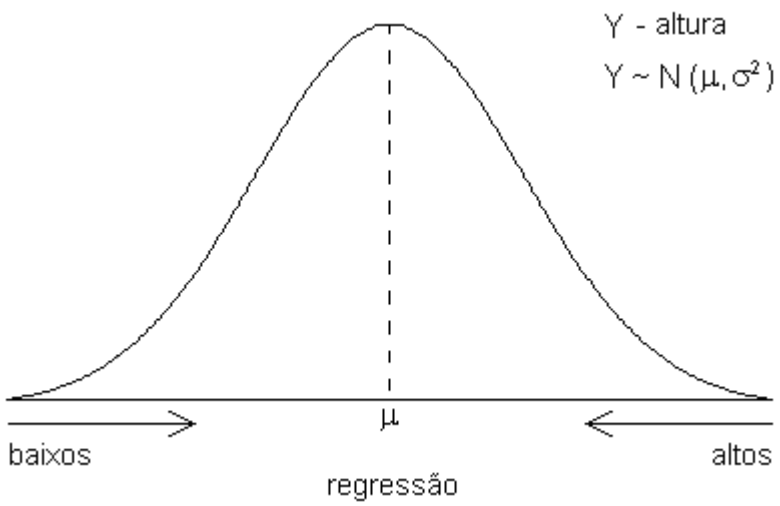
4.7. Regressão linear simples

4.7.1. Introdução

Em muitos estudos estatísticos, o objetivo do pesquisador é estabelecer relações que possibilitem predizer uma ou mais variáveis em termos de outras. Assim é que se fazem estudos para predizer as vendas futuras de um produto em função do seu preço, a perda de peso de uma pessoa em decorrência do número de dias que se submete a uma determinada dieta, a despesa de uma família com médico e remédios em função da renda, o consumo per capita de certos alimentos em função do seu valor nutritivo e do gasto com propaganda na TV, a produção de uma determinada cultura em função da quantidade de nutriente aplicada no solo, etc.

Naturalmente, o ideal seria que pudéssemos predizer uma quantidade exatamente em termos de outra, mas isso raramente é possível. Na maioria dos casos devemos contentar-nos com a predição de médias, ou valores esperados. Por exemplo, não podemos predizer exatamente quanto ganhará um bacharel nos 10 anos após a formatura, mas com base em dados adequados, é possível predizer o ganho médio de todos os bacharéis nos 10 anos após a formatura. Analogamente, podemos predizer a safra média de certa variedade de trigo em termos do índice pluviométrico de julho, e a nota média de um estudante em função do seu QI. Sendo assim, podemos dizer que a predição do valor médio de uma variável em função dos valores de outra constitui o problema principal da regressão.

A origem desse termo remonta a Francis Galton (1822 - 1911), que o empregou pela primeira vez em um estudo da relação entre as alturas de pais e filhos. Galton observou, nesse estudo, que filhos de pais muito altos, em média, não eram tão altos quanto os seus pais, da mesma forma que filhos de pais muito baixos, em média, não eram tão baixos quanto os seus pais. A partir dessas observações, concluiu que a altura dos filhos "tendia" para a média (μ) da espécie, ou seja, a cada geração a altura dos filhos convergia ou "regredia" para a média. Esse fenômeno de retorno à média foi, então, denominado *regressão*.



Por questões históricas o termo é utilizado até hoje, mas abriga uma série de técnicas estatísticas.

A expressão *regressão linear simples* é utilizada por duas razões: a regressão é *linear* porque a relação entre X e Y é expressa por uma equação de primeiro grau, representada graficamente por uma reta, e é *simples* porque envolve apenas duas variáveis.

- **Ajustamento de curvas**

Sempre que possível, procuramos expressar em termos de uma equação matemática as relações entre grandezas conhecidas e grandezas que devem ser determinadas. Isso ocorre

com frequência nas ciências naturais, onde, por exemplo, a relação entre o volume (y) e pressão (x) de um gás, a uma temperatura constante, é dada pela expressão

$$y = \frac{k}{x},$$

sendo k uma constante numérica. Outro exemplo pode ser a relação entre uma cultura de bactérias (y) e o tempo (x) em que esteve exposta a certas condições ambientais, que é dada por

$$y = ab^x,$$

onde a e b são constantes numéricas. Mais recentemente, equações como essas têm sido usadas para descrever relações também no campo das ciências do comportamento, das ciências sociais e outros.

Essa representação matemática dos fenômenos é feita "ajustando-se" uma curva aos dados observados, de tal forma que, a partir dessa "curva ajustada", possamos representar, gráfica ou analiticamente, a relação entre as variáveis. Então, ajustar uma curva é determinar uma função matemática que possa representar um conjunto de observações. Sempre que utilizamos dados observados para chegar a uma equação matemática encontramos três tipos de problema:

1. Decidir que tipo de curva e, consequentemente, que tipo de equação de predição é mais adequada.

2. Identificar a equação particular que é mais adequada para os dados.

3. Investigar possíveis problemas relativos ao mérito da equação escolhida e da predição feita a partir dela.

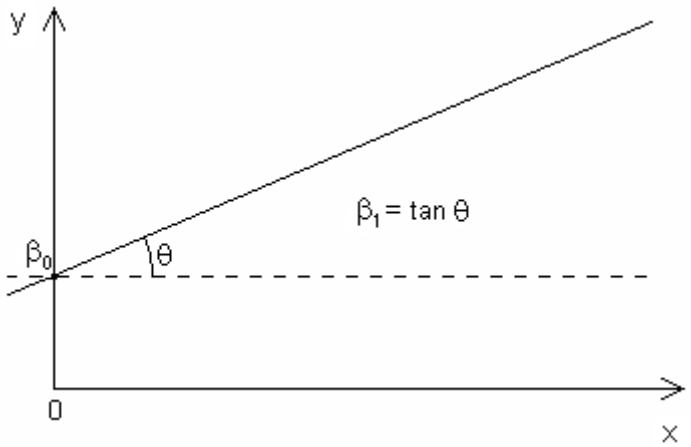
Aqui vamos restringir nosso estudo às equações lineares com duas incógnitas. Estas equações lineares são úteis e importantes não só porque muitas relações têm efetivamente esta forma, mas também porque em geral constituem boas aproximações de relações que, de outro modo, seriam difíceis de descrever em termos matemáticos.

- **Modelo estatístico**

Sendo x e y duas variáveis que se relacionam de forma linear, esta relação é expressa pela seguinte equação:

$$y = \beta_0 + \beta_1 x,$$

Na figura a seguir podemos observar a representação gráfica desta equação.



