



# REGRESSÃO LINEAR

UNIVERSIDADE ESTADUAL DE SANTA CRUZ

CURSO: AGRONOMIA

TURMA: 2023.2

DISCIPLINA: METODOLOGIA E ESTATÍSTICA EXPERIMENTAL – CET 076

DOCENTE: JOSÉ CLAUDIO FARIA

DISCENTES: JÉSSICA OLIVEIRA, JUAN MAIRO E KAIQUE FÉLIX.

# INTRODUÇÃO

- O estudo de regressão exerce um papel relevante dentro do campo da estatística experimental devido à sua ampla aplicação na interpretação de resultados experimentais. Seu objetivo é determinar a relação existente entre uma característica qualquer de interesse experimental, que é a variável dependente, e outra característica independente, quando consideradas em conjunto.
- O pesquisador seleciona os valores da variável independente e, em seguida, estabelece a relação existente entre os valores das duas variáveis. Essa relação é expressa por uma função matemática, conhecida como equação de regressão, na qual a variável dependente ( $Y$ ) é considerada uma função da variável independente ( $X$ ).

# CORRELAÇÃO

- A análise de correlação linear simples (Pearson 1896).
- Essa técnica é empregada, especificamente, para se avaliar o grau de covariância entre as variáveis aleatórias.
- Quando duas variáveis ( $Y_1$  e  $Y_2$ ), estão ligadas por uma relação estatística, dizemos que existe correlação entre elas.
- Utilizando duas variáveis dependentes, em função de ambas estão sujeitas a grandes variações e erros experimentais ponderáveis, como por exemplo:
  - Comprimento e largura de folhas de plantas.
  - Teor de potássio do solo e aumento de produção de cana-de-açúcar.
  - Ganho de peso e rendimento de carcaça em frangos de corte.

# 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:

$$r = \frac{\sum_{i=1}^n [(X_i - \bar{X}) \times (Y_i - \bar{Y})]}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2 \times \sum_{i=1}^n (Y_i - \bar{Y})^2}}$$

- Onde **n** é o número de termos

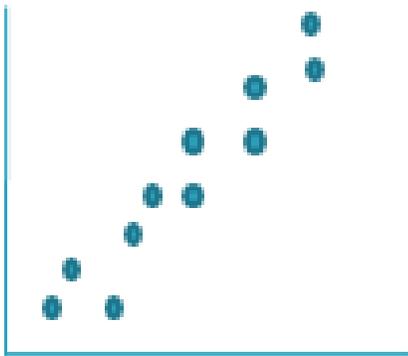
# CORRELAÇÃO

## Pressuposições de Correlação

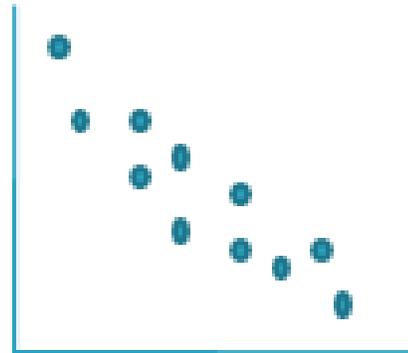
- O relacionamento entre as variáveis tem forma linear.
- As duas variáveis são aleatórias por natureza e medidas em escalas intervalares ou proporcionais, não podendo ser categóricas ou nominais.
- As variáveis apresentam distribuição normal bivariada.

# CORRELAÇÃO

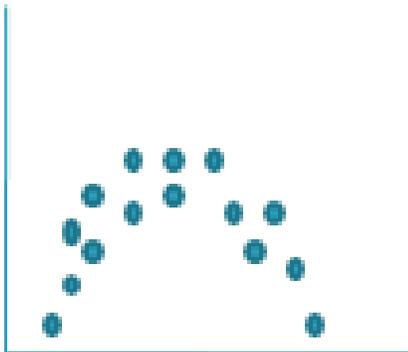
- Tipos de Diagramas de Dispersão:



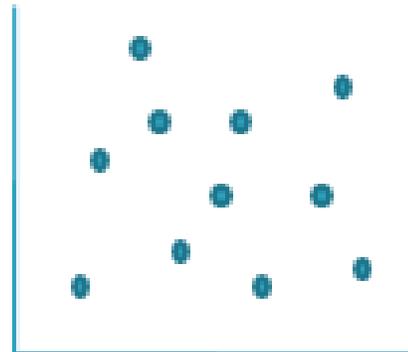
**Linear Positiva**  
(reta ascendente)



**Linear Negativa**  
(reta descendente)



**Não linear**  
(curva)



**Não há correlação**

# GRAU DE CORRELAÇÃO

- Os valores limites de **R** são de -1 até +1
- Se a correlação é perfeita e positiva o valor do **R** = +1
- Se a correlação é perfeita e negativa o valor do **R** = -1
- Se não há correlação então **R** = 0.  
Significa que as duas variáveis não estão linearmente associadas

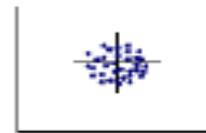
## ■ Intervalos

- Se  $|R| < 0,3$  a correlação NÃO EXISTE;
- Se  $0,3 \leq |R| \leq 0,6$  a correlação é FRACA;
- Se  $0,6 \leq |R| \leq 1$  a correlação é BOA;

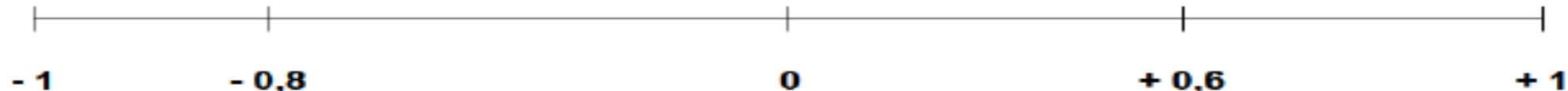
**Perfeita negativa**



**Não correlacionadas**



**Perfeita positiva**



← Aumenta grau de correlação  
negativa

Aumenta grau de correlação  
positiva →

- **Exemplo:** Considerando duas variáveis aleatórias

M: Rendimento acadêmico em matemática.

L: Rendimento acadêmico em línguas.

**Quadro 13.3 – Cálculo do coeficiente de correlação para o exemplo dado**

Obs	M	L	ML
1	36	35	1.260
2	80	65	5.200
3	50	60	3.000
4	58	39	2.262
5	72	48	3.456
6	60	44	2.640
7	56	48	2.688
8	68	61	4.148
	$\Sigma M = 480$	$\Sigma L = 400$	
n=8	$\Sigma M^2 = 30.104$	$\Sigma L^2 = 20.836$	$\Sigma ML = 24.654$
	$(\Sigma M)^2 = 230.400$	$(\Sigma L)^2 = 160.000$	

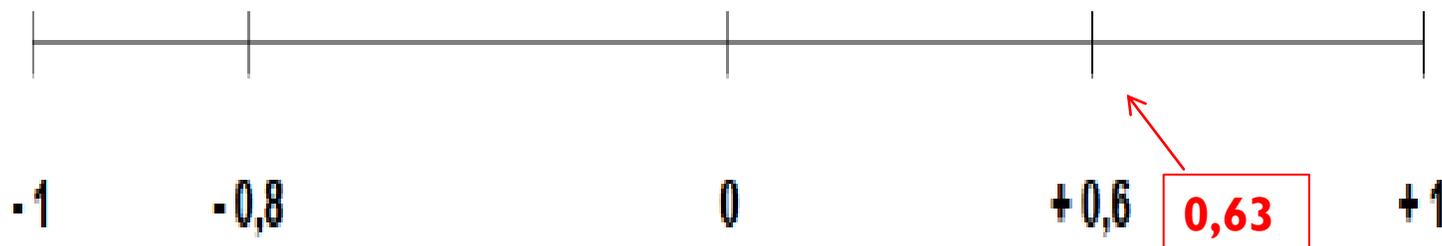
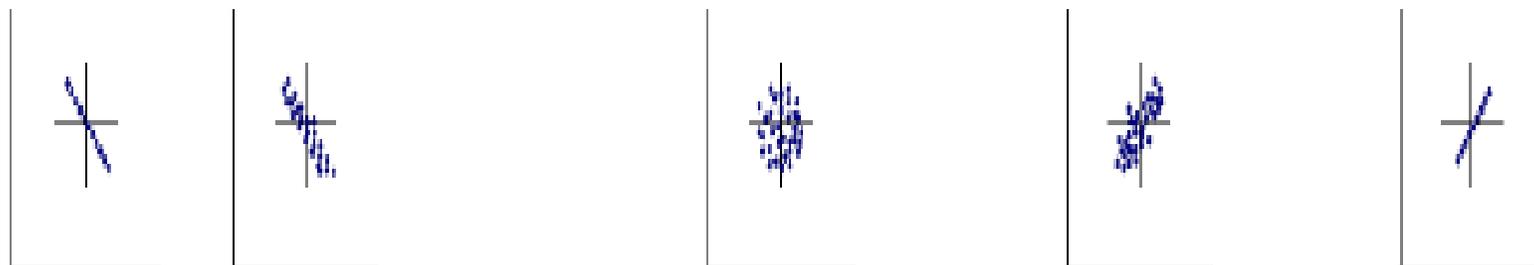
$$r(M, L) = \frac{n \cdot \sum ML - \sum M \times \sum L}{\sqrt{n \sum M^2 - (\sum M)^2} \times \sqrt{n \sum L^2 - (\sum L)^2}}$$

$$r(M, L) = \frac{8 \times 24.654 - 480 \times 400}{\sqrt{8 \times 30.104 - 230.400} \times \sqrt{8 \times 20.836 - 160.000}} = 0,63$$

Perfeita negativa

Não correlacionadas

Perfeita positiva



Aumenta grau de correlação  
negativa

Aumenta grau de correlação  
positiva

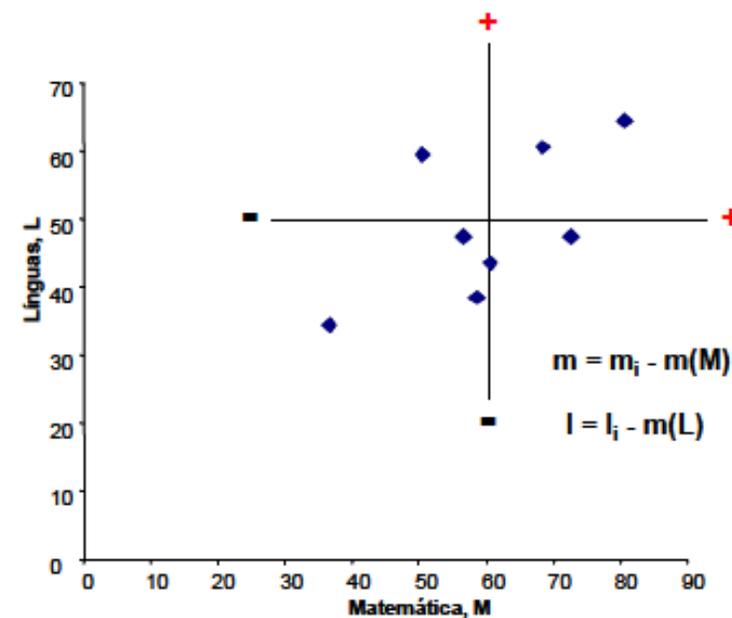


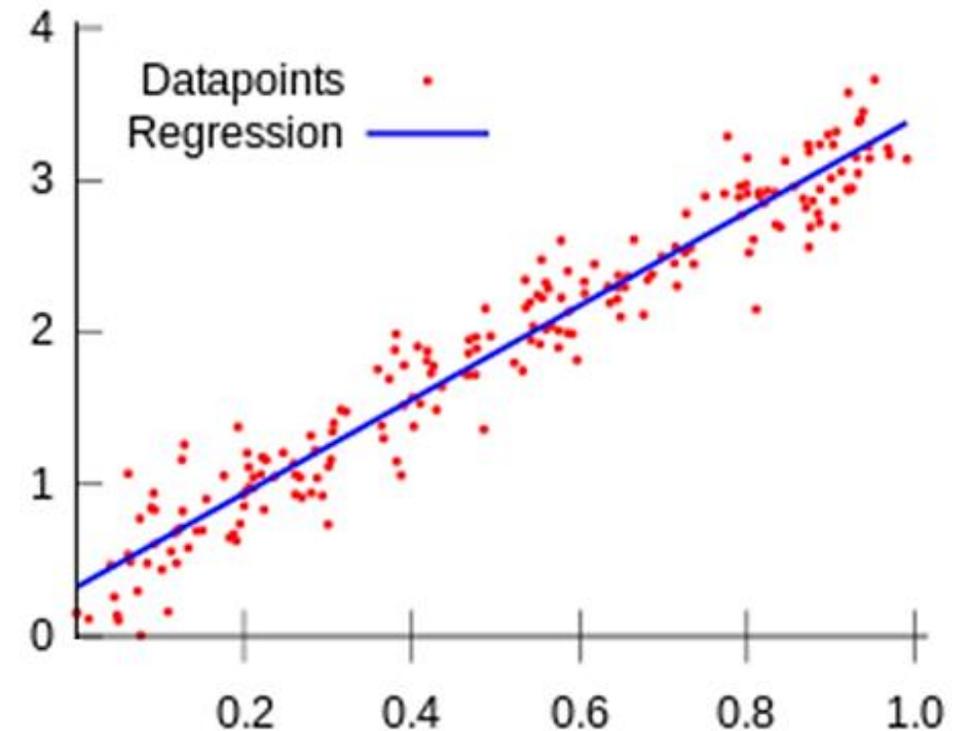
Figura 1 - Gráfico da dispersão entre  $m$  e  $l$  com as médias transladadas.

