



Serviço Geológico do Brasil – CPRM

Correlação e Regressão Linear

Aula 02 : RLS

Eber José de Andrade Pinto
Coordenador Executivo do DEHID
eber.andrade@cprm.gov.br
www.cprm.gov.br

Belo Horizonte, 20 de outubro de 2020

Livro Texto

HIDROLOGIA ESTATÍSTICA vem preencher significativa lacuna na literatura técnica especializada em língua portuguesa no campo dos recursos hídricos. O conhecimento das ferramentas de estatística é fundamental para a evolução e para a prática da Hidrologia, onde encontra diversificada gama de aplicações nas atividades rotineiras ligadas aos estudos e projetos de engenharia hidrológica, que necessitam das teorias probabilísticas para a sua solução.

Conhecer e investigar as variáveis do meio físico são atributos comuns entre os conceitos aqui registrados e o Serviço Geológico do Brasil – CPRM. O livro apresenta o material didático capaz de orientar a pesquisa, e, com essa iniciativa, a instituição amplia a visibilidade do seu papel de agente promotor dos levantamentos hidrológicos básicos no país.

HIDROLOGIA ESTATÍSTICA é publicação dirigida para os profissionais do setor, bem como para a formação de alunos de graduação e pós-graduação. Municia o leitor com princípios introdutórios, análise de dados, teoria das probabilidades, variáveis aleatórias discretas e contínuas, análise de frequência, correlação e regressão. Destaca também técnicas mais sofisticadas de tratamento, manipulação e representação de dados estatísticos, com exemplos práticos reais e selecionados da rede hidrometeorológica operada pela CPRM.

www.cprm.gov.br



Secretaria de
Geologia, Mineração e
Transformação Mineral

Ministério de
Minas e Energia



Período Contemporâneo



ANO INTERNACIONAL DO PLANETA TERRA - 2006



2007

Hidrologia Estatística

AGOSTO
DE 2007



Hidrologia Estatística

MAURO NAGHETTINI
ÉBER JOSÉ DE ANDRADE PINTO



Inferência estatística

- É uma das partes da *Estatística* que tem por objetivo a coleta, redução, análise e modelagem dos dados, a partir do que, finalmente, faz-se a inferência para uma população da qual os dados (a amostra) foram obtidos.
- Um aspecto importante da modelagem dos dados é fazer *previsões*, a partir das quais se podem tomar decisões.

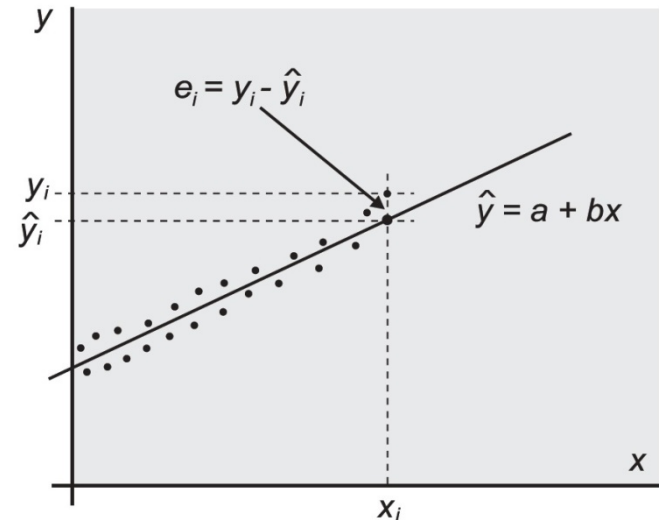
Modelos

- Quando se procede a uma análise de dados, busca-se alguma forma de *regularidade* ou *padrão* ou, ainda, *modelo*, presente nas observações.
- Os pontos da Figura não estão todos, evidentemente, sobre uma reta; essa seria o nosso padrão ou modelo. A diferença entre os dados e o modelo constitui os *resíduos*.

