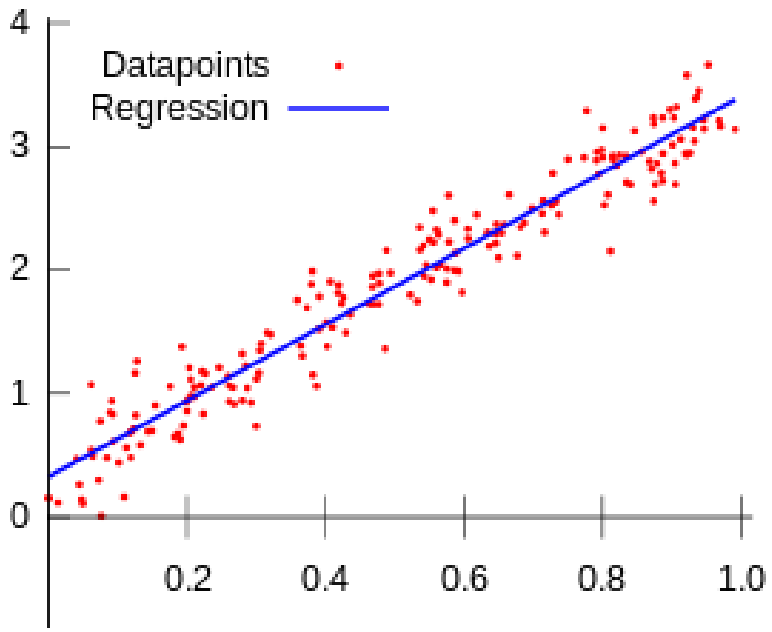
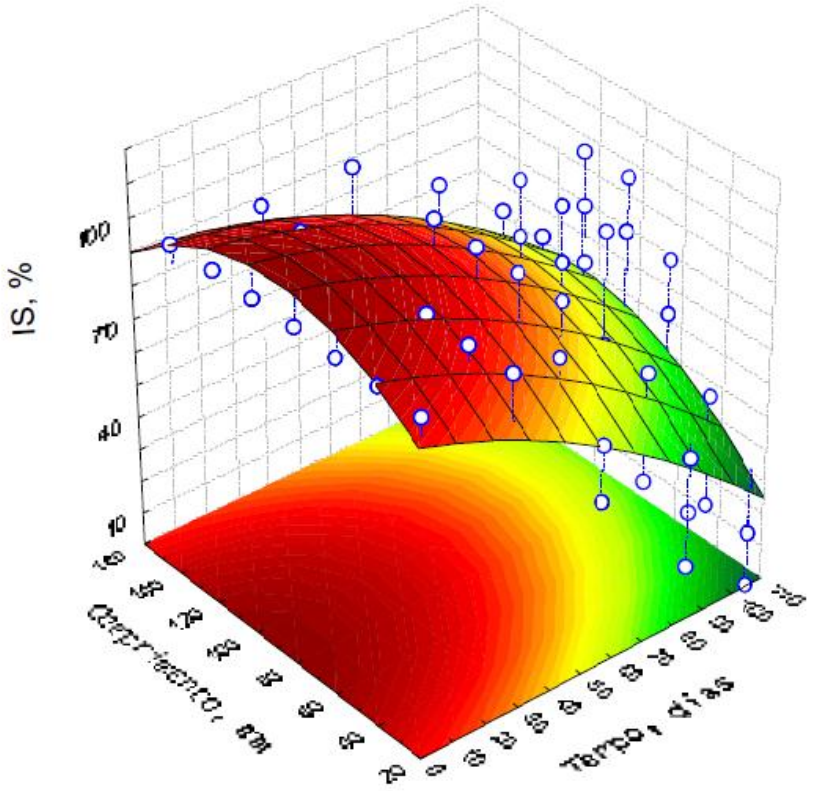
REGRESSÃO LINEAR SIMPLES E MÚLTIPLA

Curso: Agronomia
Matéria: Metodologia e Estatística Experimental
Docente: José Cláudio Faria
Discente: Michelle Alcântara e João Nascimento

UNIVERSIDADE ESTADUAL DE SANTA CRUZ
DEPARTAMENTO DE CIÊNCIAS AGRÁRIAS E AMBIENTAIS
2016.1

REGRESSÃO





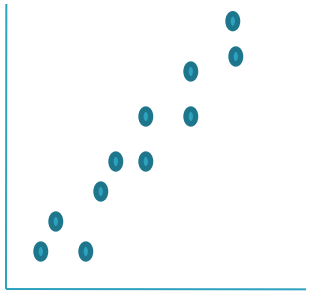
É a relação entre duas ou mais variáveis quantitativas: uma variável dependente, cujo valor deverá ser previsto e uma variável (ou variáveis) independente(s) ou explicativa(s), sobre a(s) qual(is) existe conhecimento teórico disponível.

Estimar uma equação é geometricamente equivalente a ajustar uma curva aos dados dispersos = REGRESSÃO.

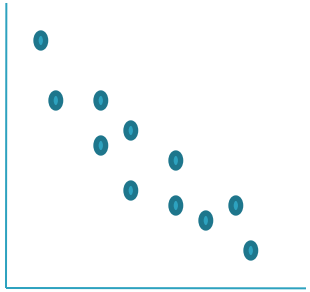
CORRELAÇÃO

- Quando duas variáveis (X e Y) estão ligadas por uma relação estatística, dizemos que existe correlação entre elas.
- Essa técnica é empregada, especificamente, para se avaliar o grau de covariação entre duas variáveis aleatórias.

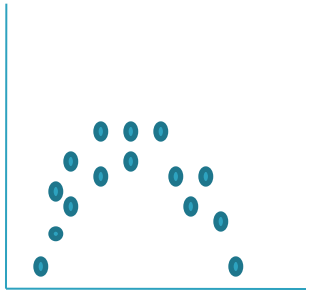
DIAGRAMA DE DISPERSÃO



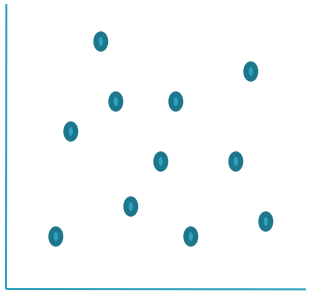
Linear Positiva
(reta ascendente)



Linear Negativa
(reta descendente)



Não linear
(curva)



Não há correlação

Coeficiente de correlação de Pearson

- É um valor que informa a intensidade e a forma da correlação linear entre duas variáveis. A partir da análise do resultado podemos determinar se é adequado ou não a utilização do modelo linear para modelagem do fenômeno.

MODELO MATEMÁTICO:

$$R = \frac{n.\Sigma xy - (\Sigma x).(\Sigma y)}{\sqrt{[n.\Sigma x^2 - (\Sigma x)^2].[n\Sigma y^2 - (\Sigma y)^2]}}$$

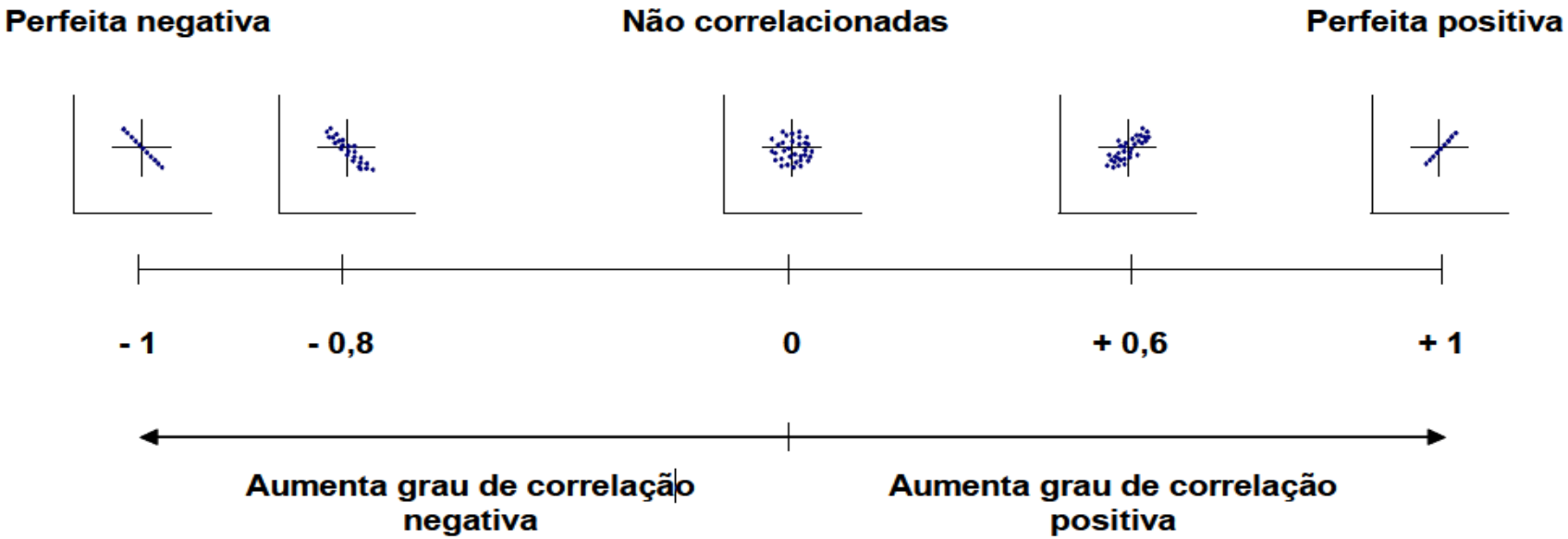
Onde **n** é o número de termos

Os valores limites de **r** são -1 e +1:

- Se correlação é perfeita positiva r=+1
- Se correlação é perfeita negativa r=-1. Isto é, se uma aumenta, a outra diminui linearmente.
- Se não há correlação então r=0. Significa que as duas variáveis não estão linearmente associadas.