Se Y é uma variável aleatória, então, está sujeita a um erro de observação. Este erro (e_i) deverá ser adicionado ao modelo, desde que se admitam como verdadeiras as seguintes pressuposições:

1. Os erros são aleatórios, têm média zero e variância constante, ou seja, $E(e_i) = 0$ e $V(e_i) = \sigma^2$.

2. Os erros têm distribuição normal e são independentes entre si.

3. O modelo é adequado para todas as observações, não podendo haver nenhum valor de X que produza um valor de Y discrepante dos demais.

4. A variável X é fixa (não aleatória).

Assim, o modelo de regressão linear simples será:

$$y_i = \beta_0 + \beta_1 x_i + e_i$$

onde:

y_i : é a variável resposta (dependente)

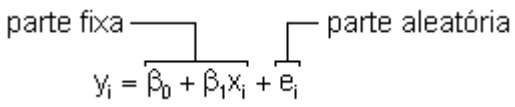
x_i : é a variável preditora (independente)

β_0 : é o intercepto ou coeficiente linear

β_1 : é o coeficiente angular ou de regressão

e_i : erro (variação aleatória não controlável)

Sendo assim, verificamos que este modelo é composto por uma parte fixa e uma parte aleatória:



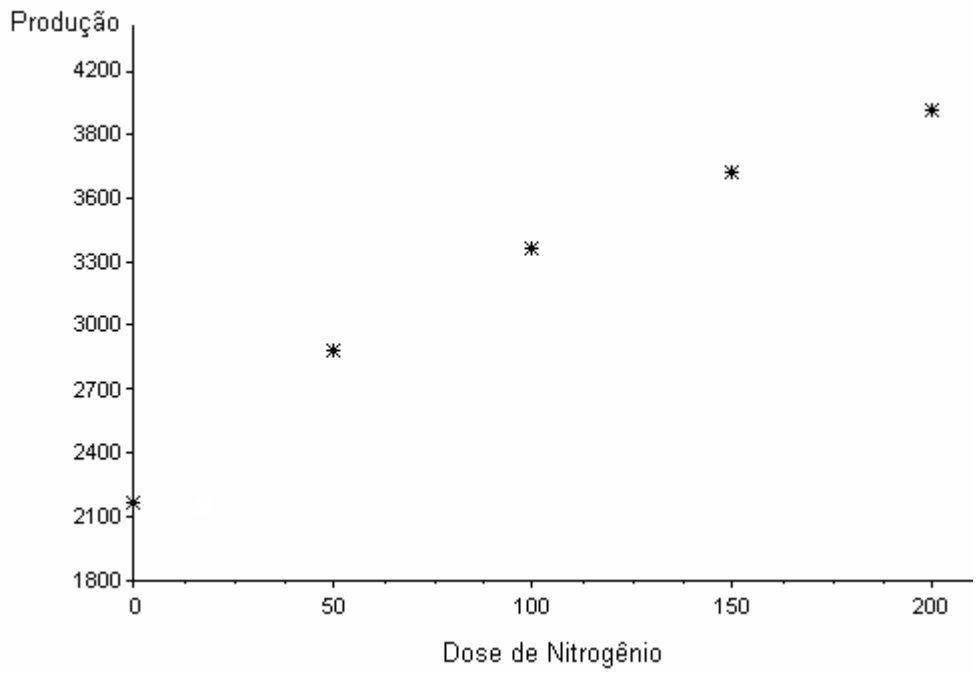
A parte fixa do modelo informa como X influencia Y e a parte aleatória mostra que Y possui uma variabilidade inerente, significando que X não é a única variável que influencia Y, embora consideremos que sua influência seja preponderante. Aliás, devemos ressaltar que este modelo será adequado quando a parte fixa for preponderante sobre a aleatória.

A título de ilustração, consideremos o exemplo a seguir.

Exemplo 1. Um experimento foi conduzido para estudar o efeito da dose de Nitrogênio aplicada no solo sobre a produção de uma espécie de forrageira. Para as cinco doses utilizadas, foram observados os seguintes resultados:

Parcela	Dose de Nitrogênio (kg/ha)	Produção de forragem (kg/ha)
1	0	2.160
2	50	2.880
3	100	3.360
4	150	3.720
5	200	4.020

De modo geral, um gráfico de dispersão de valores observados para a variável resposta já é suficiente para indicar o tipo de curva (reta, parábola, etc) que melhor descreve o padrão geral dos dados. A figura a seguir mostra a dispersão dos valores observados para a variável produção de forragem quando diferentes doses de Nitrogênio foram aplicadas. Podemos observar uma tendência linear dos dados, o que nos permite supor que a relação entre dose de Nitrogênio e produção de forragem seja linear.



Admitindo, então, o relacionamento linear entre as variáveis, vamos adotar o modelo de regressão linear simples:

$$y_i = \beta_0 + \beta_1 x_i + e_i,$$

onde:

y_i é a produção de forragem (variável resposta), em kg;

x_i é a dose de Nitrogênio (variável preditora), em kg;

β_0 é a produção de forragem quando a dose de Nitrogênio aplicada for nula (intercepto), em kg;

β_1 é a quantidade que varia na produção de forragem para cada unidade (kg) aplicada de Nitrogênio (coeficiente de regressão), em kg/kg.

e_i é o erro (variação aleatória não controlável)

4.7.2. Análise de regressão

A análise de regressão tem por objetivo determinar a equação que melhor representa a relação existente entre duas variáveis e, a partir desta equação, fazer predições para a variável resposta. Para isso, é necessário que uma sequência de passos seja seguida:

1. Obtenção das estimativas (por ponto) dos coeficientes β_0 e β_1 para ajustar a equação da regressão.

2. Aplicação dos testes de significância para as estimativas obtidas, a fim de verificar se a equação de regressão é adequada.

3. Cálculo dos intervalos de confiança para os valores estimados pela equação de regressão.

4.7.2.1. Estimação dos parâmetros do modelo

Quando temos n observações, temos n pares de valores, (x_1, y_1); (x_2, y_2); ... ; (x_n, y_n), onde os valores observados para a variável resposta (y_i) são representados pela equação da regressão:

$$y_i = \beta_0 + \beta_1 x_i + e_i \quad i = 1,2,\ldots,n \qquad \begin{cases} y_1 = \beta_0 + \beta_1 x_1 + e_1 \\ y_2 = \beta_0 + \beta_1 x_2 + e_2 \\ \quad\quad K \\ y_n = \beta_0 + \beta_1 x_n + e_n \end{cases}$$

Os coeficientes β_0 e β_1 são os parâmetros do modelo, e, portanto, constantes desconhecidas, que serão estimados a partir dos valores da amostra.

