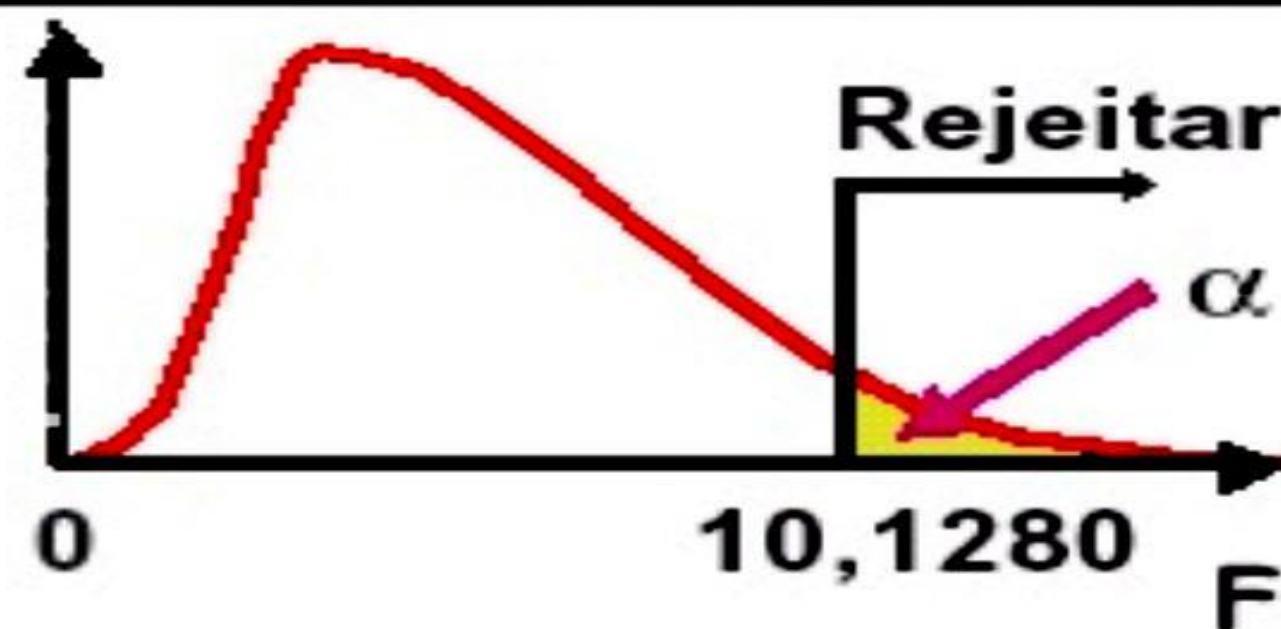
$$F_{cal} > F_{tab} \quad \text{onde}$$
$$F_{tab} = F_{1, n-2} \quad \text{Existe}$$

Regressão

Regressão Linear

$$\begin{cases} H_0 : \beta = 0 \rightarrow \text{não existe regressão} \\ H_1 : \beta \neq 0 \rightarrow \text{existe regressão} \end{cases}$$

ANOVA					
	<i>gl</i>	<i>SQ</i>	<i>MQ</i>	<i>F</i>	<i>de significação</i>
Regressão	1	49000	49000	13,3636	0,0354
Resíduo	3	11000	3666,67		
Total	4	60000			