$$\text{Dados} = \text{Modelo} + \text{Resíduos}$$

ou, ainda,

$$D = M + R$$



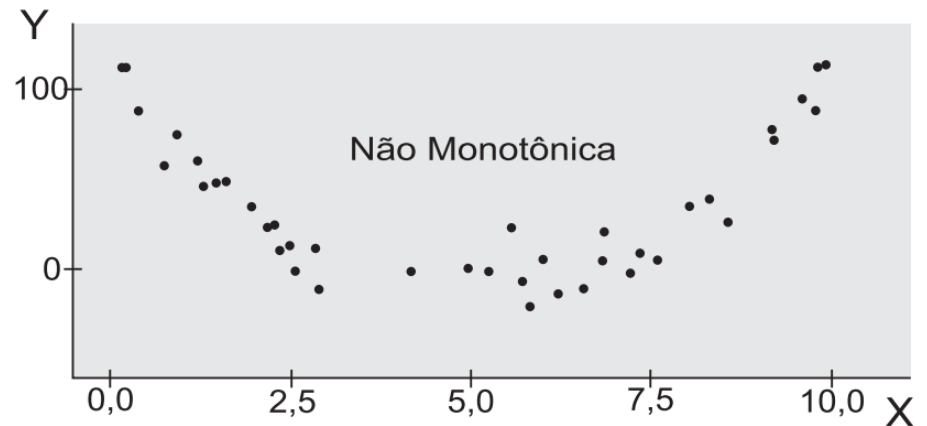
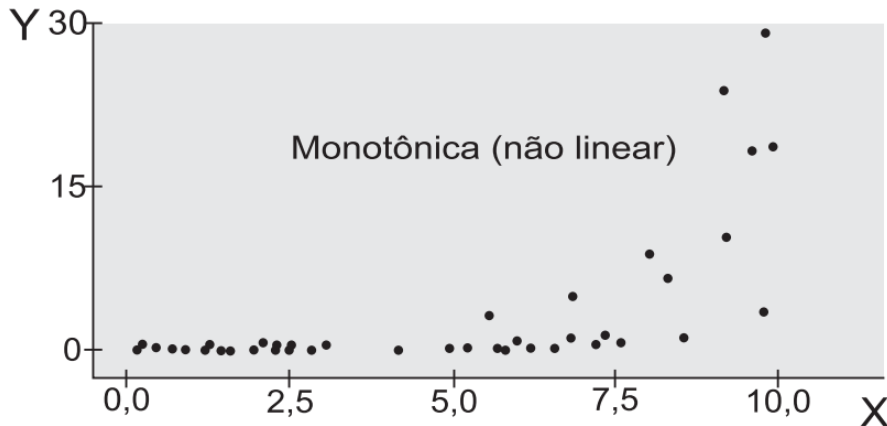
A parte M é também chamada *parte suave* dos dados, enquanto R é a *parte aleatória*.

- A parte R é tão importante quanto M , e a análise dos resíduos constitui uma parte fundamental de todo trabalho estatístico.
- Basicamente, são os resíduos que nos dizem se o modelo é adequado ou não para representar os dados.

