Serviço Geológico do Brasil – CPRM

Correlação e Regressão Linear

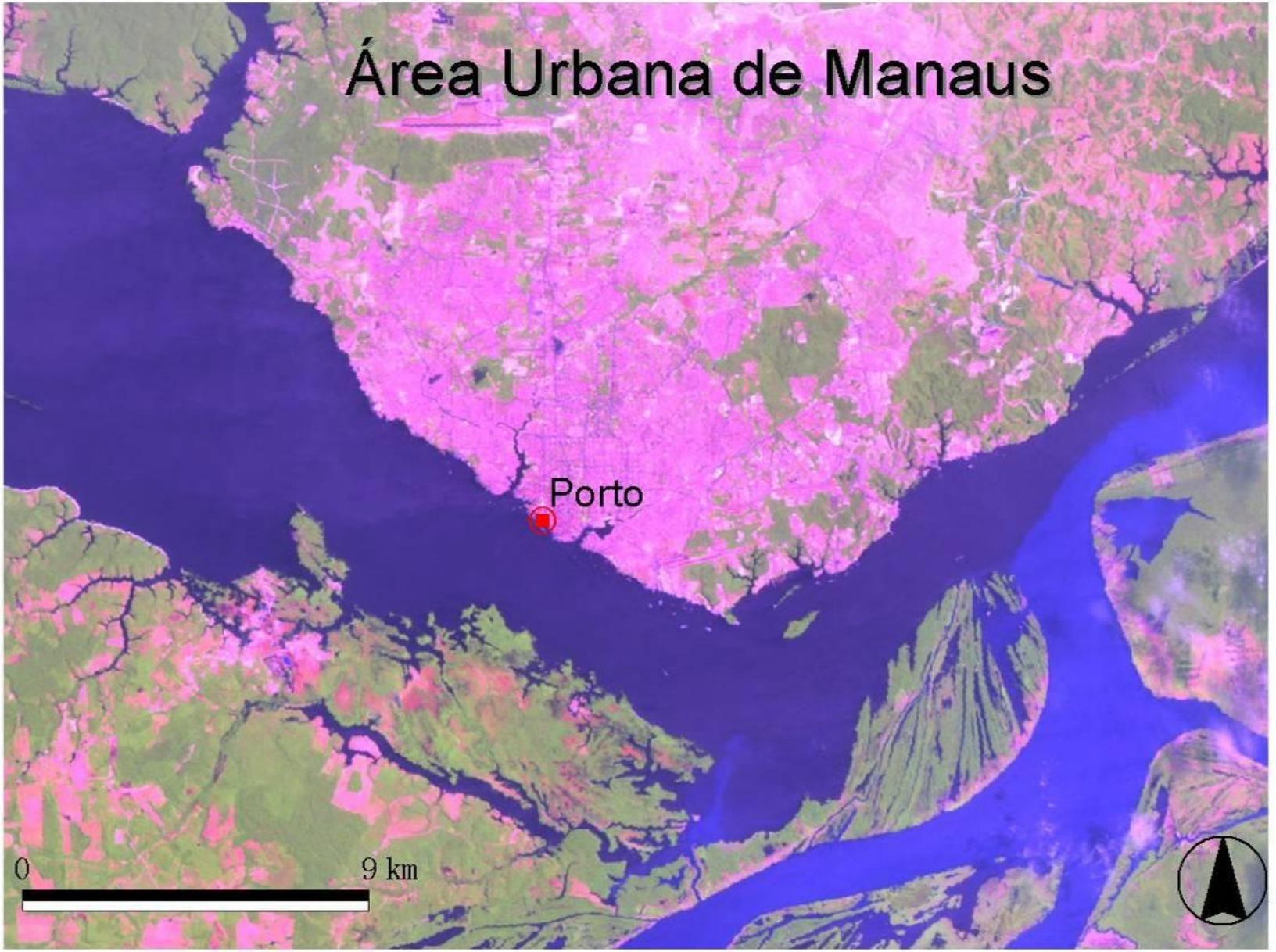
Aula 04 : Exercícios RLS

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

Belo Horizonte, 21 de outubro de 2020

Exercício Proposto

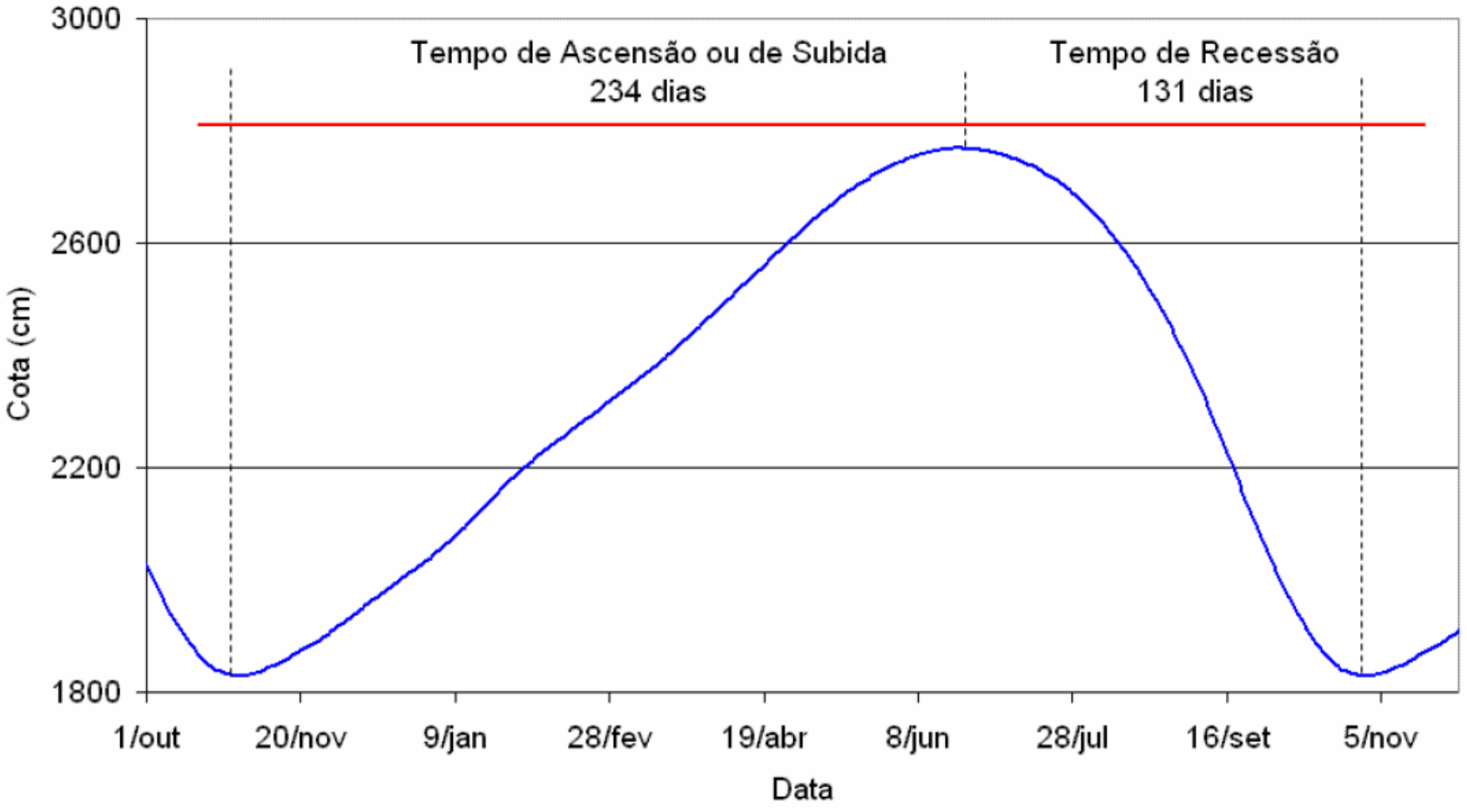
Utilizando os dados presentes na planilha Manaus.xlsx empregue o seus conhecimentos sobre RLS para definir um modelo de previsão para as cotas máximas na régua do porto de Manaus. Utilizar como variável preditora a cota registrada no dia 30 de Abril.


