CORRELAÇÃO E REGRESSÃO

CORRELAÇÃO E REGRESSÃO

1. CORRELAÇÃO

1.1. INTRODUÇÃO

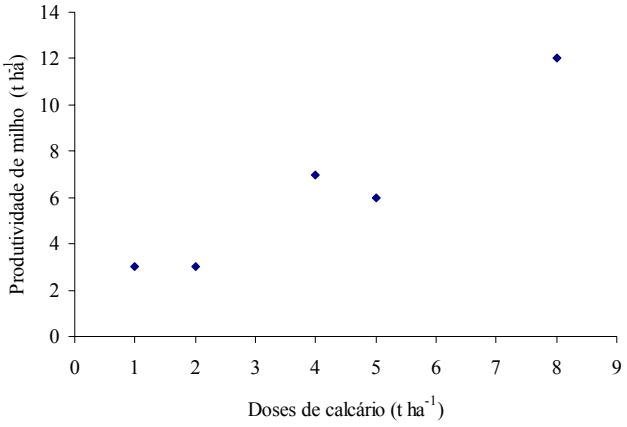
Ao se estudar uma variável o interesse eram as medidas de tendência central, dispersão, assimetria, etc. Com duas ou mais variáveis além destas medidas individuais também é de interesse conhecer se elas tem algum relacionamento entre si, isto é, se valores altos (baixos) de uma das variáveis implicam em valores altos (ou baixos) da outra variável. Por exemplo, pode-se verificar se existe associação entre a altura de planta e altura de espiga, dose de fertilizante e produtividade, taxa de infestação de uma doença e a produtividades, etc.

A associação entre duas variáveis poder ser de dois tipos: **experimental e correlacional.**

Experimental

Aplico algum tratamento.

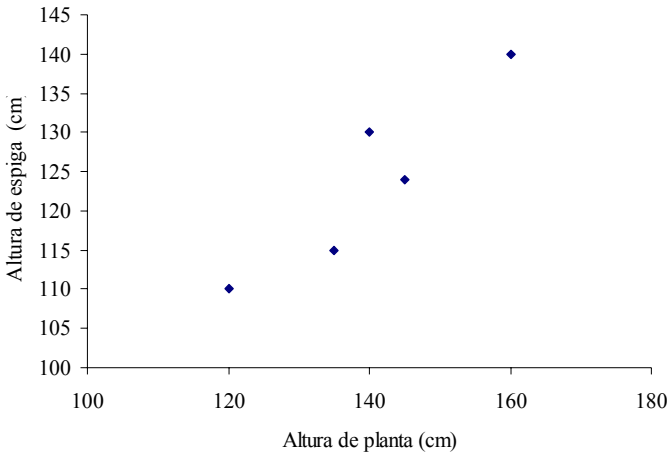
Exemplo: Aplicar doses de calcário (tratamentos) e observar as diferenças na produtividade de milho.



Correlacional

Não se tem nenhum controle sobre as variáveis sendo estudadas. Elas são observadas como ocorrem no ambiente natural, sem nenhuma interferência, isto é, as duas variáveis são aleatórias.

Exemplo: Altura de planta e alura de espiga.



Ao estudo do relacionamento entre duas ou mais variáveis denominamos de **correlação** e **regressão.**

Se o estudo tratar apenas de duas variáveis tem-se a correlação e a regressão simples, se envolver mais do que duas variáveis, tem-se a correlação e a regressão múltiplas.

A regressão e a correlação tratam apenas do relacionamento do tipo linear entre duas variáveis.

A análise de correlação fornece um número que resume o *grau de relacionamento linear* entre as duas variáveis.

Já a análise de regressão fornece uma *equação que descreve o comportamento* de uma das variáveis em função do comportamento da outra variável.

1.2. PADRÕES DE ASSOCIAÇÃO

Quando não é possível perceber uma relação sistemática entre as variáveis é dito que as variáveis são **não correlacionadas**, são **independentes** ou ainda que são **ortogonais**.

1.3. INDICADORES DE ASSOCIAÇÃO

- Tabela de contingência 2x2 – dá uma idéia muito superficial da associação entre as variáveis. Ver exemplo na apostila.

- Diagramas de dispersão – Fornecem uma idéia melhor da associação entre as variáveis.

Vamos considerar o seguinte exemplo para o estudar a correlação e regressão

linear entre duas variáveis.

Vamos supor que você fez um experimento avaliando o efeito da aplicação de doses de calcário na produtividade de milho. Estou interessado em saber se existe associação linear entre essas variáveis e se posso estimar uma regressão linear entre essas variáveis. Os dados obtidos foram os seguintes:

Tabela 1 – Doses de calcário (t ha^{-1}) e produtividade de milho (t ha^{-1}).