Se $y_i = \beta_0 + \beta_1 x_i + e_i$, então

$$E(y_i) = E(\beta_0 + \beta_1 x_i + e_i)$$
$$E(y_i) = E(\beta_0) + E(\beta_1 x_i) + E(e_i)$$
$$E(y_i) = \beta_0 + \beta_1 x_i$$

Sendo assim, se $y_i = \beta_0 + \beta_1 x_i + e_i$, então

$$y_i = E(y_i) + e_i,$$

logo,

$$e_i = y_i - E(y_i).$$

A estimação dos parâmetros β_0 e β_1 é efetuada através do método dos mínimos quadrados.

- **Método dos mínimos quadrados**

Este método tem como objetivo obter as estimativas dos parâmetros β_0 e β_1 de tal forma que a soma dos quadrados dos erros ($\sum e_i^2$) seja o menor valor possível.

Vimos que $e_i = y_i - E(y_i)$ e $E(y_i) = \beta_0 + \beta_1 x_i$,

logo,

$$\sum e_i^2 = \sum [y_i - E(y_i)]^2 = \sum [y_i - (\beta_0 + \beta_1 x_i)]^2.$$

Para encontrar os valores de β_0 e β_1 que tornam mínima a soma de quadrados dos erros, devemos, inicialmente, encontrar para a expressão acima as derivadas parciais em relação a β_0 e β_1.

$$\frac{\partial \sum e_i^2}{\partial \beta_0} = 2\sum (y_i - \beta_0 - \beta_1 x_i).(-1)$$
$$\frac{\partial \sum e_i^2}{\partial \beta_1} = 2\sum (y_i - \beta_0 - \beta_1 x_i).(-x_i)$$

Observamos que os valores de β_0 e β_1 das duas expressões acima variam de acordo com os valores que se atribui às derivadas parciais. Entretanto, para obter os pontos críticos (máximos ou mínimos), devemos igualar essas derivadas a zero, onde β_0 e β_1 assumem um valor particular, ou seja, representam as estimativas dos parâmetros de forma que a soma dos quadrados dos erros seja mínima. Deste modo, igualando a zero as derivadas parciais, temos

$-2\sum(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$, sendo $\sum(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$

e

$-2\sum(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)\, x_i$, sendo $\sum(y_i x_i - \hat{\beta}_0 x_i - \hat{\beta}_1 x_i^2) = 0$.

Podemos, então determinar os valores de $\hat{\beta}_0$ e $\hat{\beta}_1$, através de um sistema de equações normais.

$$\begin{cases} \sum(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \\ \sum(y_i x_i - \hat{\beta}_0 x_i - \hat{\beta}_1 x_i^2) = 0 \end{cases}$$

Aplicando as propriedades da soma, temos

$$\begin{cases} \sum y_i - n\hat{\beta}_0 - \hat{\beta}_1 \sum x_i = 0 \\ \sum y_i x_i - \hat{\beta}_0 \sum x_i - \hat{\beta}_1 \sum x_i^2) = 0 \end{cases}$$

e arrumando a expressão para que os termos fiquem positivos, temos

$$\begin{cases} n\hat{\beta}_0 + \hat{\beta}_1 \sum x_i = \sum y_i \\ \hat{\beta}_0 \sum x_i + \hat{\beta}_1 \sum x_i^2 = \sum y_i x_i \end{cases}$$

A resolução do sistema pode ser feita por substituição. Começamos por isolar o $\hat{\beta}_0$ na primeira equação:

$$\begin{aligned} & n\hat{\beta}_0 + \hat{\beta}_1 \sum x_i = \sum y_i \\ & \frac{n\hat{\beta}_0}{n} + \frac{\hat{\beta}_1 \sum x_i}{n} = \frac{\sum y_i}{n} \\ & \frac{n\hat{\beta}_0}{n} + \hat{\beta}_1 \frac{\sum x_i}{n} = \frac{\sum y_i}{n} \\ & \hat{\beta}_0 + \hat{\beta}_1 \bar{x} = \bar{y} \\ & \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \end{aligned}$$

Determinado o valor de $\hat{\beta}_0$, isolamos o $\hat{\beta}_1$ na segunda equação:

$$\begin{aligned} & \hat{\beta}_0 \sum x_i + \hat{\beta}_1 \sum x_i^2 = \sum y_i x_i \\ & (\bar{y} - \hat{\beta}_1 \bar{x}) \sum x_i + \hat{\beta}_1 \sum x_i^2 = \sum y_i x_i \\ & \sum x_i \bar{y} - \sum x_i \hat{\beta}_1 \bar{x} + \hat{\beta}_1 \sum x_i^2 = \sum y_i x_i \\ & \sum x_i \bar{y} - \hat{\beta}_1 \bar{x} \sum x_i + \hat{\beta}_1 \sum x_i^2 = \sum y_i x_i \\ & \sum x_i \bar{y} + \hat{\beta}_1 \sum x_i^2 - \hat{\beta}_1 \bar{x} \sum x_i = \sum y_i x_i \\ & \hat{\beta}_1 \left(\sum x_i^2 - \bar{x} \sum x_i \right) = \sum y_i x_i - \sum x_i \bar{y} \end{aligned}$$

$$\hat{\beta}_1 = \frac{\sum y_i x_i - \sum x_i \bar{y}}{\sum x_i^2 - \bar{x} \sum x_i} = \frac{\sum y_i x_i - \sum x_i \dfrac{\sum y_i}{n}}{\sum x_i^2 - \dfrac{\sum x_i}{n} \sum x_i} = \frac{\sum y_i x_i - \dfrac{\sum x_i \sum y_i}{n}}{\sum x_i^2 - \dfrac{\left(\sum x_i\right)^2}{n}}.$$

Os estimadores $\hat{\beta}_0$ e $\hat{\beta}_1$ são os pontos críticos das raízes das equações $\frac{\partial \sum e_i^2}{\partial \beta_0} = 0$ e $\frac{\partial \sum e_i^2}{\partial \beta_1} = 0$, podendo ser pontos de mínimo ou de máximo. Entretanto, demonstra-se que os pontos críticos de qualquer função que seja uma soma de quadrados serão sempre pontos de mínimo. Daí podemos concluir que $\hat{\beta}_0$ e $\hat{\beta}_1$ são pontos de mínimo, ou seja, a soma de quadrados dos erros é mínima.

Consideremos agora o experimento descrito no Exemplo 1. É importante lembrar que, sendo uma técnica de inferência, a análise de regressão linear simples tem o objetivo de determinar a equação que melhor represente o relacionamento entre as variáveis na população. No exemplo em questão, busca modelar a resposta média desta espécie de forrageira quando diferentes doses de Nitrogênio são aplicadas no solo. Sendo assim, cada parcela do experimento constitui uma amostra da população para uma determinada dose de Nitrogênio. Através da equação da reta ajustada podemos obter as estimativas dos valores médios das populações, denotados por $E(y/x_i)$ ou μ_i, para qualquer quantidade de Nitrogênio que pertença ao intervalo estudado, no exemplo, de 0 a 200 kg/ha. Vejamos agora como essas estimativas são obtidas.