Área Urbana de Manaus
Porto
0
9 km
Landsat 7 - 2002

ESTAÇÃO DO ROADWAY – PORTO DE MANAUS

Vazante 2005

Cheia 2008


Tempo de Ascensão ou de Subida
234 dias
Tempo de Recessão
131 dias
3000
2600
2200
1800
Cota (cm)
1/out
20/nov
9/jan
28/fev
19/abr
8/jun
28/jul
16/set
5/nov
Data

O procedimento para análise da RLS:

Etapa 1 Selecione a variável preditora (X) que está relacionada à variável a ser prevista (Y) por alguma relação física.

Etapa 2 Plote a variável preditora (X) em relação à variável a ser prevista (Y)

Etapa 3 Determine a forma da equação desejada; isto é, linear ou curvilíneo.

Etapa 4 Calcule o coeficiente de correlação entre as variáveis.

Etapa 5 Calcule os coeficientes de regressão.

No EXCEL: Função PROJ.LIN(), PROJ.LOG() e na ferramenta Análise de Dados/Regressão

Etapa 6 Calcule o erro padrão da estimativa, Se; desvio padrão da variável a ser prevista, Sy; e o coeficiente de determinação, r^2.

O procedimento para análise da RLS:

Etapa 7 Avalie a equação de regressão pelos seguintes métodos:

- O erro padrão da estimativa tem os limites $0 \leq Se \leq Sy$; se Se -> 0 maior parte da variância é explicada pela regressão.
- Coeficiente de determinação tem limites $0 \leq r^2 \leq 1$; quando r^2 -> 1, melhor será o “ajuste” da linha de regressão aos dados.
- Examine os resíduos para identificar deficiências na equação de regressão e verifique as suposições do modelo.

Etapa 8 Se a precisão da equação de regressão não for aceitável, reformule a equação de regressão ou transforme as variáveis. Uma solução satisfatória nem sempre é possível a partir dos dados disponíveis.

Grau de Associação entre as VA's

Coeficiente de Correlação Linear de Pearson:

base → covariância normalizada

$$r = \frac{\text{cov}(x, y)}{s_x s_y} \Rightarrow -1 \leq r \leq 0 \text{ ou } 0 \leq \text{r} \leq +1$$

$$s_y = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n-1}} \qquad s_x = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$$

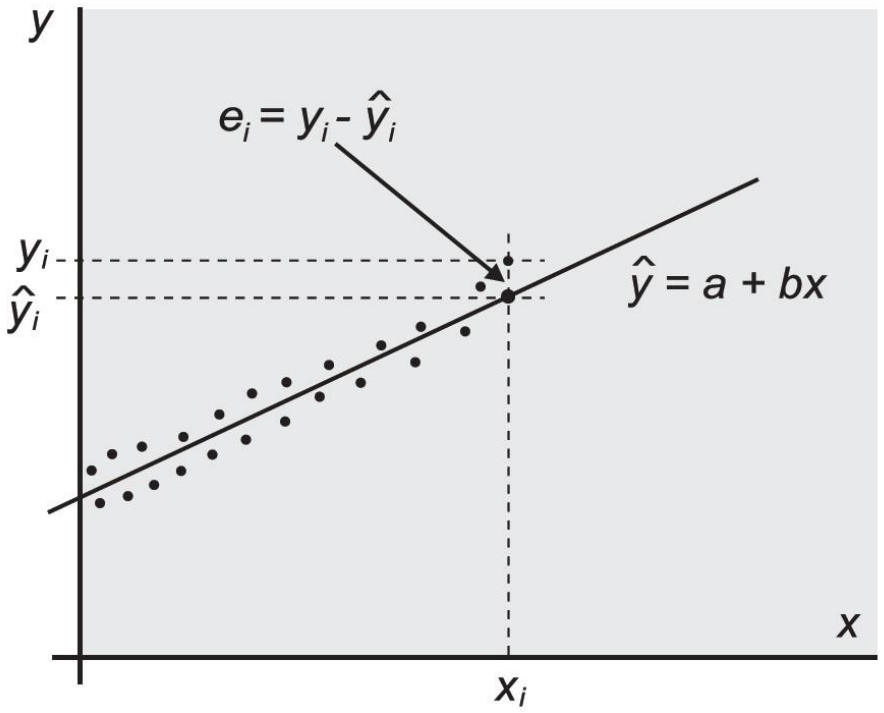
$$s_{xy} = \text{cov}(x, y) = \frac{\sum_{i=1}^{n} (x_i - \bar{x}).(y_i - \bar{y})}{n-1}$$