# REGRESSÃO LINEAR

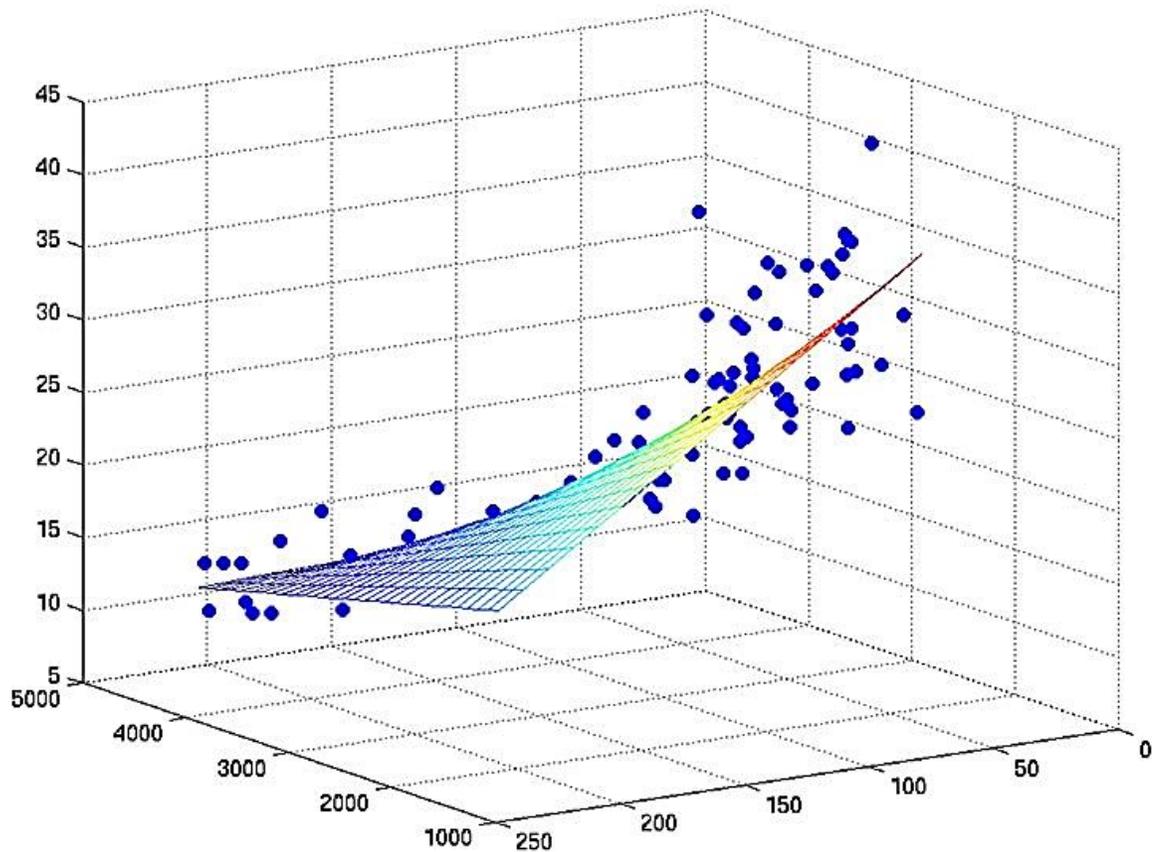
- É a relação casual entre duas ou mais variáveis quantitativas:
- Uma variável dependente ou resposta ( $Y$ ), cujo o valor deverá ser previsto.
- E uma ou mais variáveis independentes ou explicativas ( $X$ ), sobre as quais existem conhecimento teórico disponível.
- Resulta em uma equação geométrica equivalente a ajustar uma curva aos dados dispersos.

# REGRESSÃO LINEAR SIMPLES

- 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 da outra.
- Pressupõe-se alguma relação de causa e efeito, de explanação do comportamento entre as variáveis.
- Exemplos:
  - A idade e o peso de cada bezerro;
  - A alíquota de imposto e a arrecadação;
  - Preço e quantidade.



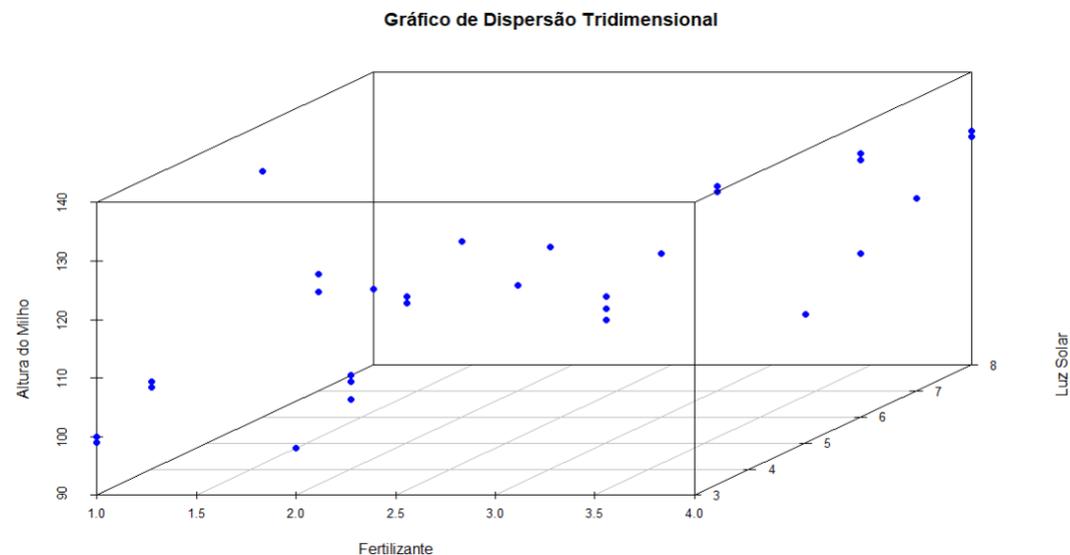
# REGRESSÃO LINEAR MÚLTIPLA



- Relação casual com mais de duas variáveis. Isto é, quando o comportamento de  $Y$  é explicado por mais de uma variável independente  $X_1, X_2, \dots, X_n$ .
- É a técnica adequada para se utilizar quando se quer investigar simultaneamente os efeitos, sobre  $Y$ , de duas ou mais variáveis preditoras.

## EXEMPLO:

- A análise de regressão múltipla permite investigar a relação entre a altura do milho e as variáveis independentes de quantidade de fertilizante e quantidade de luz solar em um estudo com 30 plantas.
- Através da análise estatística, é possível determinar se o fertilizante e a luz solar têm efeitos significativos na altura do milho, fornecendo percepções valiosas para otimizar o crescimento das plantas.
- A utilização da regressão múltipla permite construir um modelo preditivo que pode ser usado para fazer previsões sobre a altura do milho com base nos valores de fertilizante e luz solar, auxiliando no planejamento agrícola.



# EQUAÇÃO DA REGRESSÃO

- Criar um modelo de equação de reta para fazer previsões/estimativas de valores futuros através dos pontos.

A equação da Reta é uma **equação de 1º Grau**.

OBJETIVO = Ajustar uma reta

- Onde

X – é a variável explicativa ou independente;

Y – é a variável explicada ou dependente (aleatória);

$\alpha$  – coeficiente linear ou constante da regressão, representa o interceptor da reta com eixo do Y;

$\beta$  – coeficiente de regressão ou coeficiente angular da reta. Representa a variação de Y em função da variável X;

$\alpha$  e  $\beta$  são parâmetros.

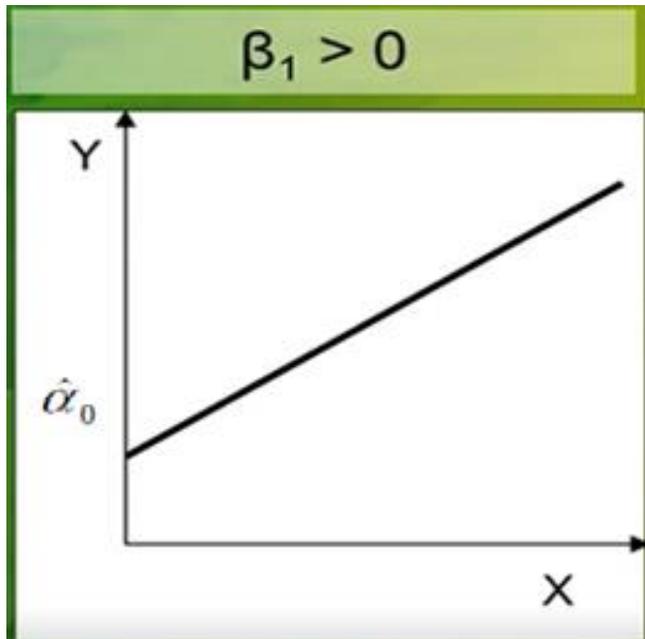
$$\hat{Y} = \hat{\alpha}_0 + \hat{\beta}X$$

# REGRESSÃO LINEAR

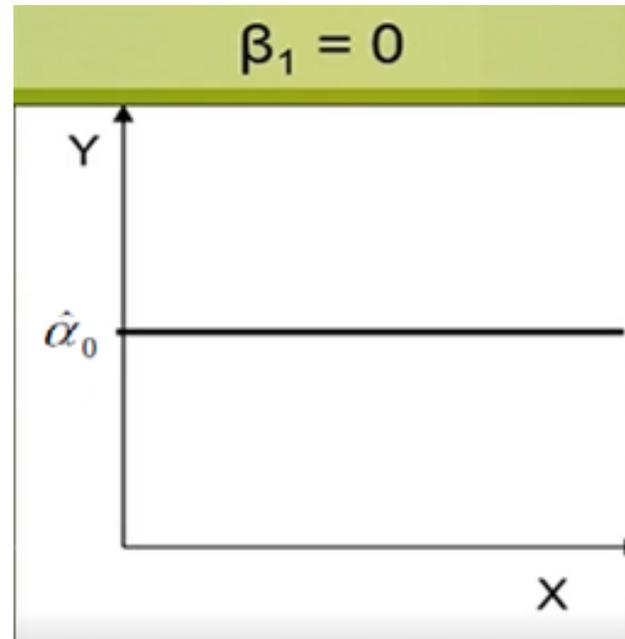
- Padrão da variação da regressão linear

