14. Introdução ao estudo de regressão linear simples

14.1. Introdução

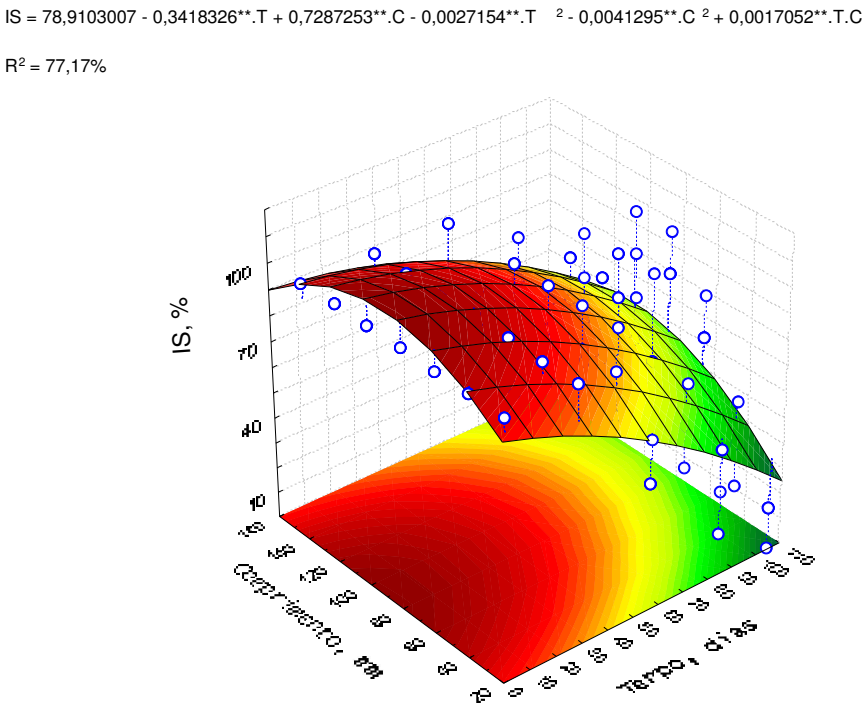


Figura 14.1 – Exemplo ilustrativo de regressão linear múltipla. O índice de sobrevivência (IS) do clone TSH 565 em função do comprimento remanescente foliar e do tempo, após preparo para propagação massal.

Nos experimentos em que os tratamentos são níveis crescentes de pelo menos um fator quantitativo, como por exemplo: adubo, herbicida, irrigação; é estritamente incorreto a utilização dos testes de comparação de médias múltiplas (TCMM), ou análise de contrastes (AC), para estudar seus efeitos sobre as variáveis aleatórias mensuradas.

Essas técnicas, TCMM e AC, são utilizadas na análise qualitativa de experimentos.

Quando os tratamentos são níveis crescentes de pelo menos um fator quantitativo, os ensaios devem ser analisados por intermédio da análise quantitativa de experimentos, isto é, regressão, e ou, correlação.

Embora as técnicas e princípios sejam comuns a ambos os métodos (regressão e correlação), existem diferenças conceituais que devem ser consideradas.

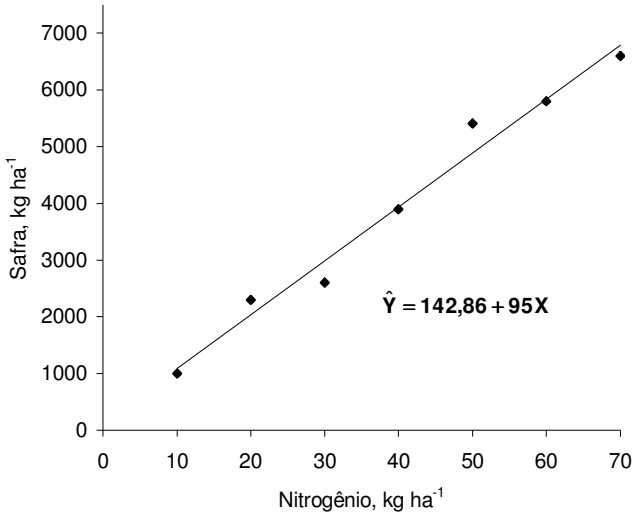
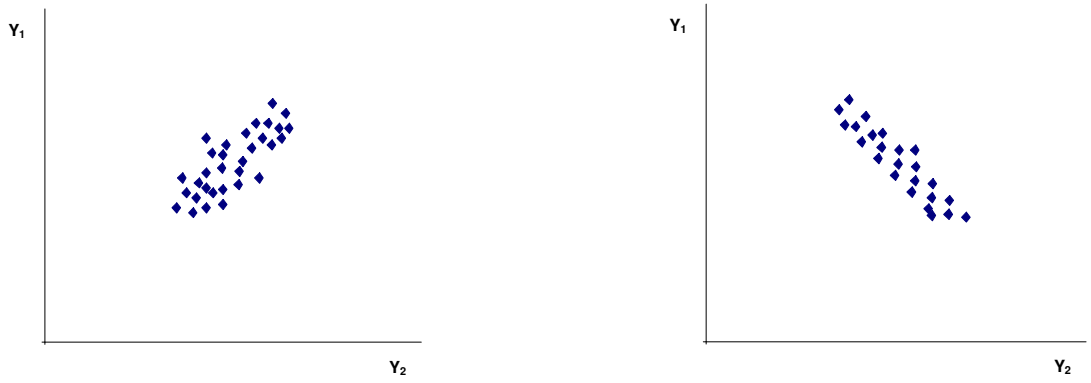


Figura 14.2 – Exemplo ilustrativo de regressão linear simples. A safra do milho em função de doses crescentes de adubo nitrogenado aplicado em cobertura.

