REGRESSÃO LINEAR

Introdução a Estatítica Aplicada a Climatologia
Programa de Pós Graduação em Geografia Física
Universidade de São Paulo

REGRESSÃO LINEAR

- "É o estudo da dependência entre duas variáveis" (WEISBERG, 2014)
  - "C*omo* alguns preditores influenciam uma resposta"
  - "Determina a relação entre a variável dependente (ou resposta, ou prevista, $\hat{y}$) e a variável independente (ou explicativa, ou previsora, x)"

REGRESSÃO LINEAR

- A correlação linear mostra *quanto* duas variáveis estão relacionadas
- A regressão linear pode mostrar ***como*** elas estão relacionadas
  - *Pode identificar uma relação de causa-efeito*
- Utilizada para realizar previsões

REGRESSÃO LINEAR

- O que se busca é a definição de uma reta que represente o ajuste entre as variáveis dependentes e a variável independente
- Permite a interpretação e análise dos efeitos de x sobre y
- Assume-se que há uma relação linear entre as variáveis

EXEMPLOS

Pode-se prever a chuva para amanhã, em mm, em função da pressão atmosférica

EXEMPLOS

Pode-se desejar prever o preço dos imóveis ou terrenos a partir da distância do imóvel ao centro da cidade

EXEMPLOS

Pode-se desejar prever o lucro que uma loja pode ter em relação ao valor gasto em propaganda

EXEMPLOS

Pode-se prever a vazão ou a chuva de uma região a partir de índices da temperatura da superfície dos mares

REGRESSÃO LINEAR

"Uma vez que especificamos *como* as variáves estão relacionadas, temos um modelo que pode ser visto como uma simplificação da realidade"

(ROGERSON, 2012, p. 201)

REGRESSÃO LINEAR SIMPLES E MÚLTIPLA

- Regressão Linear Simples: utiliza **uma** variável dependente e **uma** independente
- Regressão Linear Múltipla: utiliza **múltiplas** variáveis independentes e **uma** variável dependente

REGRESSÃO LINEAR MÚLTIPLA

Quando a regressão linear múltipla é aplicada,

é possível utilizar todas as variáveis independentes disponíveis ou selecionar as que melhor se correlacionam com a variável dependente através de métodos estatísticos adequados

As variáveis independentes podem ser selecionadas, por exemplo, a partir da correlação linear com a variável dependente, sendo selecionadas as que melhor explicam a variabilidade da variável dependente

- Isso é feito por um processo iterativo (passo-a-passo)
- Método “Forward Stepwise Selection” - mais indicado

PRINCIPAIS MÉTODOS PARA REALIZAR A REGRESSÃO LINEAR MÚLTIPLA

Seleção para frente - Forward Stepwise selection

Constitui em começar o modelo ***sem*** variáveis independentes, testando, passo a passo, a adição de uma nova variável com o uso de critérios de comparação para sua escolha (por exemplo, o teste t ou F), adicionando a variável que mais melhora o modelo e repetindo este procedimento até não conseguir mais aumentar significativamente a acurácia do modelo

PRINCIPAIS MÉTODOS PARA REALIZAR A REGRESSÃO LINEAR MÚLTIPLA

Eliminação para trás - Backward Stepwise elimination

inicia-se a elaboração do modelo ***com todas as variáveis independentes***, testando a eliminação de cada uma delas, usando um critério de comparação de escolha, eliminando as variávies que menos melhoram o modelo, e, repetindo este procedimento até não ter mais melhoria no modelo

Eliminação Bidirectional - uma combinação dos anteriores

EXEMPLOS DE REGRESSÃO LINEAR MÚLTIPLA

Pode-se prever a chuva para amanhã, em mm, em função da pressão atmosférica, da temperatura do ar e da quantidade de chuva de hoje

EXEMPLOS DE REGRESSÃO LINEAR MÚLTIPLA

Pode-se estimar o preço de imóveis ou terrenos a partir da distância ao centro da cidade, do tamanho do lote e da distância a supermercados

EXEMPLOS DE REGRESSÃO LINEAR MÚLTIPLA

Pode-se estimar o lucro de uma loja pelo valor gasto em propaganda, pela localização e pela densidade populacional do bairro

EXEMPLOS DE REGRESSÃO LINEAR MÚLTIPLA

Pode-se estimar a vazão ou a chuva de uma região a partir de índices de temperatura da superfície dos mares (SOI, PDO, AMO, NTA, CAR)

REGRESSÃO LINEAR

- Resulta em uma equação linear pela qual é possível estimar os valores de y a partir de x

(ROGERSON, 2012, p. 201)

$$\hat{y} = a + bx$$

onde:

$\hat{y}$ é o valor estimado da variável dependente

y é o valor observado da variável dependente

x é o valor observado da variável independente

a é o intercepto (valor de y quando x=0)

b é a inclinação da reta (ou seja, "a alteração esperada na variável dependente provocada pela variação de uma unidade na variável independente"