Estimação dos coeficientes de regressão

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

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

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



Estimação dos coeficientes no Excel. Função PROJ.LIN

Função Matricial: Ctrl + Shift + Enert

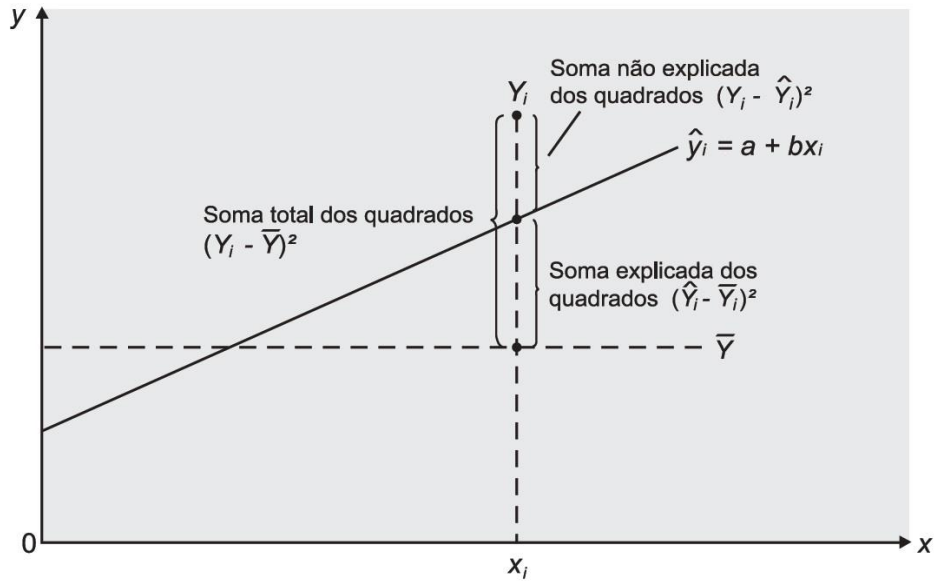
$y = mx + b$

	A	B	C	D	E	F
1	m_n	m_{n-1}	. . .	m_2	m_1	b
2	se_n	se_{n-1}	. . .	se_2	se_1	se_b
3	r_2	se_y				
4	F	d_f				
5	ss_{reg}	ss_{resid}				