Rejeita-se H_0 se
 $F_{cal} = 13,3636 > F_{tab} = 10,1280$
Existe Regressão

Regressão Linear

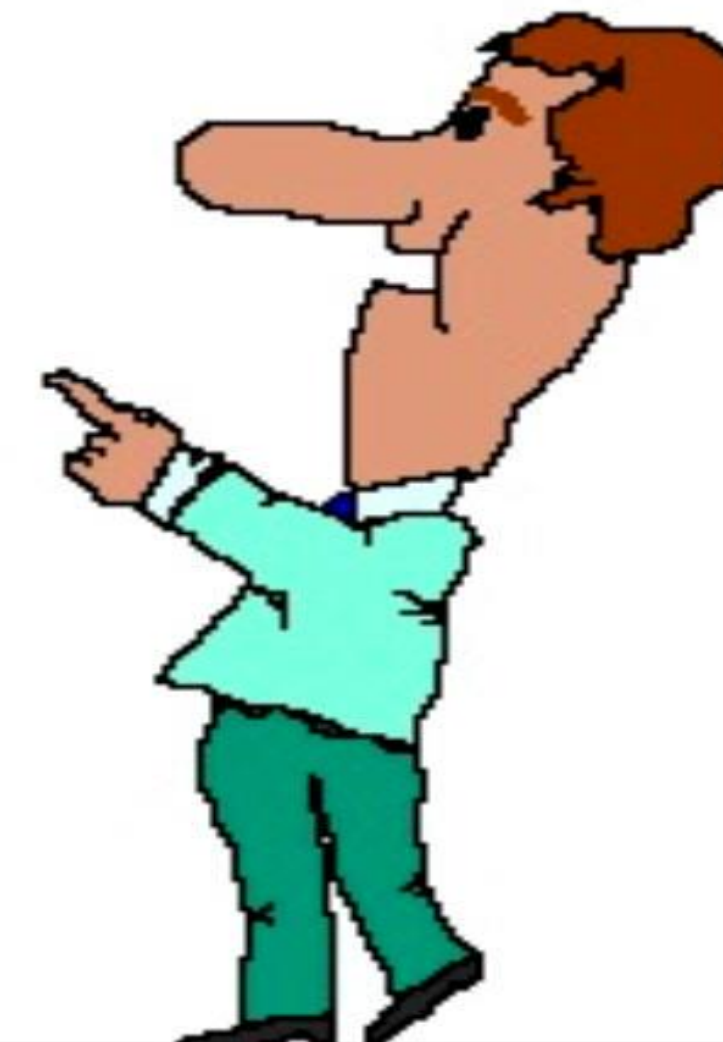
Coeficiente de Determinação

Proporção da Variação ' Explicada '
 pela relação entre X e Y

$$R^2 = \frac{\text{Variação Explicada}}{\text{Variação Total}} = \frac{SQ\text{ Reg}}{SQT}$$

$$0 \leq R^2 \leq 1$$

$$R^2 = \frac{BS_{xy}}{S_{yy}} = \frac{7.70000}{600000} = 0,82$$



Regressão Linear

Calcule o coeficiente de determinação ajustado

Leva em conta o número de variáveis e o tamanho da amostra

$$R^2_{\text{ajustado}} = 1 - \frac{(n - 1)}{[n - k - 1]} (1 - R^2)$$

n: número de observações

k: número de variáveis independentes

Regressão Linear

	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Receita		Receita Estimada pelo Modelo		Erro		MAPE	=	Mean Absolut Percentual Error					
2	100		60		40		0,4							
3	100		130		-30		0,3							
4	200		200		0		0							
5	200		270		-70		0,35							
6	400		340		60		0,15							
7				Média	0		0,24							
8														
9	Receita = -10 + 7*Vendas + erro													
10														

Regressão Linear

	A	B	C	D	E	F	G	H	I
1	RESUMO DOS RESULTADOS								
2									
3	Estatística de regressão		Receita = -10 + 7*Vendas + erro				Intervalos de Confiança		
4	R múltiplo	0,903696114	100% VAR(Y) 75,56% VAR(Y) 24,44% VAR(Y)						
5	R-Quadrado	0,816666667					P [-212.1124854 < Alfa < 192,1124854		
6	R-quadrado ajustado	0,755555556							
7	Erro padrão	60,55300708					P [0,906079277 < Beta < 13,09392072		
8	Observações	5							
9									
10	ANOVA								
11		gl	SQ	MQ	F	F de significação	O R ² ajustado é estatisticamente significativo a um nível de significância de 3,53524847% < 0,05 (5%). Portanto a regressão é estatisticamente significativa a um nível de significância de 5%.		
12	Regressão	1	49000	49000	13,36363636	0,035352847			
13	Resíduo	3	11000	3666,666667					
14	Total	4	60000						
15									
16		Coeficientes	Erro padrão	Stat t	valor-P	95% inferiores	95% superiores	Inferior 95,0%	Superior 95,0%
17	Interseção	-10	63,50852961	-0,157459164	0,88488398	-212,1124854	192,1124854	-212,1124854	192,1124854
18	Vendas	7	1,914854216	3,655630775	0,035352847	0,906079277	13,09392072	0,906079277	13,09392072
19									
20	O intercepto (-10) não é estatisticamente significativo a um nível de significância de 5% porque o seu valor-P é maior que 5%								
21	(88,488398%). O coeficiente que multiplica as vendas é estatisticamente significativa a 5% porque o seu valor-P é menor que								
22	5% (3,655630775%).								
23									

Dados

Regressão

Obrigado pela atenção!