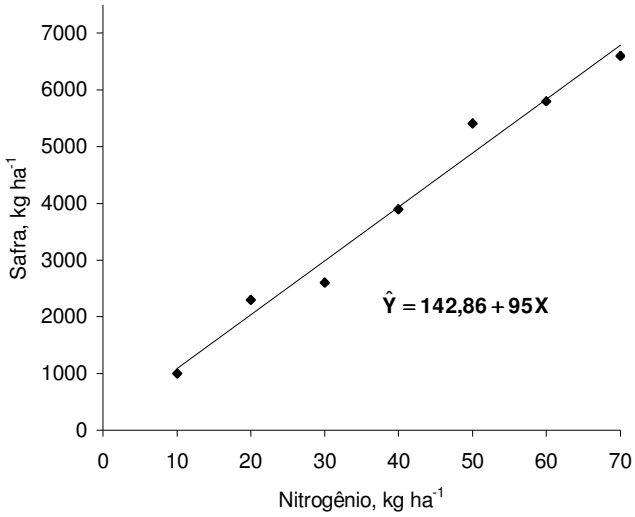
A análise de correlação é indicada para estudar o grau de associação linear entre variáveis aleatórias. Ou seja, essa técnica é empregada, especificamente, para se avaliar o grau de covariação entre duas variáveis aleatórias: se uma variável aleatória Y_1 aumenta, o que acontece com uma outra variável aleatória Y_2: aumenta, diminui ou não altera?



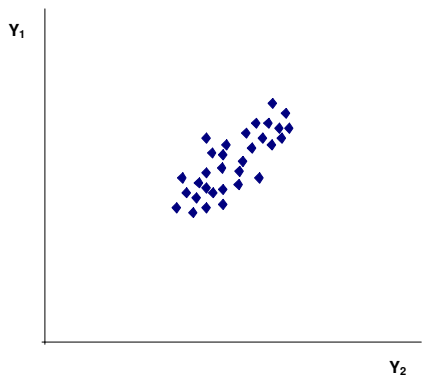
Na análise de regressão uma resposta unilateral é esperada: alterações em X (fator quantitativo) podem implicar em alterações em Y, mas alterações em Y não resultam em alterações em X.

Enquanto a análise de regressão linear nos mostra como as variáveis se relacionam linearmente, a análise de correlação vai nos mostrar apenas o grau desse mesmo relacionamento.

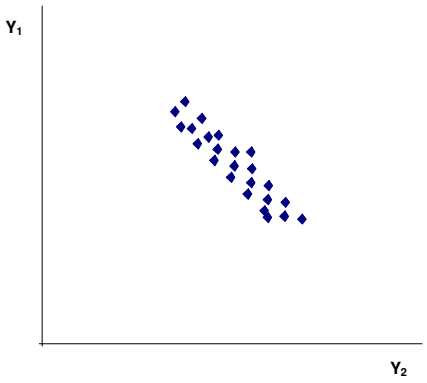
Na análise de regressão estimamos toda uma função Y = f(X), a equação de regressão:



A análise de correlação, por sua vez, nos fornece apenas um número, um índice, que quantifica o grau da associação linear entre duas variáveis aleatórias:



r = 0,6

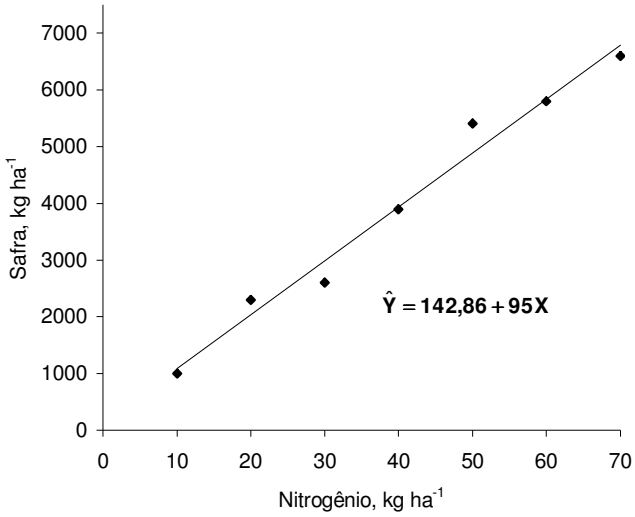


r = - 0,8

Quando se deseja verificar a existência de alguma relação estatística entre uma ou mais variáveis fixas, independentes, sobre uma variável aleatória, denominada dependente, utiliza-se a análise de regressão (embora essa análise possa, também, ser utilizada para estabelecer a relação funcional entre duas ou mais variáveis aleatórias).