Doses de calcário (t ha^{-1}) (X)	Produtividade de milho (t ha^{-1}) (Y)
1	3
2	3
4	7
5	6
8	12

Para ter uma idéia melhor, as variáveis são colocadas no que é denominado de **diagrama de dispersão.** Uma das variáveis (X) é representada no eixo horizontal e a outra variável (Y) no eixo vertical, conforme figura 1.

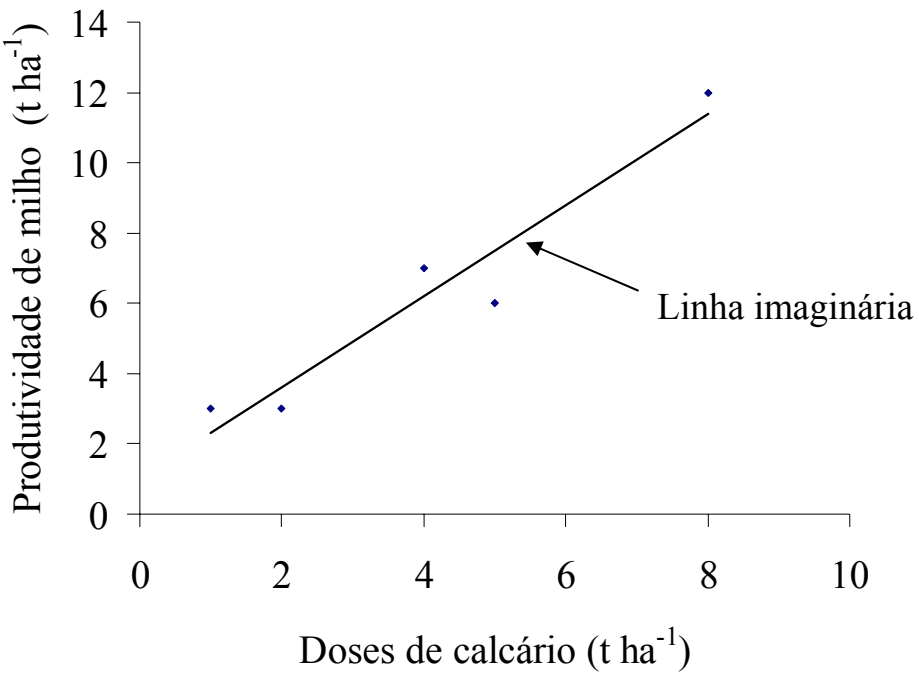


Figura 1 - Diagrama de dispersão das variáveis apresentadas na tabela 1.

Uma olhada rápida no diagrama de dispersão mostra a existência de um relacionamento entre as variáveis, com altos valores de uma das variáveis associados a altos valores da outra variável. Se não houvesse relacionamento entre elas, os pontos estariam distribuídos ao acaso no gráfico sem mostrarem alguma tendência.

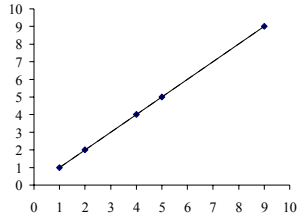
1.4. O COEFICIENTE DE CORRELAÇÃO LINEAR

ou coeficiente de correlação linear de Pearson

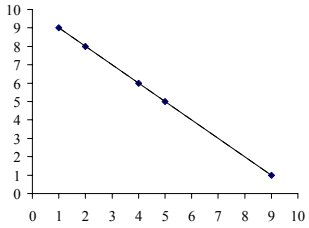
Apesar do diagrama de dispersão nos fornecer uma idéia do tipo e extensão do relacionamento entre duas variáveis X e Y, seria altamente desejável ter um número que medisse esta relação.

Esta medida existe e é denominada de **coeficiente de correlação**.

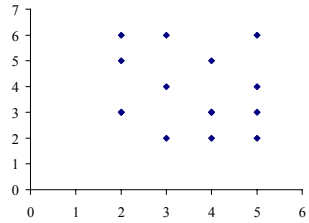
- ✓ Coeficiente de correlação amostral é indicado pela letra **r.**
- ✓ Coeficiente de correlação populacional é indicado pela letra **ρ (rho).**



Correlação linear positiva perfeita (r =1)



Correlação linear negativa perfeita (r = -1)



Ausência de relacionamento linear (r = 0)

PROPRIEDADES DO COEFICIENTE DE CORRELAÇÃO LINEAR (r)

As propriedades mais importantes do coeficiente de correlação são:

1. O intervalo de variação vai de -1 a +1. $-1 \leq r \leq 1$

2. O coeficiente de correlação é uma medida adimensional, isto é, ele é independente das unidades de medida das variáveis X e Y.

3. Quanto mais próximo de +1 for "r", maior o grau de relacionamento linear positivo entre X e Y, ou seja, se X varia em uma direção Y variará na mesma direção.

4. Quanto mais próximo de -1 for "r", maior o grau de relacionamento linear negativo entre X e Y, isto é, se X varia em um sentido Y variará no sentido inverso.