Vimos que os valores observados são expressos por $y_i = \beta_0 + \beta_1 x_i + e_i$ e os valores esperados por $E(y_i/x_i) = \beta_0 + \beta_1 x_i$. As estimativas destes valores esperados são denotadas por $\hat{\mu}_i$ e podem ser obtidas através da equação ajustada:

$$\hat{\mu}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

A partir daí podemos obter também as estimativas dos erros. Sendo $e_i = y_i - E(y_i/x_i)$, as estimativas dos erros são obtidas por

$$\hat{e}_i = y_i - \hat{\mu}_i .$$

Utilizando os dados do Exemplo 1, vamos estimar os parâmetros do modelo de regressão linear simples. Inicialmente, devemos construir uma tabela auxiliar que inclua todos os cálculos intermediários para a obtenção das estimativas dos parâmetros, através do modelo $y_i = \beta_0 + \beta_1 x_i + e_i$.

Tabela auxiliar:

i	Dose de Nitrogênio (x_i)	Produção de forragem (y_i)	x_i^2	$x_i y_i$
1	0	2.160	0	0
2	50	2.880	2.500	144.000
3	100	3.360	10.000	336.000
4	150	3.720	22.500	558.000
5	200	4.020	40.000	804.000
Σ	500	16.140	75.000	1.842.000
Média	100	3.228	-	-

Obtidas a soma de quadrados X e a soma de produtos de X e Y, podemos calcular as estimativas de β_1 e β_0.

$$\hat{\beta}_1 = \frac{\sum y_i x_i - \frac{\sum x_i \sum y_i}{n}}{\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}} = \frac{1842000 - \frac{500 \times 16140}{5}}{75000 - \frac{500^2}{5}} = \frac{228000}{25000} = 9{,}12$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 3228 - 9{,}12 \times 100 = 2316$$

Podemos obter agora as estimativas das médias de produção de forragem e dos erros para cada dose de Nitrogênio.

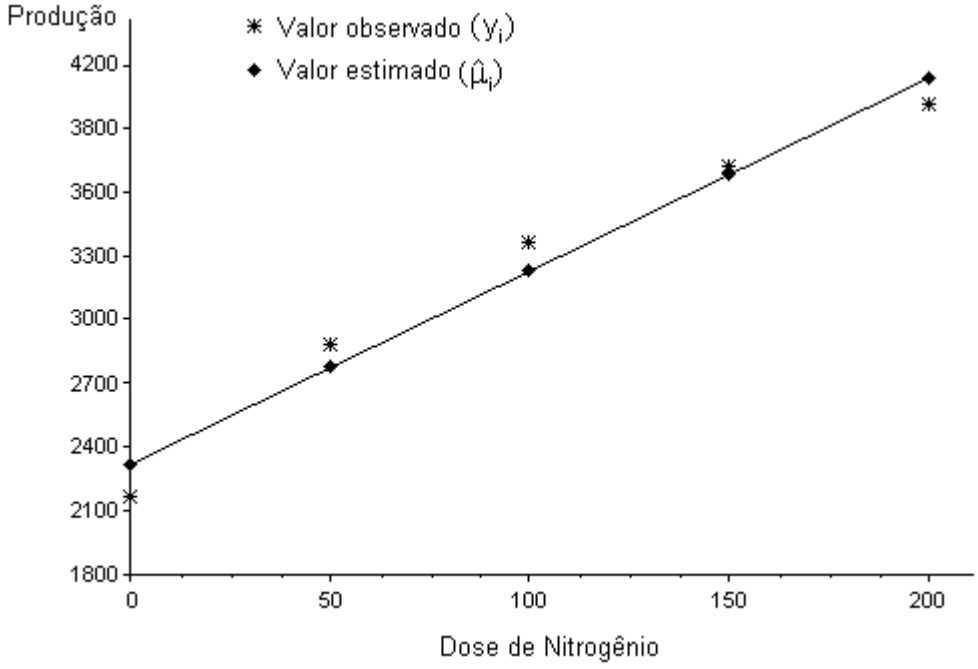
Sendo $\hat{\mu}_i = 2316 + 9{,}12 x_i$, temos

$\hat{\mu}_1 = 2316 + 9{,}12\, x_1 = 2316 + 9{,}12 \times 0 = 2.316$ kg/ha;
$\hat{\mu}_2 = 2316 + 9{,}12\, x_2 = 2316 + 9{,}12 \times 50 = 2.772$ kg/ha;
$\hat{\mu}_3 = 2316 + 9{,}12\, x_3 = 2316 + 9{,}12 \times 100 = 3.228$ kg/ha;
$\hat{\mu}_4 = 2316 + 9{,}12\, x_4 = 2316 + 9{,}12 \times 150 = 3.684$ kg/ha;
$\hat{\mu}_5 = 2316 + 9{,}12\, x_5 = 2316 + 9{,}12 \times 200 = 4.140$ kg/ha;