Para exemplificar, vamos considerar que conduzimos um experimento submetendo plantas de milho a doses crescentes de nitrogênio.

Naturalmente, a produção será dependente da quantidade aplicada desse fertilizante, X:



Assim, o fertilizante nitrogenado aplicado é a variável independente, e cada uma das quantidades aplicadas são seus níveis, x_i (10 ... 70 kg ha^{-1}).

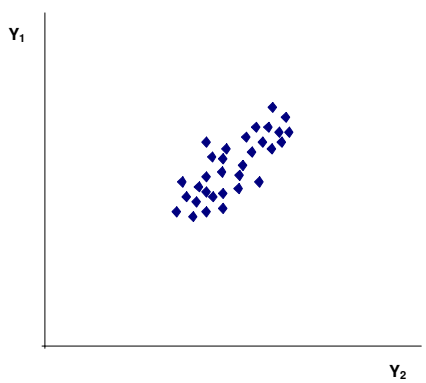
Cada variável aleatória mensurada na cultura do milho, sujeita a influência dos níveis x_i da variável independente, ou seja, das doses de nitrogênio, é chamada "variável dependente" ou "fator resposta".

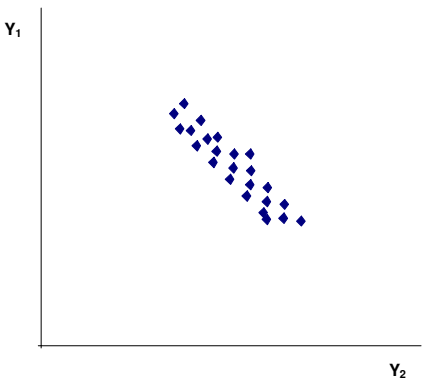
Poderia-se medir, por exemplo, o número de espigas por planta (Y_1), a altura média das plantas (Y_2), o peso de 1.000 grãos (Y_3), o teor de proteínas dos grãos (Y_4), o teor de gordura dos grãos (Y_5), etc.

Como a aplicação do fertilizante não depende da safra, sendo, ao contrário, determinada independentemente pelo pesquisador, designamo-la "variável independente" ou "regressor".

Podemos estudar via análise de regressão o efeito da variável, neste caso, fixa, independente, X (dose de nitrogênio), sobre as variáveis aleatórias, ou dependentes, Y_i (produção de matéria seca, teor de proteínas dos grãos, teor de gordura dos grãos, etc.). Diz-se regressão de Y sobre X.

Posteriormente, caso seja de interesse, podemos utilizar a análise de correlação para estudar o grau de associação linear, por exemplo, entre o teor de proteínas e o teor de gordura dos grãos, sendo ambas variáveis aleatórias:





Ou seja, poderemos estudar via correlação linear simples o grau de associação entre um par qualquer (Y_i,Y_i). Por exemplo, se o teor de proteínas aumenta, o que acontece com o teor de gordura (aumenta, diminui ou não altera). Estaremos, então, interessados em averiguar a covariação entre estas duas variáveis aleatórias.

Nada impede, entretanto, que o estudo entre o teor de proteínas e teor de gordura seja feito, por meio da análise de regressão. Nesses casos, seria indiferente a posição ocupada por cada uma das variáveis aleatórias, ou seja, a posição Y_i (dependente) ou X_i (independente).

O incorreto seria estudar via análise de correlação o efeito do nitrogênio (variável fixa) sobre a produção de matéria seca dos grãos de milho (variável aleatória), ou sobre os teores de proteína, gordura, etc.

Em síntese, o método da análise de regressão pode ser utilizado sempre que existir uma relação funcional entre uma variável chamada dependente e uma outra chamada independente (regressão linear simples) ou entre uma variável dependente e duas ou mais variáveis independentes (regressão linear múltipla).

Ajustamento

Se precisarmos considerar como a safra depende de diferentes quantidades de nitrogênio, deveremos definir a aplicação do nitrogênio segundo uma escala numérica.

Se grafarmos a safra, Y, decorrente das diversas aplicações, X, de nitrogênio, poderemos observar uma dispersão análoga a Figura 14.3:

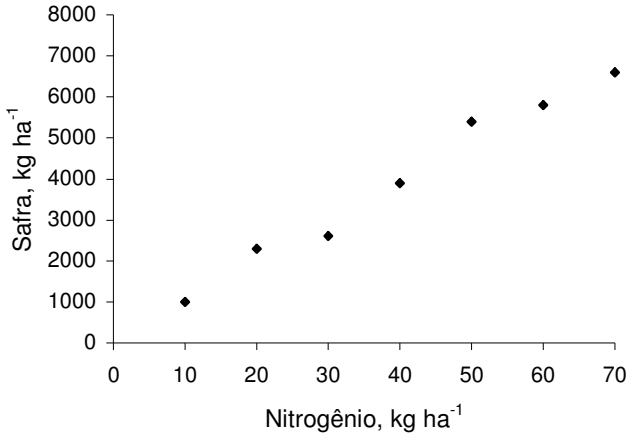


Figura 14.3 - Relação observada entre a safra e a aplicação de nitrogênio.

A aplicação de nitrogênio afeta a safra.

