Regressão Linear

PNV-3421 – Processos Estocásticos

Prof. Dr. João Ferreira Netto

Bibliografia Principal

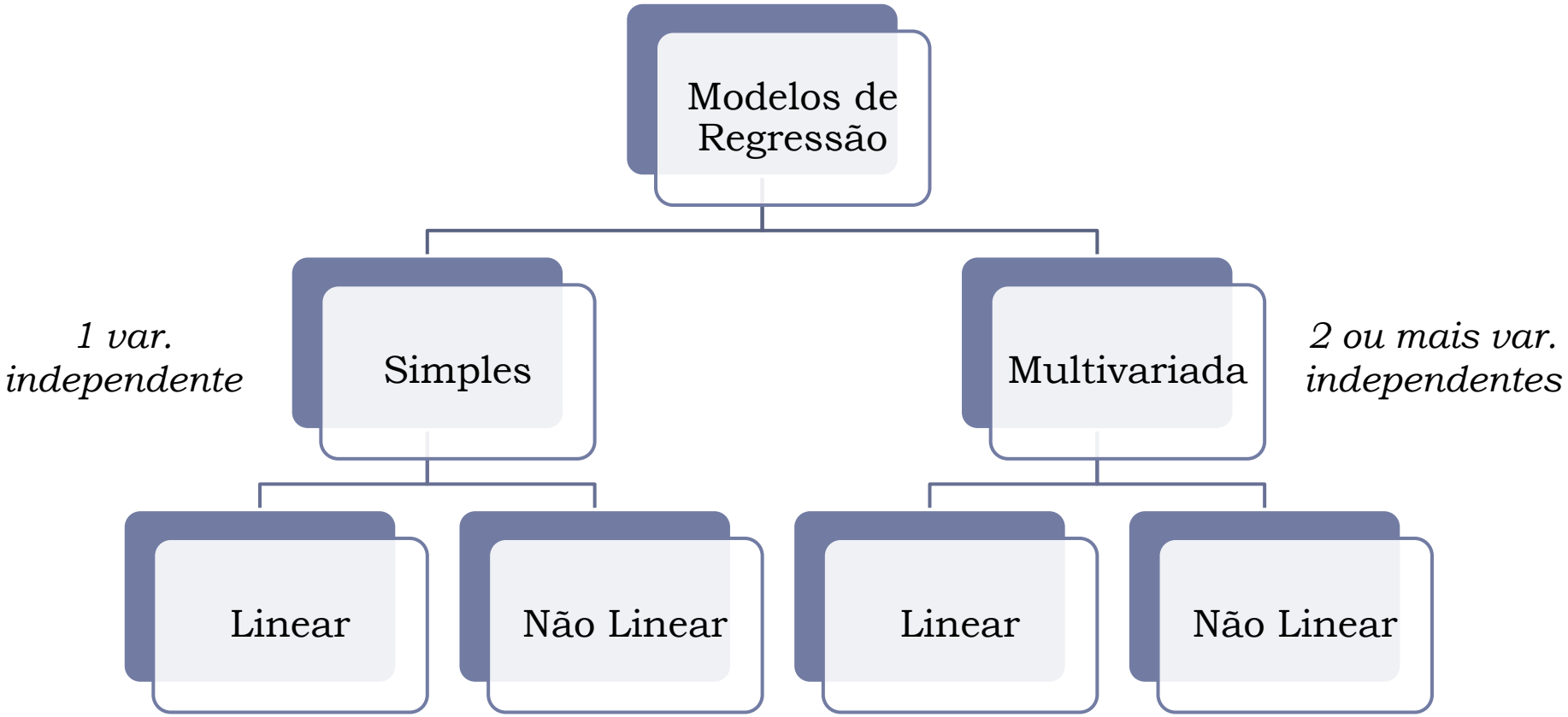
- Hanke, J.E & Reitsch A.G. (1998) Business Forecasting. 6th Edition, Prentice Hall, Upper Sadle River, NJ.

Modelos de Regressão

- Expressar a relação entre uma variável dependente e variáveis explanatórias (independentes), por meio de uma equação, com o objetivo de previsão.
- As variáveis podem ser numéricas ou indicativas de uma categoria.