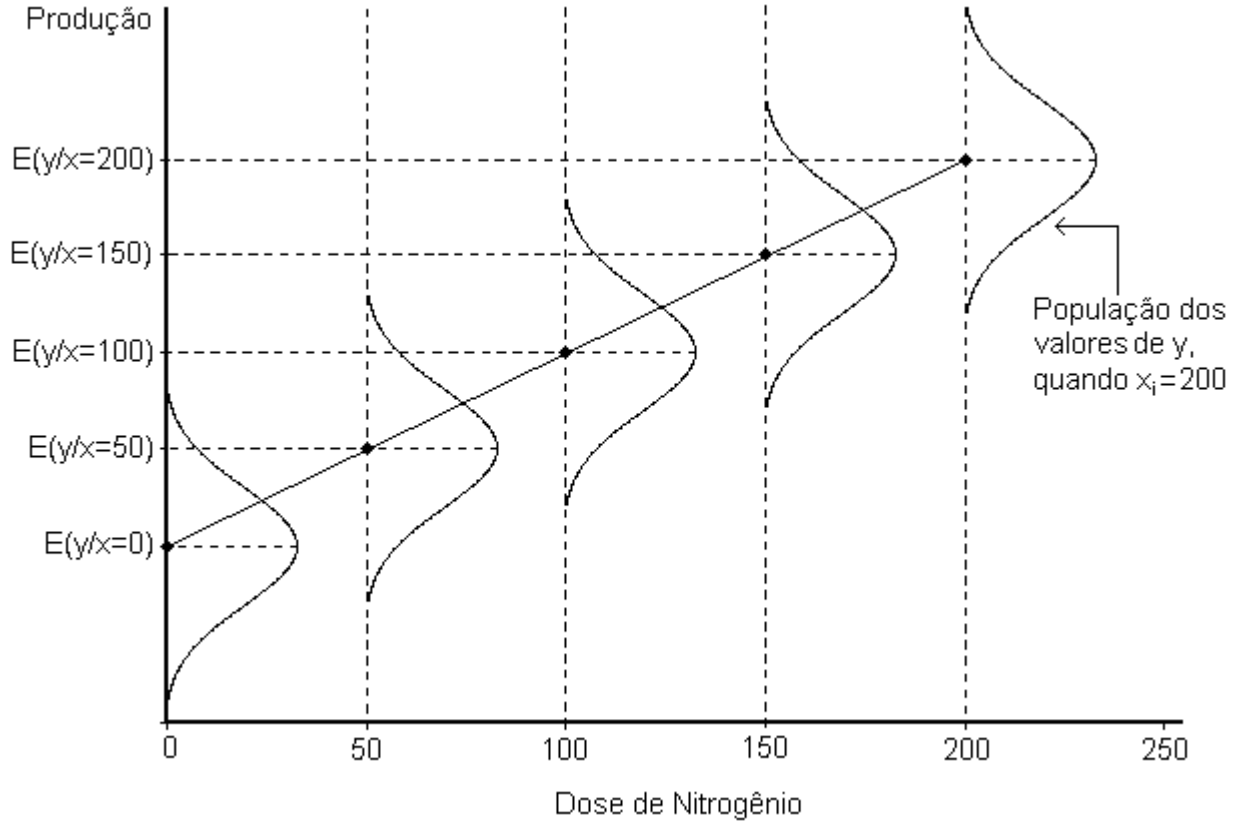
Sendo $\hat{e}_i = y_i - \hat{\mu}_i$, temos

$\hat{e}_1 = y_1 - \hat{\mu}_1 = 2160 - 2316 = -156$ kg/ha;
$\hat{e}_2 = y_2 - \hat{\mu}_2 = 2880 - 2772 = 108$ kg/ha;
$\hat{e}_3 = y_3 - \hat{\mu}_3 = 3360 - 3228 = 132$ kg/ha;
$\hat{e}_4 = y_4 - \hat{\mu}_4 = 3720 - 3684 = 36$ kg/ha;
$\hat{e}_5 = y_5 - \hat{\mu}_5 = 4020 - 4140 = -120$ kg/ha.

Na figura abaixo podemos observar o gráfico de dispersão dos valores de Y com a reta ajustada.



Admitindo que a variável resposta tem distribuição normal, os valores $\hat{\mu}_i$ estimam a produções médias populacionais $E(y/x_i)$ correspondentes às cinco doses de Nitrogênio aplicadas. O valor $y_5 = 4.020$ kg/ha , por exemplo, é o valor observado na parcela que recebeu 200 kg/ha de Nitrogênio e que constitui uma amostra aleatória da população que recebe esta dose, enquanto o valor $\hat{y}_5 = 4.140$ kg/ha é a estimativa da média desta população $E(y/x_i) = 200$, conforme podemos observar na figura a seguir.



É importante destacar também que o modelo de regressão linear simples pressupõe que as variâncias das populações de valores de Y são iguais para quaisquer valores de X. Essa homogeneidade de variâncias é representada na figura 4.4 pelas curvas de mesmo formato.

4.7.2.2. Testes de significância para a estimativa de β_1

Devemos considerar que as estimativas de β_0 e β_1, obtidas até agora, são estimativas por ponto, de modo que não sabemos o quão próximas elas estão dos parâmetros. Dentre os parâmetros do modelo de regressão linear simples, o coeficiente de regressão (β_1) é considerado o mais importante, pois é ele quem define a declividade da reta. Sendo assim, quando estimamos o β_1, devemos verificar se esta estimativa difere significativamente de zero. Esta verificação é feita através de um teste de hipóteses, cujas hipóteses de interesse são:

$$\begin{cases} H_0 : \beta_1 = 0 \\ H_A : \beta_1 \neq 0 \end{cases}$$

Se o β_1 não diferir estatisticamente de zero significa que o efeito linear de X sobre Y não é significativo. Para testar H_0 podemos utilizar dois procedimentos: a análise da variância e o teste t, já estudado anteriormente.

- **Análise da variância**

A análise da variância consiste em decompor a variação total das observações, representada pelos desvios $(y_i - \bar{y})$, em duas partes:

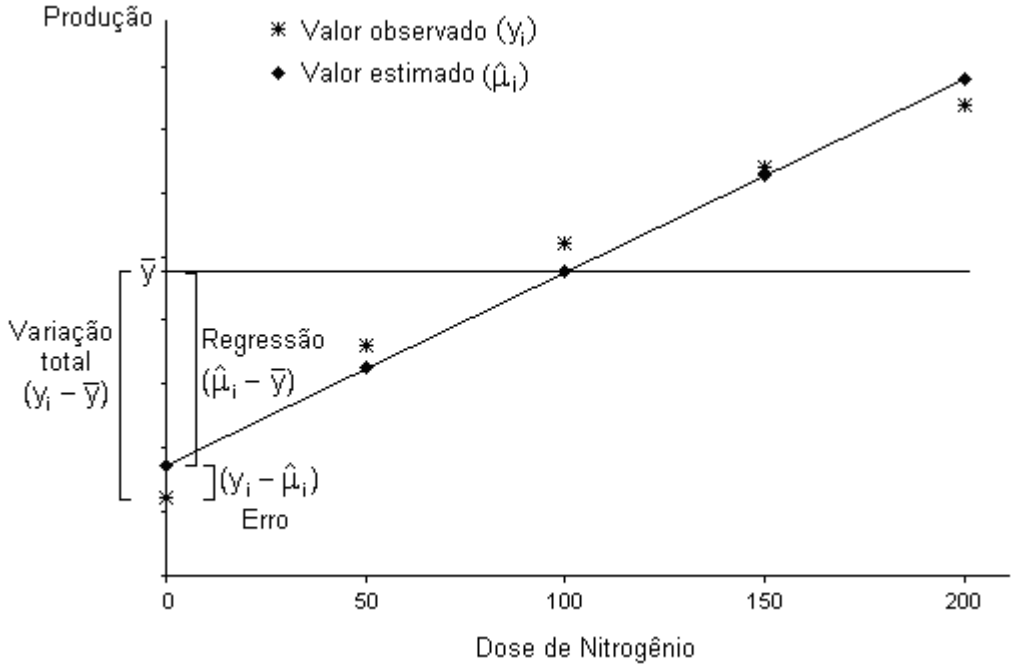
- a variação explicada pela reta da regressão, representada pelos desvios $(\hat{\mu}_i - \bar{y})$.