Podemos, por meio de uma equação, relacionando X e Y, descrever como afeta.

Estimar uma equação é geometricamente equivalente a ajustar uma curva àqueles dados dispersos, isto é, a "regressão de Y sobre X".

Esta equação será útil como descrição breve e precisa de predizer a safra Y para qualquer quantidade X de nitrogênio.

Como safra depende do nitrogênio, a safra é chamada "variável dependente" ou "fator resposta", Y.

A aplicação do nitrogênio não depende da safra, sendo, ao contrário, determinada independentemente pelo pesquisador, é chamada a "variável independente" ou "regressor", X.

Vamos considerar um estudo sobre a influência do N (nitrogênio) aplicado em cobertura sobre a safra do milho.

Suponhamos que só dispomos de recursos para fazer sete observações experimentais.

O pesquisador fixa então sete valores de X (sete níveis do regressor), fazendo apenas uma observação Y (fator resposta), em cada caso, tal como se vê na Figura 14.4:

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

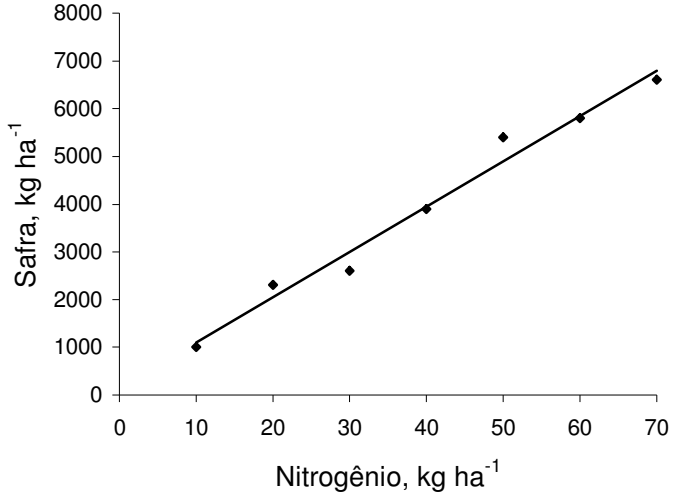
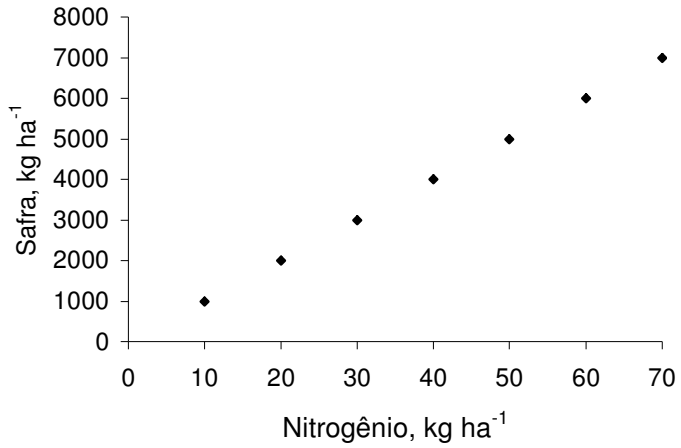


Figura 14.4 - Dados e reta ajustada a olho aos dados apresentados.

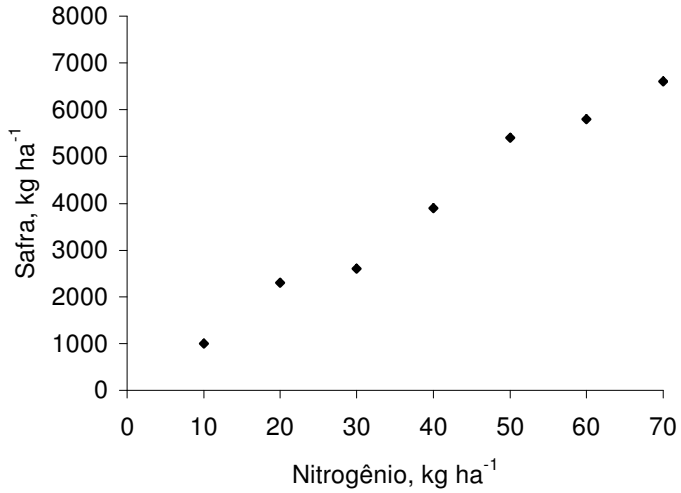
Até onde é bom um ajustamento feito a olho, tal como o da Figura 14.4?

Verificar a ilustração de vários graus de dispersão (Figura 14.5).

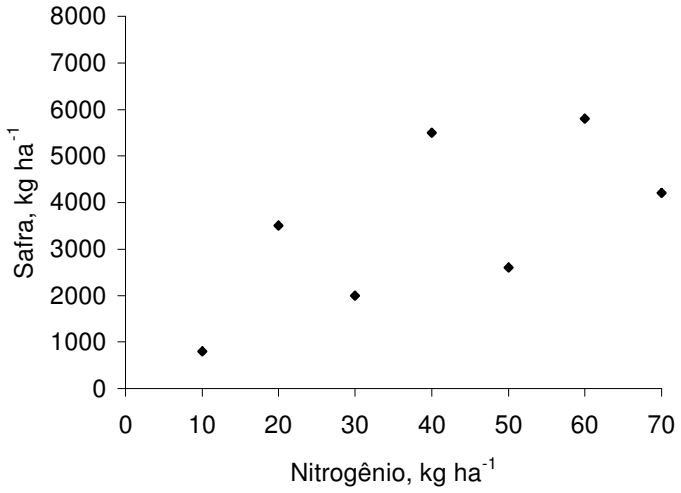
Necessitamos então de um método objetivo, que possa ser estendido ao maior número de situações, onde o ajustamento a olho esteja fora de questão.



a.



b.



c.