Dados estatísticos	Descrição
se1.se2.....sem	Os valores de erro padrão para os coeficientes m1.m2.....mn.
seb	O valor de erro padrão para a constante b (seb = #N/D quando constante é FALSO).
r^2	O coeficiente de determinação. Compara os valores y estimados e reais e os intervalos no valor de 0 a 1. Se for 1, existe uma correlação perfeita no exemplo — não há diferença entre o valor y estimado e o valor y real. No outro extremo, se o coeficiente de determinação for 0, a equação de regressão não será útil na previsão de um valor y. Para obter informações sobre como os^2 são calculados, consulte "Comentários" mais adiante neste tópico.
Sey	O valor de erro para a estimativa de y.
F	A estatística F, ou o valor de F observado. Use a estatística F para determinar se a relação observada entre as variáveis dependentes e independentes ocorre por acaso.
Df	Os graus de liberdade. Use os graus de liberdade para ajudar a encontrar os valores F críticos em uma tabela estatística. Compare os valores encontrados na tabela com a estatística F retornada por PROJ.LIN de modo a determinar um nível de confiança para o modelo. Para obter informações sobre como df é calculado, consulte "Comentários", mais adiante neste tópico. O Exemplo 4 mostra o uso de F e df.
Ssreg	A soma dos quadrados da regressão.
Ssresid	A soma residual dos quadrados. Para obter informações sobre como ssreg e ssresid são calculados, consulte "Comentários" mais adiante neste tópico.

Coeficiente de Determinação

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



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

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

$$SQT = SQ\,Re\,s + SQ\,Re\,g$$

$$r^2 = \frac{Variância\quad Explicada}{Variância\quad Total} = \frac{SQ\,\mathrm{Re}\,g}{SQT} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \text{ ou}$$

$$r^2 = \frac{SQT - SQ\,\mathrm{Re}\,s}{SQT} = 1 - \frac{SQ\,\mathrm{Re}\,s}{SQT} = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

RLS:

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

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

Erro Padrão da Estimativa

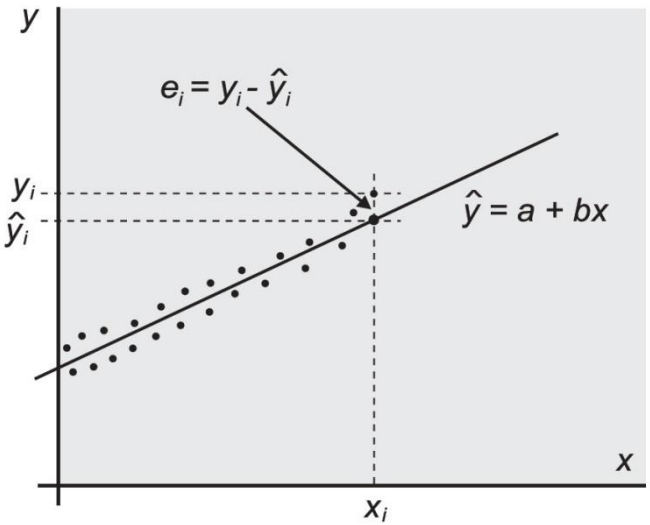
Variância dos Resíduos: $Var(e_i) = \sigma_e^2 = E(e_i^2) - E^2(e_i) = E(e_i^2)$

Ver capítulo 3, equação 3.21, página 75

$E(e_i) = 0$

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

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



Intervalos de Confiança para os Coeficientes da RLS

Variabilidade amostral → a reta de regressão estimada é uma das muitas retas possíveis.

Parâmetros *a* e *b* → estimadores pontuais dos parâmetros populacionais α e β.

$$a - t_{1-\frac{\alpha}{2},n-2}.s_a \leq \alpha \leq a + t_{1-\frac{\alpha}{2},n-2}.s_a \qquad s_a = \sqrt{s_e^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)}$$

$$b - t_{1-\frac{\alpha}{2},n-2}.s_b \leq \beta \leq b + t_{1-\frac{\alpha}{2},n-2}.s_b \qquad s_b = \sqrt{\frac{s_e^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

$$s_e^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2}$$

$t_{1-\frac{\alpha}{2},n-2}$ é valor da distribuição *t* de Student, para um nível de significância α e (n – 2) graus de liberdade

Avaliação da RLS

Linearidade → gráfico de dispersão e TH sobre o coeficiente angular β

Hipótese Nula: $H_0 : \beta = 0$ (não há relação linear)

Hipótese Alternativa: $H_1 : \beta \neq 0$ (há relação linear)