- a variação aleatória, não explicada pela reta, representada pelos desvios $(y_i - \hat{\mu}_i)$.

Assim, a variação de cada observação pode ser representada pela seguinte expressão:

$$(y_i - \bar{y}) = (\hat{\mu}_i - \bar{y}) + (y_i - \hat{\mu}_i)$$
$$(y_i - \bar{y}) = (\hat{\mu}_i - \bar{y}) + \hat{e}_i$$

Esses desvios podem ser observados na figura abaixo, onde temos o gráfico de dispersão dos pontos e a reta ajustada para os dados do experimento com Nitrogênio (Exemplo 1).



Considerando que a soma de desvios em relação à média é sempre zero, para obtermos a variação total das observações, devemos somar os quadrados dos desvios, o que resulta

$$\sum (y_i - \bar{y})^2 = \sum (\hat{\mu}_i - \bar{y})^2 + \sum (y_i - \hat{\mu}_i)^2$$

variação total | desvio explicado pela reta | desvio não explicado pela reta (erro)

Ao dividirmos as somas de quadrados (SQ) pelos graus de liberdade obtemos as variâncias (S^2), também denominadas quadrados médios (QM).

Os graus de liberdade e as variâncias (quadrados médios) são obtidos da seguinte forma:

– Grau de liberdade total: $\nu_{Total} = n–1$, onde n é o número de observações.

– Grau de liberdade da regressão: $\nu_{Reg} = p–1$, onde p é o número de parâmetros do modelo.

– Grau de liberdade do erro: $\nu_{Erro} = n–p$

– Variância da regressão: $S^2_{Reg} = \dfrac{SQ_{Reg}}{\nu_{Reg}}$

– Variância do erro: $S^2 = \dfrac{SQ_{Erro}}{\nu_{Erro}}$

A variância do erro (S^2) e a variância da regressão (S^2_{Reg}) são utilizados para testar a hipótese de interesse ($H_0 : \beta_1 = 0$). A S^2 estima a variação aleatória (σ^2), enquanto a S^2_{Reg} estima a variação da regressão (σ^2_{Reg}) que é composta pela variação aleatória (σ^2) mais o efeito linear de X sobre Y (ϕ_{Reg}), ou seja, $\sigma^2_{Reg} = \sigma^2 + \phi_{Reg}$. Assim, temos um conjunto de hipóteses a respeito das variâncias que corresponde ao conjunto de hipóteses a respeito do β_1:

$$\begin{cases} H_0 : \sigma^2_{Reg} = \sigma^2 \rightarrow \text{efeito linear de X sobre Y não é significativo} \\ H_A : \sigma^2_{Reg} > \sigma^2 \end{cases}$$

$$\begin{cases} H_0 : \beta_1 = 0 \rightarrow \text{efeito linear de X sobre Y não é significativo} \\ H_A : \beta_1 \neq 0 \end{cases}$$

Para testar H_0, utilizamos a estatística F, que é definida como a razão entre duas variâncias e tem distribuição F, com parâmetros ν_1 e ν_2:

$$F = \frac{S^2_{Reg}}{S^2}$$

Se esta razão for significativamente maior do que 1 (um), concluímos que a variação da regressão é significativamente maior que a variação do erro e que, portanto, esta diferença se deve ao efeito linear de X sobre Y. Vale lembrar que o modelo só é adequado para explicar o relacionamento entre as duas variáveis quando a parte fixa do modelo (Regressão) é preponderante sobre a parte aleatória (Erro).

Em geral, a análise da variância é apresentada na forma de tabela, conforme o esquema abaixo.

Tabela da análise da variância:

Fonte de variação	ν	SQ	E(S^2)	F
Regressão	p – 1	$\sum(\hat{\mu}_i - \bar{y})^2$	$\sigma^2 + \phi_{Reg}$	$\dfrac{S^2_{Reg}}{S^2}$
Erro	n – p	$\sum e_i^2 = \sum(y_i - \hat{\mu}_i)^2$	σ^2	
Total	n – 1	$\sum(y_i - \bar{y})^2$		

Para facilitar o processo de cálculo na obtenção das somas de quadrados, as seguintes fórmulas práticas podem ser utilizadas:

$$SQ_{Total} = \sum y_i^2 - \frac{(\sum y_i)^2}{n};$$

$$SQ_{Reg} = \hat{\beta}_1^2 \sum (x_i - \bar{x})^2;$$

$SQ_{Erro} = SQ_{Total} - SQ_{Reg}$ (por diferença).

A decisão a respeito de H_0 será tomada comparando o valor da estatística F com o valor crítico encontrado na tabela de F.

Rejeitamos H_0, ao nível α de significância, se $f = \frac{S_{Reg}^2}{S^2} > f_{\alpha(\nu_1,\nu_2)}$.

Não rejeitamos H_0, ao nível α de significância, se $f = \frac{S_{Reg}^2}{S^2} < f_{\alpha(\nu_1,\nu_2)}$.

Para o Exemplo 1 vamos testar a hipótese de interesse a respeito do β_1. Inicialmente, obtemos as somas de quadrados, através das fórmulas práticas. Temos então:

$$SQ_{Total} = \sum y_i^2 - \frac{(\sum y_i)^2}{n} = 54248400 - \frac{260499600}{5} = 2148480$$

$$SQ_{Reg} = \hat{\beta}_1^2 \sum (x_i - \bar{x})^2 = 9{,}12^2 \times 25000 = 2079360$$

$SQ_{Erro} = SQ_{Total} - SQ_{Reg} = 2148480 - 2079360 = 69120$

Obtidas as somas de quadrados, os demais resultados podem ser apresentados diretamente na tabela da análise da variância.

Tabela da análise da variância:

Fonte de variação	ν	SQ	S^2	F
Regressão	1	2.079.360	2.079.360	90,25
Erro	3	69.120	23.040	
Total	4	2.148.480		

Como o valor calculado f = 90,25 foi maior que o valor crítico $f_{0,01(1,3)} = 34{,}12$, concluímos, ao nível α = 0,01, que o efeito linear da dose de Nitrogênio sobre a produção desta forrageira é significativo, sendo que essa relação pode ser expressa pela equação $\hat{\mu}_i = 2316 + 9{,}12x_i$. Isto significa que para cada kg/ha de Nitrogênio aplicado no solo a produção de forragem aumenta, em média, 9,12 kg/ha.