Figura 14.5 - Ilustração de diversos graus de dispersão.

14.1.1. Critérios para se ajustar uma reta

Precisamente, o que é um bom ajustamento?

A resposta óbvia seria: um ajustamento que acusa pequeno erro total.

A Figura 14.6 ilustra um erro típico (desvio).

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, $\left(Y_i - \hat{Y}_i\right)$:

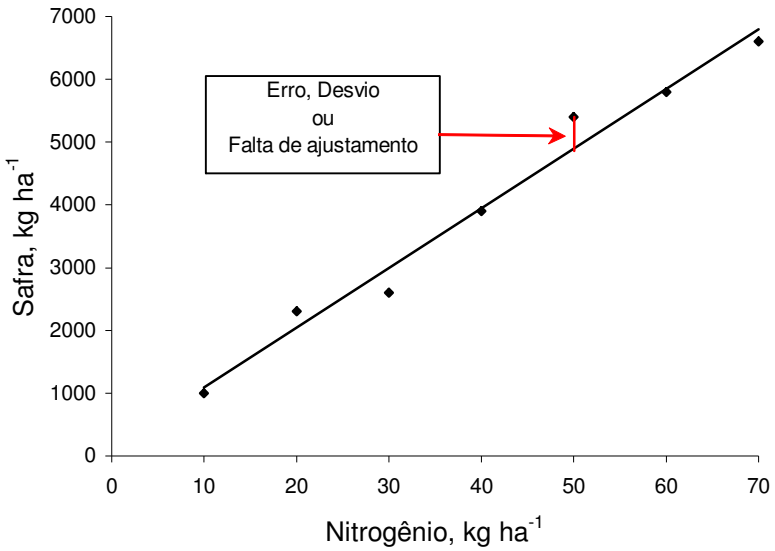


Figura 14.6 - Erro típico no ajustamento de uma reta.

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

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

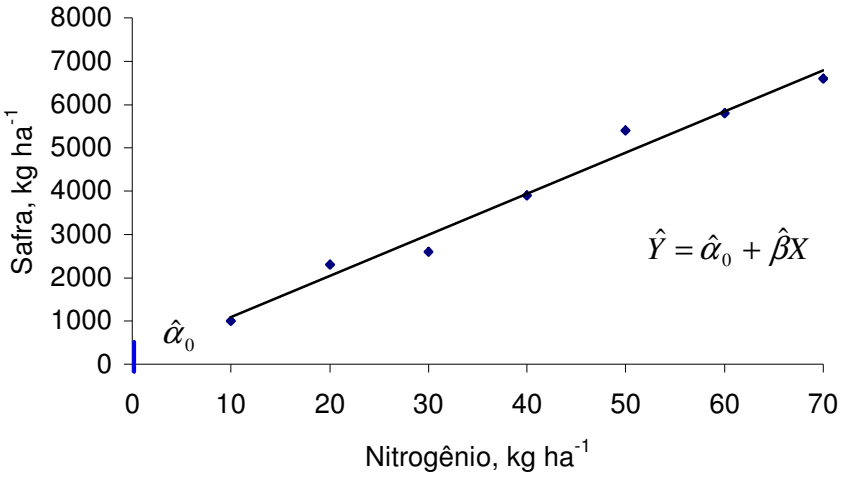
conhecido como critério dos "mínimos quadrados" ou "mínimos quadrados dos erros". Sua justificativa inclui as seguintes observações:

O quadrado elimina o problema do sinal, pois torna positivos todos os erros.

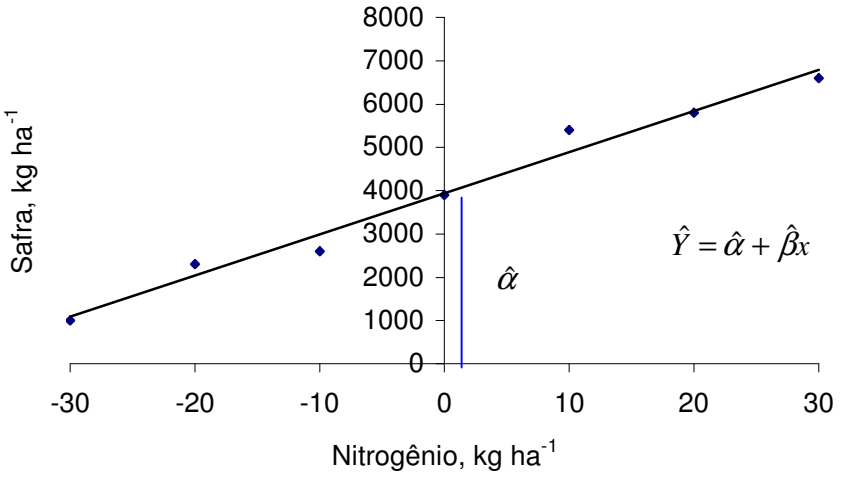
A álgebra dos mínimos quadrados é de manejo relativamente fácil.