Objetivo 2 : Regressão

Definir a forma da associação entre as VA's

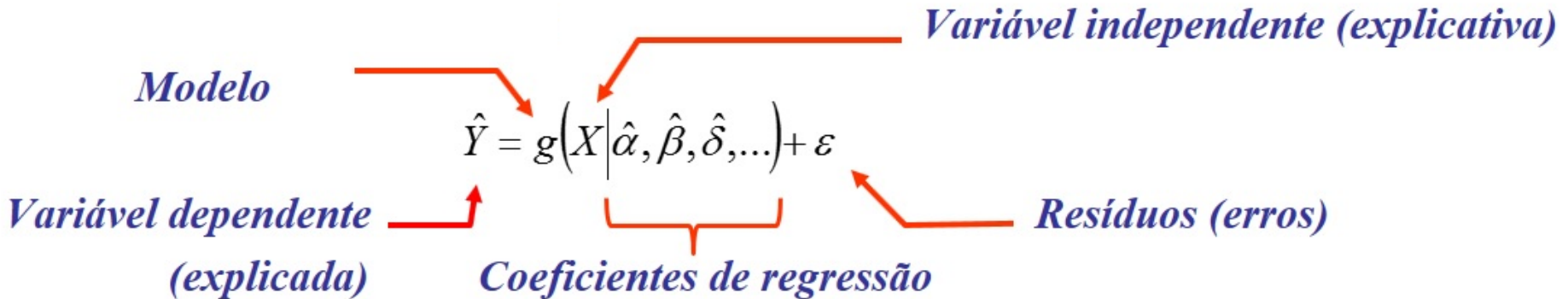
Abordagem Exploratória: Associações Monotônicas Lineares e Não-Lineares



Abordagem Quantitativa: estabelecer a forma matemática (modelo de regressão) da associação entre as VA's, por meio do uso sistemático de algum critério de minimização dos 'resíduos' (erros) entre os os valores observados e os valores previstos.

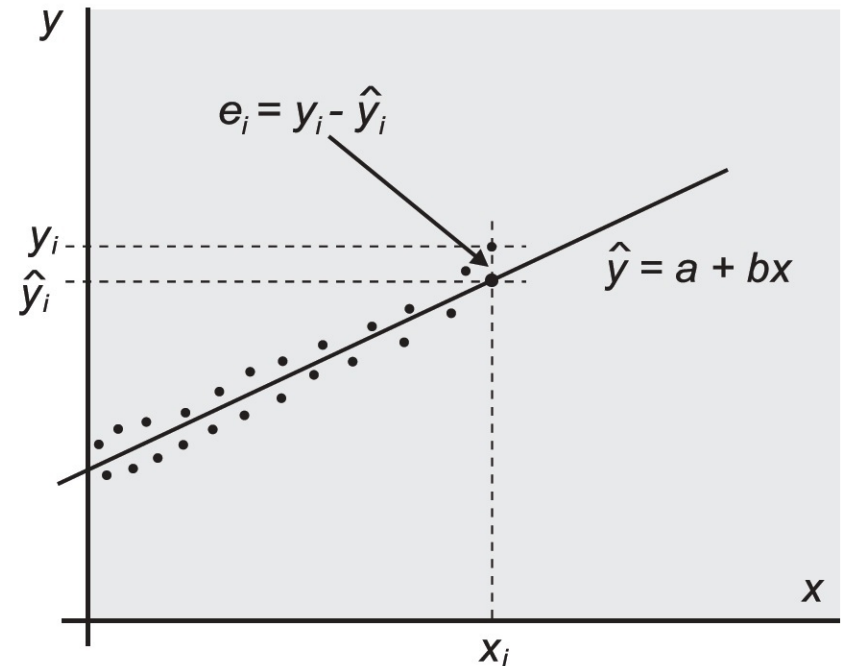
Critério mais usado: **método dos mínimos quadrados.**