◆ **Teste t**

Outro procedimento que pode ser utilizado para testar $H_0 : \beta_1 = 0$ é o teste t. Como já visto em seções anteriores, utilizamos a estatística T que tem distribuição t de Student quando H_0 é verdadeira. Nesse caso, temos $\theta = \beta_1 = 0$, resultando:

$$T = \frac{\hat{\theta} - \theta}{S(\hat{\theta})} = \frac{\hat{\theta} - 0}{S(\hat{\theta})} = \frac{\hat{\theta}}{S(\hat{\theta})} \sim t(\nu),$$

onde:

$\hat{\theta} = \hat{\beta}_1$;

$S(\hat{\theta}) = S(\hat{\beta}_1)$;

$\nu = n - 2$;

A estimativa do erro padrão do estimador do coeficiente de regressão, $S(\hat{\beta}_1)$, é obtida da seguinte forma:

$$V(\hat{\beta}_1) = V\left(\frac{\sum y_i(x_i - \overline{x})}{\sum (x_i - \overline{x})^2}\right)$$

$$V(\hat{\beta}_1) = \left(\frac{1}{\sum (x_i - \overline{x})^2}\right)^2 V\left[\sum y_i(x_i - \overline{x})\right]$$

$$V(\hat{\beta}_1) = \frac{1}{\left[\sum (x_i - \overline{x})^2\right]^2} (\sum x_i - \overline{x})^2 V(y_i)$$

$$V(\hat{\beta}_1) = \frac{V(y_i)}{\sum (x_i - \overline{x})^2}$$

$$V(\hat{\beta}_1) = \frac{\sigma^2}{\sum (x_i - \overline{x})^2}$$

Sendo σ^2 um parâmetro desconhecido, utilizamos o seu estimador

$$S^2 = \frac{\sum (y_i - \hat{\mu}_i)^2}{n - 2} = \frac{\sum \hat{e}_i^2}{n - 2}$$

para obter a estimativa da variância do estimador do coeficiente de regressão

$$S^2(\hat{\beta}_1) = \frac{S^2}{\sum (x_i - \overline{x})^2} = \frac{\frac{\sum \hat{e}_i^2}{n - 2}}{\sum (x_i - \overline{x})^2}.$$

Daí resulta que

$$S(\hat{\beta}_1) = \sqrt{S^2(\hat{\beta}_1)} = \sqrt{\frac{\frac{\sum \hat{e}_i^2}{n - 2}}{\sum (x_i - \overline{x})^2}}.$$

Assim, sob H_0 verdadeira, temos

$$T = \frac{\hat{\beta}_1}{S(\hat{\beta}_1)} = \frac{\hat{\beta}_1}{\sqrt{\frac{\sum \hat{e}_i^2}{\frac{n-2}{\sum (x_i - \bar{x})^2}}}} \sim t(\nu)$$

No exemplo, temos

$$t = \frac{\hat{\beta}_1}{s(\hat{\beta}_1)} = \frac{9,12}{\sqrt{\frac{23040}{25000}}} = \frac{9,12}{0,96} = 9,5$$

Como o valor calculado t = 9,5 foi maior que o valor crítico valor $t_{\alpha/2(3)}$ = 5,841, concluímos, ao nível α = 0,01, que o efeito linear da dose de Nitrogênio sobre a produção desta forrageira é significativo. Podemos verificar também a correspondência entre os valores das estatísticas F e T. O valor da estatística F deve ser igual ao quadrado do valor da estatística T $(f = t^2)$. Para esse exemplo temos $f = 90,25 = 9,5^2 = t^2$.

Vimos em seções anteriores que o teste t bilateral e o intervalo de confiança, para um mesmo nível α, são procedimentos estatísticos equivalentes de modo que conduzem aos mesmos resultados. Sendo assim, o intervalo de confiança também pode ser utilizado para verificar se β_1 difere significativamente de zero ou não. Utilizando as mesmas expressões acima deduzidas, podemos obter o intervalo de confiança para o β_1. Partindo da expressão geral para intervalos de confiança

$$IC(\theta; 1-\alpha) : \hat{\theta} \pm t_{\alpha/2} S(\hat{\theta}),$$

e fazendo as substituições referentes ao parâmetro em questão, temos

$$IC(\beta_1; 1-\alpha) : \hat{\beta}_1 \pm t_{\alpha/2} S(\hat{\beta}_1)$$

$$IC(\beta_1; 1-\alpha) : \hat{\beta}_1 \pm t_{\alpha/2} \sqrt{\frac{\sum \hat{e}_i^2}{\frac{n-2}{\sum (x_i - \bar{x})^2}}}$$

No exemplo, temos

$$IC(\beta_1; 1-\alpha) : \hat{\beta}_1 \pm t_{\alpha/2} \sqrt{\frac{\sum \hat{e}_i^2}{\frac{n-2}{\sum (x_i - \bar{x})^2}}}$$

$$IC(\beta_1; 0,99) : 9,12 \pm 5,841 \sqrt{\frac{23040}{25000}}$$

$$IC(\beta_1; 0,99) : 9,12 \pm 5,61$$

$$\text{Limite inferior} : 9,12 - 5,61 = 3,51$$

$$\text{Limite superior} : 9,12 + 5,61 = 14,63$$

$$P(3,51 < \beta_1 < 14,63) = 0,99$$

Assim, concluímos que probabilidade de os limites 3,51 e 14,63 conterem o verdadeiro valor do coeficiente de regressão β_1 é de 0,99. Portanto, o efeito linear da dose de Nitrogênio sobre a produção da forrageira é significativo.

O teste de significância e o intervalo de confiança para o parâmetro β_0 são feitos de maneira análoga. Nesse caso, a estatística

$$T = \frac{\hat{\theta}}{S(\hat{\theta})} \sim t(\nu)$$

é utilizada considerando o seguinte:

$\theta = \beta_0$;

$\hat{\theta} = \hat{\beta}_0$;

$$S(\hat{\theta}) = S(\hat{\beta}_0) = \sqrt{\left[\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i - \bar{x})^2}\right] S^2} = \sqrt{\left[\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i - \bar{x})^2}\right] \frac{\sum e_i^2}{n-2}};$$

$\nu = n - 2$.

4.7.2.3. Intervalos de confiança para as médias das populações μ_i

Como vimos anteriormente, μ_i é um parâmetro e $\hat{\mu}_i$ é a estimativa pontual desse parâmetro. Vejamos agora como construir um intervalo de confiança para μ_i. Consideremos a expressão geral do intervalo de confiança:

$$IC(\theta; 1-\alpha): \hat{\theta} \pm t_{\alpha/2}\, S(\hat{\theta}),$$

onde:

$\theta = \mu_i$

$\hat{\theta} = \hat{\mu}_i$

$S(\hat{\theta}) = S(\hat{\mu}_i)$

$\nu = n - 2$.