RETA DE REGRESSÃO

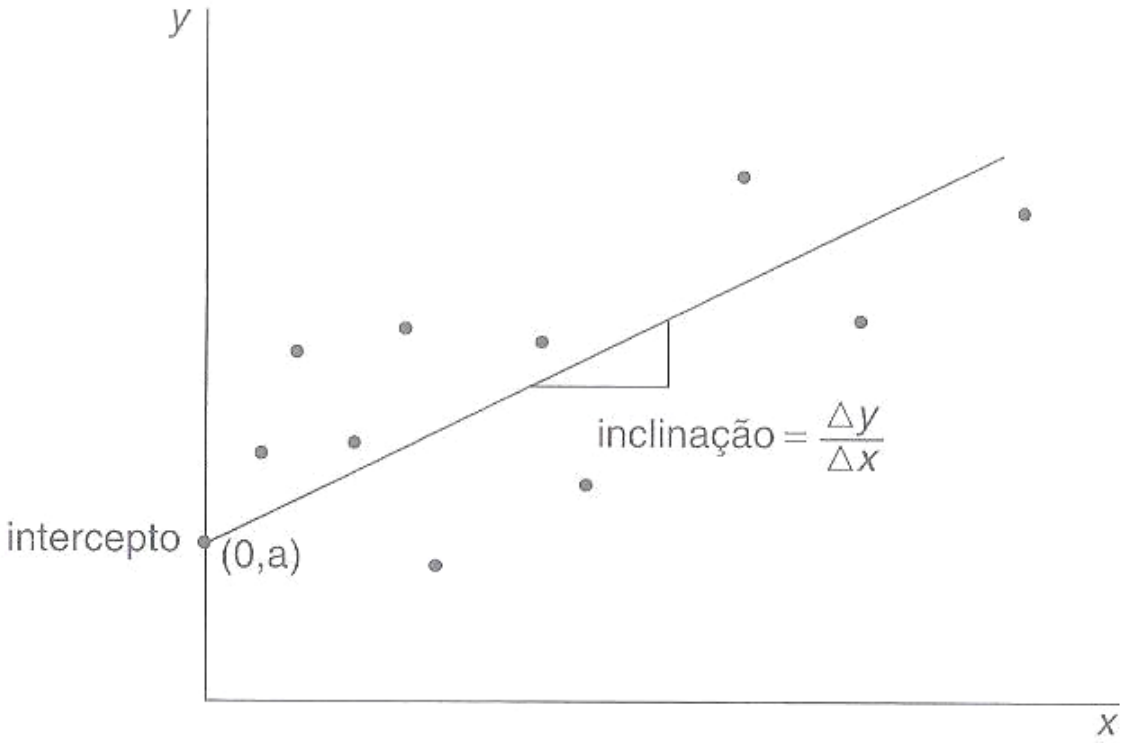
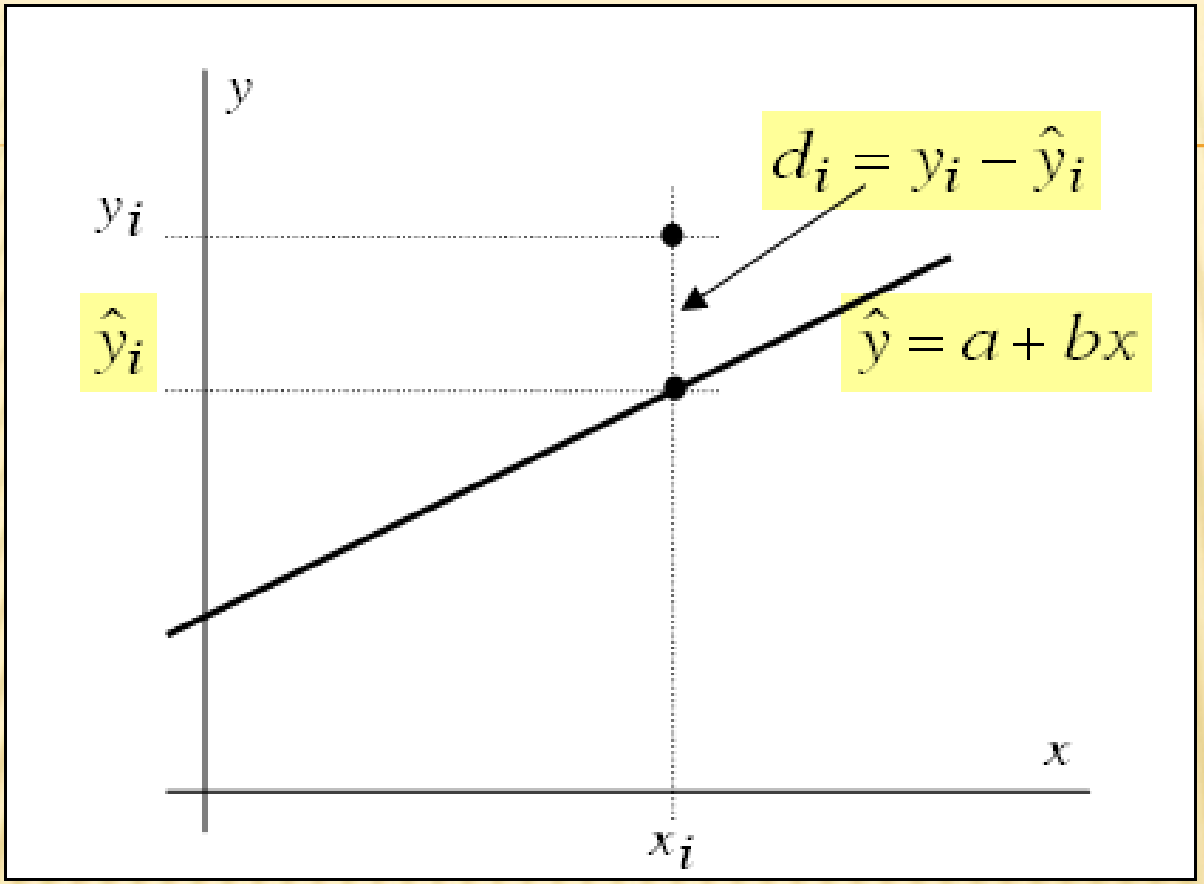


FIGURA 8.1 **Reta de regressão através de um conjunto de pontos.**

Fonte: ROGERSON, 2012, p. 202.



d_i = erro

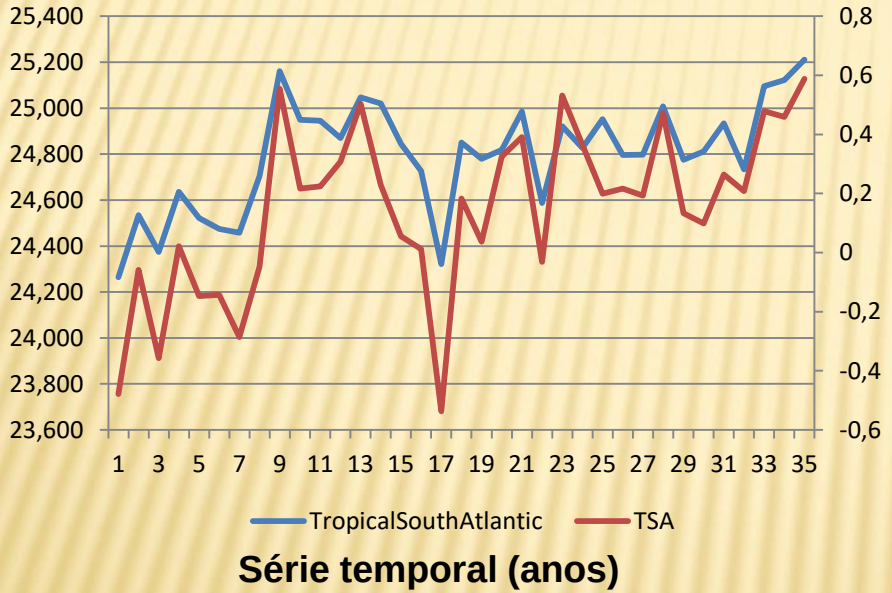
y_i = valor observado

$\hat{y}_i$ = valor estimado

Quanto menor for a soma dos desvios, mais ajustada é a reta aos valores de y e maior é a explicação da variável dependente pela variável independente

DIAGRAMA DE DISPERSÃO (SCATTERPLOT)

- Como o objetivo é ver como os valores de y variam com a variação do x, o scatterplot é uma ótima ferramenta.