Temos também:

- Se 0,6 ≤|r|≤1 boa correlação;
- Se 0,3 ≤|r|≤ 0,6 correlação fraca;
- Se |r|< 0,3 praticamente não existe correlação.



REGRESSÃO LINEAR

- A análise de regressão linear tem como resultado uma regressão matemática que descreve o relacionamento entre duas variáveis.
- Utiliza-se a Regressão Linear para estimar o valor de uma variável com base em valores conhecidos de outro. Pressupõe-se alguma relação de causa e efeito, de explanação do comportamento entre as variáveis. Ex. a idade e o peso de cada bezerro; a alíquota de imposto e a arrecadação; preço e quantidade.

- Regressão Linear Simples: Relação casual entre duas variáveis, e pode ser descrita por uma reta; Uma variável chamada dependente, e uma outra chamada independente. Também tem por objetivo determinar a equação da reta ajustada(modelo matemático linear).

- Regressão Linear múltipla: Relação casual com mais de duas variáveis. Isto é, quando o comportamento de **Y** é explicado por mais de uma variável independente **X**1, **X**2, ...**X**n. É a técnica adequada para se utilizar quando se quer investigar simultaneamente os efeitos, sobre Y, de 2 ou mais variáveis preditoras.

AJUSTAMENTO DE RETA

X Nitrogênio kg ha^{-1}	Y Safra kg ha^{-1}
10	1.000
20	2.300
30	2.600
40	3.900
50	5.400
60	5.800
70	6.600

Figura1. Relação observada entre a safra e a aplicação de nitrogênio.

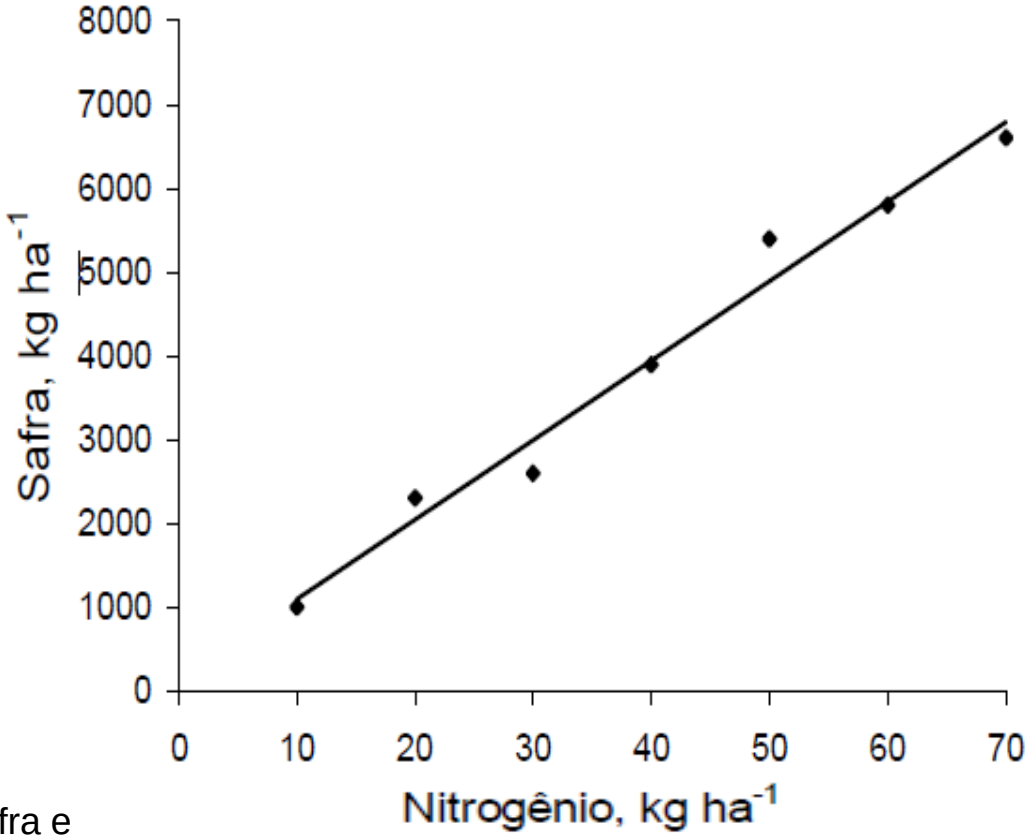


Figura 2. Dados e reta ajustada a olho aos dados apresentados.

Critérios para o ajustamento da reta

- O que é um bom ajustamento? Um ajustamento que causa pequeno erro total.
- O erro ou a falta de ajustamento é definido como a distância vertical entre o valor observado Y_i e o valor ajustado $\hat{Y}_i$ na reta, isto é, $\left(Y_i - \hat{Y}_i\right)$:

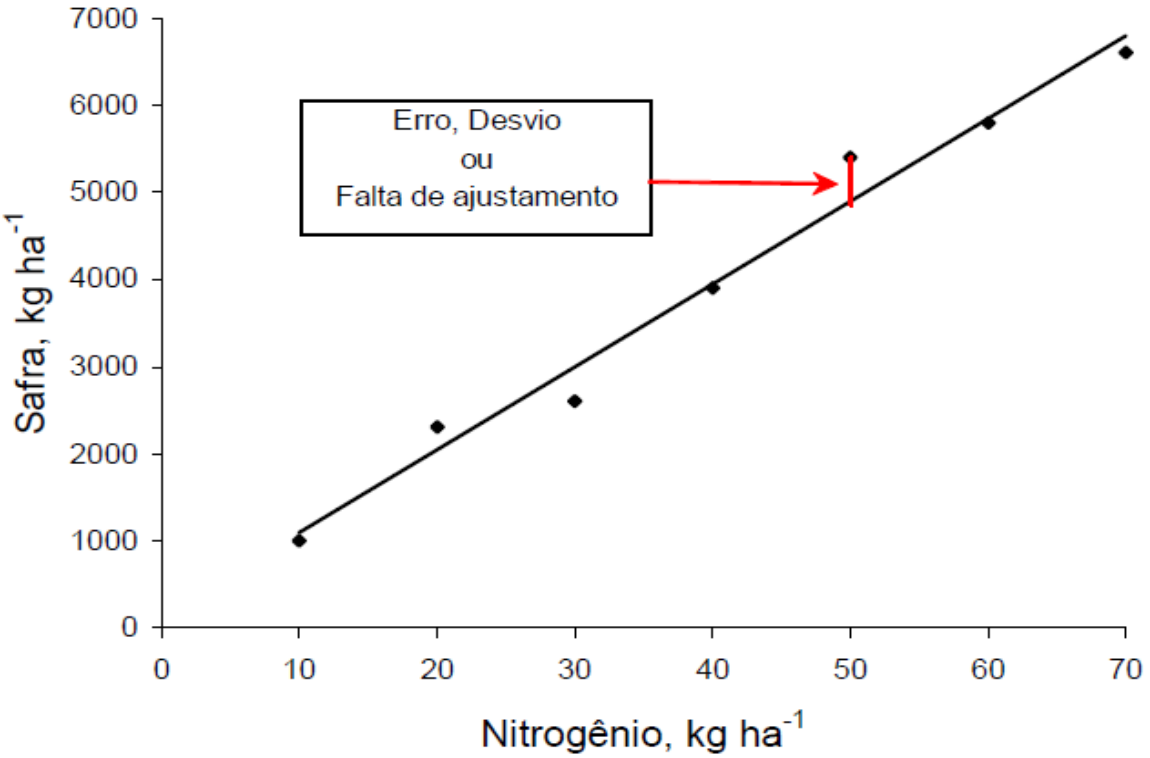


Figura 3. Erro típico no ajustamento de uma reta.

Método dos Mínimos Quadrados

- O método mais utilizado para ajustar uma reta aos pontos dispersos é o que minimiza a soma de quadrados dos erros:

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

- O quadrado elimina o problema do sinal, pois torna positivos todos os erros;
- A álgebra dos mínimos quadrados é de manejo relativamente fácil;
- O método dos mínimos quadrados permite encontrar as estimativas de α e β;
- Minimizando a soma do quadrado de erros, encontraremos α e β, que trarão a menor diferença entre a previsão de Y_i e o $\hat{Y}_i$.