14.1.2. Ajustando uma reta

O conjunto de valores X e Y observados na Figura 14.4 é grafado novamente na Figura 14.7(a):



a.



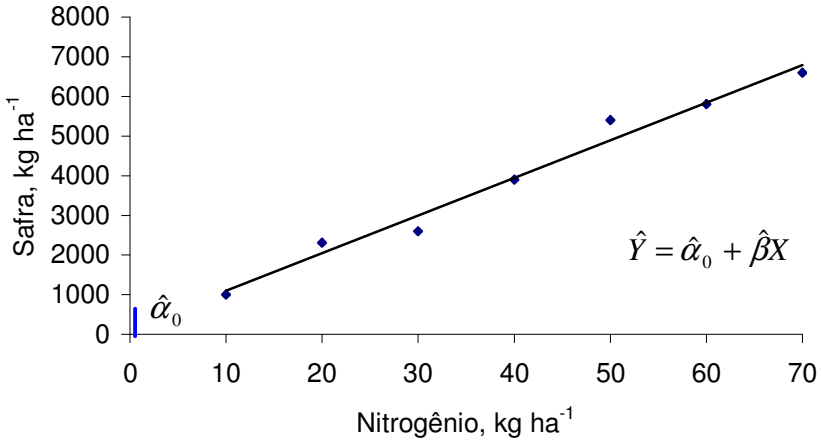
b.

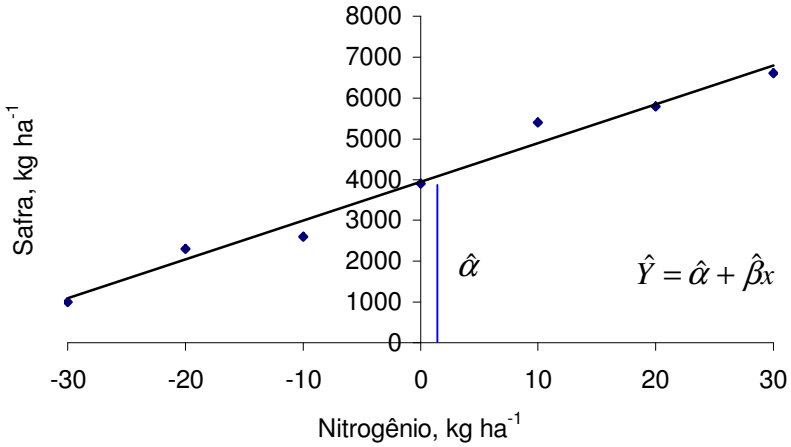
Figura 14.7 - Translação de eixos. (a) Regressão utilizando os valores originais. (b) Regressão após transladar Y.

Estágio 1: 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 translação 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 $\overline{X}$.

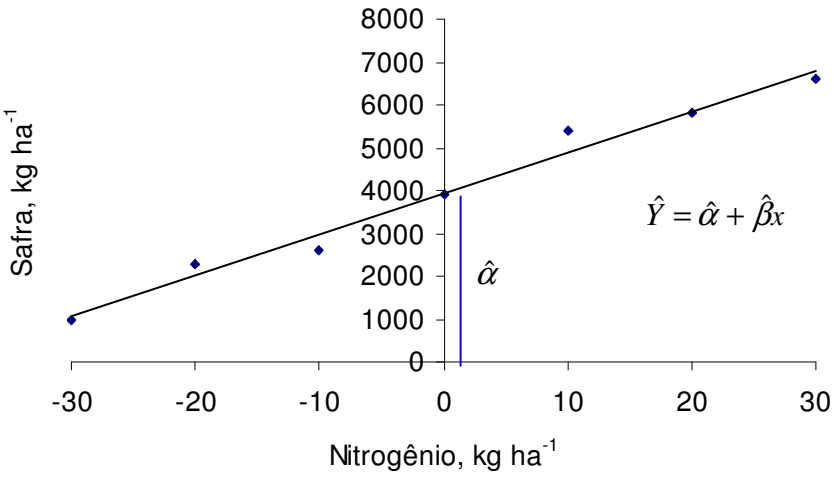
Não há modificação nos valores de Y.