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

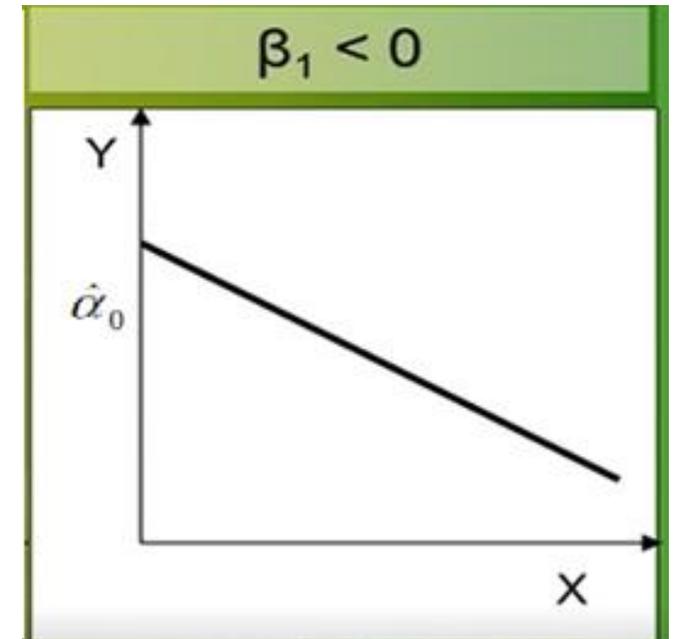
- Reta Ascendente



- Reta Constante

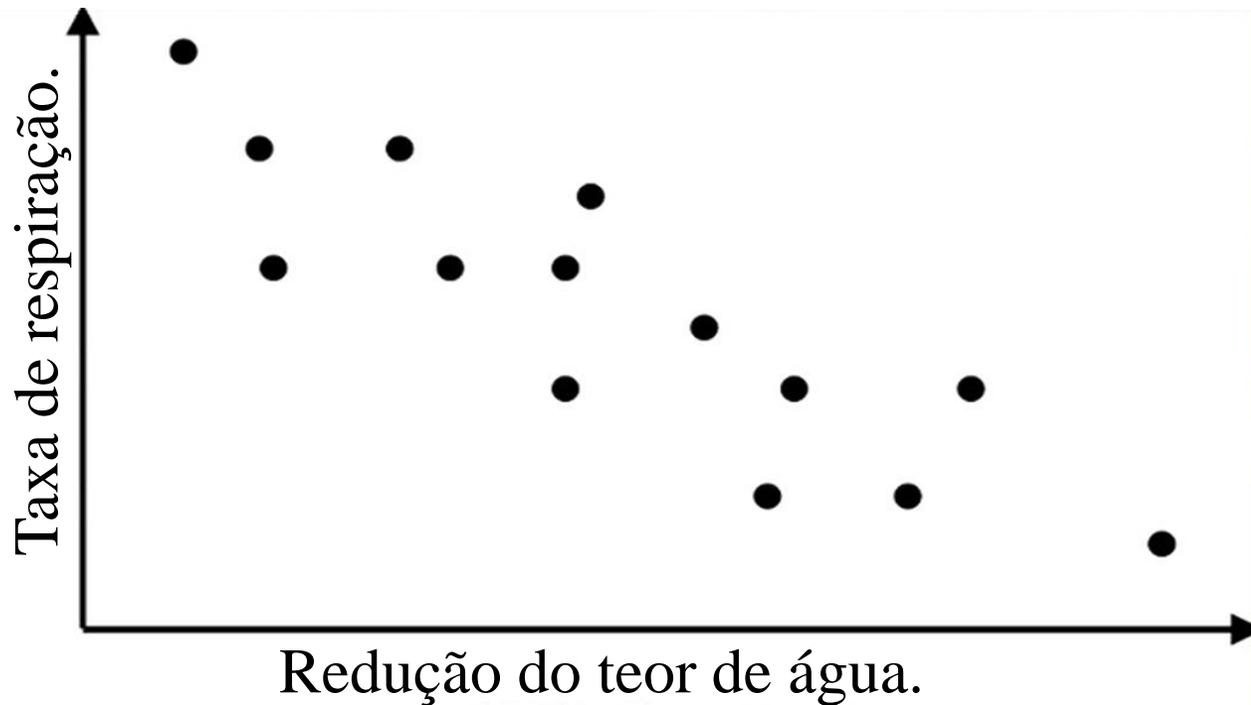


- Reta Descendente



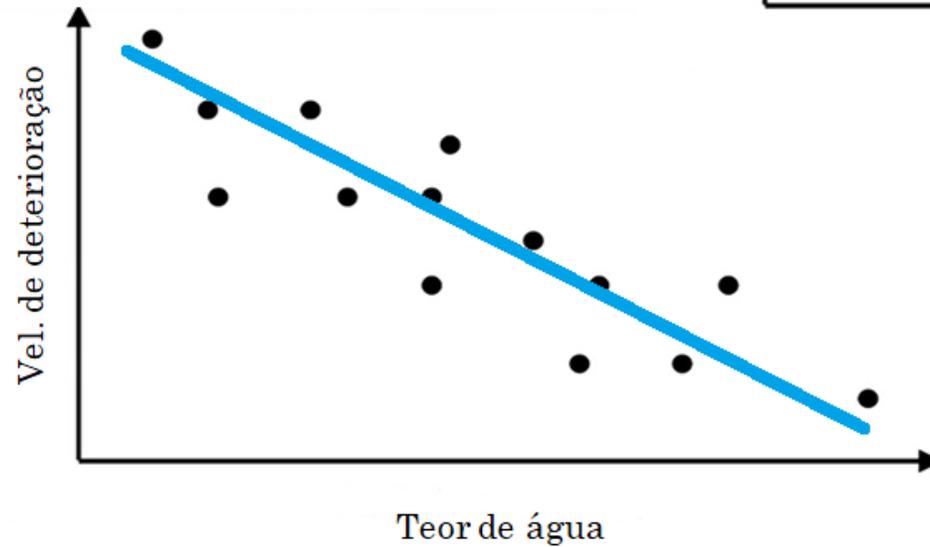
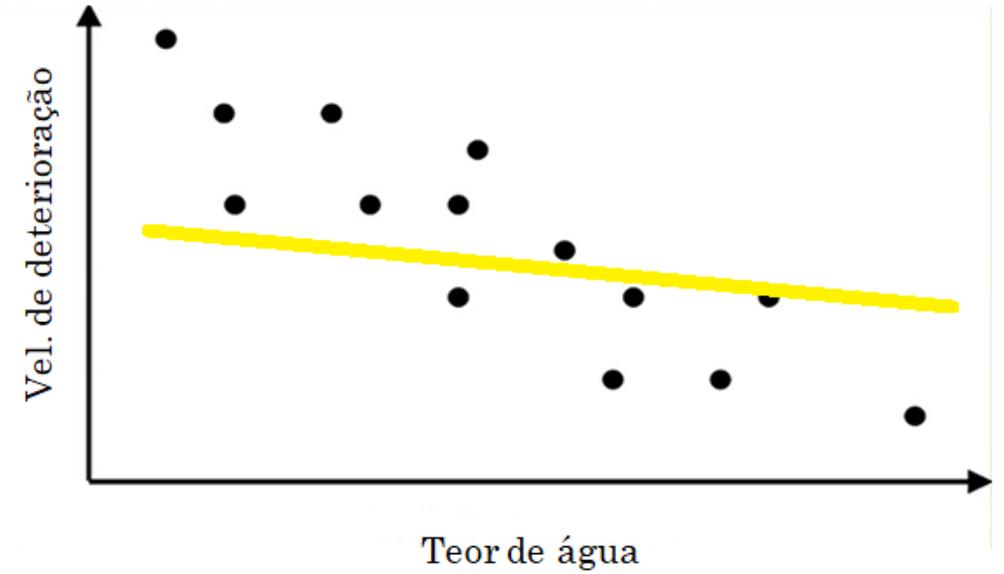
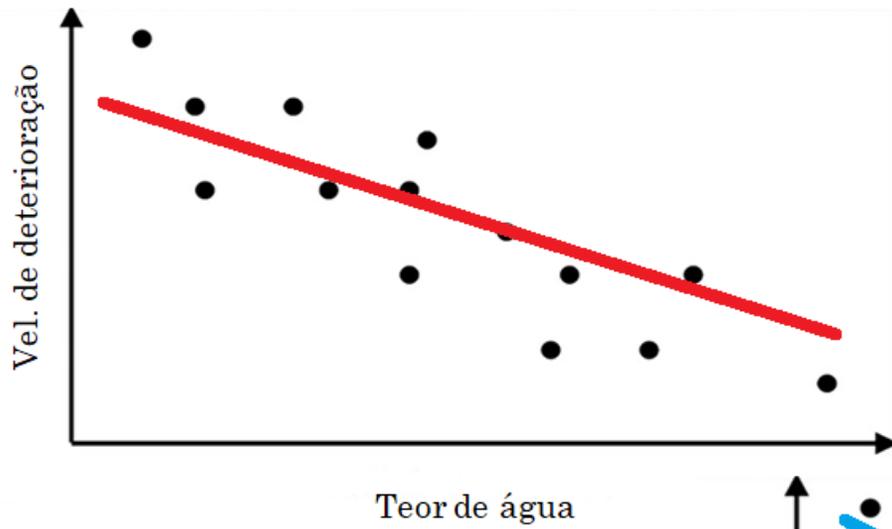
## ■ Exemplo

Um dos destinos das sementes após a colheita são as UBSs, nelas os lotes serão avaliadas para determinar se vai haver necessidade de submeter as sementes a algum tipo de beneficiamento, com o objetivo de melhorar suas qualidade físicas e capacidade de armazenamento.

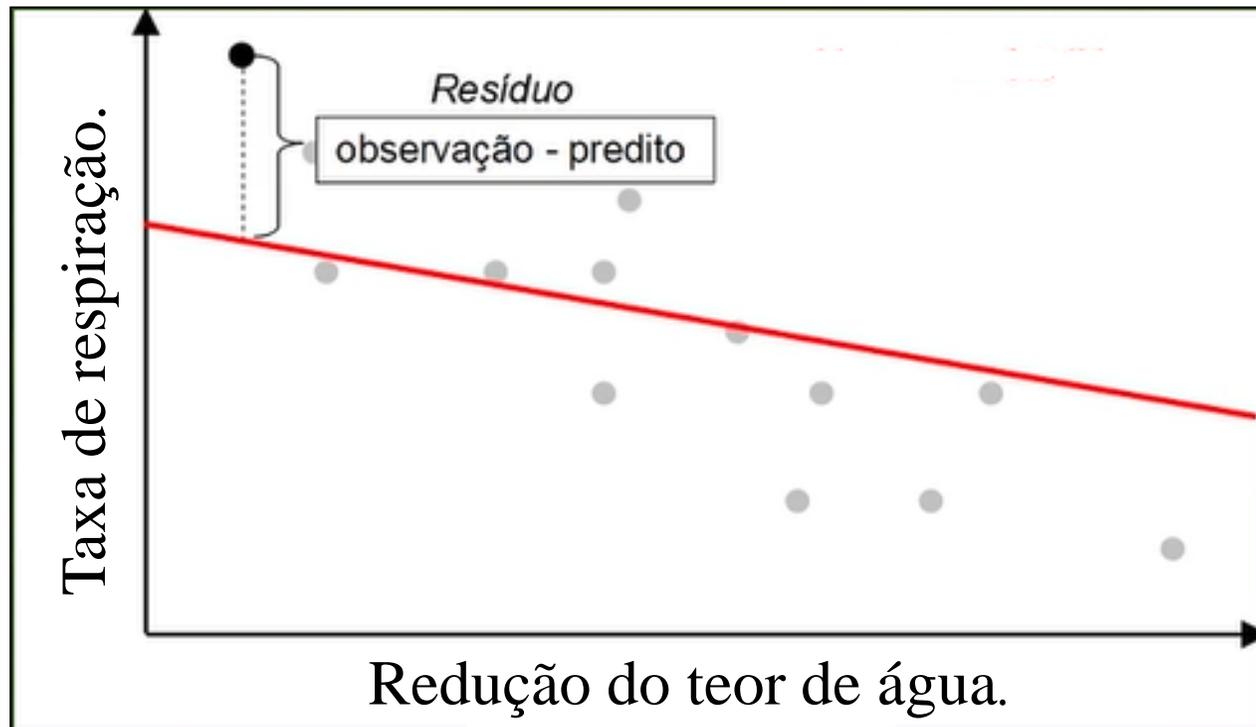


(Y) Variável dependente – Taxa de Respiração  
(X) Variável Independente – Teor de água