Regressão Linear Simples



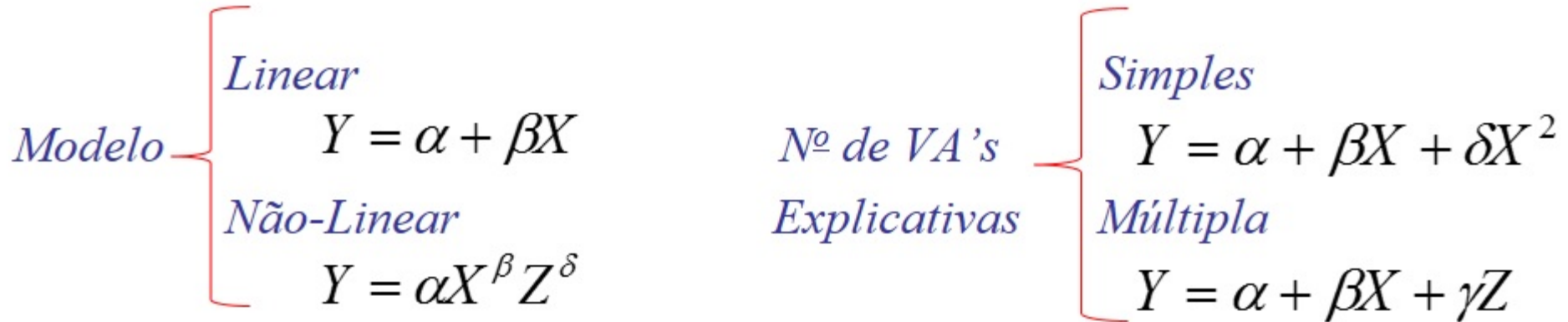
Objetivo das regressões: achar o lugar geométrico dos pontos (ou achar as estimativas dos coeficientes de regressão) que minimiza o conjunto dos erros.

$$\varepsilon_i = Y_i - \hat{Y}_i = Y_i - g(X_i | \hat{\alpha}, \hat{\beta}, \dots)$$



Regressão Linear Simples

Classificação das regressões

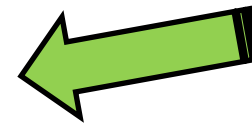


Critérios de Minimização dos Resíduos:

$$\text{Min} \sum_{i=1}^N \varepsilon_i$$

$$\text{Min} \sum_{i=1}^N |\varepsilon_i|$$

$$\text{Min} \sum_{i=1}^N \varepsilon_i^2$$



Método dos
Mínimos
Quadrados

Estimação dos coeficientes de regressão

Modelo: $Y = \alpha + \beta X + \varepsilon$

Estimativa: $\hat{Y}_i = \hat{\alpha} + \hat{\beta}X_i + \hat{\varepsilon}_i = a + bX_i + e_i$

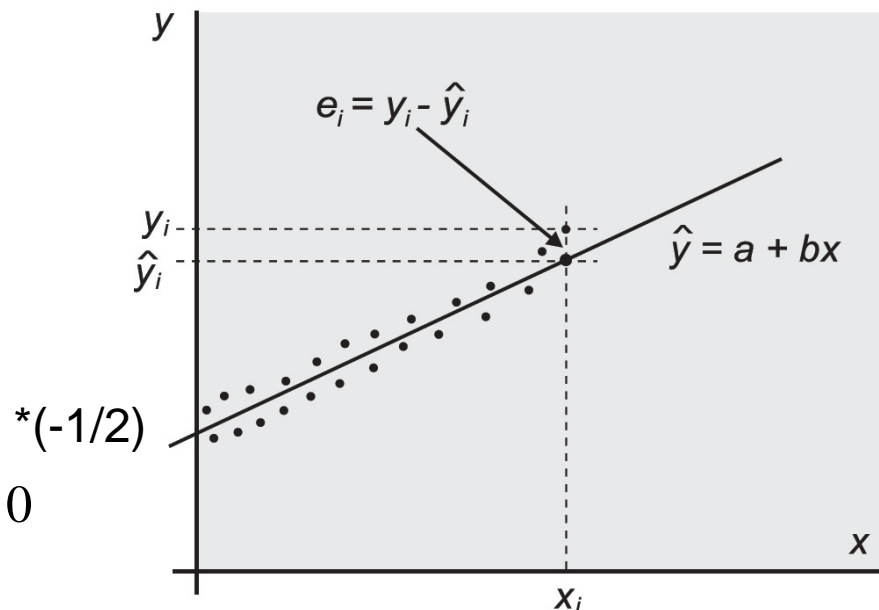
$$e_i = y_i - (a + bx_i) = y_i - a - bx_i$$

$$e_i^2 = (y_i - a - b.x_i)^2 = y_i^2 - 2.y_i.a - 2.y_i.b.x_i + a^2 + 2.a.b.x_i + b^2.x_i^2$$

$$Z = \sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n y_i^2 - 2.a.\sum_{i=1}^n y_i - 2.b.\sum_{i=1}^n x_i.y_i + na^2 + 2.a.b\sum_{i=1}^n x_i + b^2.\sum_{i=1}^n x_i^2$$