- REGRESSÃO = Criar um modelo de *equação de reta* para fazer previsões/estimativas de valores futuros através dos pontos.
- Equação de Reta = **Equação de 1ºGrau**

OBJETIVO = Ajustar uma reta

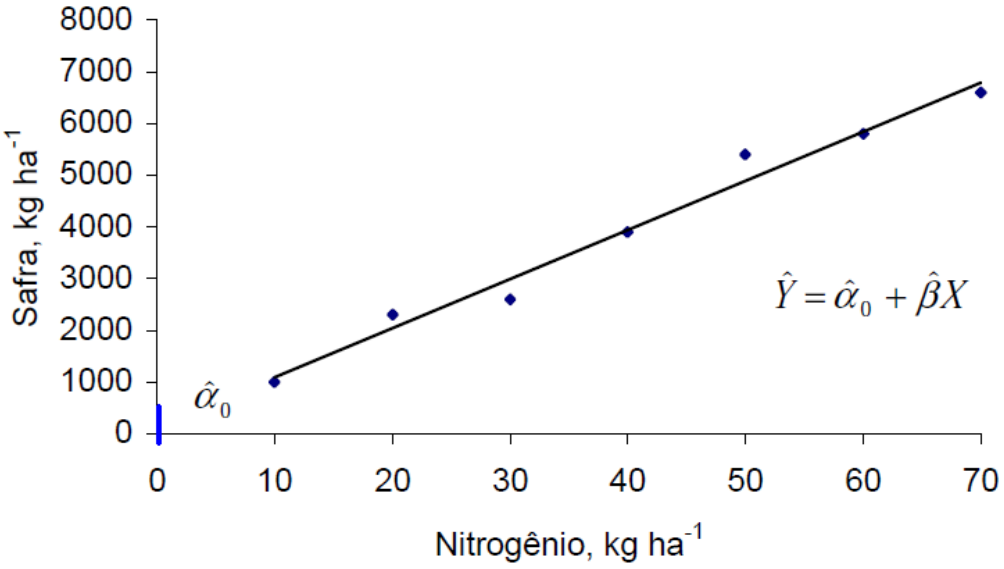
$$\hat{Y} = \hat{\alpha}_0 + \hat{\beta}X$$

Onde:

- X- variável explicativa ou independente;
- Y- variável explicada ou dependente (aleatória);
- $\hat{\alpha}_0$ - coeficiente linear ou constante da regressão, representa o interceptor da reta com o eixo do Y;
- β- coeficiente de regressão ou coeficiente angular da reta. Representa a variação de Y em função da unitária variável x;
- α e β são parâmetros.

❖Ajustando uma reta : 3 estágios

a.



b.

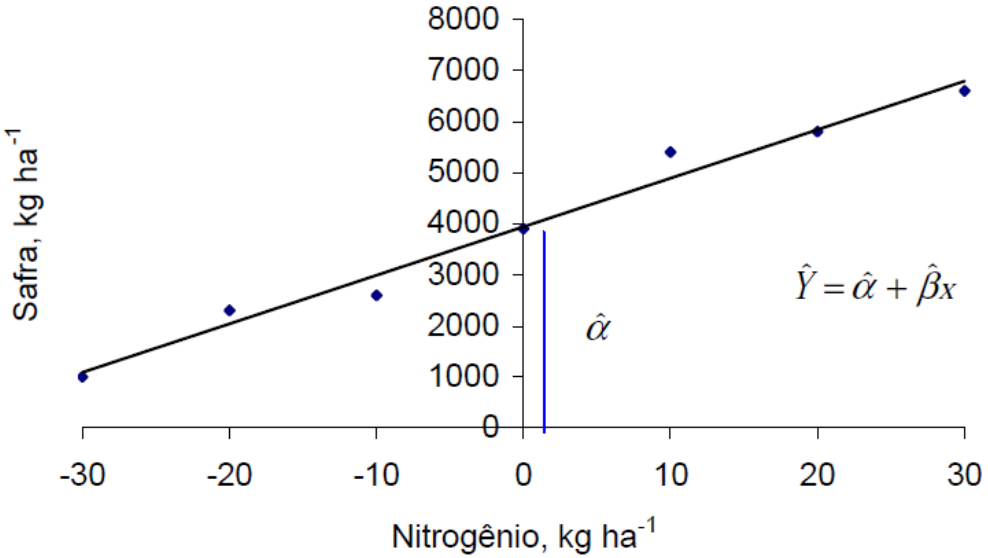
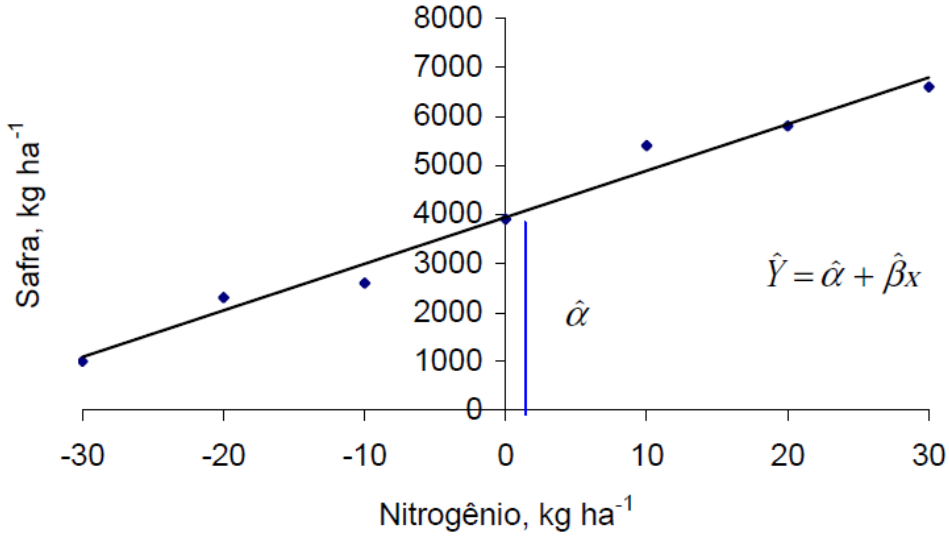


Figura 4. Translação de eixos. (a) Regressão utilizando os valores originais. (b) Regressão após transladar Y.



X	$x = X - \overline{X}$ $x = X - 40$
10	- 30
20	- 20
30	- 10
40	0
50	10
60	20
70	30

$$\sum X = 280$$

$$\overline{X} = \frac{1}{N}\sum X \qquad \sum x = 0$$

$$\overline{X} = \frac{280}{7} = 40$$

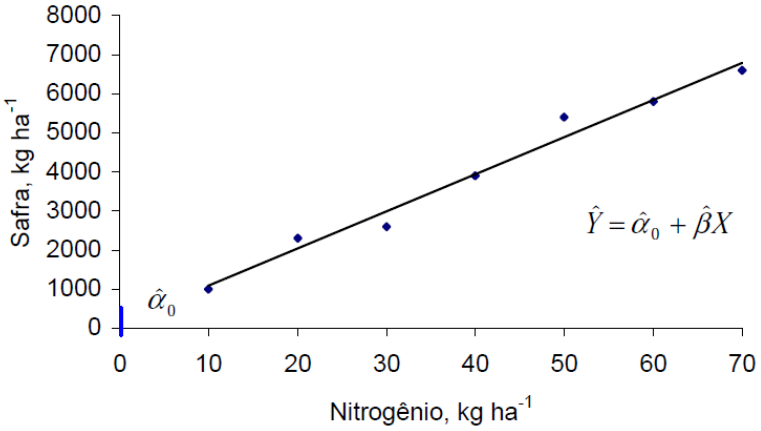
Figura 5. Calculo para encontrar os valores negativos.

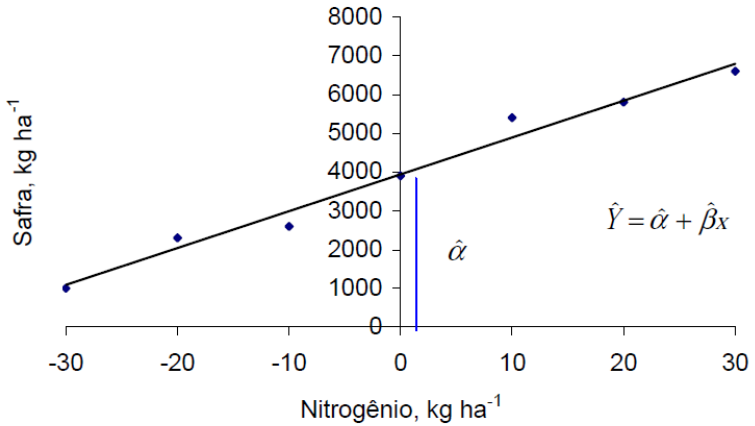
1º Estágio

- Exprimir X em termos de desvios a contar de sua média, isto é, definir uma nova variável x (minúsculo), tal que:

$$x = X - \overline{X}$$

- Isto equivale a uma relação geométrica de eixos:





- Observa-se que o eixo Y foi deslocado para a direita, de 0 a $\overline{X}$
- O novo valor x torna-se positivo, ou negativo, conforme X esteja a direita ou a esquerda de $\overline{X}$.
- Não há modificação nos valores de Y.
- O intercepto $\hat{\alpha}$ difere do intercepto original, $\hat{\alpha}_0$,mas o coeficiente angular, $\hat{\beta}$,permanece o mesmo.