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

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

Estágio 2: Ajustar a reta da Figura 14.7(b), isto é, a reta: $\hat{Y} = \hat{\alpha} + \hat{\beta}x$



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 $\hat{Y}_i$ 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 as Equações 01 e 02, isto pode ser expresso algebricamente como:

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

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

Utilizou-se $S(\hat{\alpha},\hat{\beta})$ para enfatizar que esta expressão depende de $\hat{\alpha}$ e $\hat{\beta}$. Ao variarem $\hat{\alpha}$ e $\hat{\beta}$ (quando se tentam 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:

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

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

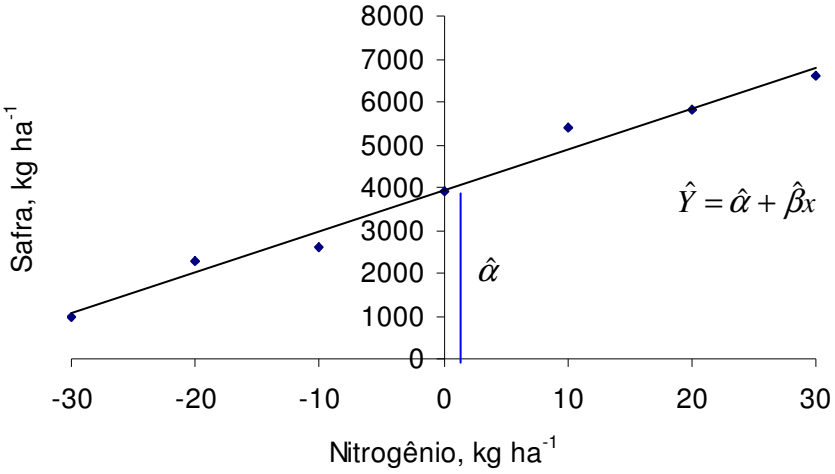
$$\sum Y_i-n\hat{\alpha}=0$$

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

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

Assim, a estimativa de mínimos quadrados para $\hat{\alpha}$ é simplesmente o valor médio de Y.

Verifica-se que isto assegura que a reta de regressão ajustada deve passar pelo ponto $\left(x,\bar{Y}\right)$, 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:

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

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

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

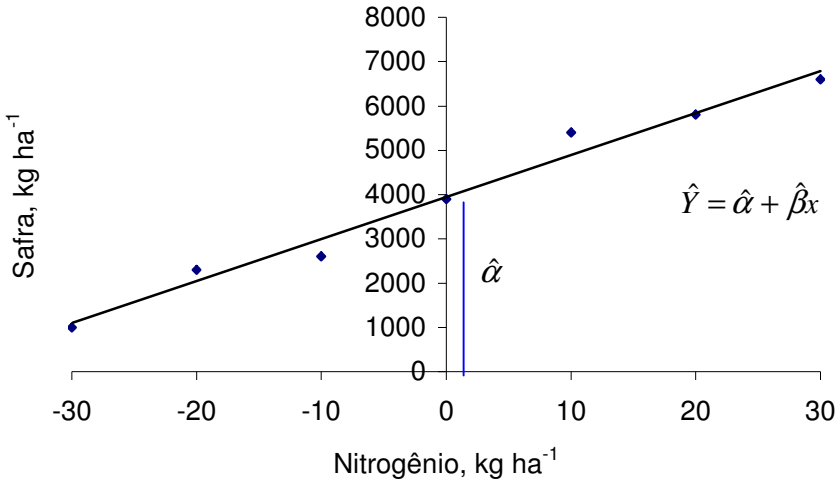
$$\hat{\beta} \sum x_i^2 = \sum x_i Y_i$$

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

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:

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



Para os dados da Figura 14.4, $\hat{\alpha}$ e $\hat{\beta}$ acham-se calculados no Quadro 14.1.

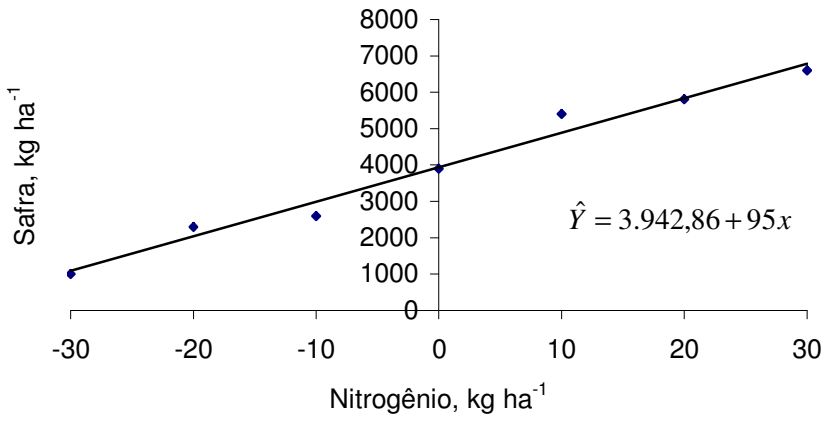
Quadro 14.1 - Cálculos dos valores necessários

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

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

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

$$\hat{Y} = 3.942,86 + 95x \qquad \text{Equação 03}$$



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

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

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

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

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

$$\hat{Y} = 142{,}86 + 95X \qquad \text{Equação 04}$$

$$\hat{Y} = 3.942{,}86 + 95x \qquad \text{Equação 03}$$

Comparando as Equações 03 e 04, observa-se que:

- 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.