Série temporal (anos)

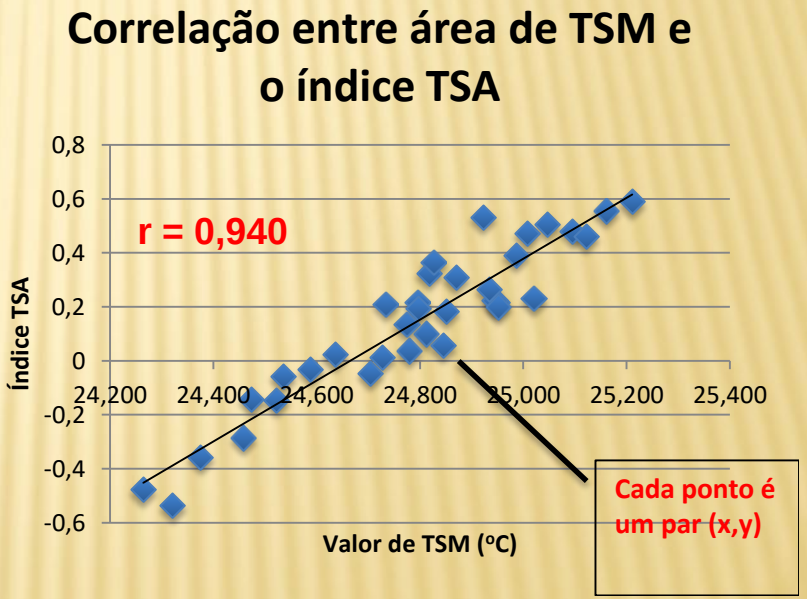
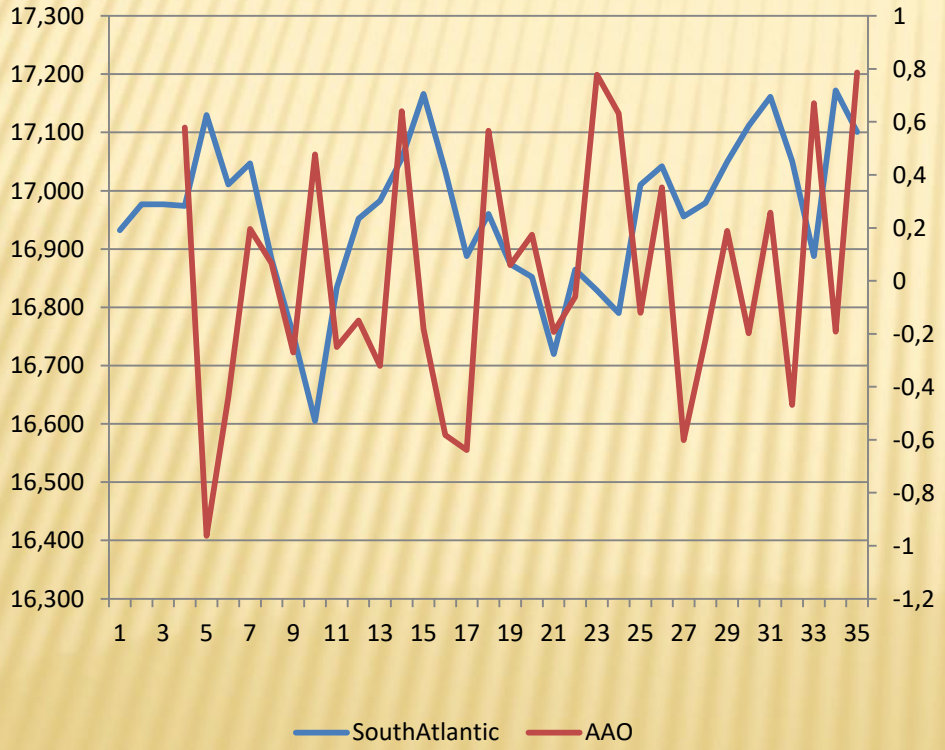
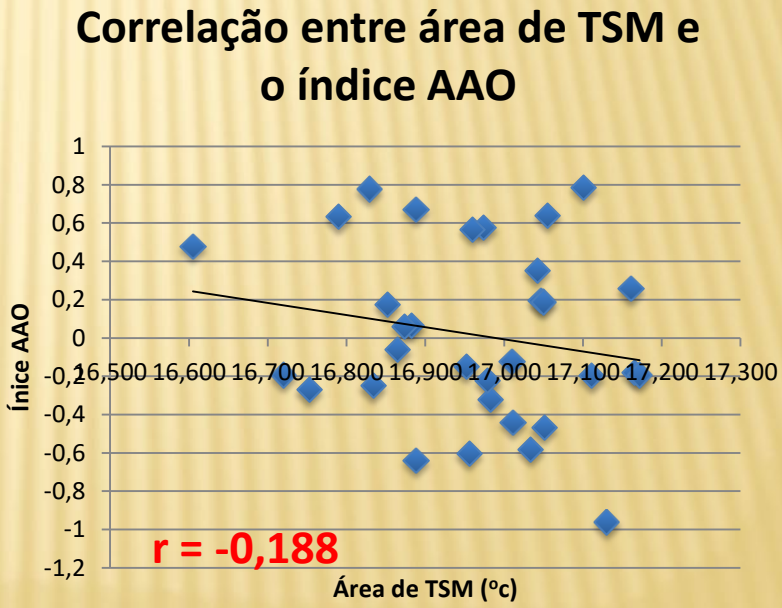


DIAGRAMA DE DISPERSÃO (SCATTERPLOT)

Pode-se verificar que a diferença média entre y e ŷ é maior que no slide anterior





- Como calcular a e b se temos apenas x e y?