- Medir X como desvio a contar de $\bar{X}$ *simplifica os cálculos porque a soma dos* novos valores x é igual a zero, isto é:

$$\sum x_i = 0 \qquad \therefore \qquad \sum x_i = \sum (X_i - \bar{X}) = \sum X_i - n\bar{X} = n\bar{X} - n\bar{X} = 0$$

2º Estágio

- Devemos ajustar a reta aos dados, escolhendo valores para $\hat{\alpha}$ e $\hat{\beta}$, que satisfaçam o critério dos mínimos quadrados. Ou seja, escolher valores de $\hat{\alpha}$ e $\hat{\beta}$ que minimizem:

$$\sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2 \qquad \text{Equação 01}$$

- Cada valor ajustado de $\hat{Y}_i$ estará sobre a reta estimada:

$$\hat{Y}_i = \hat{\alpha} + \hat{\beta}x_i \qquad \text{Equação 02}$$

- Assim, estamos diante da seguinte situação: devemos encontrar os valores $\hat{\alpha}$ e $\hat{\beta}$ de modo a minimizar a soma de quadrados dos erros.
- Considerando a equação 1 e 2, isto pode ser expresso algebricamente como:

$$\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 \;\therefore\; \hat{Y}_i = \hat{\alpha} + \hat{\beta}x_i$$

$$S(\hat{\alpha},\hat{\beta}) = \sum\left(Y_i - (\hat{\alpha} + \hat{\beta}X_i)\right)^2 = \sum\left(Y_i - \hat{\alpha} - \hat{\beta}X_i\right)^2$$

- Usamos a notação $S(\hat{\alpha},\hat{\beta})$ para enfatizar que esta expressão depende de $\hat{\alpha}$ e $\hat{\beta}$. Ao variarem $\hat{\alpha}$ e $\hat{\beta}$ (quando se tem várias retas), $S(\hat{\alpha},\hat{\beta})$ variará também.

- *Pergunta-se então, para que valores de* $\hat{\alpha}$ *e* $\hat{\beta}$ *haverá um mínimo de erros?*

A resposta a esta pergunta nos fornecerá a reta “ótima” (de mínimos quadrados dos erros).

- A técnica de minimização mais simples é fornecida pelo *cálculo*. A minimização de $S(\hat{\alpha}, \hat{\beta})$ *exige o anulamento simultâneo de suas derivadas parciais.*

Igualando a zero a derivada parcial em relação a $\hat{\alpha}$:

$$\frac{\partial}{\partial \hat{\alpha}} \sum \left(Y_i - \hat{\alpha} - \hat{\beta} X_i\right)^2 = \sum 2(-1)\left(Y_i - \hat{\alpha} - \hat{\beta} X_i\right)^1 = 0$$

- Dividindo ambos os termos por (-2) e reagrupando:

$$\sum Y_i - n\hat{\alpha} - \hat{\beta}\sum x_i = 0 \qquad \therefore \qquad \sum x_i = 0$$

$$\sum Y_i - n\hat{\alpha} - 0 = 0$$

$$\sum Y_i - n\hat{\alpha} = 0$$

$$n\hat{\alpha} = \sum Y_i$$

$$\hat{\alpha} = \frac{\sum Y_i}{n} = \overline{Y}$$

Verifica-se que isto assegura que a reta de regressão ajustada deve passar pelo ponto (x, $\overline{Y}$), *que pode ser interpretado como o centro de gravidade da amostra de n* pontos.

- É preciso também anular a derivada parcial em relação a $\hat{\beta}$:

$$\frac{\partial}{\partial \hat{\beta}} \sum \left(Y_i - \hat{\alpha} - \hat{\beta} x_i\right)^2 = \sum 2(-x_i)\left(Y_i - \hat{\alpha} - \hat{\beta} x_i\right)^1 = 0$$

Dividindo ambos os termos por (-2):

$$\sum x_i \left(Y_i - \hat{\alpha} - \hat{\beta} x_i\right) = 0$$

- Reagrupando:

$$\sum x_i Y_i - \hat{\alpha}\sum x_i - \hat{\beta}\sum x_i^2 = 0 \qquad \therefore \qquad \sum x_i = 0$$

$$\sum x_i Y_i - 0 - \hat{\beta}\sum x_i^2 = 0$$

$$\sum x_i Y_i - \hat{\beta}\sum x_i^2 = 0$$

$$\hat{\beta}\sum x_i^2 = \sum x_i Y_i$$

$$\hat{\beta} = \frac{\sum x_i Y_i}{\sum x_i^2}$$

- Podemos sintetizar da seguinte forma :

Com os valores x medidos como desvios a contar de sua média, os valores $\hat{\alpha}$ e $\hat{\beta}$ de mínimos quadrados dos erros são:

$$\hat{\alpha} = \frac{\sum Y_i}{n} = \bar{Y}$$

$$\hat{\beta} = \frac{\sum x_i Y_i}{\sum x_i^2}$$

❖Dados do exemplo

X	$x = X - \overline{X}$ $x = X - 40$	Y	xY	x^2
10	- 30	1.000	- 30.000	900
20	- 20	2.300	- 46.000	400
30	- 10	2.600	- 26.000	100
40	0	3.900	0	0
50	10	5.400	54.000	100
60	20	5.800	116.000	400
70	30	6.600	198.000	900

$$\sum X = 280 \qquad \sum x = 0 \qquad \sum Y = 27.600 \qquad \sum xY = 266.000 \qquad \sum x^2 = 2.800$$

$$\overline{X} = \frac{1}{N}\sum X \qquad \overline{Y} = \frac{1}{N}\sum \overline{Y}$$

$$\overline{X} = \frac{280}{7} = 40 \qquad \overline{Y} = \frac{27.600}{7}$$

$$\overline{Y} = 3.942{,}86$$

$$\hat{\alpha} = \frac{\sum Y_i}{n} = \overline{Y} \;\therefore\; \hat{\alpha} = \frac{27.600}{7} = 3.942{,}86$$

$$\hat{\beta} = \frac{\sum x_i Y_i}{\sum x_i^2} \;\therefore\; \hat{\beta} = \frac{266.000}{2.800} = 95{,}00$$

$$\hat{Y} = 3.942{,}86 + 95x$$

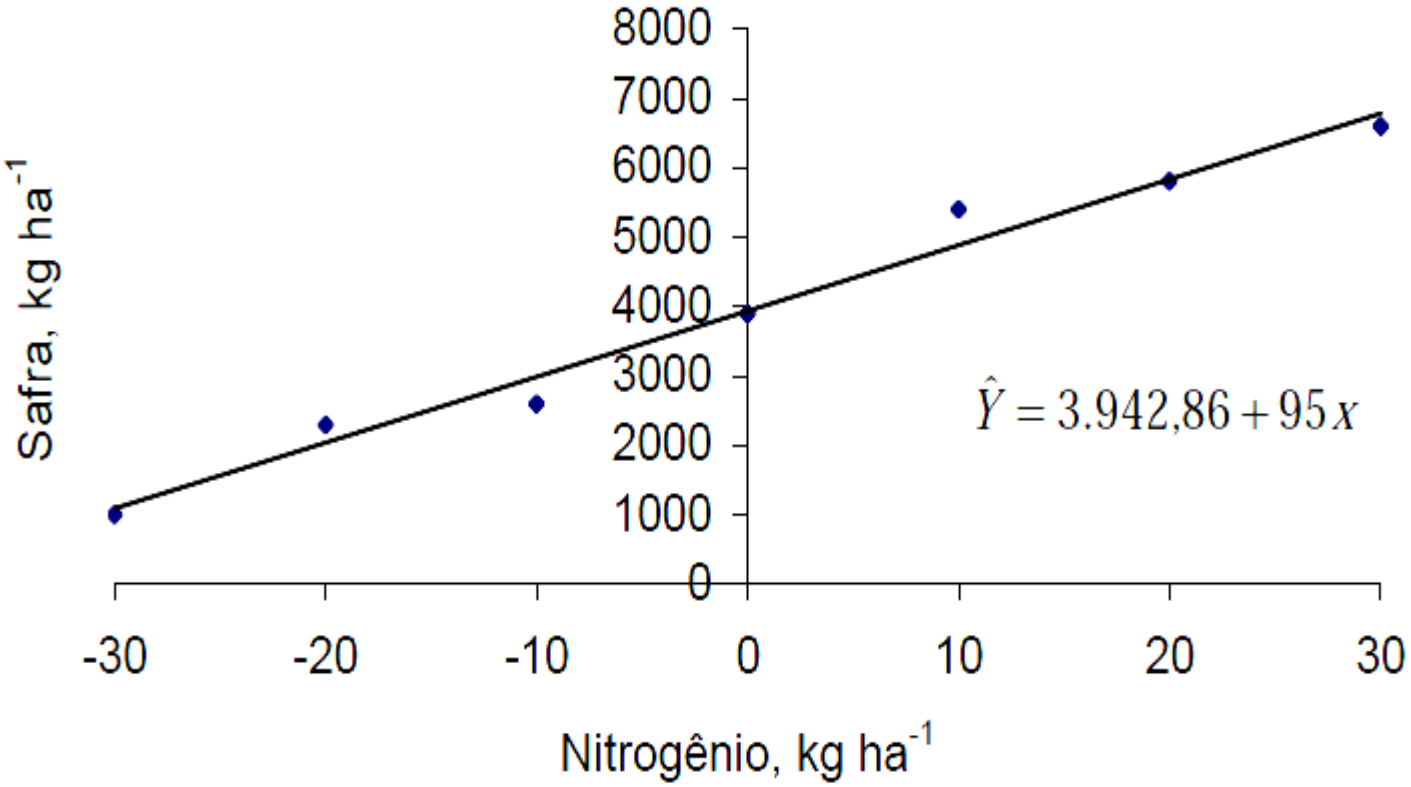


Figura 6. Equação da reta translocada.