MMQ: $\text{Min } Z = \text{Min} \sum_{i=1}^n \varepsilon_i^2 \quad \begin{cases} \frac{\partial Z}{\partial a} = 0 \\ \frac{\partial Z}{\partial b} = 0 \end{cases}$

$$\begin{cases} \frac{\partial Z}{\partial a} = -2.\sum_{i=1}^n y_i + 2.n.a + 2.b.\sum_{i=1}^n x_i = 0 \\ \frac{\partial Z}{\partial b} = -2.\sum_{i=1}^n x_i.y_i + 2.a.\sum_{i=1}^n x_i + 2.b.\sum_{i=1}^n x_i^2 = 0 \end{cases}$$

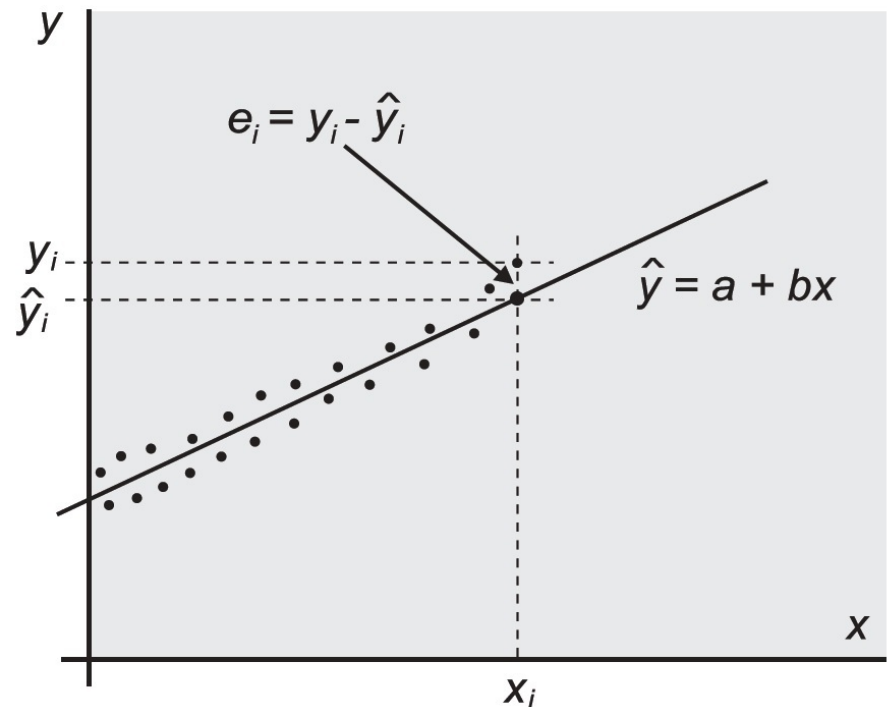


Estimação dos coeficientes de regressão

$$\begin{cases} \sum_{i=1}^n y_i - n.a - b.\sum_{i=1}^n x_i = 0 \\ \sum_{i=1}^n x_i y_i - a.\sum_{i=1}^n x_i - b.\sum_{i=1}^n x_i^2 = 0 \end{cases}$$