Estatística de Teste: $t = (b - \beta)/s_b$ ou, sob H_0, $t = b/s_b \sim t$ Student com $n-2$ gl

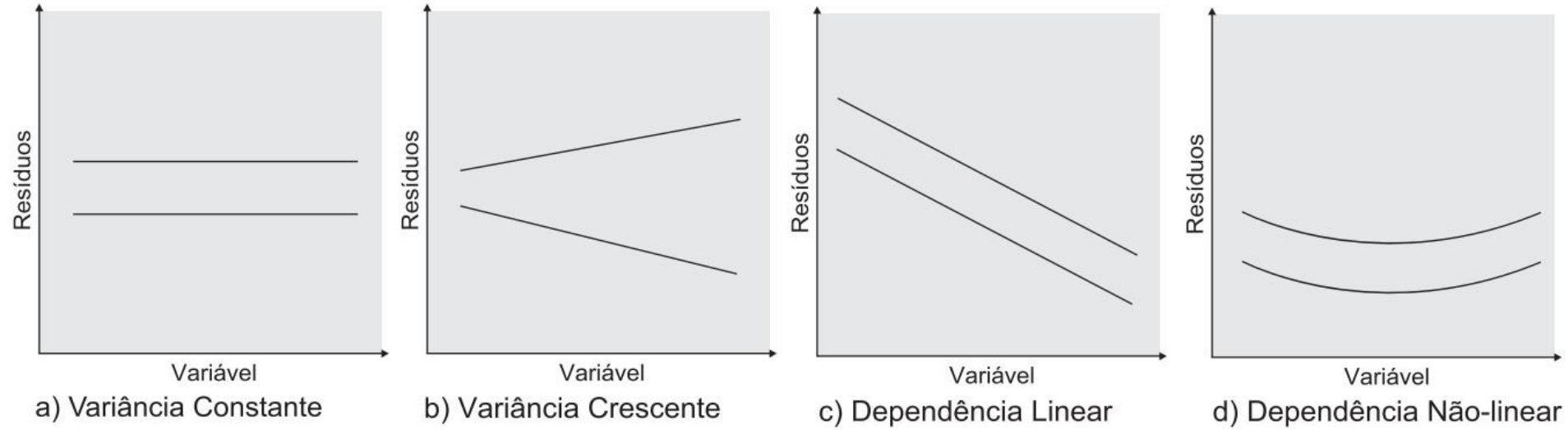
Decisão: rejeitar H_0 se $|t| > t_{1-\alpha/2,n-2}$

Avaliação da RLS

Normalidade dos Resíduos → testes de aderência, papel de probabilidade

Resíduos com Média Nula → OK pela estimação pelo MMQ

Resíduos Homocedásticos (Variância Constante) →



Posição de Plotagem

Critérios para estimativa das posições de plotagem Gumbel (1958):

- Deve ser tal que todas as observações possam ser plotadas, **evitando os valores 0 e 1**;
- A posição de plotagem deve estar compreendida entre $(i-1)/n$ e i/n, onde i denota a ordem de classificação (max→decrescente; min→crescente) de uma amostra de tamanho n;
- Para as séries anuais, o tempo de retorno de um valor maior ou igual à maior observação (ou menor ou igual à menor observação) deve convergir para n.
- As observações devem ser igualmente espaçadas na escala de frequências;
- A posição de plotagem deve ser intuitiva, analiticamente simples e fácil de usar.