5. Quanto mais próximo de zero estiver "r" menor será o relacionamento linear entre X e Y.

1.5. HIPÓTESES BÁSICAS

A suposição básica sobre o coeficiente de correlação é que o relacionamento entre as duas variáveis seja linear. Isto é, o coeficiente de correlação é adequado para avaliar somente o relacionamento linear.

Uma segunda hipótese é que as variáveis envolvidas sejam aleatórias e que sejam medidas no mínimo em escala de intervalo. Ele não se aplica a variáveis em escala nominal ou ordinal ou quando uma das variáveis é manipulada experimentalmente, pois neste caso, a escolha dos valores experimentais vai influenciar o valor de r obtido.

Uma terceira hipótese é que as duas variáveis tenham uma distribuição conjunta normal bivariada. Isto é equivalente a dizer que para cada x dado a variável y é normalmente distribuída.

1.6. DEFINIÇÃO

O coeficiente de correlação amostral (r) pode ser calculado através da seguinte expressão:

$$r = \frac{n(\sum XY) - (\sum X)(\sum Y)}{\sqrt{n(\sum X^2) - (\sum X)^2}\sqrt{n(\sum Y^2) - (\sum Y)^2}}$$

Nosso exemplo

X = Doses de calcário (t ha^{-1})
Y = Produtividade de milho (t ha^{-1})

	X	**Y**	**X^2**	**Y^2**	**XY**
	1	3	1	9	3
	2	3	4	9	6
	4	7	16	49	28
	5	6	25	36	30
	8	12	64	144	96
somatório	20	31	110	247	163
	ΣX=20	ΣY=31	ΣX^2=110	ΣY^2=247	ΣXY=163

$$r = \frac{5(163) - (20)(31)}{\sqrt{5(110) - (20)^2}\sqrt{5(247) - (31)^2}} = 0{,}9618$$

1.7. TESTE DE HIPÓTESE PARA O COEFICIENTE LINEAR (r)

Para o nosso exemplo queremos testar se existe ou não correlação linear entre X = Doses de calcário (t ha^{-1}) e Y = produtividade de milho (t ha^{-1}). Vamos verificar se existe relacionamento linear entre as duas variáveis ao nível de 5% de significância.

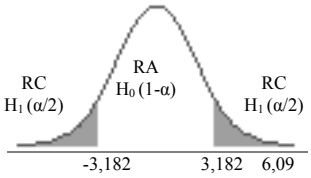
a) As hipóteses a serem testadas são:
H_0: $\rho = 0$ (Não existe relacionamento linear na população)
H_1: $\rho \neq 0$ (Existe relacionamento linear na população)
(teste bilateral)

b) Nível de significância do teste, ou erro tipo I ($\alpha = 5\%$)

c) Teste estatístico

$$t_{calc} = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} \rightarrow t_{calc} = \frac{0{,}9618}{\sqrt{\frac{1-0{,}9618^2}{5-2}}} = 6{,}09$$ Comparar com o valor de $t_{tab\ com\ n-2\ graus\ de\ liberdade}$ (Tabela 2)

d) Comparar com o valor de $t_{tab\ 5\%\ bilateral}$ ou $t_{tab\ 2{,}5\%\ unilateral\ com\ (5-2)\ =\ 3\ gl}$ ➔ $t_{tab} = 3{,}182$



Lado esquerdo ➔ de menos infinito até -3,182 acumula uma área de 2,5% (Tabela 2).
Lado direito ➔ de 3,182 até mais infinito acumula uma área de 2,5% (Tabela 2).

e) Conclusão: t_{calc} (6,09) está dentro de H_1, logo rejeitamos H_0 e podemos afirmar, ao nível de 5% de significância, ou 5% de probabilidade de erro, que existe relacionamento linear na população, ou seja, o aumento da dose de calcário ocasiona um aumento na produtividade de milho. Ou ainda a correlação linear entre essas duas variáveis é significativa.

Dado que há fortes evidências de que as duas variáveis possuem um relacionamento linear pode-se então ajustar uma linha de regressão entre elas.

Exemplo de como apresentar uma tabela de correlação em um artigo, resumo, etc.

Tabela 2 - Matriz de coeficientes de correlação genotípica de Pearson entre os caracteres produtividade de grãos (PROD), número de vagens por planta (NVP), número de sementes por vagem (NSV), massa de cem grãos (MCG), população final de plantas (POP), número de dias da emergência ao florescimento (FLOR), número de dias da emergência à colheita (CICLO), altura de inserção da primeira vagem (A1V), altura de inserção da última vagem (AUV), grau de acamamento (ACA) e coloração do tegumento dos grãos (COR) de 14 cultivares de feijão, avaliadas em nove experimentos. (1) Santa Maria-RS, UFSM, 2005.