Para obter o erro padrão do estimador $\hat{\mu}_i$, partimos do modelo

$$\hat{\mu}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i,$$

Sendo $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$, temos

$$\hat{\mu}_i = \bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x_i$$

$$\hat{\mu}_i = \bar{y} + \hat{\beta}_1 (x_i - \bar{x}).$$

Para este modelo temos

$$V(\hat{\mu}_i) = V\left[\bar{y} + \hat{\beta}_1 (x_i - \bar{x})\right]$$

$$V(\hat{\mu}_i) = V(\bar{y}) + V\left[\hat{\beta}_1 (x_i - \bar{x})\right]$$

$$V(\hat{\mu}_i) = V(\bar{y}) + (x_i - \bar{x})^2\, V(\hat{\beta}_1)$$

Sendo $\hat{\beta}_1 = \dfrac{\sum y_i (x_i - \bar{x})}{\sum (x_i - \bar{x})^2}$, $V(\bar{y}) = \dfrac{\sigma^2}{n}$ e $\sigma^2 = \dfrac{\sum e_i^2}{n-2}$, temos

$$V(\hat{\mu}_i) = \frac{\sigma^2}{n} + (x_i - \bar{x})^2 \, V\left(\frac{\sum y_i (x_i - \bar{x})}{\sum (x_i - \bar{x})^2}\right)$$

$$V(\hat{\mu}_i) = \frac{\sigma^2}{n} + (x_i - \bar{x})^2 \frac{\sigma^2}{\sum (x_i - \bar{x})^2}$$

$$V(\hat{\mu}_i) = \left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum (x_i - \bar{x})^2}\right)\sigma^2$$

$$V(\hat{\mu}_i) = \left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum (x_i - \bar{x})^2}\right)\left(\frac{\sum e_i^2}{n-2}\right)$$

Sendo σ^2 um valor desconhecido, utilizamos o seu estimador

$$S^2 = \frac{\sum \hat{e}_i^2}{n-2}$$

para obter a estimativa da variância do estimador $\hat{\mu}_i$

$$S^2(\hat{\mu}_i) = \left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum (x_i - \bar{x})^2}\right) S^2 = \left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum (x_i - \bar{x})^2}\right)\left(\frac{\sum \hat{e}_i^2}{n-2}\right).$$

Daí resulta que

$$S(\hat{\mu}_i) = \sqrt{S^2(\hat{\mu}_i)} = \sqrt{\left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum (x_i - \bar{x})^2}\right)\left(\frac{\sum \hat{e}_i^2}{n-2}\right)}.$$

O intervalo de confiança para μ_i é obtido pela expressão

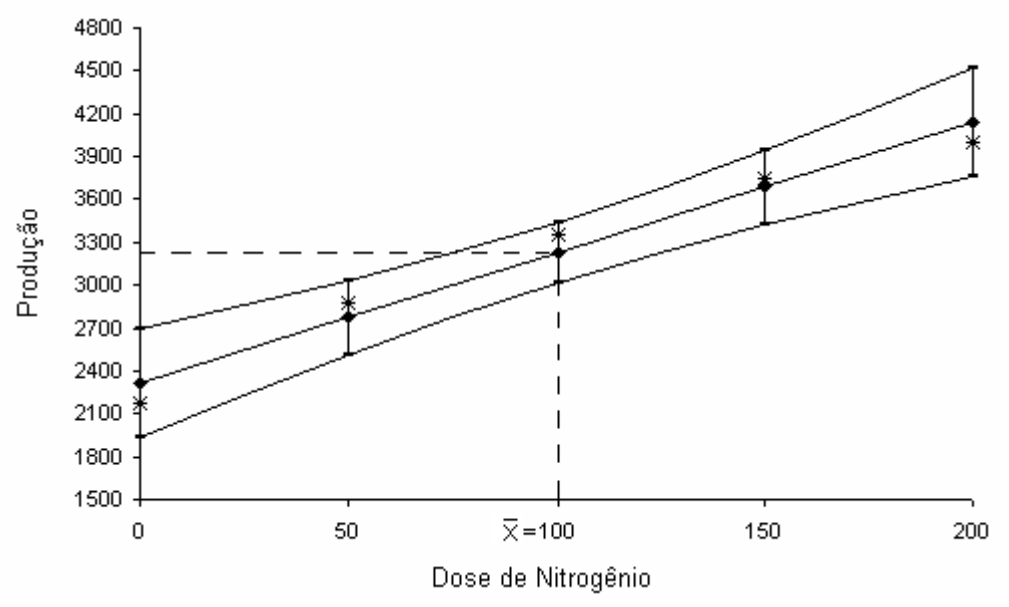
$$IC(\mu_i; 1-\alpha) : \hat{\mu}_i \pm t_{\alpha/2} \, S(\hat{\mu}_i)$$

$$IC(\mu_i; 1-\alpha) : \hat{\mu}_i \pm t_{\alpha/2} \sqrt{\left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum (x_i - \bar{x})^2}\right)\left(\frac{\sum \hat{e}_i^2}{n-2}\right)}.$$

Utilizando a expressão acima, vamos construir os intervalos de confiança para as médias do Exemplo 1. Na tabela auxiliar abaixo temos os cálculos intermediários e os valores obtidos para os limites dos intervalos, considerando α = 0,05 e o valor $t_{\alpha/2(3)}$ = 3,183.

i	x_i	y_i	$(x_i - \bar{x})^2$	$\hat{\mu}_i$	e_i^2	$s(\hat{\mu}_i)$	$t_{\alpha/2}\, s(\hat{\mu}_i)$	Limite inferior	Limite superior
1	0	2160	10000	2.316	24.336	117,58	374,26	1.941,76	2.690,24
2	50	2880	2500	2.772	11.664	83,14	264,63	2.507,37	3.036,63
3	100	3360	0	3.228	17.424	67,88	216,06	3.011,93	3.444,07
4	150	3720	2500	3.684	1.296	83,14	264,63	3.419,37	3.948,63
5	200	4020	10000	4.140	14.400	117,58	374,26	3.765,76	4.514,24
Σ	500	16.140	25.000	16.140	69.120	-	-	-	-

A figura a seguir apresenta o gráfico de dispersão dos valores de Y com os intervalos ao nível de 95% de confiança estimados para as médias μ_i. Podemos observar que o intervalo de confiança tem maior precisão no ponto $x_i = \bar{x}$, onde o desvio $(x_i - \bar{x})$ é igual a zero. À medida que se distancia da média, o intervalo de confiança aumenta sua amplitude, ou seja, diminui a precisão.



Dispersão dos valores de Y com os intervalos ao nível de 95% de confiança estimados para as médias μ_i.