Estágio 3 - A regressão pode agora ser transformada para o sistema original de referência:

$$\hat{Y} = 3.942{,}86 + 95\,x \quad \therefore \quad x = (X - \overline{X})$$

$$\hat{Y} = 3.942{,}86 + 95(X - \overline{X})$$

$$\hat{Y} = 3.942{,}86 + 95(X - 40)$$

$$\hat{Y} = 3.942{,}86 + 95X - 3.800$$

$$\hat{Y} = 142{,}86 + 95X$$

O coeficiente angular da reta de regressão ajustada ($\hat{\beta}$ = 95X) permanece inalterado.

A única diferença é o intercepto, $\hat{\alpha}$, onde a reta tangencia o eixo Y.

O intercepto original foi facilmente reobtido.

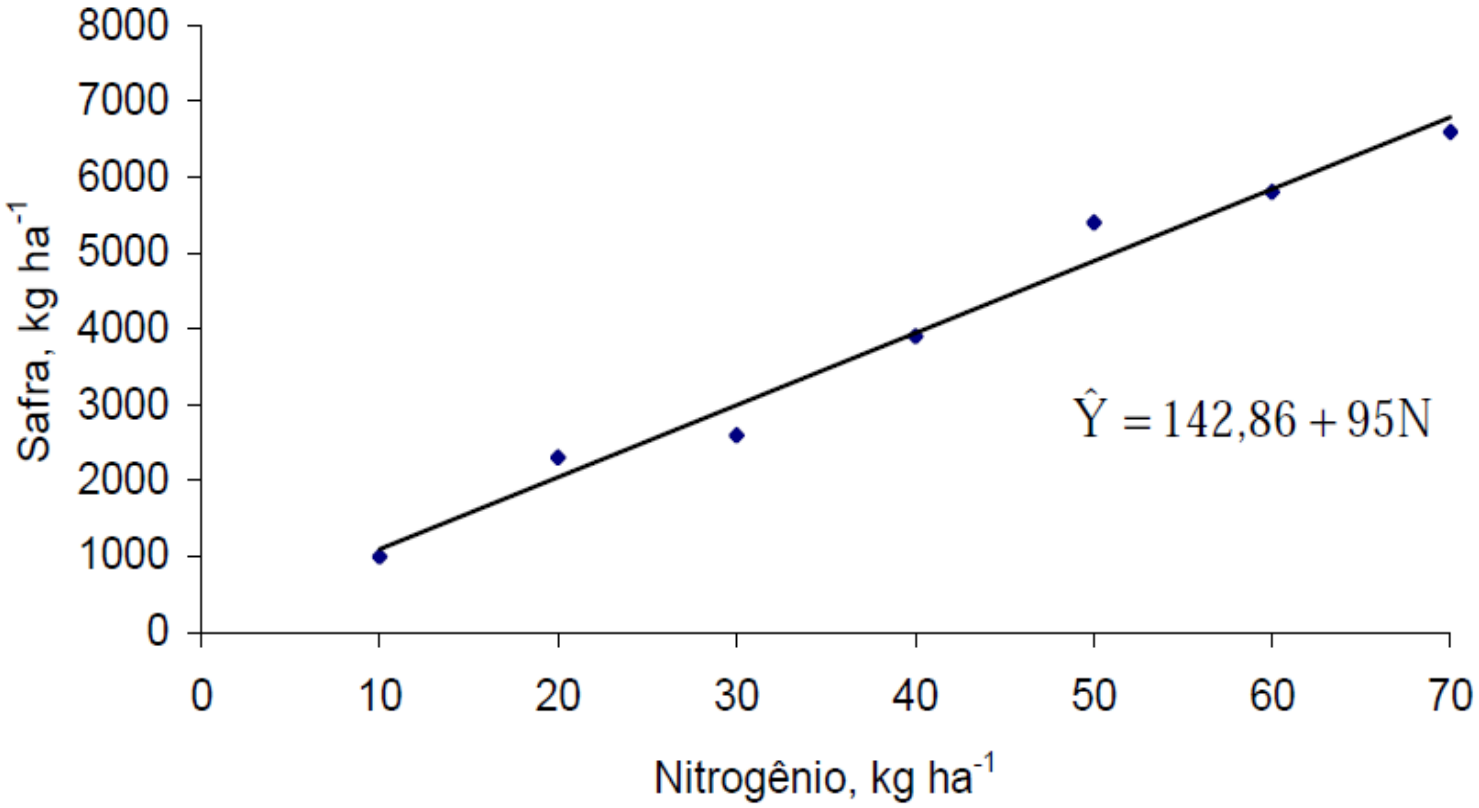


Figura7. Gráfico dos pontos dispersos com a reta ajustada.

Análise de Variância da Regressão

Para se decidir quão bem o modelo ajustado é adequado à natureza dos dados experimentais, pode-se lançar mão da análise de variância da regressão (ANOVAR).

Para o caso em estudo, a ANOVAR irá particionar a variação total (SQDtot) da variável dependente - ou fator resposta - em função das variações nos níveis da variável independente - ou regressor, em duas partes:

- Uma parte associada ao modelo ajustado (SQDDreg): soma de quadrados dos desvios devido à regressão, que quantifica o quanto da variação total da safra, provocada pela variação das doses de nitrogênio, é explicada pelo modelo ajustado.
- Uma outra parte associada à falta de ajuste (SQDDerr): soma de quadrados dos desvios devido ao erro, que quantifica o montante da variação total da safra, provocada pela variação da dose de nitrogênio, que não é explicada pelo modelo ajustado.

Para o exemplo em análise a ANOVAR teria a seguinte estrutura:

Hipóteses:

$H_o: \lvert\beta_i\rvert = 0$	ou	$H_o: Y \neq \alpha_0 + \beta X$
$H_1: \lvert\beta_i\rvert > 0$	ou	$H_1: Y = \alpha_0 + \beta X$

- Significado de H_o: A equação de regressão não explica a variação da variável dependente Y, em decorrência da variação da variável independente X, ao nível de ...% de probabilidade.
- Significado de H_1: A equação de regressão explica a variação da variável dependente Y, em decorrência da variação da variável independente X, ao nível de ...% de probabilidade.

A análise de variância é esquematizada como:

F.V.	G.L.	S.Q.	Q.M.	F	*p-value*
Modelo	k	SQ(Mod.)	QM(Mod.)	QM(Mod.) / QM(Res.)	p
Resíduo	N-k-1	SQ(Res.)	QM(Res.)		
Total	N-1	SQ(Tot.)			

F.V. – Fontes de Variação, G.L. – Graus de Liberdade, S.Q. – Somas de Quadrados, Q.M. – Quadrados Médios, N – Número de observações, k – número de variáveis independentes.

A estatística F testa a hipótese: H_0: $B_1=B_2= \ldots =B_k=0$ *vs* H_1: $B_i \neq B_{i'}$, para algum $i \neq i'$.

O valor p (***p-value***) é obtido supondo que a estatística F tem uma distribuição F central com K e N-k-1 graus de liberdade. Essa pressuposição é válida se os erros forem iid - independentes e identicamente distribuídos, com distribuição normal $N(0,\sigma^2)$.

Para exemplo:

ANOVAR

Causa da variação	GL	SQD	QMD	F_{cal}	Pr
Regressão	1	25.270.000,00	25.270.000,00	239,69	< 0,0001
Erro	5	527.142,86	105.428,57		
Total	6	25.797.142,86			

Conclusão: rejeita-se H0 ao nível de 5% de probabilidade pelo teste F.

ANÁLISE DE RESÍDUOS

- É importante, após a análise de regressão, testar se os pressupostos do modelo linear se aplicam aos dados estudados;
- Resíduos representam a diferença entre o valor observado de y e o que foi predito pelo modelo de regressão;
- A primeira forma de se avaliar resíduos é plotar um gráfico no qual os resíduos (y- $\hat{Y}$) são colocados no eixo vertical (y) e os valores esperados de y (β y) no eixo horizontal (x).