Caráter	NVP	NSV	MCG	POP	FLOR	CICLO	A1V	AUV	ACA (2)	COR
PROD (kg ha^{-1})	0,69**	-0,20**	0,31**	0,02 ns	-0,50**	-0,61**	-0,68**	-0,43**	-0,49**	0,04 ns
NVP		0,34**	-0,37**	-0,31**	0,03 ns	-0,29**	-0,54**	-0,24**	0,12 *	-0,40**
NSV			-0,94**	-0,25**	0,83**	0,69**	0,43**	0,60**	0,40**	-0,73**
MCG (g)				0,25**	-0,82**	-0,60**	-0,36**	-0,46**	-0,58**	0,71**
POP (plantas 3m^{-2})					-0,50**	-0,36**	-0,10 ns	-0,48**	-0,62**	-0,32**
FLOR (dias)						0,88**	0,70**	0,79**	0,65**	-0,31**
CICLO (dias)							0,91**	0,86**	0,43**	-0,16**
A1V (cm)								0,89**	0,30**	-0,08 ns
AUV (cm)									0,46**	-0,12*
ACA (2)										-0,09 ns

(1) ns Não-significativo. **, * significativo a 1 e 5% de probabilidade, respectivamente, pelo teste t com 376 graus de liberdade. (2) Escala de notas entre um a nove (1 = planta ereta; 9 = planta acamada).

2. REGRESSÃO

Uma vez constatado que existe correlação linear entre duas variáveis, pode-se tentar prever o comportamento de uma delas em função da variação da outra.

A variável **X** (denominada variável controlada, explicativa ou independente) com valores observados X1, X2, ..., Xn e a variável **Y** (denominada variável dependente ou explicada) com valores Y1, Y2, ..., Yn.

Desta forma pode-se considerar que o modelo para o relacionamento linear entre as variáveis X e Y seja representado por uma equação do tipo:

- ✓ População (parâmetros): $Y = \alpha + \beta X + U_i$,
- ✓ Amostra (estimadores): $\hat{Y} = a + bX + e_i$,

onde "U" é o termo erro, isto é, "U" representa as outras influências na variável Y além da exercida pela variável "X".

Esperança de erro é zero ➔ E(Ui) = 0; logo Y = a + bX

2.1. ESTIMATIVA DOS PARÂMETROS DE REGRESSÃO

Se fosse conhecido toda a população de valores (Xi, Yi) então seria possível determinar os valores exatos dos parâmetros α e β. Como, em geral, se trabalha com amostras se faz necessário, então, estimar estes parâmetros com base nos valores da amostra.

Existem alguns métodos para ajustar uma linha entre as variáveis X e Y o mais utilizado é o denominado **método dos mínimos quadrados (MMQ).**

O método dos mínimos quadrados exige que os estimadores **a** e **b** sejam escolhidos de tal forma que a **soma dos quadrados dos desvios dos mesmos em relação à reta de regressão ajustada seja mínima**.

Os valores de "a" e "b" são obtidos através das seguintes expressões:

$$b = \frac{n(\sum XY) - (\sum X)(\sum Y)}{n(\sum X^2) - (\sum X)^2}$$

$$a = \frac{\sum Y - b(\sum X)}{n}$$

Nosso exemplo

X = Doses de calcário (t ha^{-1})
Y = Produtividade de milho (t ha^{-1})

	X	Y	X^2	Y^2	XY
	1	3	1	9	3
	2	3	4	9	6
	4	7	16	49	28
	5	6	25	36	30
	8	12	64	144	96
somatório	20	31	110	247	163
	ΣX=20	ΣY=31	ΣX^2=110	ΣY^2=247	ΣXY=163