# QUAL A MELHOR REPRESENTAÇÃO DA RETA?

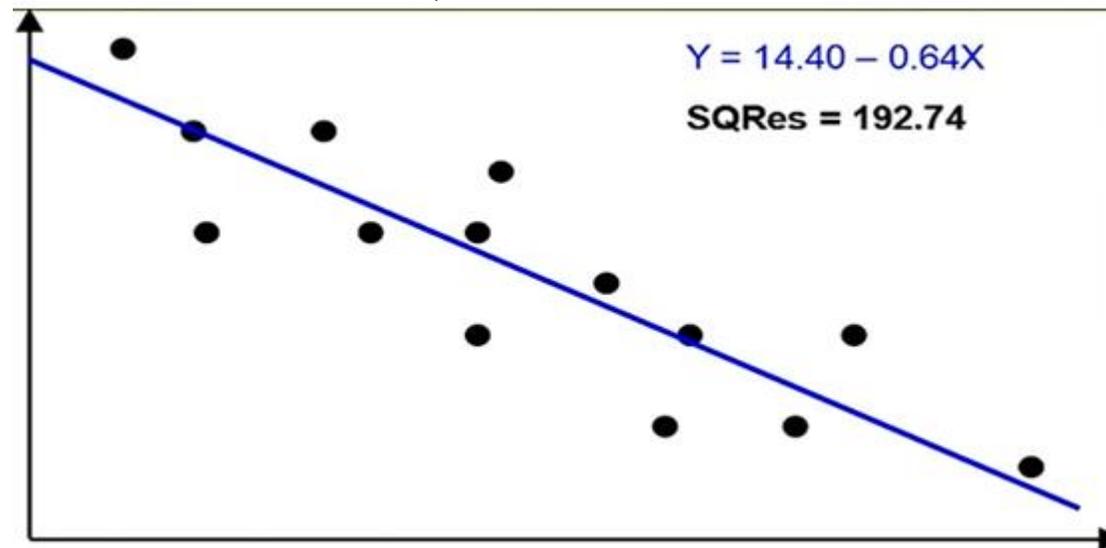
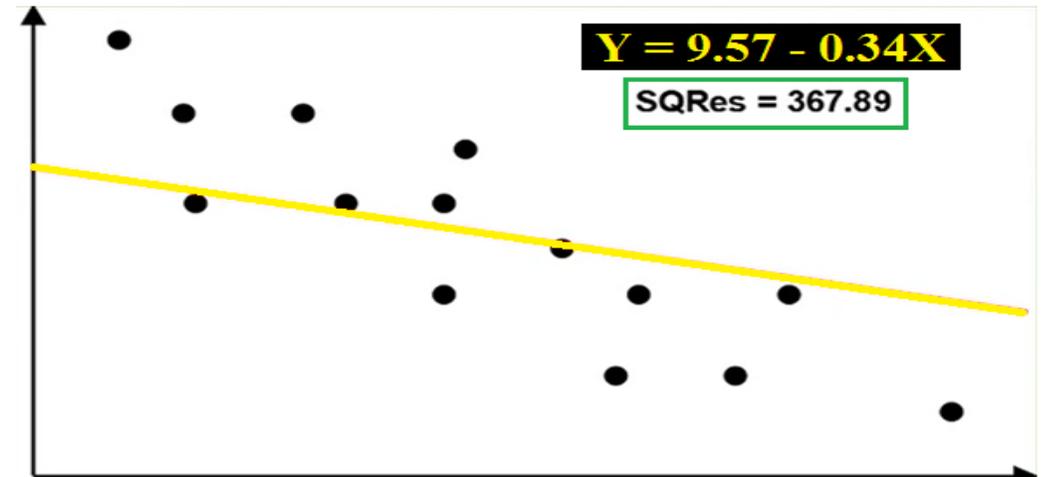
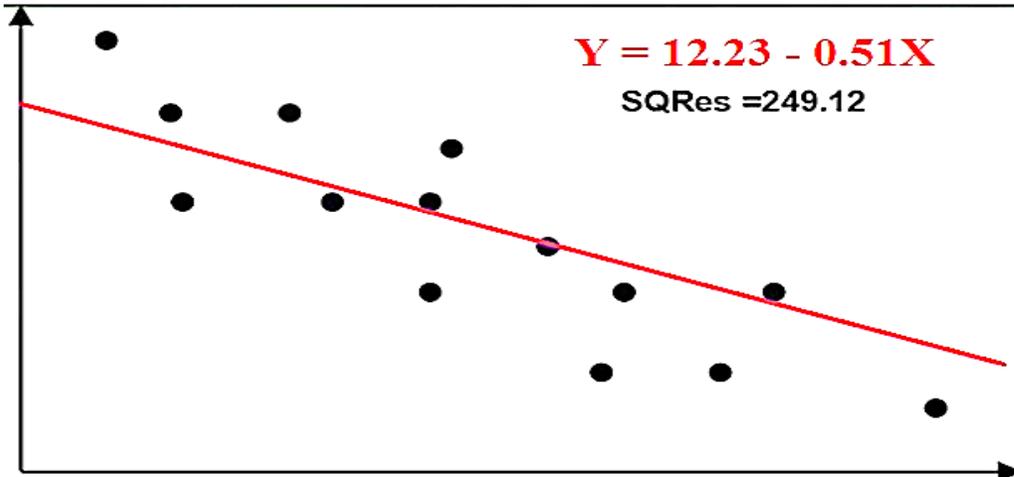


# PORQUE DEVO AJUSTAR A RETA?



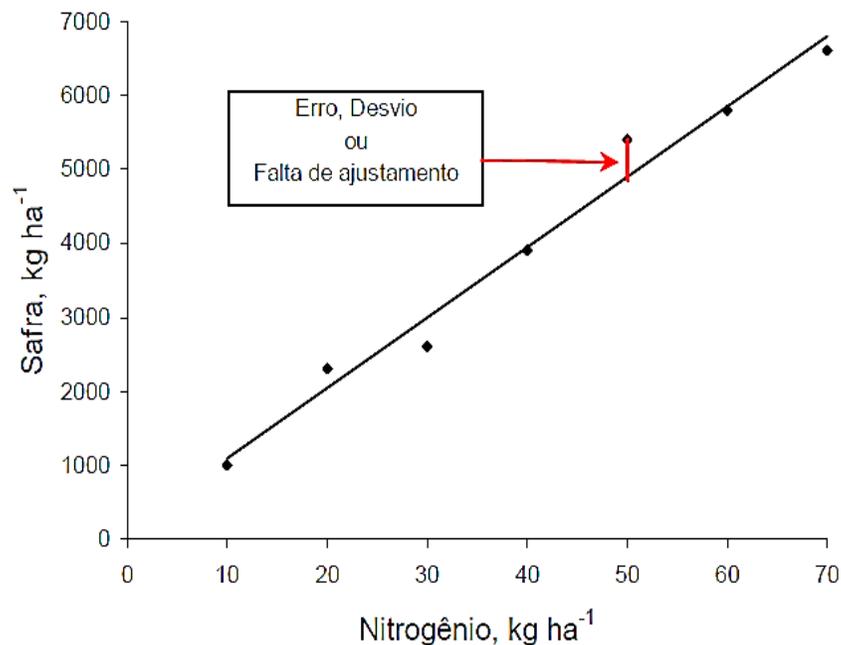
- Temos que reduzir a distância entre as observações e a reta.

# REGRESSÃO LINEAR



# CRITÉRIOS PARA O AJUSTAMENTO DA RETA

- O que é um bom ajustamento?  
É aquele ajuste que causa pequeno erro total.



O erro ou 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:  $(Y_i - \hat{Y}_i)$  (erro)

Figura 2 - 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:

Soma de quadrado dos erros  $\sum_{i=1}^n (Y_i - \hat{Y}_i)^2$

- 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  $\alpha$  e  $\beta$ ;
- Minimizando a soma do quadrado de erros, encontraremos  $\alpha$  e  $\beta$  que trarão a menor diferença entre a previsão de  $Y_i$  e  $\hat{Y}_i$ .

# EXEMPLO:

Quadro 1. Relação observada entre a safra e a aplicação de Nitrogênio

X Nitrogênio kg ha <sup>-1</sup>	Y Safra kg ha <sup>-1</sup>
10	1.000
20	2.300
30	2.600
40	3.900
50	5.400
60	5.800
70	6.600

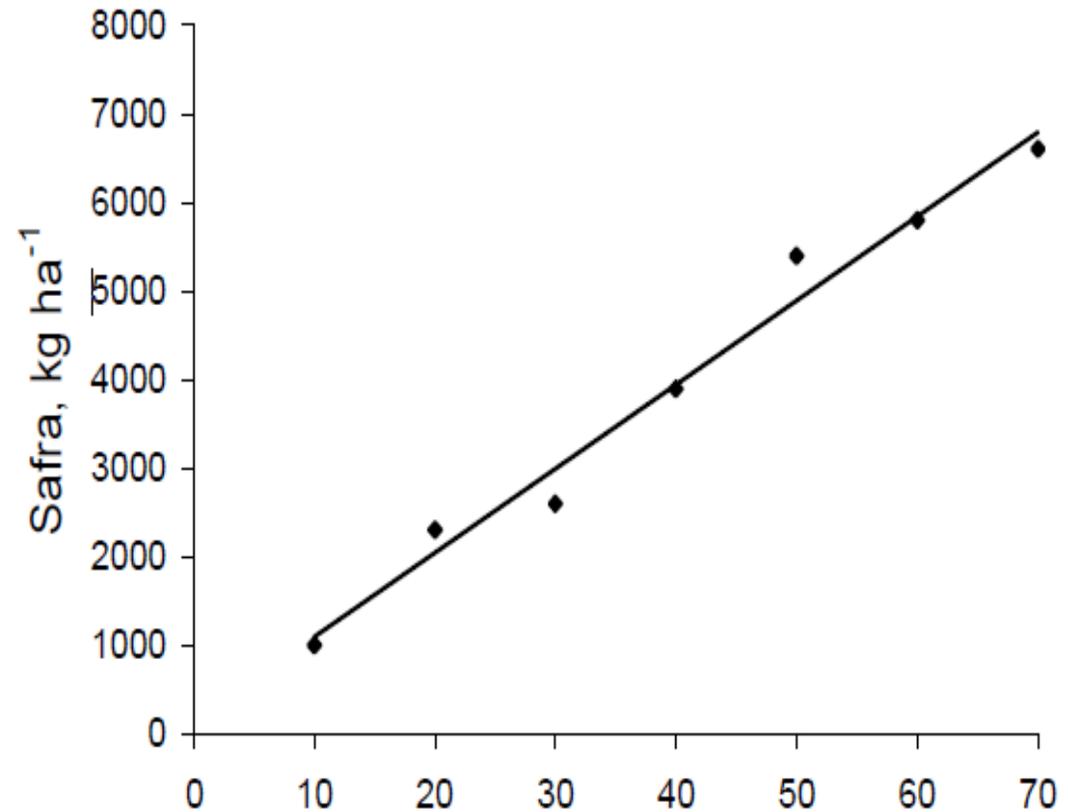
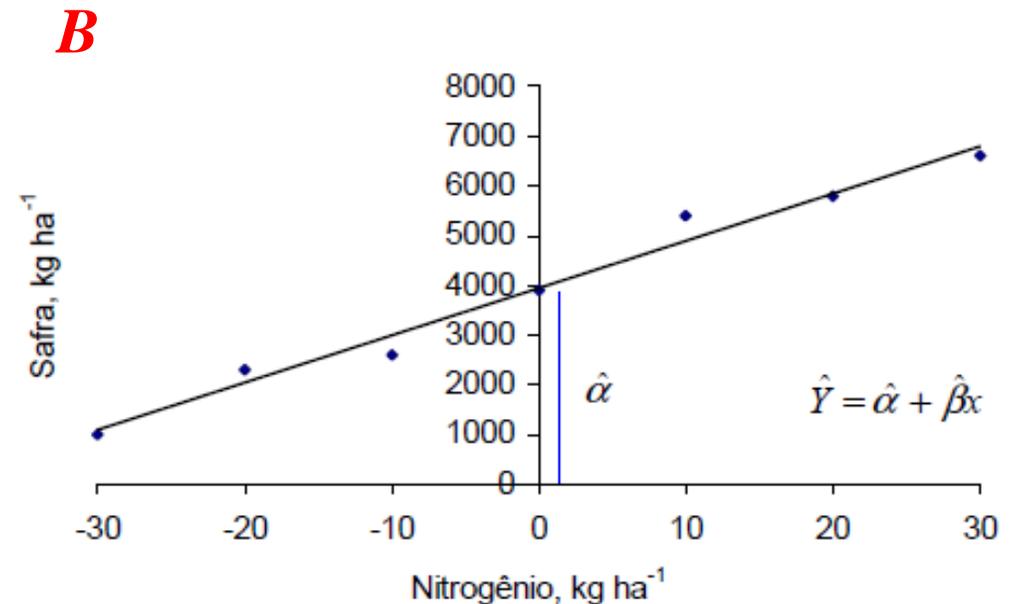
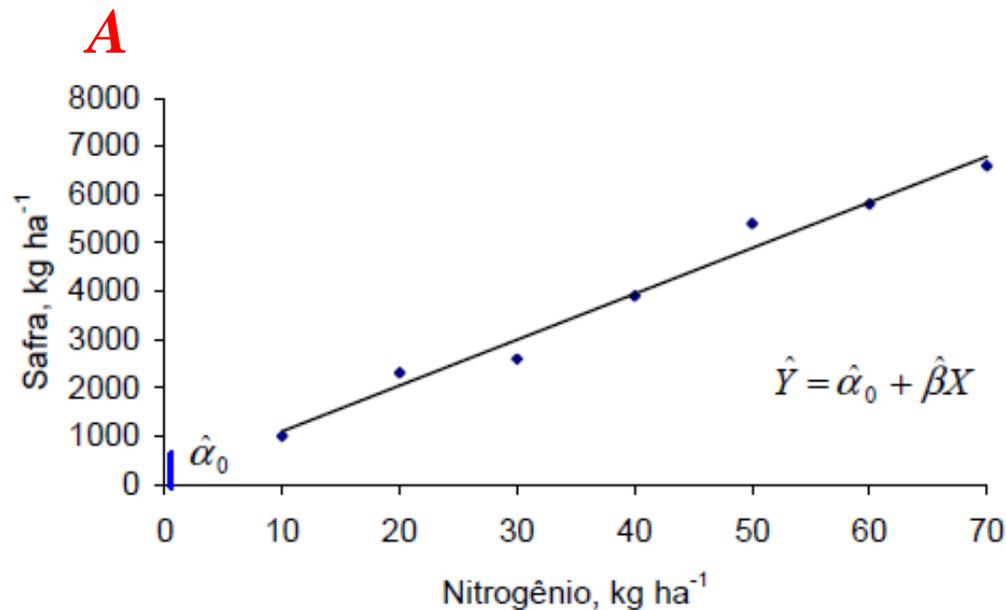


Figura 3 - Dados e reta ajustada a olho aos dados apresentados.