Método dos Mínimos Quadrados

$$\Sigma(d_i^{\,2}) = \Sigma\,(y_i - \hat{y}_i)^{\,2} \longrightarrow \text{mínimo}$$

- O desvio ou resíduo em cada tempo i é representado por:

$$d_i = y_i - \hat{y}_i$$

- Substituindo a equação da reta de regressão ($\hat{y}_i = a + bx_i$) na equação acima, tem-se:

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

Ou

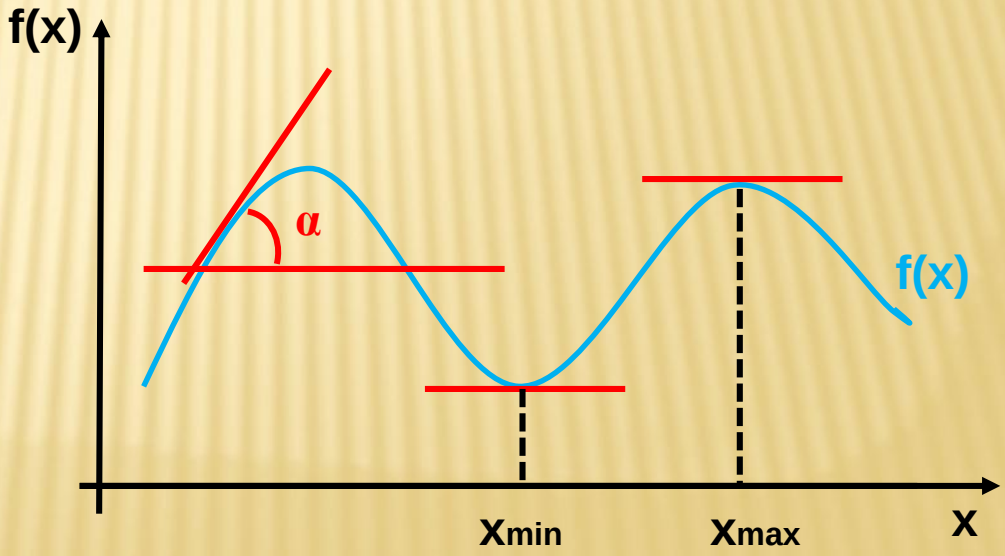
$$d_i = y_i - a - bx_i$$

- Do cálculo diferencial infinitesimal, sabe-se que o mínimo ou máximo de uma função f(x) é encontrado quando igualamos a derivada desta função, em relação a x, a zero.

$$\text{mín } (f(x)) = \frac{df(x)}{dx} = 0$$

ou

$$\text{máx } (f(x)) = \frac{df(x)}{dx} = 0$$



- Voltando ao desvio mínimo quadrático da estimativa de y

  Considerando que f(a,b) = $\sum d_i^{\,2} = \sum (y_i - \hat{y}_i)^2$

- Igualando a derivada de f(a,b) em relação a *a* e *b* a zero, pode-se encontrar *a* e *b*

Lembrando que f(a,b) = $\sum d_i^{\,2} = \sum (y_i - \hat{y}_i)^{\,2}$

- $\frac{df(a,b)}{da} = 0$

$\frac{df(a,b)}{db} = 0$

e substituindo $\hat{y}_i$ em *f(a,b)*, tem-se:

$$f(a,b) = \sum (y_i - (a + bx_i))^2$$
$$f(a,b) = \sum (y_i - a - bx_i)^2$$

- A derivada de f(a,b) em relação a a é escrita como:

$$\frac{\partial f(a,b)}{\partial a} = 2 \bullet \sum(y_i - a - bx_i) * (-1)$$

- A derivada de f(a,b) em relação a b é escrita como:

$$\frac{\partial f(a,b)}{\partial b} = 2 \bullet \sum(y_i - a - bx_i) * (-x)$$

- Igualando cada uma das equações a zero, tem-se:

$$\begin{cases} -2 \cdot \Sigma(y_i - a - bx_i) = 0 \\ -2 \cdot \Sigma\, x_i (y_i - a - bx_i) = 0 \end{cases}$$

Tem-se, pois, um sistema de duas equações com duas incógnitas a e b, que pode ser resolvido.

Assim,

$$\begin{cases} a = \bar{y} - b\bar{x} \\ b = \dfrac{\sum_i^n x_i \left(y_i - \bar{y}\right)}{\sum_i^n x_i \left(x_i - \bar{x}\right)} \end{cases}$$

ou

$$\begin{cases} a = \bar{y} - b\bar{x} \\ b = \dfrac{\sigma_{xy}}{\sigma_y^2} = r\dfrac{\sigma_x}{\sigma_y} \end{cases}$$

EXEMPLO DE REGRESSÃO LINEAR SIMPLES

- Um supermercado está interessado em saber como os níveis de renda (x) podem afetar a quantidade de dinheiro gasto por semana por seus clientes (y)

Quantia gasta por semana (R$) (y)	Renda mensal x 100 (R$) (x)
$120	65
$68	35
$35	30
$60	44
$100	80
$91	77
$44	32
$71	39
$89	44
$113	77

- Para determinar a reta de regressão, deve-se calcular as seguintes variáveis:

$\overline{x} = 52,3$ $\quad s_x = 20,20$

$\overline{y} = 79,1$ $\quad s_y = 28,34$

r = ?? (slide 31)

- $\sum_{n}^{i=1} (x_i - \overline{x})\,(y_i - \overline{y}) = 4301,7$

$$r = \sum_{n}^{i=1} (x_i - \overline{x})\,(y_i - \overline{y}) \,/\, (n-1)\, s_x\, s_y =$$

$$= \frac{4301,7}{9 \cdot (20,20) \cdot (28,34)} = \mathbf{0,835}$$

- Para encontrar a inclinação da reta b:

$$b = r \frac{s_x}{s_y} = 0{,}835 \times \frac{28{,}34}{20{,}20} = 1{,}171$$

- Para encontrar o intercepto a:

$$a = \overline{y} - b\,\overline{x} = 79{,}1 - 1{,}171 \times (52{,}3) = 17{,}8$$

Assim, a reta de regressão linear pode ser expressa como:

$$\hat{y} = R\$\ 17{,}8 + 1{,}171x$$

- Como calcular o valor estimado para cada “cliente pesquisado”?