$$b = \frac{5(163) - (20)(31)}{5(110) - (20)^2} = \frac{195}{150} = 1{,}3$$

$$a = \frac{31 - 1{,}3(20)}{5} = 1{,}0$$

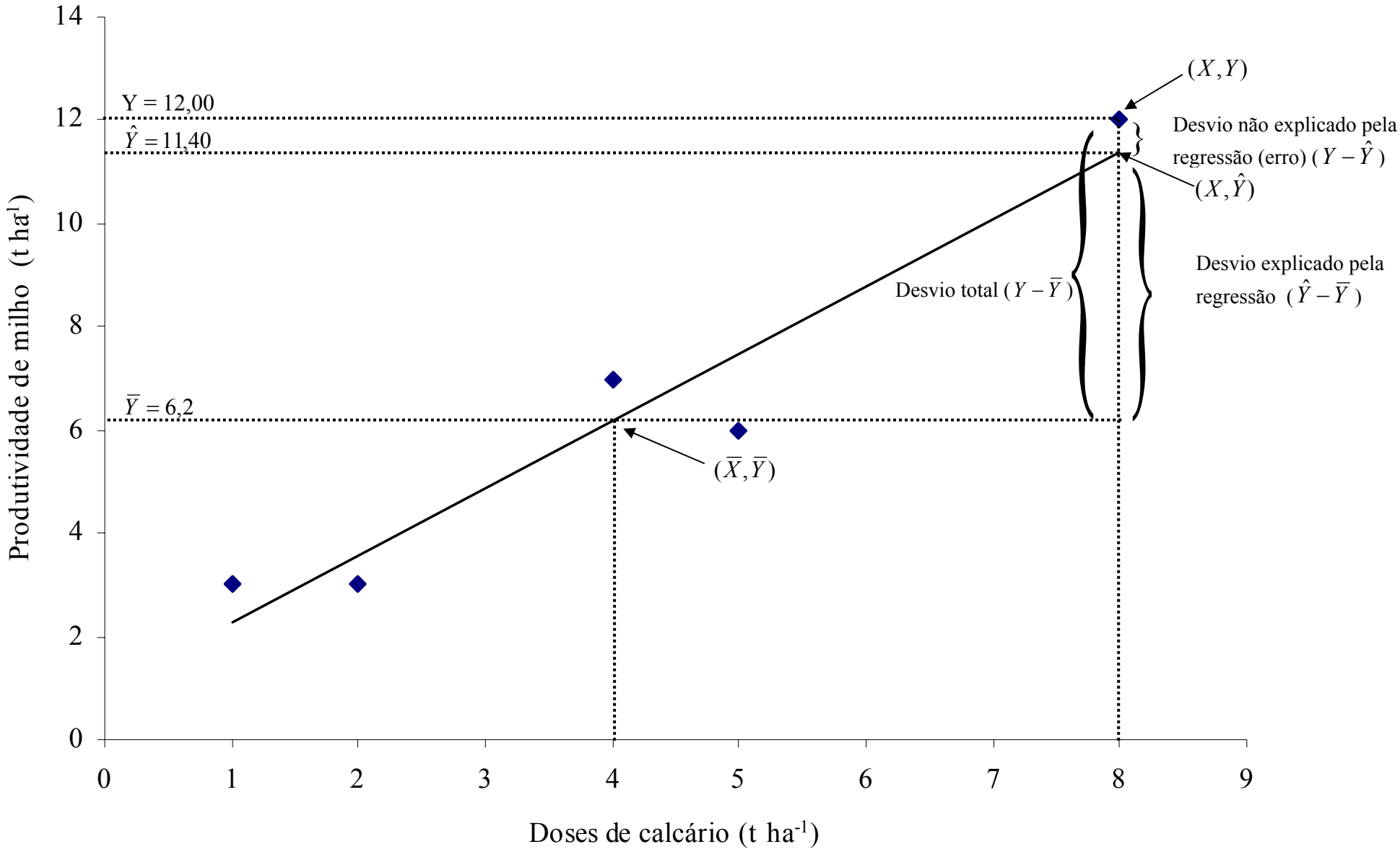
Então a linha estimada será: $\hat{Y}$= 1,0 + 1,3X

Esta reta é o "melhor" ajustamento para estes dados e seria diferente para cada amostra das variáveis X e Y, retiradas desta mesma população. Esta reta pode ser considerada uma estimativa da verdadeira linha de regressão onde 1,3 seria uma estimativa do valor β (parâmetro angular) e 1,0 uma estimativa do valor α (parâmetro linear), que são os verdadeiros coeficientes de regressão.

2.2. DECOMPOSIÇÃO DA SOMA DOS QUADRADOS

2.2.1. DECOMPOSIÇÃO DOS DESVIOS

Pela figura abaixo, pode-se perceber que o desvio em relação a Y (desvio total), pode ser decomposto em dois outros desvios (não explicado pela regressão (erro) e explicado pela regressão. Somando cada um dos desvios ao quadrado vamos obter as somas de quadrados total, do erro e da regressão.


14
12
10
8
6
4
2
0
0
1
2
3
4
5
6
7
8
9
Produtividade de milho (t ha-1)
Doses de calcário (t ha-1)
Y = 12,00
$\hat{Y}$ = 11,40
$\overline{Y}$ = 6,2
(X, Y)
$(X, \hat{Y})$
$(\overline{X}, \overline{Y})$
Desvio não explicado pela regressão (erro) $(Y - \hat{Y})$
Desvio total $(Y - \overline{Y})$
Desvio explicado pela regressão $(\hat{Y} - \overline{Y})$

2.2.2. CÁLCULO DAS VARIAÇÕES (somas de quadrados)

X = Doses de calcário (t ha^{-1})

Y = Produtividade de milho (t ha^{-1}) $\hat{Y}$ =1+1,3X Erro SQ_{erro} SQ_{total} $SQ_{regressão}$