# TRANSLAÇÃO DE EIXOS

- Para iniciarmos o ajuste de reta devemos fazer a translação de eixos, conservando os valores originais.



**Como isso é feito?**

# ESTÁGIO 1

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

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

média

tratamento

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

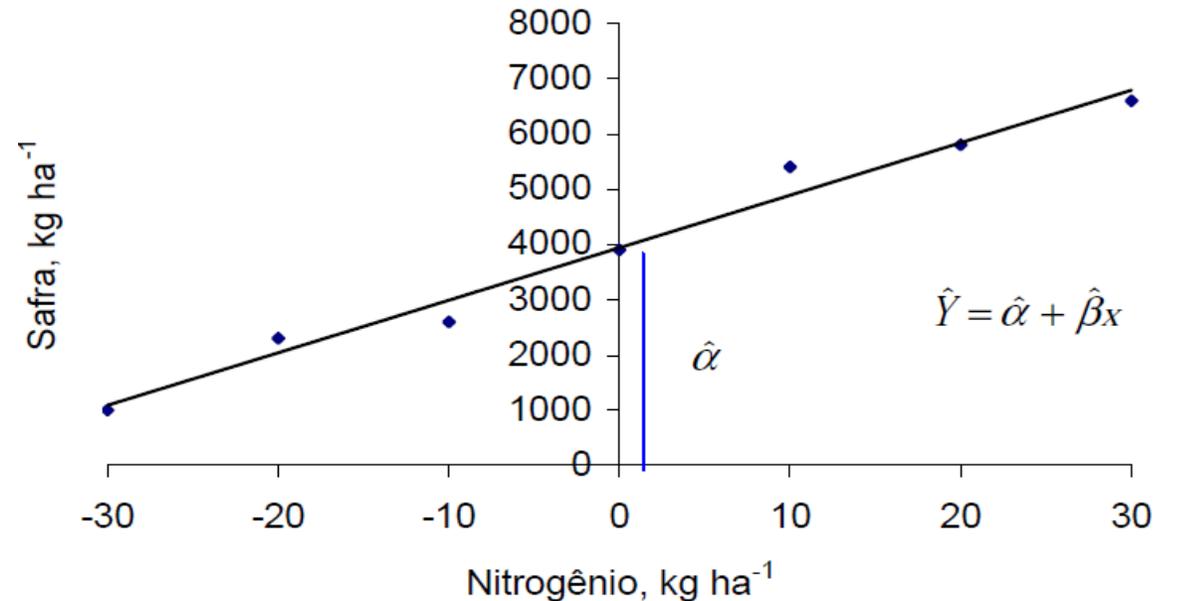
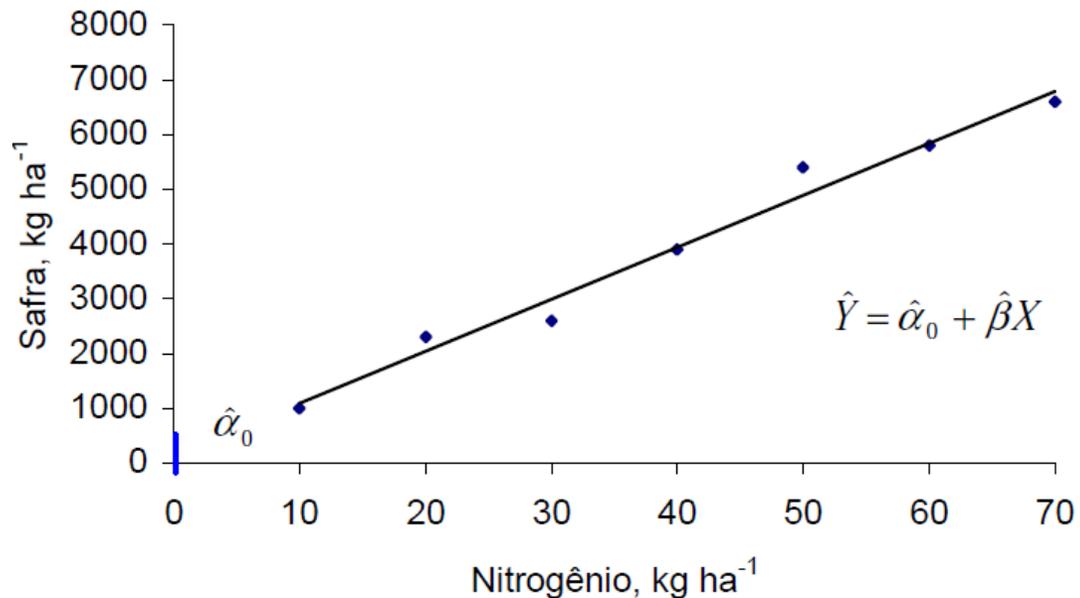
$$\sum X = 280$$

$$\bar{X} = \frac{1}{N} \sum X$$

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

$$\sum x = 0$$

# EQUIVALE A UMA TRANSLAÇÃO GEOMÉTRICA DE EIXOS



Observa-se que o eixo Y foi deslocado para a direita, de 0 a  $\bar{X}$  (média)

O novo valor x torna-se positivo, ou negativo, conforme X esteja a direita ou a esquerda de  $\bar{X}$  (média)

Não há modificação nos valores de  $\hat{Y}_i$ .

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  $X$  simplifica os cálculos porque a soma dos novos valores  $x$  é igual a zero, isto é:

$$\sum x_i = 0 \quad \therefore \quad \sum x_i = \sum (X_i - X) = \sum X_i - nX = nX - nX = 0$$

Somatório de  $x$

# ESTÁGIO 2

- 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:

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

Equação 01

Soma de quadrado dos erros

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

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

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

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

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

- Usa-se a notação  $S(\hat{\alpha}, \hat{\beta})$ , para enfatizar que a expressão tem dependência dos valores de  $\hat{\alpha}$  e  $\hat{\beta}$ .

- Pergunta-se então, para que valores de  $\hat{\alpha}$  e  $\hat{\beta}$  haverá um mínimo de erros?  
A resposta a essa pergunta nos fornecerá a reta “ótima” (de mínimos quadrados dos erros).
- A técnica de minimização mais simples é fornecida pelo cálculo.
- A  $S(\hat{\alpha}, \hat{\beta})$  minimização de exige o anulamento simultâneo de suas derivadas parciais.
- Igualando a zero a derivada parcial em relação a  $\hat{\alpha}$  :

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

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

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

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

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

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

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

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

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

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

- Dividindo ambos os termos por (-2):

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

➤ Reagrupando:

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

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

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

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

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

❖ 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:

$X$	$x = X - \bar{X}$ $x = X - 40$	$Y$	$xY$	$x^2$
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

$$\sum X = 280$$

$$\bar{X} = \frac{1}{N} \sum X$$

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

$$\sum x = 0$$

$$\sum Y = 27.600$$

$$\bar{Y} = \frac{1}{N} \sum Y$$

$$\bar{Y} = \frac{27.600}{7}$$

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

$$\sum xY = 266.000$$

$$\sum x^2 = 2.800$$

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

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

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

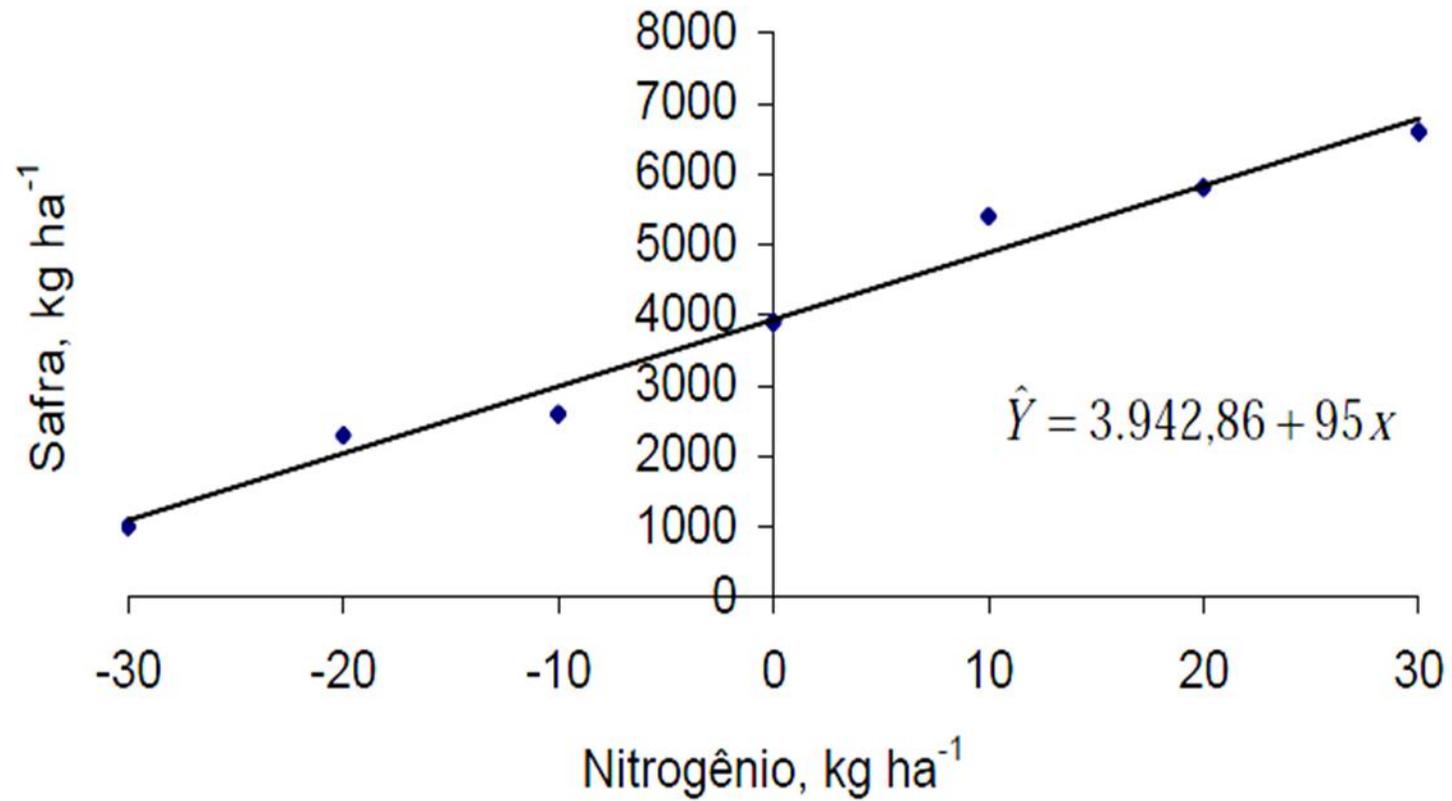


Figura 6 - Equação da reta com o eixo X transladado.