- Vamos pegar o exemplo do cliente número 1, que tem renda de R$ 65

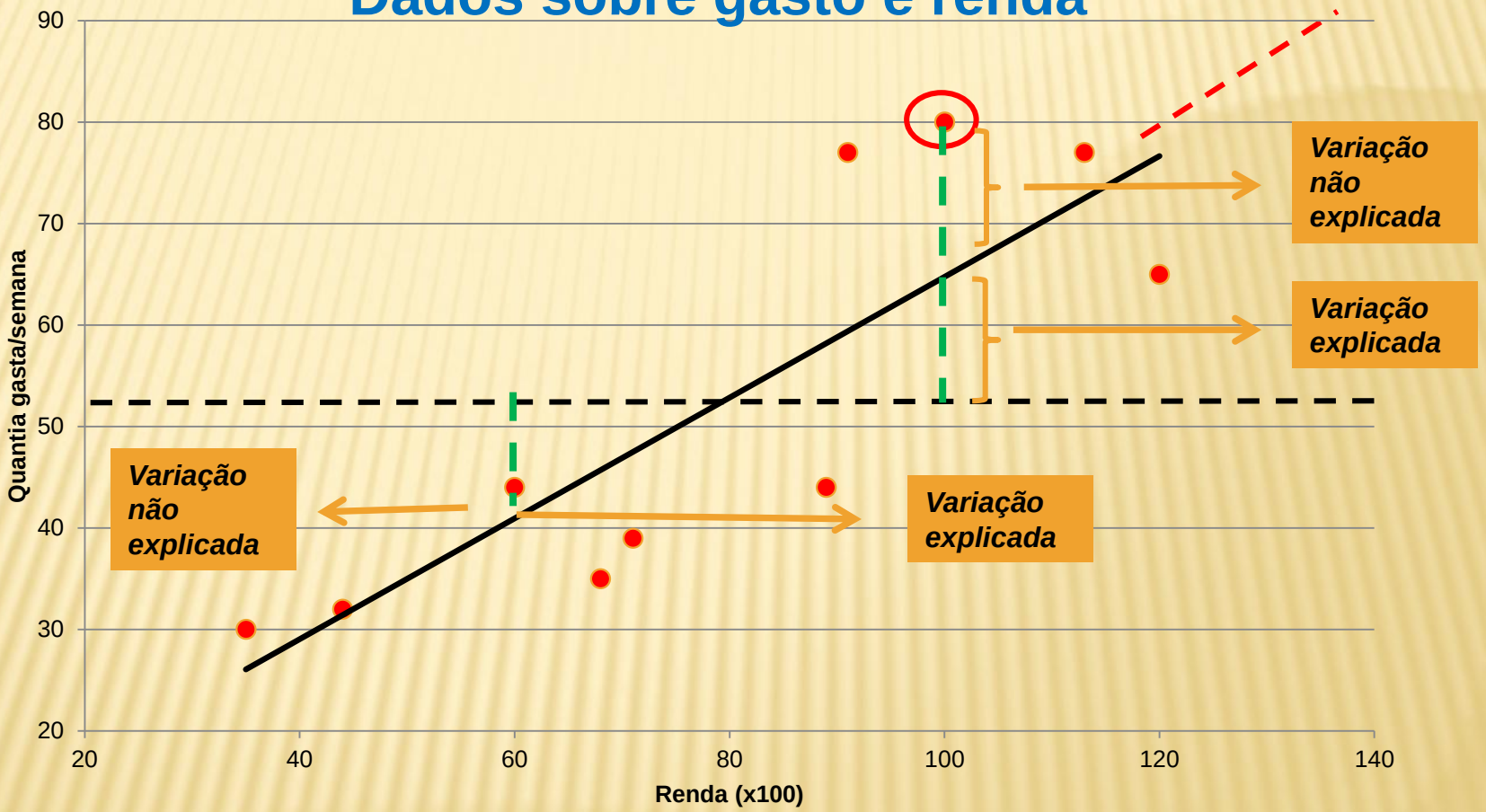
$$\hat{y}_1 = \text{R\$ } 17{,}8 + (1{,}171) \times (\text{R\$ } 65) = \text{R\$ } 93{,}90$$

- O desvio para esse caso é:

$$(y_i - \hat{y}_i) = 120 - 93 = 26{,}1$$

  - obs est

Dados sobre gasto e renda

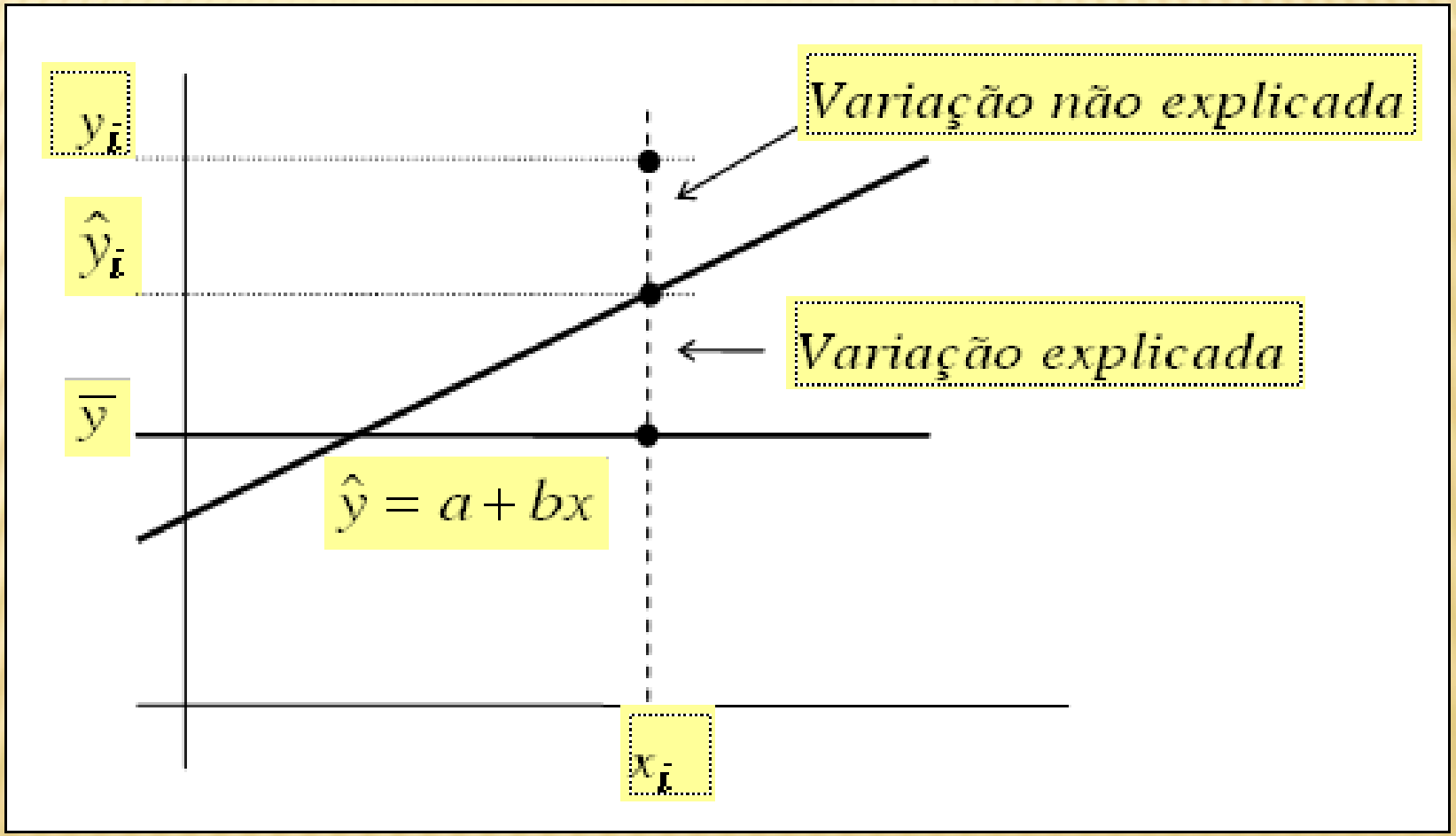


Média da quantia gasta: R$ 52,30

Clientes e valores preditos associados com os dados de gasto e renda

Valor gasto/semana (y)	Renda mensal x 100 (x)	Predito $\hat{y}$	Resíduo e
120	65	93,9	26,1
68	35	58,8	9,2
35	30	52,9	-17,9
60	44	69,3	-9,3
100	80	111,5	-111,5
91	77	108,0	-17
44	32	55,3	-11,3
71	39	63,5	7,5
89	44	69,3	19,7
113	77	108,0	5

QUÃO BOM É O AJUSTE?



- ***Variação total*** é o resultado da soma dos quadrados dos desvios dos valores de *y observado* em relação à média de y:

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

- ***Variação explicada*** pela variável independente é o resultado da soma dos quadrados dos desvios dos valores estimados em relação à média:

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

- ***Variação não-explicada*** é o resultado da soma dos quadrados dos desvios de *y observado* em relação aos valores estimados:

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

- Pode-se demonstrar que:

$$SST = SSR + SSE$$

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

variação total = variação explicada + variação não explicada

COEFICIENTE DE DETERMINAÇÃO R^2

R^2: coeficiente de determinação do modelo [0,1]

$$r^2 = \frac{\text{Variação explicada pela variável independente}}{\text{Variação total da variável dependente}} = \frac{SSR}{SST}$$

ou

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

COEFICIENTE DE DETERMINAÇÃO R^2

- R^2 : é interpretado como a fração da variabilidade da variável dependente explicada pela variável independente utilizada no modelo

  Quanto mais próximo de 1, melhor é o modelo

- O problema quanto à utilização do R^2 é que ele não leva em consideração o número de variáveis utilizadas no modelo

R^2 AJUSTADO

- O coeficiente de determinação ajustado (r^2*ajustado*) é uma medida utilizada em regressão linear múltipla
- Importante quando o n é relativamente pequeno
- Partindo da regressão linear simples, com uma única variável independente, o significado do coeficiente de determinação é a porcentagem de explicação dessa regressão

R2 AJUSTADO

- Ao adicionar uma ou mais variáveis independentes na regressão, demonstra-se que o r^2 não deverá diminuir, devendo aumentar em alguns casos. O r^2 ajustado tenta compensar o aumento natural de explicação provocado pelo aumento do número de variáveis independentes e o tamanho da amostra, sendo calculado com a expressão:

$$R_a^2 = 1 - \left[\frac{(n-1)}{n-(p+1)}\right](1-R^2)$$

n: tamanho da amostra

p: quantidade de variáveis explicativas usadas no modelo

- Outra forma de determinar se a regressão foi bem sucedida em explicar uma parcela significativa da variância de y é testar a hipótese nula de que a proporção da variabilidade de y explicada por x é igual a zero. Esse trabalho é realizado com um **teste F**, análogo ao teste F utilizado na análise de variância. Para a regressão simples, em específico:

$$F_0 = \frac{variância\ explicada}{variância\ não\ explicada}$$

$$F_0 = \frac{\frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{(k-1)}}{\frac{\sum_{i=1}^{n}(\ y_i - \hat{y}_i)^2}{(n-k)}}$$

k = número de variáveis independentes

- A estatística F é o quadrado da estatística t. Os testes são idênticos no sentido de que sempre produzem conclusões e valores de *p* idênticos

  (ROGERSON, 2012, p. 210)

- Estatística F (de Fisher ou Fisher-Snedecor): é a razão entre o modelo e seu erro (razão entre a variância explicada e a não explicada). Quanto maior, melhor.

REGRESSÃO LINEAR NO GRADS

- Existem as funções

tregr → regressão temporal

sgregr → regressão espacial

ltrend → tendência linear

EXERCÍCIO REGRESSÃO LINEAR - GRADS

1) Calcule a tendência linear da precipitação mensal na América do Sul para o período de **1950 a 2015**, com os dados do CRU.