	X	Y	X^2	Y^2	XY	$\hat{Y}$	E=Y-$\hat{Y}$	E^2	$(Y-\overline{Y})^2$	$(\hat{Y}-\overline{Y})^2$
	1	3	1	9	3	2,3	0,7	0,49	10,24	15,21
	2	3	4	9	6	3,6	-0,6	0,36	10,24	6,76
	4	7	16	49	28	6,2	0,8	0,64	0,64	0,00
	5	6	25	36	30	7,5	-1,5	2,25	0,04	1,69
	8	12	64	144	96	11,4	0,6	0,36	33,64	27,04
somatório	20	31	110	247	163	31	0	4,1	54,8	50,7
	ΣX	ΣY	ΣX^2	ΣY^2	ΣXY	Σ$\hat{Y}$	ΣE	ΣE^2	$\Sigma(Y-\overline{Y})^2$	$\Sigma(\hat{Y}-\overline{Y})^2$

ESTIMATIVA do "erro-padrão da estimativa" ou "erro-padrão amostral da regressão" (S)

$$S = \sqrt{\frac{\sum E^2}{n-2}} = \sqrt{\frac{4{,}10}{5-2}} = 1{,}17$$

ESTIMATIVA da soma de quadrados dos desvios de X (Sxx)

$$S_{XX} = \sum (X_i - \overline{X})^2$$

$$S_{XX} = (1-4)^2 + (2-4)^2 + (4-4)^2 + (5-4)^2 + (8-4)^2 = 30$$

2.3. INTERVALOS DE CONFIANÇA

2.3.1. INTERVALO PARA O COEFICIENTE LINEAR (α)

1 - α = 95% ➔ probabilidade de que o intervalo contenha o parâmetro, no caso o coeficiente linear (α)
α = 5% ➔ probabilidade de que o intervalo não contenha o parâmetro (probabilidade de erro), no caso o coeficiente linear (α)

$$P\left(a - t_{\alpha/2}.S\sqrt{\frac{1}{n}+\frac{\bar{X}^2}{Sxx}} \leq \alpha \leq a + t_{\alpha/2}.S\sqrt{\frac{1}{n}+\frac{\bar{X}^2}{Sxx}}\right) = 1 - \alpha$$

Ver valor de $t_{\alpha/2}$. com n-2 graus de liberdade na tabela t de Student.

O valor de $t_{tab\ 5\%\ bilateral\ ou\ t_{tab\ 2,5\%\ unilateral\ com\ (5-2)\ =\ 3\ gl}}$ é 3,182
Logo:

$$P\left(1 - 3{,}182.1{,}17\sqrt{\frac{1}{5}+\frac{4^2}{30}} \leq \alpha \leq 1 + 3{,}182.1{,}17\sqrt{\frac{1}{5}+\frac{4^2}{30}}\right) = 95\%$$

$$P(1 - 3{,}188 \leq \alpha \leq 1 + 3{,}188) = 95\%$$

$$P(-2{,}188 \leq \alpha \leq 4{,}188) = 95\%$$

Posso afirmar com 95% de confiança que o o coeficiente linear (α) vai estar entre o limite inferior (-2,188) e o limite superior (4,188).

2.3.2. INTERVALO PARA O COEFICIENTE ANGULAR (β)

1 - α = 95% - probabilidade de que o intervalo contenha o parâmetro – no caso o coeficiente linear (α)
α = 5% - probabilidade de que o intervalo não contenha o parâmetro (probabilidade de erro) – no caso o coeficiente linear (α)

$$P\left(b - t_{\alpha/2}.\frac{S}{\sqrt{Sxx}} \leq \beta \leq b + t_{\alpha/2}.\frac{S}{\sqrt{Sxx}}\right) = 1 - \alpha$$

Ver valor de $t_{\alpha/2}$. com n-2 graus de liberdade na tabela t de Student.

O valor de $t_{tab\ 5\%\ bilateral\ ou\ t_{tab\ 2,5\%\ unilateral\ com\ (5-2) = 3\ gl}}$ é 3,182
Logo:

$$P\left(1,3 - 3,182\ \frac{1,17}{\sqrt{30}} \leq \beta \leq 1,3 + 3,182\frac{1,17}{\sqrt{30}}\right) = 95\%$$

$$P(1,3 - 0,68 \leq \beta \leq 1,3 + 0,68) = 95\%$$

$$P(0,62 \leq \beta \leq 1,98) = 95\%$$

Posso afirmar com 95% de confiança que o o coeficiente angular (β) vai estar entre o limite inferior (0,62) e o limite superior (1,98).