# Estágio 3

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

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

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

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

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

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

- 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.

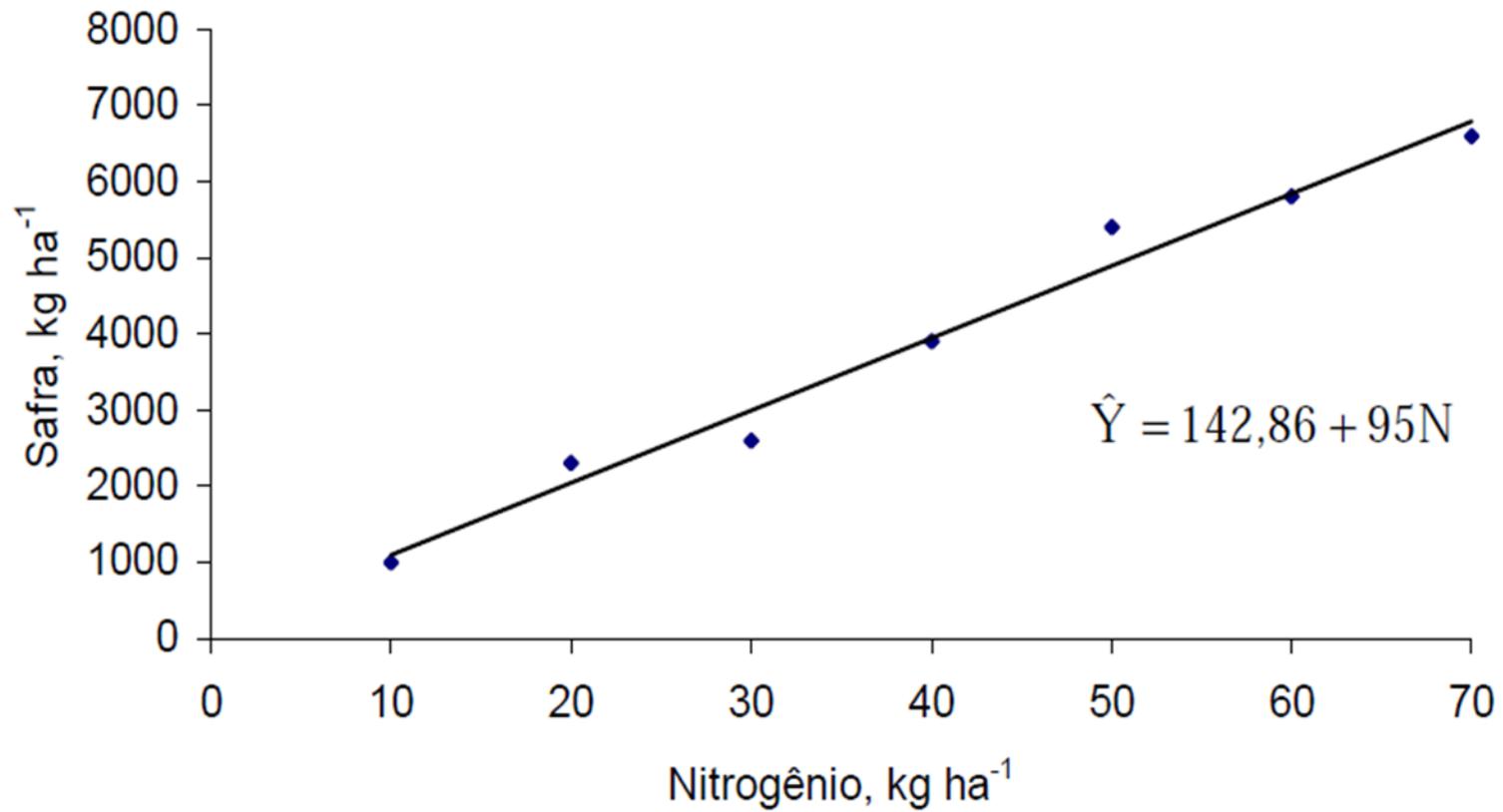
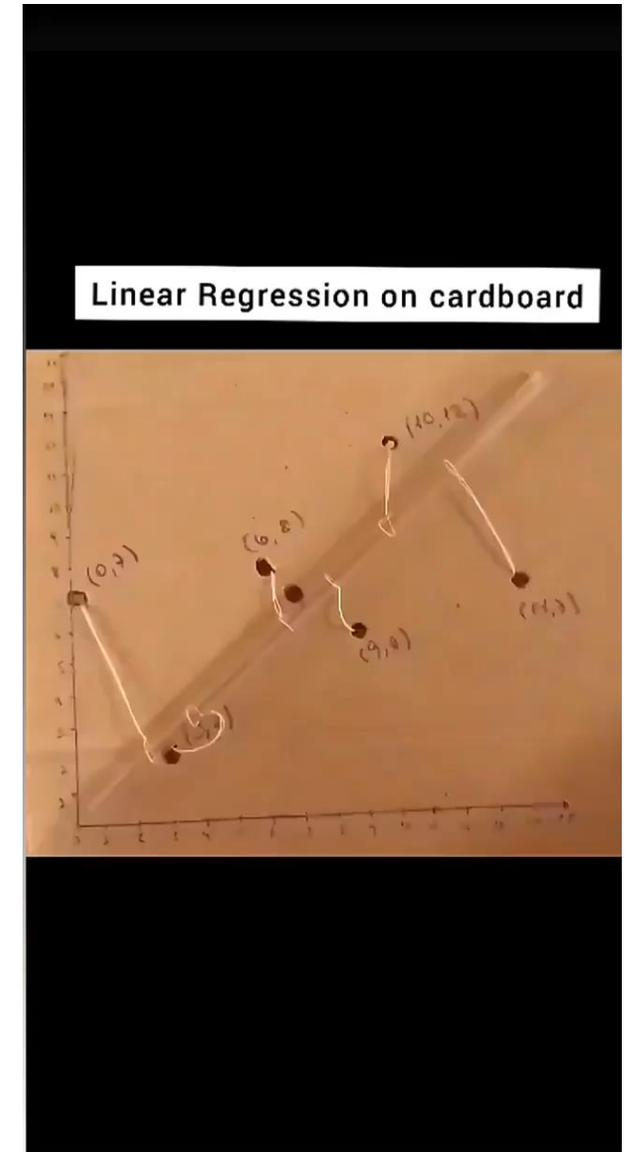


Figura 7 - Gráfico dos pontos dispersos com a reta ajustada.



# 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 ( $SQD_{tot}$ ) 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 ( $SQDD_{reg}$ ): 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 ( $SQDD_{err}$ ): 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.

# ANÁLISE DE VARIÂNCIA DA REGRESSÃO

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

Hipóteses:

$H_0:  \beta_i  = 0$	ou	$H_0: Y \neq \alpha_0 + \beta X$
$H_1:  \beta_i  > 0$	ou	$H_1: Y = \alpha_0 + \beta X$

- Significado de  $H_0$ : 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	<i>p-value</i>
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:  $H_0: B_1=B_2= \dots =B_k=0$  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  $N(0, \sigma^2)$ .

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			

# 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 previsto 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  $Y$  ( $\beta Y$ ) no eixo horizontal ( $x$ ).

# OBSERVAÇÕES A RESPEITO DA REGRESSÃO

- Quando os dados provêm de um delineamento experimental, onde são observadas repetições, e por conseguinte existe um erro experimental, além do erro devido a falta de ajuste do modelo:
- O ajustamento segue os mesmos princípios, ou seja, geralmente, é realizado observando-se as médias de cada tratamento;
- A análise de variância sofre ligeiras alterações;

## 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<sup>-1</sup>

N kg.ha <sup>-1</sup>	Repetições					Totais	Rep.	Médias
	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

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

$$SQD_{tot} = [(1.000)^2 + (916)^2 + \dots + (6.645)^2] - C = 129.112.384,29$$

$$SQD_{tra_t} = 1/5 [(5.000)^2 + (11.510)^2 + \dots + (33.000)^2] - C = 128.985.714,29$$

$$SQD_{res} = SQD_{tot} - SQD_{tra} = 129.112.384,29 - 128.985.714,29 = 126.670,00$$

### Hipóteses:

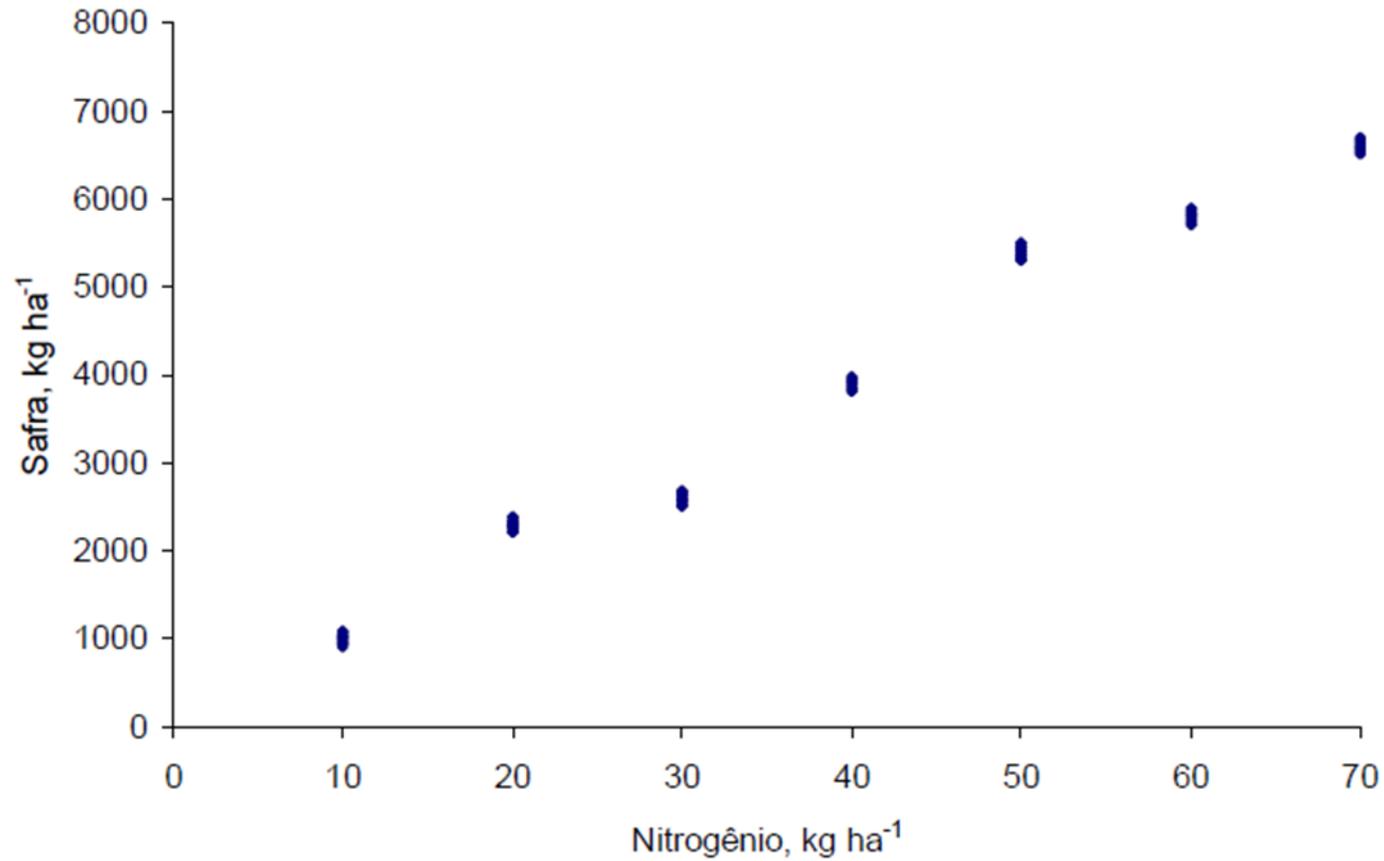
$$H_0: \mu_{10} = \dots = \mu_{70}$$

$H_1$ : 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  $H_0$  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 A RETA

- Ajustando o modelo linear:  $\hat{Y} = \alpha_0 + \beta_1 X$
- Valores necessários para o ajustamento do modo linear:

X	$x = X - \bar{X}$ $x = X - 40$	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
$\sum X = 280$		$\sum Y = 27.600$		
$\bar{X} = \frac{1}{N} \sum X$	$\sum x = 0$	$\bar{Y} = \frac{1}{N} \sum Y$	$\sum xY = 266.000$	$\sum x^2 = 2.800$
$\bar{X} = \frac{280}{7} = 40$		$\bar{Y} = \frac{27.600}{7}$		
		$\bar{Y} = 3.942,86$		

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

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

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

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

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

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

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

Equação da reta ajustada:

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

# ANÁLISE DE VARIÂNCIA DA REGRESSÃO: ANOVAR

## ANOVAR

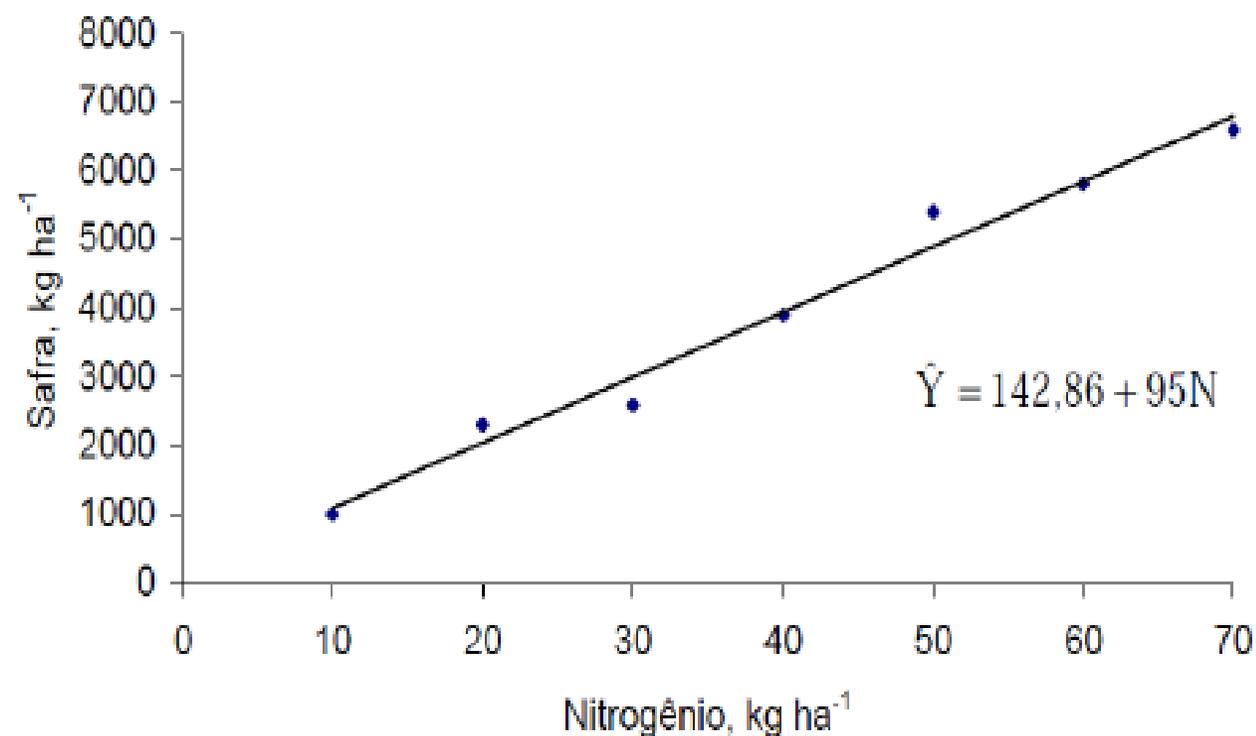
Causa da variação	GL
Regressão	1
Erro	5
Total	6

- 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 (SQD<sub>tot</sub>) 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 (SQDD<sub>reg</sub>) – Soma do quadrado dos desvio devido à regressão
  - Uma outra parte associada à falta de ajuste (SQDD<sub>err</sub>) – Soma do quadrado dos desvios devido ao erro

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

Vejamos<sup>1</sup>:

N , kg ha <sup>-1</sup>	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)	$m_{(\text{Erro})}$	Erro- $m_{(\text{Erro})}$	$[\text{Erro}-m_{(\text{Erro})}]^2$
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
					<u>527.142,86</u>

## ANOVAR

Causa da variação	GL	SQD	QMD	$F_{\text{cal}}$	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:

$$\text{SQD}_{\text{tot}} = \sum Y_i^2 - \frac{(\sum Y_i)^2}{n}$$

$$\text{SQD}_{\text{reg}} = \hat{\alpha}_0 \sum Y_i + \hat{\beta} \sum X_i Y_i - \frac{(\sum Y_i)^2}{n}$$

$$\text{SQD}_{\text{err}} = \text{SQD}_{\text{tot}} - \text{SQD}_{\text{reg}}$$

X	Y	Y <sup>2</sup>	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

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

$$\text{SQD}_{\text{reg}} = \hat{\alpha}_0 \sum Y_i + \hat{\beta} \sum X_i Y_i - \frac{(\sum Y_i)^2}{n}$$

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

$$\text{SQD}_{\text{reg}} = 25.270.000$$

$$\text{SQD}_{\text{err}} = \text{SQD}_{\text{tot}} - \text{SQD}_{\text{reg}}$$

$$\text{SQD}_{\text{err}} = 25.797.142,86 - 25.270.000$$

$$\text{SQD}_{\text{err}} = 527.142,86$$

# ILUSTRAÇÃO DA ANOVAR APENAS PARA EFEITO DE COMPARAÇÃO COM A ANOVA

## 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			

## 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

- O coeficiente de determinação do modelo de regressão,  $r^2$ , é uma medida do grau de ajuste do modelo aos dados experimentais:

$$r^2 = \frac{SQD_{reg}}{SQD_{tot}} \quad \therefore \quad 0 \leq r^2 \leq 1$$

- Este coeficiente, nos dá uma informação do quão bem, ou não, o modelo utilizado se ajusta a natureza dos dados experimentais. Para o exemplo em análise:

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

- Interpretação: 97,96% da variação total da safra, em decorrência da variação da dose de nitrogênio, é explicada pelo modelo de regressão ( $\hat{Y} = 142,86 + 95N$ ) ajustado.

# CONSIDERAÇÕES SOBRE A ANOVAR

- Observa-se que a soma de quadrados, e os respectivos graus de liberdade, associados aos tratamentos foram desdobrados em duas partes:
  - Uma parte associada ao modelo de regressão utilizado (SQDDreg) ( $\hat{Y} = 142,86 + 95N$ ).
  - Uma parte associada à falta de ajuste ou erro de ajustamento (SQDDerr)
- Na análise de regressão em um experimento com repetições utiliza-se a ANOVA, sendo que:
  - Para a obtenção da soma de quadrados do desvio devido à regressão e ao independente da regressão tem-se duas opções:
    - a) 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.
    - b) Utilizar o teorema do limite central (que facilita bastante os cálculos).

# TEOREMA DO LIMITE CENTRAL

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

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

$$SQDD_{\text{reg}} = 25.270.000,00 \overset{\text{(regressão)}}{\times 5} = 126.350.000,00$$

$$SQDD_{\text{ireg}} = 527.142,86 \overset{\text{(resíduo)}}{\times 5} = 2.635.714,29$$

ANOVA

Causa da variação	GL	SQD	QMD	F <sub>cal</sub>	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			

➤ Critérios para decisão de um modelo ajustado e considerações finais

ANOVA

Causa da variação	GL	SQD	QMD	F <sub>cal</sub>	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			

- 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?

## ➤ Critérios para decisão de um modelo ajustado e considerações finais

- Nem sempre se consegue respostas favoráveis a todo o conjunto destes pontos (a ... d).
- Quanto mais próximo da situação ideal: melhor o modelo ajustado.
- É necessário bom censo e muita prática para se realizar bons ajustes de modelos de regressão aos dados experimentais.
- Individualmente, a análise de regressão é um dos mais amplos tópicos da estatística e da estatística experimental.
- A abordagem utilizada, embora não seja a usual para trabalhos do dia a dia, é a mais simples, prática e objetiva para um estudo introdutório, possibilitando um entendimento inicial claro aos modelos de regressão linear.

# 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, 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

- Obter uma equação para prever valores de  $Y$  a partir dos valores de várias variáveis  $X_1, X_2, \dots, X_k$ .
- Explorar as relações entre múltiplas variáveis ( $X_1, X_2, \dots, X_k$ ) para determinar que variáveis influenciam  $Y$ .

# MODELO MATEMÁTICO

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + \varepsilon$$

■ Onde:

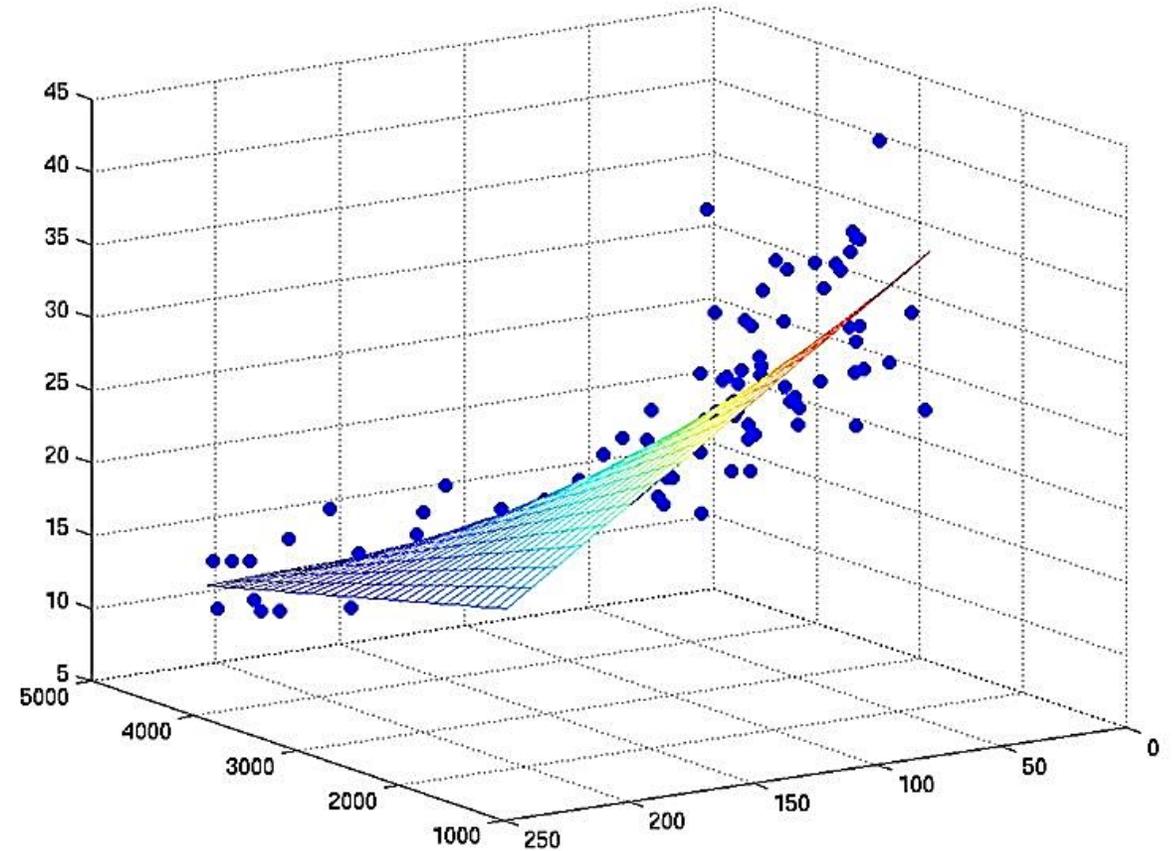
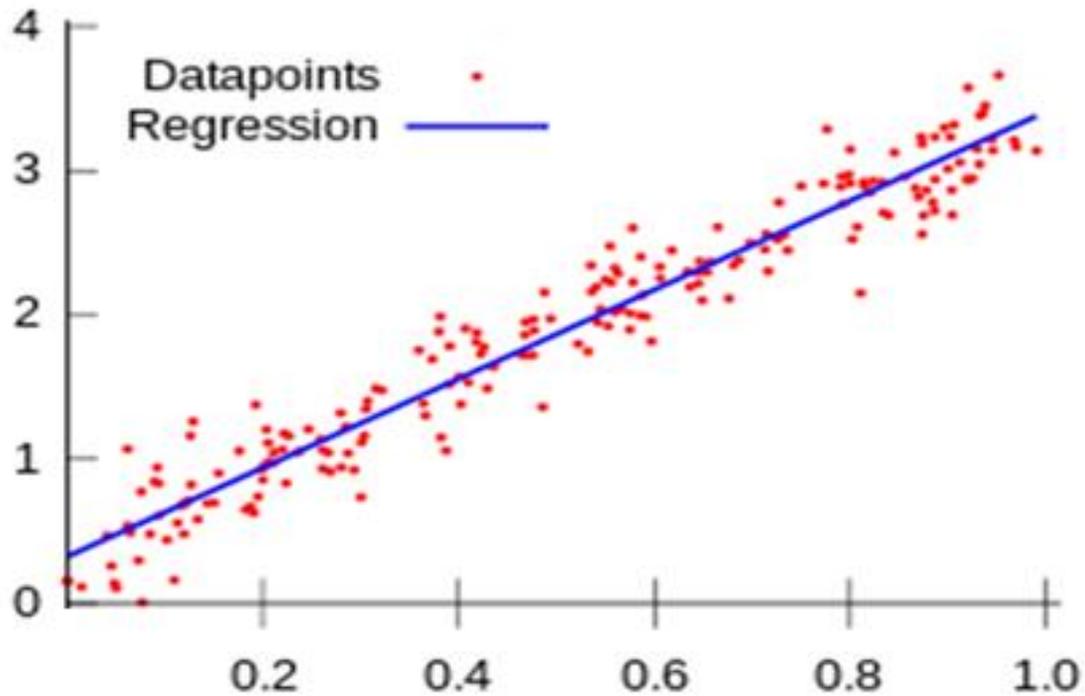
$y$  = é a variável dependente;