REGRESSÃO LINEAR DA PRECIPITAÇÃO NO GRADS

```
'c'
'reinit'
'set display color white'
'c''set grads off‘

'sdfopen cru_ts3.20.1901.2011.pre.dat.nc'
'set y 1
''set z 1'
'set t 601 1332'
'define AS = aave(pre, lon=-90, lon=-30, lat=-60, lat=20)'
'set lon -90 -30''set lat -60 20‘
'set z 1‘
'set t 601'
'set gxout shaded‘

'set clevs 0 5 7.5 10 12.5 15 17.5 20 22.5'
'set ccols 49 47 45 42 41 23 24 25 27 29‘

'set rgb 49  20 100 210‘
'set rgb 47  40 130 240'
'set rgb 45  80 165 245‘
'set rgb 42 180 240 250'
'set rgb 41 225 255 255‘
'set rgb 23 255 192  60'
'set rgb 24 255 160   0‘
```

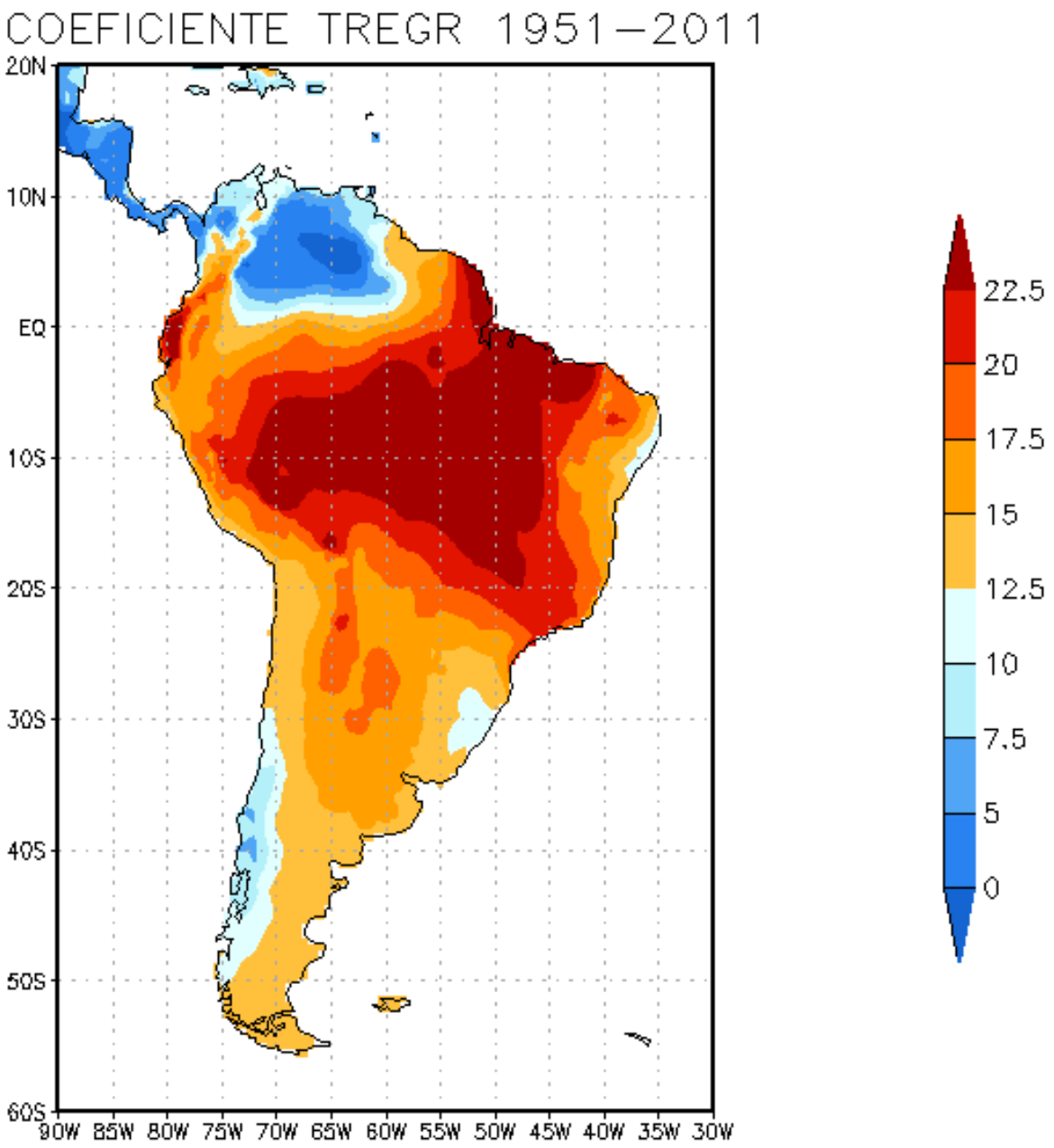
```
'set rgb 25 255  96   0'
'set rgb 27 225  20   0‘
'set rgb 29 165   0   0‘

*'d tregr(AS, pre, t=601, t=1332)*10‘

'define coeff = tregr(AS, pre, t=601, t=1332)'
'define preave = ave(AS, t=601, t=1332)'
'define ASave = ave(AS, t=601, t=1332)'
'd (coeff * (AS - ASave) + preave)/10'

'set gxout bar'
'cbarn'
'set strsiz .20'
'set string 1 c 5 0'

'draw string 5.5 8 COEFICIENTE TREGR 1951-2011'
'printim tregr-shaded.png'
```


COEFICIENTE TREGR 1951–2011
20N
10N
EQ
10S
20S
30S
40S
50S
60S
90W 85W 80W 75W 70W 65W 60W 55W 50W 45W 40W 35W 30W
22.5
20
17.5
15
12.5
10
7.5
5
0

- A **tendência linear** no tempo pode ser calculada a partir da regressão linear simples da observação em relação ao tempo, e pode ser matematicamente expressa como:

$$\hat{y}(t) = b + a * t$$

ŷ: valor estimado da observação y(t) (coordenada y do diagrama cartesiano)

b: intercepto (valor de y quando a reta ajustada cruza o eixo das ordenadas, em t=0)

a: coeficiente angular (inclinação da reta, tendência linear)

t: contador temporal de cada observação

- A tendência pode ser expressa na unidade da variável em um determinado intervalo de tempo

mm de chuva/10 anos

- O coeficiente angular indica a variação da variável no espaço de tempo unitário ($\Delta\hat{y}/\Delta t$)