Exemplo de Análise completa:

Os dados abaixo são provenientes de um ensaio experimental em que foram utilizadas sete doses de nitrogênio aplicado em cobertura sobre a produtividade de milho. O Experimento foi montado no delineamento inteiramente casualizado, DIC, com cinco repetições. Os dados são fornecidos abaixo:

Quadro 14.2 – Produção de milho, kg ha^{-1}

N	Repetições					Totais	Rep.	Médias
kg.ha^{-1}	1	2	3	4	5			
10	1.000	916	958	1.084	1.042	5.000	5	1.000
20	2.340	2.220	2.300	2.260	2.380	11.500	5	2.300
30	2.559	2.518	2.682	2.641	2.600	13.000	5	2.600
40	3.976	3.900	3.862	3.938	3.824	19.500	5	3.900
50	5.448	5.304	5.352	5.400	5.496	27.000	5	5.400
60	5.843	5.886	5.800	5.714	5.757	29.000	5	5.800
70	6.600	6.555	6.690	6.510	6.645	33.000	5	6.600
						138.000	35	3.942,86

$C = (138.000)^2 / 35 = 544.114.285,71$

$SQDtot = [(1.000)^2 + (916)^2 + ... + (6.645)^2] - C = 129.112.384,29$

$SQDtra_t = 1/5\ [(5.000)^2 + (11.510)^2 + ... + (33.000)^2] - C = 128.985.714,29$

$SQDres = SQDtot - SQDtra = 129.112.384,29 - 128.985.714,29 = 126.670,00$

Hipóteses:

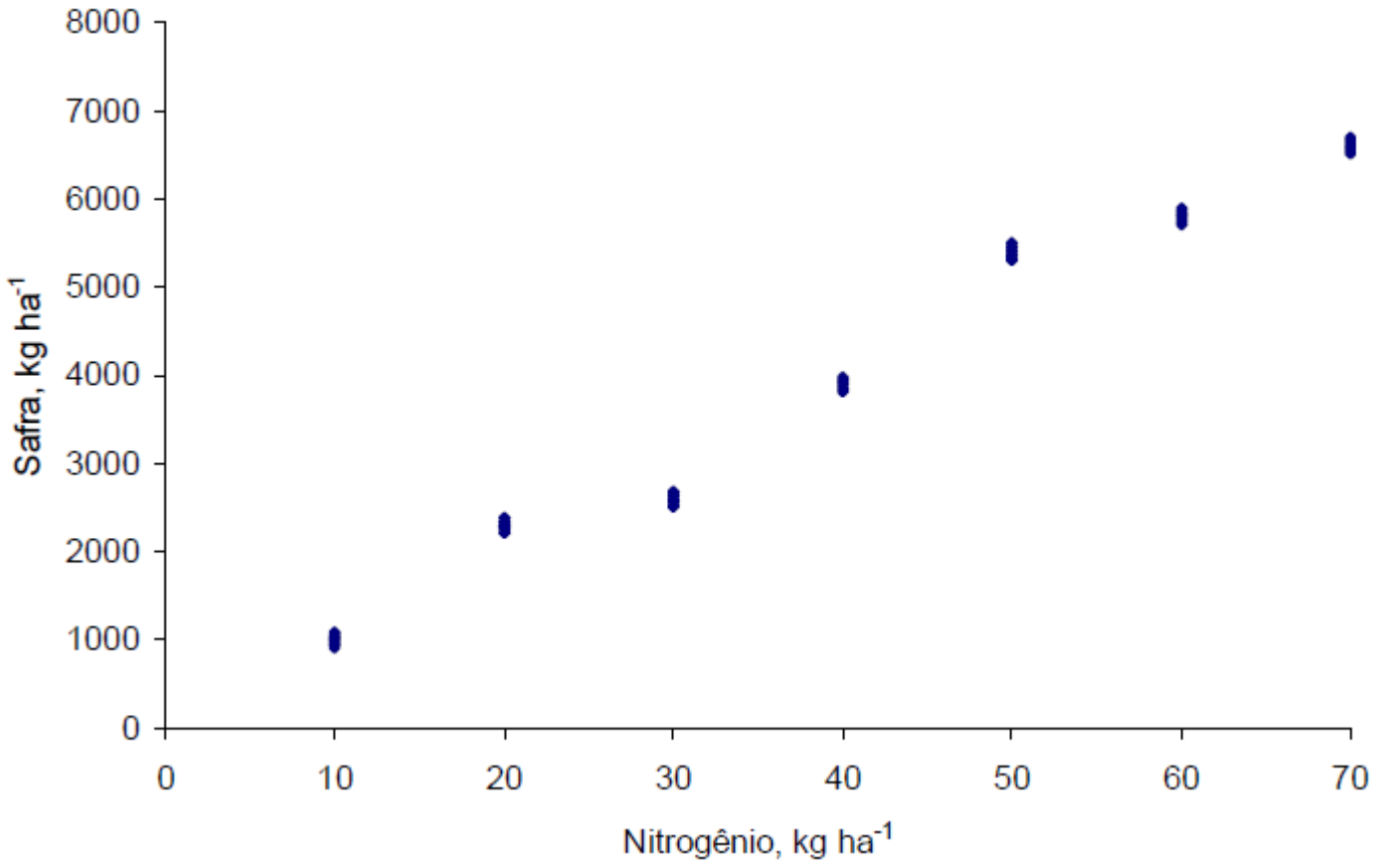
H_0: $\mu_{10} = ... = \mu_{70}$

H_1: Nem todas as médias são iguais

ANOVA

Causa da variação	GL	SQD	QMD	F_{cal}	Pr
Tratamentos	6	128.985.714,29	21.497.619,05	4.751,98	< 0,0001
Resíduo	28	126.670,00	4.523,93		
Total	34	129.112.384,29			

Conclusão: rejeita-se Ho ao nível de significância de 5% pelo teste F.



A visualização dos dados experimentais em um gráfico de dispersão auxilia na escolha do modelo a ser ajustado.

Ajustando uma Reta

Ajustando um modelo linear: $\hat{Y} = \alpha_0 + \beta_1 X$

Valores necessários para o ajustamento do modelo linear.

X	$x = X - \overline{X}$ $x = X - 40$	Y	xY	x^2
10	- 30	1.000	- 30.000	900
20	- 20	2.300	- 46.000	400
30	- 10	2.600	- 26.000	100
40	0	3.900	0	0
50	10	5.400	54.000	100
60	20	5.800	116.000	400
70	30	6.600	198.000	900
$\sum X = 280$ $\overline{X} = \frac{1}{N}\sum X$ $\overline{X} = \frac{280}{7} = 40$	$\sum x = 0$	$\sum Y = 27.600$ $\overline{Y} = \frac{1}{N}\sum \overline{Y}$ $\overline{Y} = \frac{27.600}{7}$ $\overline{Y} = 3.942{,}86$	$\sum xY = 266.000$	$\sum x^2 = 2.800$

Recomenda-se trabalhar com o máximo possíveis de casas decimais.

$$\hat{\alpha} = \frac{\sum Y_i}{n} = \overline{Y} \quad \therefore \quad \hat{\alpha} = \frac{27.600}{7} = 3.942,86$$

$$\hat{\beta} = \frac{\sum x_i Y_i}{\sum x_i^2} \quad \therefore \quad \hat{\beta} = \frac{266.000}{2.800} = 95,00$$

$$\hat{Y} = 3.942,86 + 95x$$

$$\hat{Y} = 3.942,86 + 95x \qquad \therefore \qquad x = (X - \overline{X})$$

$$\hat{Y} = 3.942,86 + 95(X - \overline{X})$$

$$\hat{Y} = 3.942,86 + 95(X - 40)$$

$$\hat{Y} = 3.942,86 + 95X - 3.800$$

Equação da reta ajustada: $\hat{Y} = 142,86 + 95X$

Análise de Variância da Regressão: ANOVAR

ANOVAR

Causa da variação	GL
Regressão	1
Erro	5
Total	6

Para se decidir quão bem o modelo ajustado é adequado à natureza dos dados experimentais, pode-se lançar mão da análise de variância da regressão (ANOVAR).

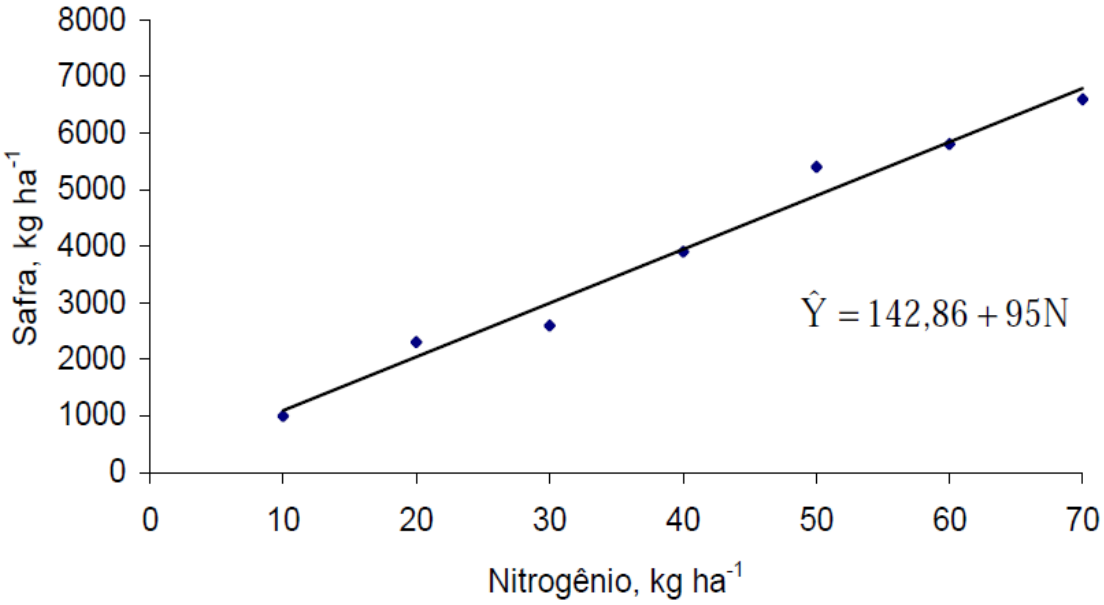
Para o caso em estudo, a ANOVAR irá particionar a variação total (SQDtot) da variável dependente - ou fator resposta - em função das variações nos níveis da variável independente - ou regressor, em duas partes:

- Uma parte associada ao modelo ajustado (SQDDreg)
- Uma outra parte associada à falta de ajuste (SQDDerr)

$$Quadrado\ \ médio\ \ dos\ \ desvios = s^2 = \frac{SQD}{n-1} \ \therefore \qquad SQD = \sum (Y_i - m)^2$$

Vejamos[1]:

N , kg ha^{-1}	Safra_Obs	Safra_Est
10	1.000	1092,86
20	2.300	2042,86
30	2.600	2992,86
40	3.900	3942,86
50	5.400	4892,86
60	5.800	5842,86
70	6.600	6792,86



SQDtot

Obs	$m_{(Obs)}$	Obs-$m_{(Obs)}$	[Obs-$m_{(Obs)}$]^2
1.000	3.942,86	-2.942,86	8.660.408,16
2.300	3.942,86	-1.642,86	2.698.979,59
2.600	3.942,86	-1.342,86	1.803.265,31
3.900	3.942,86	-42,86	1.836,73
5.400	3.942,86	1.457,14	2.123.265,31
5.800	3.942,86	1.857,14	3.448.979,59
6.600	3.942,86	2.657,14	7.060.408,16
			25.797.142,86

SQDreg

Est	$m_{(Est)}$	Est-$m_{(Est)}$	[Est-$m_{(Est)}$]^2
1.093	3.942,86	-2.850,00	8.122.500,00
2.043	3.942,86	-1.900,00	3.610.000,00
2.993	3.942,86	-950,00	902.500,00
3.943	3.942,86	0,00	0,00
4.893	3.942,86	950,00	902.500,00
5.843	3.942,86	1.900,00	3.610.000,00
6.793	3.942,86	2.850,00	8.122.500,00
			25.270.000,00

SQDerr

Obs	Est	Erro(Obs-Est)	$m_{(Erro)}$	Erro-$m_{(Erro)}$	[Erro-$m_{(Erro)}$]^2
1.000	1.092,86	-92,86	0,00	-92,86	8.622,45
2.300	2.042,86	257,14	0,00	257,14	66.122,45
2.600	2.992,86	-392,86	0,00	-392,86	154.336,73
3.900	3.942,86	-42,86	0,00	-42,86	1.836,73
5.400	4.892,86	507,14	0,00	507,14	257.193,88
5.800	5.842,86	-42,86	0,00	-42,86	1.836,73
6.600	6.792,86	-192,86	0,00	-192,86	37.193,88
					527.142,86

ANOVAR

Causa da variação	GL	SQD	QMD	F_{cal}	Pr
Regressão	1	25.270.000,00	25.270.000,00	239,69	< 0,0001
Erro	5	527.142,86	105.428,57		
Total	6	25.797.142,86			

Cálculos alternativos da soma de quadrados dos desvios

$$SQDtot = \sum Y_i^2 - \frac{\left(\sum Y_i\right)^2}{n}$$

$$SQDreg = \hat{\alpha}_0 \sum Y_i + \hat{\beta} \sum X_i Y_i - \frac{\left(\sum Y_i\right)^2}{n}$$

$$SQDerr = SQDtot - SQDreg$$

X	Y	Y^2	XY
10	1.000	1.000.000	10.000
20	2.300	5.290.000	46.000
30	2.600	6.760.000	78.000
40	3.900	15.210.000	156.000
50	5.400	29.160.000	270.000
60	5.800	33.640.000	348.000
70	6.600	43.560.000	462.000
	27.600	134.620.000	1.370.000

$$SQDtot = \sum Y_i^2 - \frac{(\sum Y_i)^2}{n} = 134.620.000 - \frac{(27.600)^2}{7} = 25.797.142,86$$

$$SQDreg = \hat{\alpha}_o \sum Y_i + \hat{\beta} \sum X_i Y_i - \frac{(\sum Y_i)^2}{n}$$

$$SQDreg = 142,85714286 \times 27.600 + 95 \times 1.370.000 - \frac{(27.600)^2}{7}$$

$$SQDreg = 25.270.000$$

$$SQDerr = SQDtot - SQDreg$$

$$SQDerr = 25.797.142,86 - 25.270.000$$

$$SQDerr = 527.142,86$$

Ilustração da ANOVAR apenas para efeito de comparação com a ANOVA:

ANOVAR

Causa da variação	GL	SQD	QMD	F_{cal}	Pr
Regressão	1	25.270.000,00	25.270.000,00	239,69	< 0,0001
Erro	5	527.142,86	105.428,57		
Total	6	25.797.142,86			

Coeficiente de determinação da Regressão:

$$r^2 = \frac{SQDreg}{SQDtot} \quad \therefore \quad 0 \leq r^2 \leq 1$$

$$r^2 = \frac{25.270.000,00}{25.797.142,86} = 0,9796 = 97,96\%$$

Observa-se que a soma de quadrados, e os respectivos graus de liberdade, associados a tratamentos foram desdobrados em duas partes:

Uma parte associada ao modelo de regressão utilizado (*Yˆ = 142,86 + 95N).*

Uma parte associada à falta de ajuste ou erro de ajustamento:

Para a obtenção da soma de quadrados do devido à regressão e ao independente da regressão tem-se duas opções:

•Realizar todos os cálculos das somas de quadrados dos desvios considerando agora todas as repetições, o que embora possa ser feito, é um processo mais trabalhoso.

•Utilizar o teorema do limite central (que facilita bastante os cálculos):

Teorema do limite central:

$$Var(m) = \frac{\sigma^2}{n} \quad \therefore \quad \sigma^2 = Var(m) \times n$$

$$SQD(m) = \frac{SQD}{n} \quad \therefore \quad SQD = SQD(m) \times n \quad \therefore \quad Como\ n = r$$

$$SQDDreg = 25.270.000,00 \times 5 = 126.350.000,00$$

$$SQDDireg = 527.142,86 \times 5 = 2.635.714,29$$

ANOVA

Causa da variação	GL	SQD	QMD	F_{cal}	Pr
Tratamentos	(6)	(128.985.714,29)			
Dev. regressão	1	126.350.000,00	126.350.000,00	27.929,26	< 0,0001
Ind. regressão	5	2.635.714,29	527.142,86	116,52	< 0,0001
Resíduo	28	126.670,00	4.523,93		
Total	34	129.112.384,29			

Conclusão: rejeita-se H0 ao nível de significância de 5% pelo teste F.

Critérios para decisão de um modelo ajustado e considerações finais

ANOVA

Causa da variação	GL	SQD	QMD	F_{cal}	Pr
Tratamentos	(6)	(128.985.714,29)			
Dev. regressão	1	126.350.000,00	126.350.000,00	27.929,26	< 0,0001
Ind. regressão	5	2.635.714,29	527.142,86	116,52	< 0,0001
Resíduo	28	126.670,00	4.523,93		
Total	34	129.112.384,29			

Conclusão: rejeita-se H0 ao nível de significância de 5% pelo teste F.

- O modelo é adequado à natureza do fenômeno em estudo, ou adequado ao que se sabe sobre o fenômeno?
- O coeficiente de determinação (r^2) é elevado?
- No quadro final da análise de variância o efeito do devido a regressão é significativo?
- No quadro final da análise de variância o efeito do devido ao independente da regressão é não significativo?

REGRESSÃO LINEAR MÚLTIPLA

- A análise de uma regressão múltipla segue, basicamente, os mesmos critérios da análise de uma regressão simples.

- A regressão múltipla envolve três ou mais variáveis, portanto, estimadores. Ou seja, ainda uma única variável dependente, porém duas ou mais variáveis independentes .

- A finalidade das variáveis independentes adicionais é melhorar a capacidade de predição em confronto com a regressão linear simples.

USOS DA REGRESSÃO MÚLTIPLA

- Ajustar dados estudando o efeito de uma variável X, levando em conta outras variáveis independentes.
- Obter uma equação para predizer valores de Y a partir dos valores de várias variáveis X1, X2, ...,Xk .
- Explorar as relações entre múltiplas variáveis (X1, X2,...,Xk)para determinar que variáveis influenciam Y.