Fórmula	Autor	Atributos de aplicação
$q_i = \frac{i}{n+1}$	Weibull	Probabilidades de excedência não enviesadas para todas as distribuições
$q_i = \frac{i-0{,}44}{n+0{,}12}$	Gringorten	Otimizada para os quantis das distribuições de Gumbel e GEV
$q_i = \frac{i-0{,}375}{n+0{,}25}$	Blom	Quantis não-enviesados para as distribuições Normal e Log-Normal
$q_i = \frac{i-0{,}5}{n}$	Hazen	Quantis da distribuição Gama de 3 parâmetros (PIII ou LPIII)
$q_i = \frac{i-0{,}40}{n+0{,}20}$	Cunnane	Quantis aproximadamente não enviesados para todas as distribuições

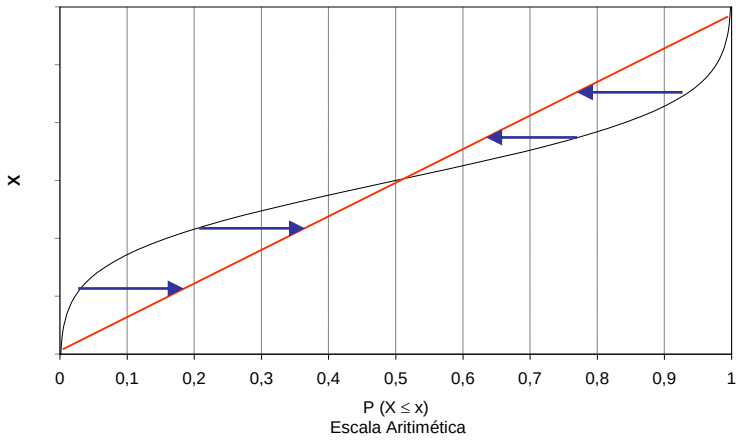
Cunnane (1978)

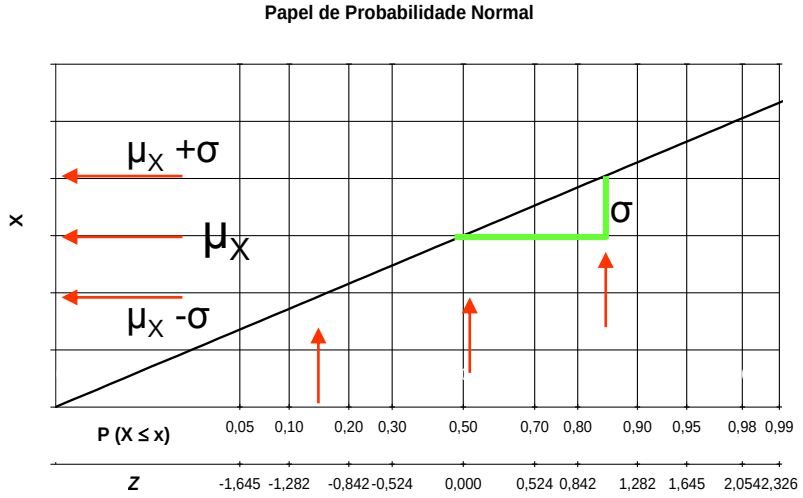
$$q_i = \frac{i-a}{n+1-2.a}$$

Construção de papéis de probabilidade

São gráficos para plotagem de observações amostrais e suas respectivas probabilidades empíricas, cujas escalas **são previamente transformadas** de modo a linearizar a relação entre $F_X(x)$ (ou $[1-F_X(x)]$ ou ainda T) e X.

Exemplo: papel Normal





Exercício Desafio

Um pesquisador da CPRM trabalhando com 50 pares de pontos (x, y) estabeleceu a seguinte regressão:

$$Y = 30 + 0{,}04X$$

Também obteve os seguintes resultados:

$$S_a = 12{,}5$$

$$S_b = 0{,}20$$

$$R^2 = 0{,}06$$

$$SQresiduos = 4000$$

Suponha que os valores de X mudaram de unidade, sendo que a nova variável independente é $X_{NU} = \gamma . X$. Onde γ é fator de conversão de unidade.

1) Deduza de forma analítica como estimar o novo intercepto e o novo coeficiente de X.

Considerando que $\gamma = 4$, responda:

2) Qual será o novo valor do coeficiente de X ?

3) Qual será o novo valor do intercepto ?

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, 21 de outubro de 2020