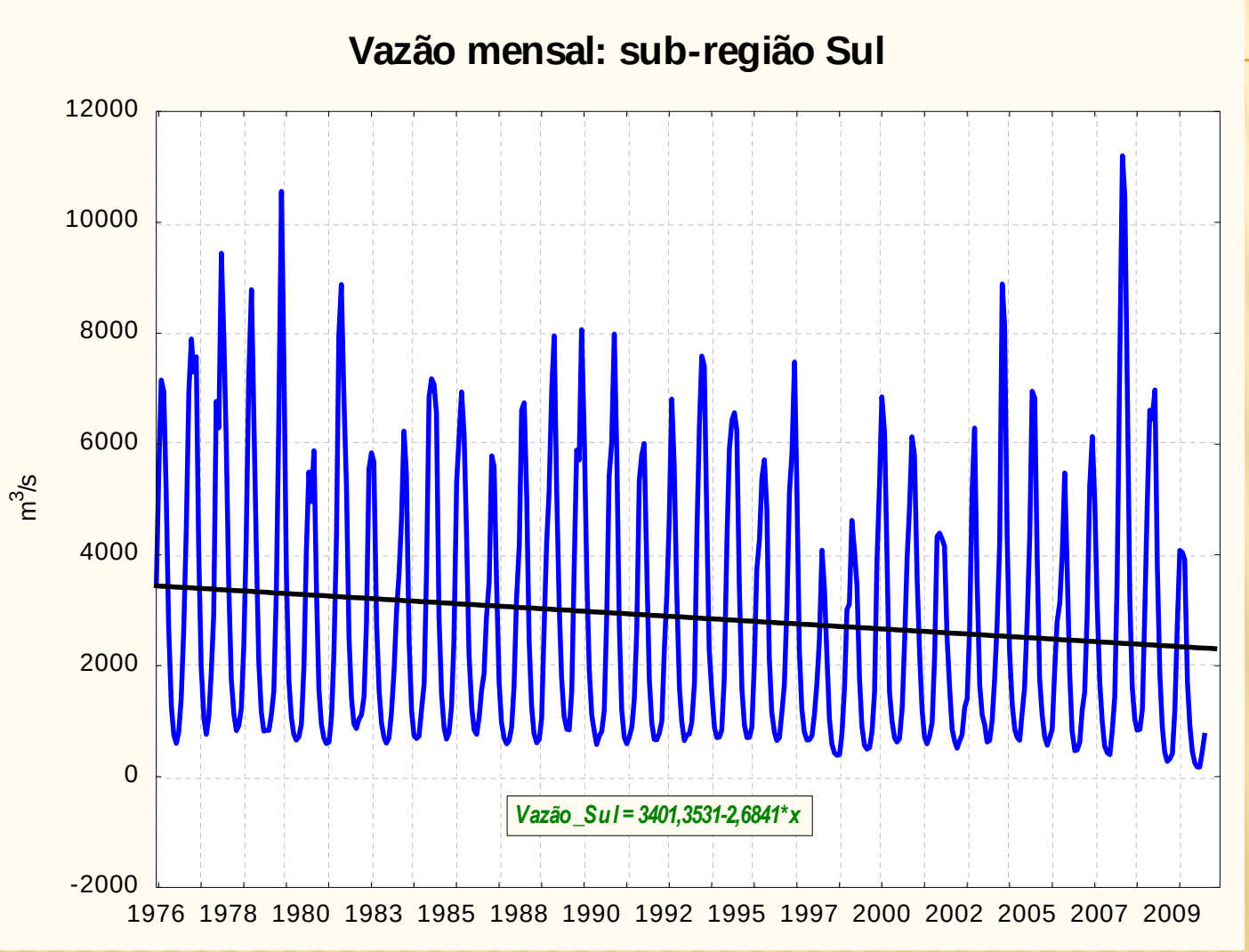
- Pode ser também expressa em porcentagem, em relação ao valor climatológico mensal ou anual, como expresso a seguir:

tendência (%) = (a * Δt / média climatológica) * 100%

a expressa o coeficiente angular ($\Delta\hat{y}/\Delta t$)

Δt, a quantidade de observações

TENDÊNCIA POR REGRESSÃO LINEAR



***Δy/Δt* =**
- 2,6841 $m^3\ s^{-1}$

vazão média =
2836 $m^3\ s^{-1}$

***tendência (%)* =** *(a * Δt / média climatológica) * 100%*
tendência (%) = (2,6481) *(420/2836)*100%
tendência (%) = -39,75%

TENDÊNCIA LINEAR NO GRADS (LTREND)

script constitui um arquivo texto com um nome qualquer;
em geral é um conjunto de comandos que podem ser digitados diretamente no terminal do GrADS.

```
'c'
'reinit'
'set display color white'
'c'
'set grads off'

'sdfopen cru_ts3.20.1901.2011.pre.dat.nc'
'set t 601 1332'
'set lon -100 -10'
'set lat -60 23'

'set gxout shaded'
'set clevs  -100 -75 -50 -25 0 25 50 75 100'
'set ccols 49 47 45 42 41 23 24 25 27 29'

'set rgb 49  20 100 210'
'set rgb 47  40 130 240'
'set rgb 45  80 165 245'
```

- 'set rgb 42 180 240 250'
- 'set rgb 41 225 255 255'
- 'set rgb 21 255 250 170'
- 'set rgb 22 255 232 120'
- 'set rgb 23 255 192 60'
- 'set rgb 24 255 160   0'
- 'set rgb 25 255 96   0'
- 'set rgb 27 225 20   0'
- 'set rgb 29 165   0   0'
- 'ltrend pre precip1 s e'

- 'set t 1201'
- 'set xlopts 1 5 .20'
- 'set ylopts 1 5 .20'
- 'set xaxis -100 -10 20'
- 'set yaxis -50 20 20'
- 'd s*10*731' * mudar para mm; 731= 61*12 (anos*meses)
- 'd precip''set gxout bar'
- 'cbarn'
- 'set strsiz .20'
- 'set string 1 c 5 0'
- 'draw string 5.5 8 TENDENCIA 1951-2011'
- 'printim testei.png x1000 y800'

Tendência linear da Precipitação mensal – Itrend (mm/60 anos)