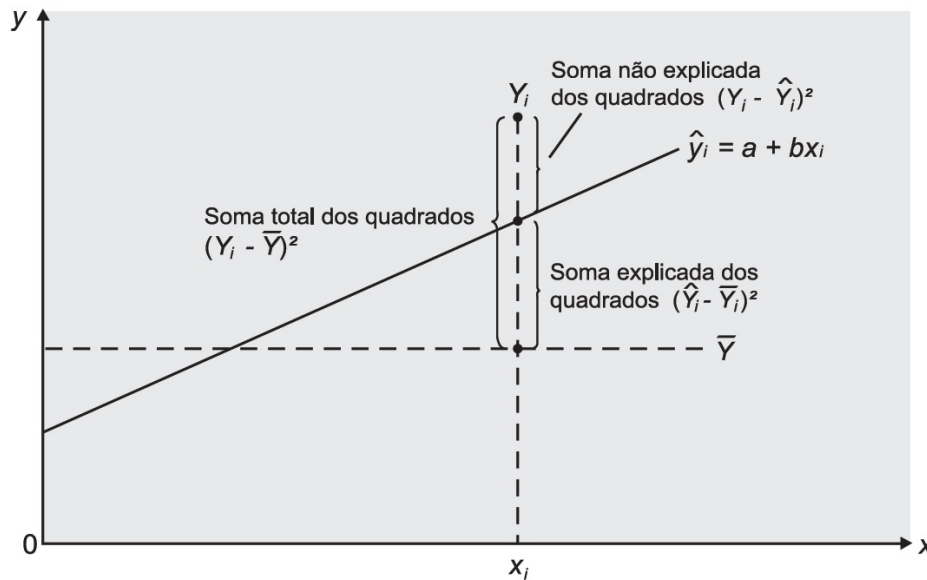
$$a = \frac{\sum_{i=1}^n y_i}{n} - b.\frac{\sum_{i=1}^n x_i}{n} = \bar{y} - b.\bar{x}$$

$$b = \frac{n.\sum_{i=1}^n x_i.y_i - \sum_{i=1}^n y_i.\sum_{i=1}^n x_i}{n.\sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2}$$



Coeficiente de Determinação

Qual é a parcela da variância total de Y que foi explicada pela regressão com X ?



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

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$$

$$SQT = SQRes + SQReg$$

$$r^2 = \frac{\text{Variância Explicada}}{\text{Variância Total}} = \frac{SQReg}{SQT} = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \text{ ou}$$

$$r^2 = \frac{SQT - SQRes}{SQT} = 1 - \frac{SQRes}{SQT} = 1 - \frac{\sum_{i=1}^n (\hat{y}_i - y_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