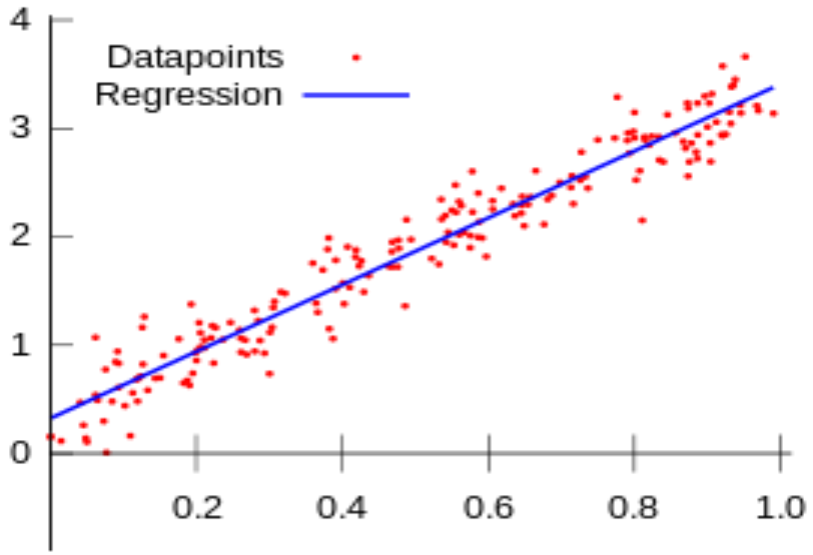
MODELO MATEMÁTICO

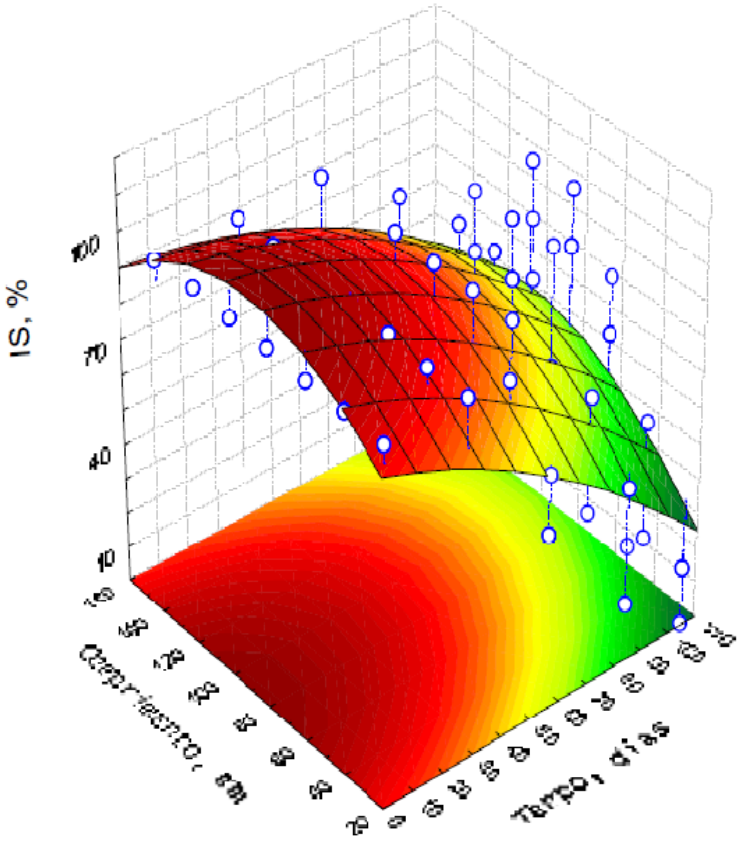
$$Yc = a + b_1x_1 + b_2x_2 + ... + b_kx_k$$

Onde:

- Yc = variável dependente;
- a = intercepto do eixo y;
- b = coeficiente angular da i-ésima variável;
- k = número de variáveis independentes.

- Enquanto uma regressão simples de duas variáveis resulta na equação de uma reta, um problema de três variáveis implica num plano, e um problema de k variáveis implica em um hiperplano;





- Também na regressão múltipla, as estimativas dos mínimos quadrados são obtidas pela escolha dos estimadores que minimizam a soma dos quadrados dos desvios entre os valores observados Yi e os valores ajustados Yc.

SOLUÇÃO DOS MÍNIMOS QUADRADOS

- A solução dos mínimos quadrados é a que minimiza a soma dos quadrados dos desvios entre os valores observados e a superfície de regressão ajustada.

$$\sum_{I=1}^{n}(Y_i-\hat{Y}_i)^2=\sum_{i=1}^{n}(Y_i-(\hat{\beta}_0+\hat{\beta}_1X_{1i}+\hat{\beta}_2X_{2i}+\ldots+\hat{\beta}_k\hat{X}_{ki})^2$$

COEFICIENTE DE DETERMINAÇÃO

- O **coeficiente de determinação múltipla** é uma medida de quão bem a equação e regressão múltipla se ajusta aos dados amostrais:
  - Ajuste perfeito: $r^2 = 1$.
  - Ajuste bom: r^2= prox. de 1.
  - Ajuste pobre: r^2= prox. 0.
- Defeito: Na medida em que mais variáveis são incluídas, r^2 cresce (pela simples inclusão de todas as variáveis disponíveis);
- Por causa dessa falha, a comparação de diferentes equações é feita mais adequadamente com o ajuste do coeficiente de determinação para o número de variáveis e o tamanho amostral.

REGRESSÃO NÃO LINEAR

- Os dados são modelados por uma função que é uma combinação não-linear de parâmetros do modelo e depende de uma ou mais variáveis independentes.

- Pode a partir de suposições importantes sobre o problema trabalhar no sentido de obter uma relação teórica entre as variáveis observáveis de interesse.

- Diferentemente do caso linear, é que os parâmetros entram na equação de forma não linear, assim, nós não podemos simplesmente aplicar fórmulas para estimar os parâmetros do modelo.

Exemplo:

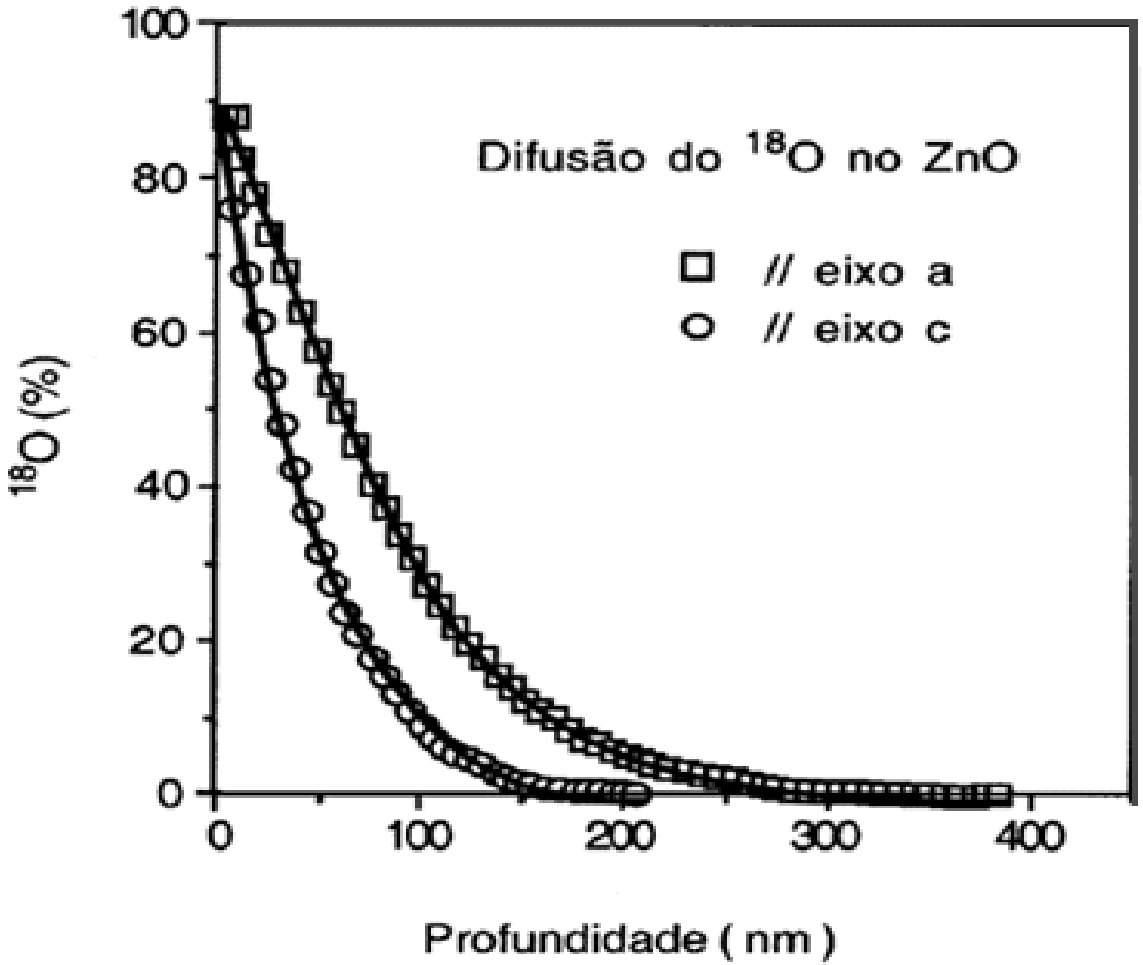


Figura 5: Anisotropia da difusão do oxigênio no ZnO monocristalino [14].

REFERÊNCIAS BIBLIOGRÁFICAS

- FARIA, J.C. **Notas de aulas expandidas**. Ilhéus, UESC, 2006.
- Wonnacott, Thomas H. **Estatistica aplicada a economia e a administracao** / Thomas H Wonnacott e Ronald J. Wonnacott, 1981

Obrigado !