2.4. TESTES DE HIPÓTESES

2.4.1. TESTE PARA A EXISTÊNCIA DA REGRESSÃO

Testar a existência da regressão é testar se o parâmetro β é diferente de zero. Desta forma o que se quer testar é:

a) As hipóteses a serem testadas são:
H_0: β = 0 (Não existe regressão linear)
H_1: β ≠ 0 (existe regressão linear)
(teste bilateral)

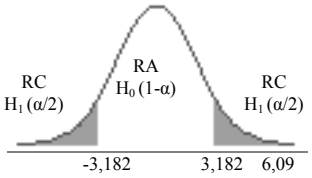
b) Nível de significância do teste, ou erro tipo I (α = 5%)

c) Teste estatístico

$t_{calc} = \frac{b}{\frac{S}{\sqrt{Sxx}}}$ ➔ $t_{calc} = \frac{1,3}{\frac{1,17}{\sqrt{30}}} = 6,09$ ou pode usar o mesmo teste de hipótese da correlação

$t_{calc} = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$ ➔ $t_{calc} = \frac{0,9618}{\sqrt{\frac{1-0,9618^2}{5-2}}} = 6,09$ Comparar com o valor de $t_{tab\ com\ n-2\ graus\ de\ liberdade}$ (Tabela 2).

d) Comparar com o valor de $t_{tab\ 5\%\ bilateral}$ ou $t_{tab\ 2,5\%\ unilateral\ com\ (5-2)\ =\ 3\ gl}$ ➔ $t_{tab} = 3,182$



Lado esquerdo ➔ de menos infinito até -3,182 acumula uma área de 2,5% (Tabela 2).
Lado direito ➔ de 3,182 até mais infinito acumula uma área de 2,5% (Tabela 2).

e) Conclusão: t_{calc} (6,09) está dentro de H_1, logo rejeitamos H_0 e podemos afirmar, ao nível de 5% de significância, ou 5% de probabilidade de erro, que existe regressão linear.

2.4.2. TESTE PARA O COEFICIENTE LINEAR (α)

Testar se a regressão passa pela origem.

a) As hipóteses a serem testadas são:
H_0: α = 0 (regressão linear passa pela origem)
H_1: α ≠ 0 (regressão linear não passa pela origem)
(teste bilateral)

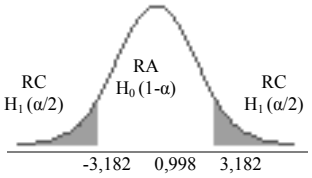
b) Nível de significância do teste, ou erro tipo I (α = 5%)

c) Teste estatístico

$$t_{calc} = \frac{a}{S\left(\sqrt{\frac{1}{n} + \frac{\overline{X}^2}{Sxx}}\right)} \rightarrow t_{calc} = \frac{1}{1{,}17\left(\sqrt{\frac{1}{5} + \frac{4^2}{30}}\right)} = 0{,}998$$

Comparar com o valor de $t_{tab\ com\ n\text{-}2\ graus\ de\ liberdade}$ (Tabela 2).

d) Comparar com o valor de $t_{tab\ 5\%\ bilateral}$ ou $t_{tab\ 2{,}5\%\ unilateral\ com\ (5\text{-}2) = 3\ gl}$ ➔ $t_{tab} = 3{,}182$



Lado esquerdo ➔ de menos infinito até -3,182 acumula uma área de 2,5% (Tabela 2).
Lado direito ➔ de 3,182 até mais infinito acumula uma área de 2,5% (Tabela 2).

e) Conclusão: t_{calc} (0,998) está dentro de H_0, logo não rejeitamos H_0 e não podemos afirmar, ao nível de 5% de significância, ou 5% de probabilidade de erro, que a regressão linear não passa pela origem.

2.5. COEFICIENTE DE DETERMINAÇÃO OU DE EXPLICAÇÃO (R^2)

$$R^2 = \frac{SQ_{regressão}}{SQ_{total}} \rightarrow R^2 = \frac{50,7}{54,8} = 0,925$$

O coeficiente de determinação R^2 também pode ser obtido elevando o coeficiente de correlação (r) ao quadrado ➔ r = 0,9618 ➔ $R^2 = 0,9618^2 = 0,925$

O coeficiente de determinaçao, nesse caso, mostra que 92,5% da variação de Y (Produtividade de milho - t ha^{-1}) é explicada pela variação de X (Doses de calcário - t ha^{-1}). O restante (7,5%) é explicada por outras variáveis (erro) como exemplo precipitação pluvial, época de semeadura, espaçamento, etc.

$$\mathbf{0 \leq R^2 \leq 1}$$

Se R^2 for igual a 1, isto significa que todos os pontos observados se situam "exatamente" sobre a reta de regressão. Tendo-se, neste caso, um ajuste perfeito. As variações da variável Y são 100% explicadas pelas variações da variável X, não ocorrendo desvios em torno da função estimada.

Por outro lado, se $R^2 = 0$, isto quer dizer que as variações de Y são exclusivamente aleatórias e explicadas pelas variações de outros fatores que não X.

EXERCÍCIOS SOBRE CORRELACAO E REGRESSAO

(01) Para cada uma das situações abaixo, diga o que é mais adequado: a análise de regressão ou a análise de correlação. Por quê?