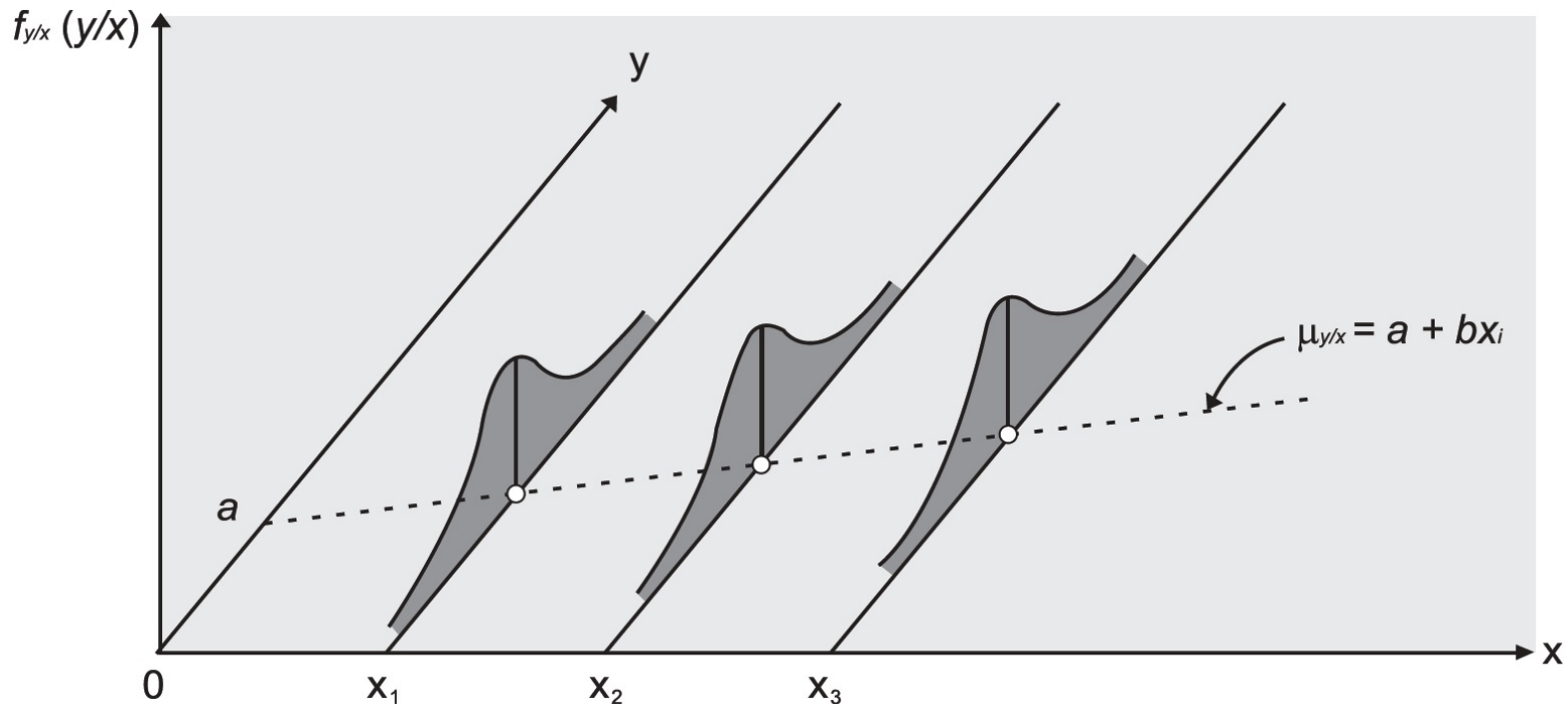
RLS:

$$r^2 = b^2 \frac{s_X^2}{s_Y^2} > 0$$

$$r = \pm \sqrt{r^2} = \langle \text{ sinal de } b \rangle \sqrt{r^2}$$

Hipóteses Fundamentais da Regressão (TH+IC)

A linearidade, a normalidade e a homoscedasticidade dos resíduos



$$e_i = y_i - (\alpha + \beta \cdot x_i) \Leftrightarrow \varepsilon \sim N(0, \sigma_e)$$

Implicações:

- O valor esperado do erro é zero $E(e_i) = 0$
- A correlação serial entre os resíduos é nula
- A variância dos resíduos é constante ao longo da variação de X (Homocedasticidade)

Erro Padrão da Estimativa

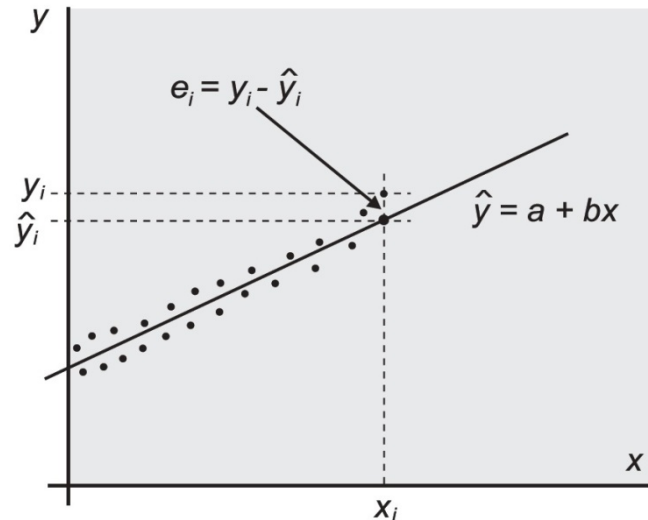
Variância dos Resíduos: $Var(e_i) = \sigma_e^2 = E(e_i^2) - E^2(e_i) = E(e_i^2)$