$\beta_1, \beta_2, \dots, \beta_k$  = são parâmetros a serem estimados;

$X_1, X_2, \dots, X_k$  = são as variáveis independentes;

$\varepsilon$  = é o erro aleatório referente a variabilidade em  $y$  quem não pode ser explicada pelas variáveis  $x$ 's.

- Enquanto uma regressão simples de duas variáveis resulta na equação de uma reta, um problema de três variáveis implica numa superfície de resposta, e um problema de  $k$  variáveis implica em um volume de resposta;



# SOLUÇÃO DOS MÍNIMOS QUADRADOS

- 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  $Y_i$  e os valores ajustados  $\hat{Y}_i$ .
- 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.

$$\sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^n (Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \dots + \hat{\beta}_k X_{ki}))^2$$

# 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:  $R^2 = \text{prox. de } 1$ .
- Ajuste pobre:  $R^2 = \text{prox. } 0$ .

# COEFICIENTE DE DETERMINAÇÃO

- Defeito: Na medida em que mais variáveis são incluídas,  $R^2$  cresce (pela simples inclusão das variáveis  $X$  independentes);
- 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.

$$R_{ajust}^2 = 1 - \left( \frac{n - 1}{n - p - 1} \cdot (1 - R^2) \right)$$

# 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-se 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, os parâmetros entram na equação de forma não linear, assim, nós não podemos simplesmente aplicar fórmulas para estimar os parâmetros do modelo.

# REGRESSÃO NÃO LINEAR

## ➤ MODELO LINEAR

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

Modelo Linear  $\eta(x, \theta_0, \theta_1, \theta_2) = \theta_0 + \theta_1 x + \theta_2 x^2$

Derivada  $\frac{\partial \eta}{\partial \theta_0} = 1, \quad \frac{\partial \eta}{\partial \theta_1} = x, \quad \frac{\partial \eta}{\partial \theta_2} = x^2,$

Não envolve os parâmetros

## ➤ MODELO NÃO-LINEAR

$$y = ae^{bx} U$$

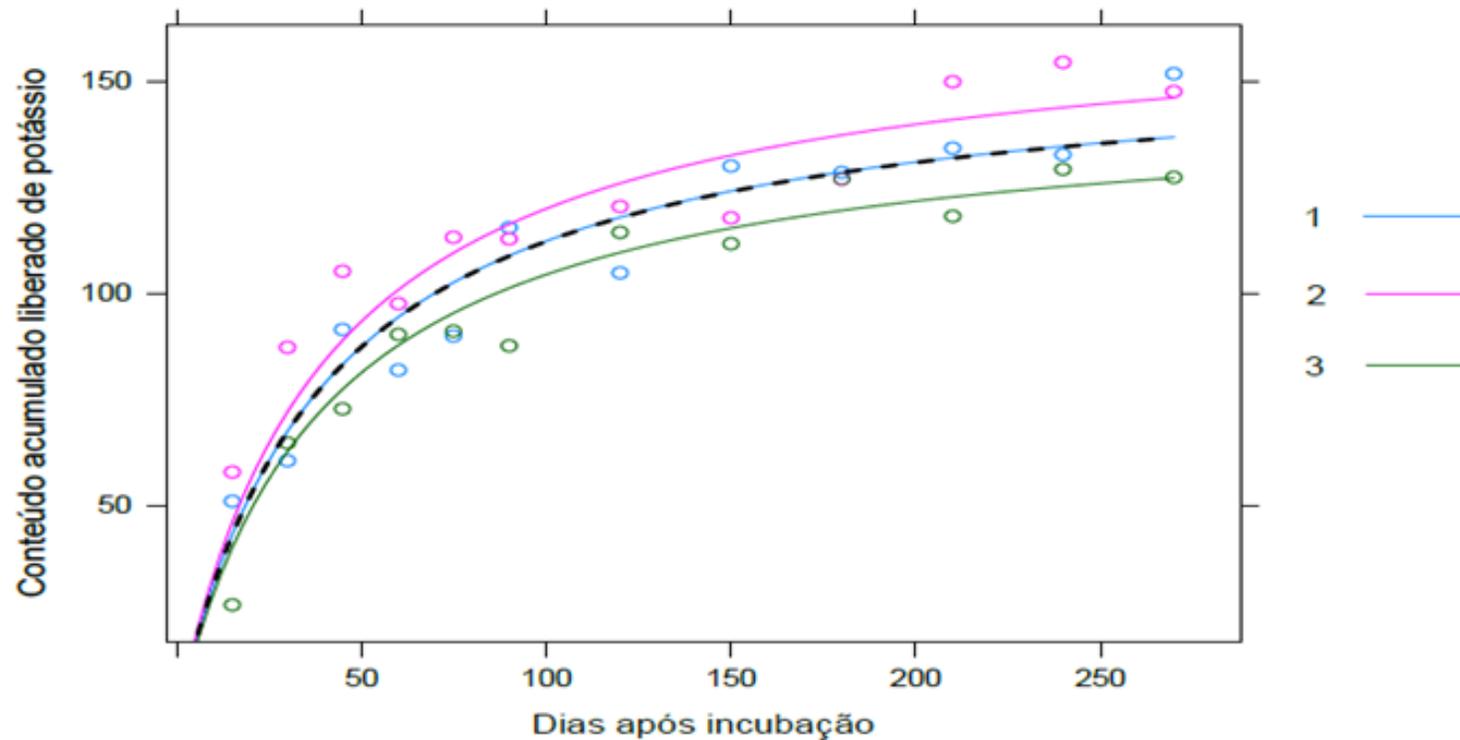
Modelo não linear

$$\eta(x, \theta_a, \theta_e, \theta_c) = \theta_a(1 - \exp\{-\theta_e(x - \theta_c)\})$$

Derivadas  $\frac{\partial \eta}{\partial \theta_a} = 1 - \exp\{-\theta_e(x - \theta_c)\}$   
 $\frac{\partial \eta}{\partial \theta_e} = -\theta_a(\theta_c - x) \exp\{-\theta_e(x - \theta_c)\}$   
 $\frac{\partial \eta}{\partial \theta_c} = -\theta_a \theta_e \exp\{-\theta_e(x - \theta_c)\},$

Derivada envolve os parâmetros

# EXEMPLO



- Valores observados e preditos para o conteúdo acumulado liberado de potássio em função dos dias após a incubação do esterco de codorna em LVdf com destaque para a predição a nível populacional e de unidade experimental.

# REGRESSÃO NÃO LINEAR

Métodos de Estimação:

- Método dos Mínimos Quadrados
- Método dos Mínimos Quadrados Generalizados
- Método da Máxima Verossimilhança
- Métodos Iterativos de Gauss-Newton

# MODELOS DE REGRESSÃO NÃO-LINEAR

Modelo	Função	Modelo	Função
A Schnute	$y_i = \frac{\beta_1}{\left(1 + \beta_4 e^{(\beta_3 \beta_2 - x_i)}\right)^{\frac{1}{\beta_4}}} + e_i$	H Meloun II	$y_i = \beta_1 - e^{(-\beta_2 - \beta_3 x_i)} + e_i$
B Mitscherlich	$y_i = \beta_1 \left(1 - e^{(\beta_3 \beta_2 - \beta_3 x_i)}\right) + e_i$	N Brody	$y_i = \beta_1 \left(1 - \beta_2 e^{-\beta_3 x_i}\right) + e_i$
C Richards	$y_i = \frac{\beta_1}{\left(1 + e^{(\beta_2 - \beta_3 x_i)}\right)^{\frac{1}{\beta_4}}} + e_i$	O von Bertalanffy	$y_i = \beta_1 \left(1 - \beta_2 e^{-\beta_3 x_i}\right)^3 + e_i$
D Gompertz	$y_i = \beta_1 e^{(-e^{(\beta_2 - \beta_3 x_i)})} + e_i$	P Michaelis-Menten	$y_i = \frac{\beta_1 x_i}{x_i + \beta_2} + e_i$
E Logístico	$y_i = \frac{\beta_1}{\left(1 + e^{(\beta_2 - \beta_3 x_i)}\right)} + e_i$	Q Michaelis-Menten Modificado	$y_i = \frac{\beta_2 \beta_3^{\beta_4} + \beta_1 x_i^{\beta_4}}{\beta_3^{\beta_4} + x_i^{\beta_4}} + e_i$
F Meloun I	$y_i = \beta_1 - \beta_2 e^{(-\beta_3 x_i)} + e_i$		

# VANTAGENS

- Sua escolha têm sustentação baseada em teoria ou princípios mecanísticos (físicos, químicos ou biológicos) ou qualquer outra informação prévia.
- Certos parâmetros são quantidade de interesse para o pesquisador.
- Podem ser feitas predições fora do domínio observado de  $x$ .
- São parcimoniosos pois tipicamente possuem menos parâmetros.
- Partem do conhecimento do pesquisador sobre o fenômeno alvo.

# DESVANTAGENS

- Requerem procedimentos iterativos de estimação baseados no fornecimento de valores iniciais para os parâmetros.
- Métodos de inferência são aproximados.
- Exigem conhecimento do pesquisador sobre o fenômeno alvo.

# REFERÊNCIAS BIBLIOGRÁFICAS

- FARIA, J.C. **Notas de aulas expandidas**. Ilhéus, UESC, 2006.
- Wonnacott, Thomas H. **Estatística aplicada a economia e a administração** / Thomas H Wonnacott e Ronald J. Wonnacott, 1981.
- ALLAMAN, Ivan Bezerra. **Regressão linear múltipla**. UESC.  
<[http://nbcgib.uesc.br/lec/download/material\\_didatico/pdf\\_files/est\\_infer/reg\\_linear\\_multipla.pdf](http://nbcgib.uesc.br/lec/download/material_didatico/pdf_files/est_infer/reg_linear_multipla.pdf)>. Acesso dia 23 de novembro de 2020.
- THOMAS, Gustavo. **Regressão Não Linear**. Piracicaba-SP, ESALQ, 2019.  
<[https://edisciplinas.usp.br/pluginfile.php/2340838/mod\\_resource/content/0/Gustavo\\_Relatorio.pdf](https://edisciplinas.usp.br/pluginfile.php/2340838/mod_resource/content/0/Gustavo_Relatorio.pdf)>. Acesso dia 23 de novembro de 2020.
- ZEVIANI, W.M.; JÚNIOR, P.J.R.; BONAT, W.H. **Curso: Modelos de regressão não linear**. 58° RBRAS E 15° SEAGRO, Universidade Federal do Paraná, 2013.
- SILVEIRA, Fernanda Gomes da *et al.* **Análise de agrupamento na seleção de modelos de regressão não-lineares para curvas de crescimento de ovinos cruzados**. *Ciência Rural*, v. 41, n. 4, p. 692-698, abr. 2011. Disponível em: <https://doi.org/10.1590/s0103-84782011000400024>. Acesso em: 29 jun. 2023.

---

MUITO

OBRIIGADO!