(01.1) Uma equipe de pesquisadores deseja determinar se o rendimento na Universidade sugere êxito na profissão escolhida. *Resposta: Correlação. Apenas verificar que se existe associação.*
(01.2) Deseja-se estimar o número de quilômetros que um pneu radial pode rodar antes de ser substituído. *Resposta: Regressão. Estimar algo.*
(01.3) Deseja-se prever quanto tempo será necessário para executar uma determinada tarefa por uma pessoa, com base no tempo de treinamento. *Resposta: Regressão. Prever algo.*
(01.4) Deseja-se verificar se o tempo de treinamento é importante para avaliar o desempenho na execução de uma dada tarefa. *Resposta: Correlação. Apenas verificar que se existe associação.*
(01.5) Um gerente deseja estimar as vendas semanais com base nas vendas das segundas e terças-feiras. *Resposta: Regressão. Estimar algo.*

(02) Um coeficiente de correlação linear (r), baseado em uma amostra de tamanho n = 18, foi calculado entre as variáveis X (produção em toneladas) e Y (custo total em milhões de R$) como 0,54. Pode-se concluir ao nível de significância de 0,05 que o coeficiente de correlação, correspondente na população é diferente de zero?
Resposta:
Tcal = 2,56 Ttab 5% bilateral com 16 gl = 2,120
Logo rejeito H_0, e concluo ao nível de 5% de probabilidade de erro, que existe correlação linear entre as variáveis.

(03) Os dados abaixo foram obtidos em cinco plantas de alface.

Altura de planta (X)	Número de folhas (Y)
8	16
4	9
6	13
5	9
12	20

a) Faça o diagrama de dispersão.

b) Calcule e teste se o coeficiente de correlação pode ser zero, ao nível de 5% de significância.

c) Calcular a função linear (Y = a + bX) para o número de folhas e o coeficiente de determinação.

d) Calcular qual o valor predito pela reta, em relação ao número de folhas, quando a altura de planta é 5.

(04) Os dados do exercício anterior foram analisados no Excel e os resultados estão abaixo.

Analise de correlação linear

	Altura de planta (X)	*Número de folhas (Y)*
Altura de planta (X)	1	
Número de folhas (Y)	0.970991819	1

Analise de regressão linear

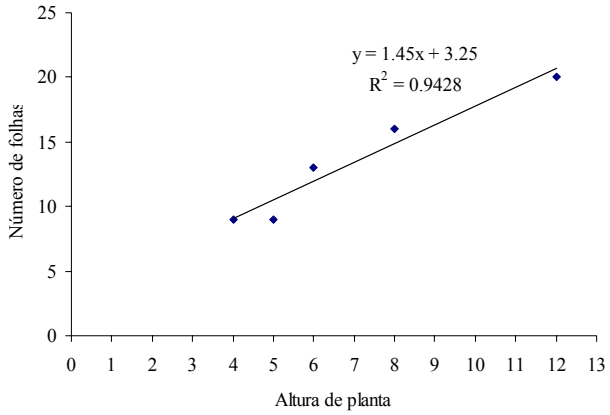
Estatística de regressão	
R múltiplo	0.970991819
R-Quadrado	0.942825112
Erro padrão	1.303840481
Observações	5

ANOVA

	Gl	*SQ*	*MQ*	*F*	*F de significação*
Regressão	1	84.1	84.1	49.47058824	0.005904945
Resíduo	3	5.1	1.7		
Total	4	89.2			

	Coeficientes	*Erro padrão*	*Stat t*	*valor-P*	*95% inferiores*	*95% superiores*
Interseção	3.25	1.556	2.088	0.128	-1.703	8.203
Altura (X)	1.45	0.206	7.034	0.006	0.794	2.106

Gráfico da regressão linear



(04.1) Qual e o coeficiente de correlação linear entre as variáveis? É significativo a 5% de significância?
(04.2) Qual e a função linear (Y = a + bX) para o número de folhas?
(04.3) Qual é o valor e o significado do coeficiente de determinação?
(04.4) Qual é o intervalo de confiança para o coeficiente linear (α) e o angular (β) e o seu significado?

(05) Interprete os resulatdos da tabela abaixo.

Tabela 1 - Matriz de coeficientes de correlação genotípica de Pearson entre os caracteres produtividade de grãos (PROD), número de vagens por planta (NVP), número de sementes por vagem (NSV) e massa de cem grãos (MCG) de 14 cultivares de feijão, avaliadas em nove experimentos. [(1)] Santa Maria-RS, UFSM, 2005.

Caráter	NVP	NSV	MCG
PROD (kg ha^{-1})	0,69**	-0,20**	0,31**
NVP		0,34**	-0,37**
NSV			-0,94**
MCG (g)			

[(1)] ns Não-significativo. ** significativo a 1% de probabilidade, pelo teste t com 376 graus de liberdade.