Ver capítulo 3, equação 3.21, página 75

$$E(e_i) = 0$$

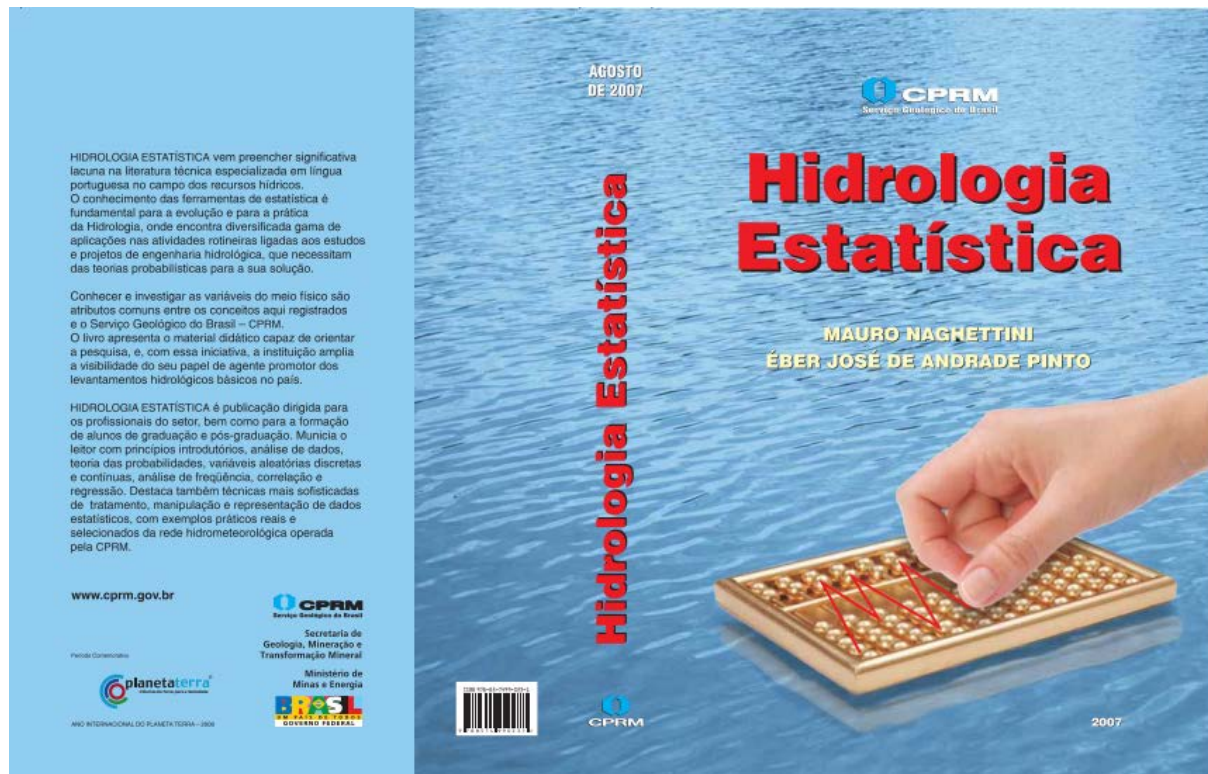
Estimador sem Viés: $\hat{\sigma}_e^2 = s_e^2 = \frac{\sum_{i=1}^n e_i^2}{n-2} = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n-2}$

Erro Padrão da Estimativa: $\hat{\sigma}_e = s_e = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n-2}}$



Recomendações

Para consolidar conhecimentos estudar no livro texto até o item 9,4, inclusive





Serviço Geológico do Brasil – CPRM

Departamento de Hidrologia da CPRM

Eber José de Andrade Pinto
Coordenador Executivo do DEHID
eber.andrade@cprm.gov.br
www.cprm.gov.br

Belo Horizonte, 20 de outubro de 2020