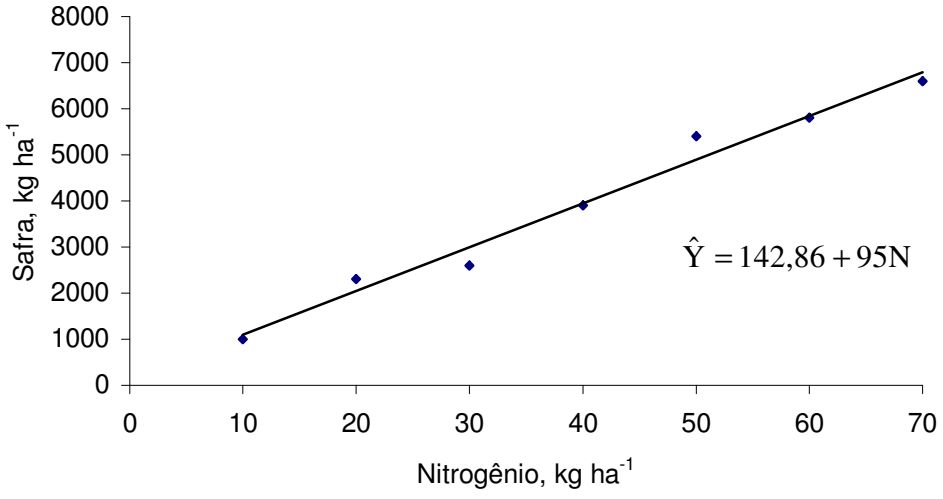


Figura 14.8 - Gráfico dos pontos dispersos com a reta ajustada.

Esta equação é útil como descrição breve e precisa de predizer a safra, em kg ha^{-1}, para qualquer quantidade de nitrogênio, também em kg ha^{-1}, aplicada.

Observar que:

- Se nenhum nitrogênio for aplicado à cultura, a safra estimada será de 142,86 kg.
- Esta safra se deve a absorção pela cultura do N disponível no solo, possivelmente associado ao ciclo orgânico.
- No intervalo das doses aplicadas (10 a 70 kg), considerando-se um hectare, para cada kg de nitrogênio aplicado, a cultura responde com 95 kg de grãos.

14.2. 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: $\|\beta_i\| = 0$	ou	H_o: $Y \neq \alpha_0 + \beta X$
H_1: $\|\beta_i\| > 0$	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.

ANOVAR

Causa da variação	GL
Regressão	1
Erro	5
Total	6

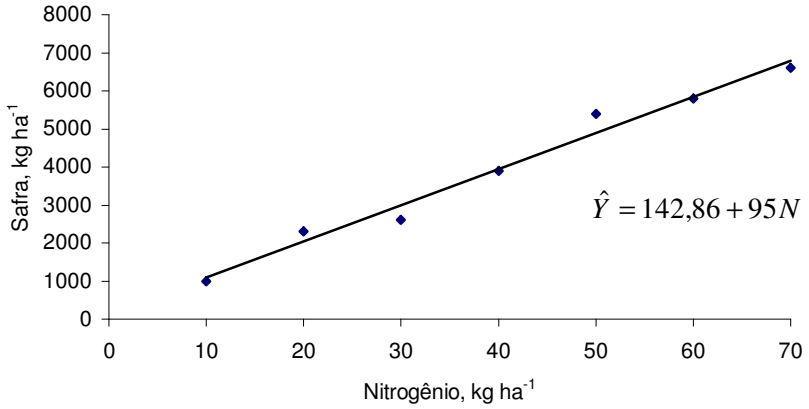
Existem várias formas de realizar estes cálculos.

Objetivando clareza de idéias e conceitos, a forma que será empregada utilizará o conceito mais elementar da estatística, ou seja, a variância:

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

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



[1] Obs = Observado: valores observados de Y

Est = Estimado: valores estimados para Y a partir da equação de regressão.

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			

Conclusão: rejeita-se H_0 ao nível de 5% de probabilidade pelo teste F.

Ou seja, a equação de regressão ajustada explica a variação da safra, em decorrência da variação das doses de nitrogênio, ao nível de 5% de probabilidade pelo teste F.

14.2.1. Cálculos alternativos da soma de quadrados dos desvios

É possível demonstrar algebricamente que:

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

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

$$\text{SQDerr} = \text{SQDtot} - \text{SQDreg}$$

Esta forma de realizar os cálculos da soma de quadrados dos desvios, embora menos compreensível a primeira vista, é a mais prática e deve ser a preferencialmente utilizada.

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

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

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

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

$$SQDreg = 25.270.000$$

$$SQDerr = SQDtot - SQDreg$$

$$SQDerr = 25.797.142,86 - 25.270.000$$

$$SQDerr = 527.142,86$$

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			

14.2.2. 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:

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

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.

14.2.3. Relação entre o coeficiente de determinação e o coeficiente de correlação

Se análise de regressão linear simples for realizada entre duas variáveis aleatórias, a relação existente entre o o coeficiente de determinação da regressão, r^2, e o coeficiente de correlação, r, é a seguinte:

$$r = \sqrt{r^2}$$

Nos casos da regressão ter sido realizada entre uma variável aleatória e uma variável fixa, esta relação não possui significado estatístico.

14.2.4. Observações a respeito da regressão

Quando os dados não provêm de um delineamento experimental, como no exemplo analisado, a ANOVAR pode ser realizada da forma apresentada, e se terá chegado ao fim da análise.

Entretanto, 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.