Variáveis explicativas

Modelos de Regressão



Variáveis Correlacionadas

- Covariância entre duas variáveis aleatórias é uma medida da variabilidade conjunta.

$$\operatorname{cov}(x,y) = \frac{\sum_{i=1}^{n}(x_i - \bar{X})(y_i - \bar{Y})}{n-1}$$

n - dados na amostra

- $\operatorname{cov}(x,y) > 0 \rightarrow x$ e y são correlacionadas positivamente
- $\operatorname{cov}(x,y) < 0 \rightarrow x$ e y são correlacionadas negativamente
- $\operatorname{cov}(x,y) = 0 \rightarrow x$ e y são independentes

Variáveis Correlacionadas

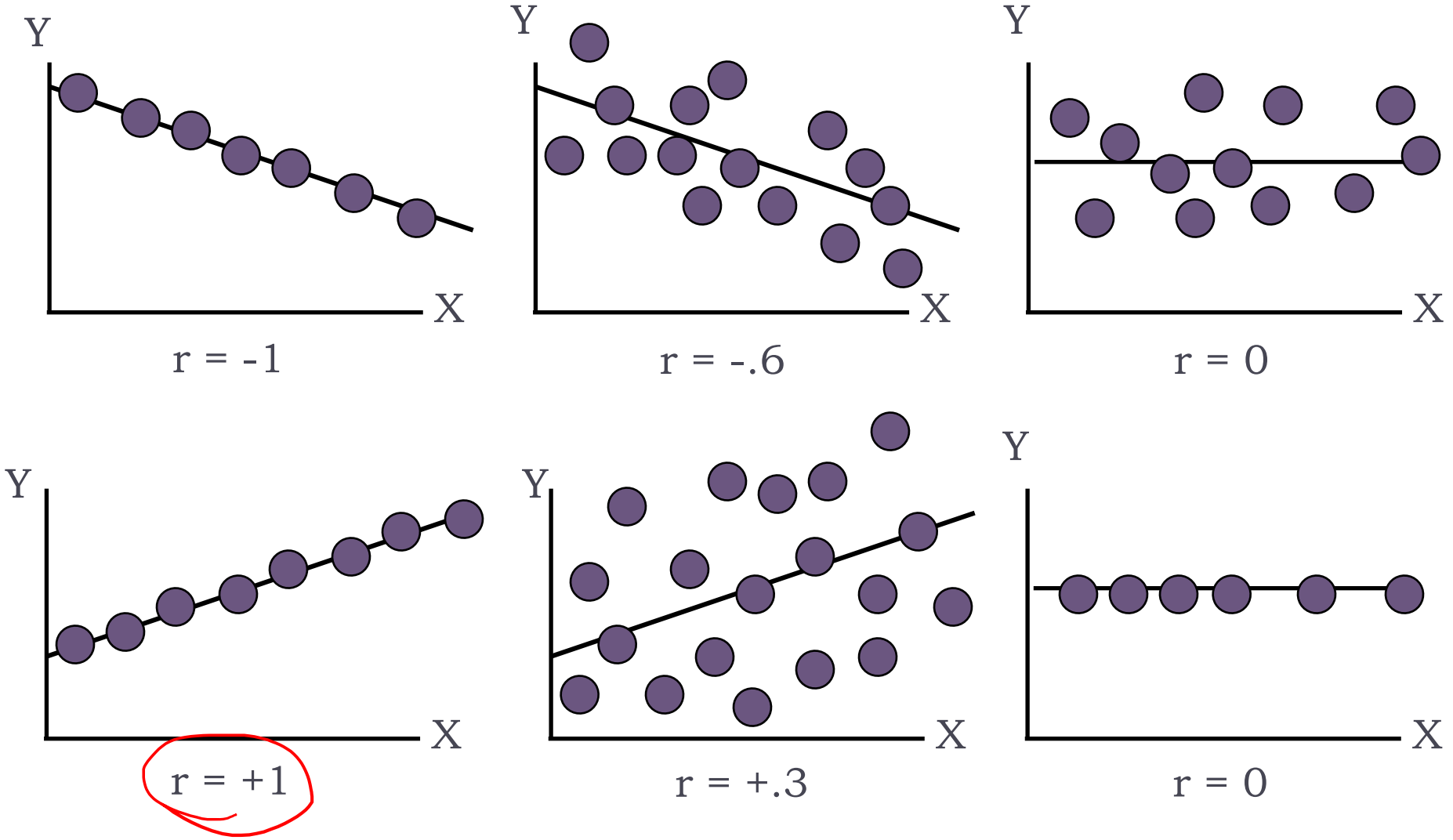
- O índice de correlação de Pearson é uma medida normalizada de covariância.

$$r = \frac{\operatorname{cov}(x, y)}{\sqrt{\operatorname{var} x}\sqrt{\operatorname{var} y}}$$

Adimensional

- Varia entre -1 e +1.
- Quanto mais próximo de -1, maior a relação linear negativa.
- Quanto mais próximo de +1, maior a relação linear positiva.
- Quanto mais próximo de 0, maior a ausência de relação linear.

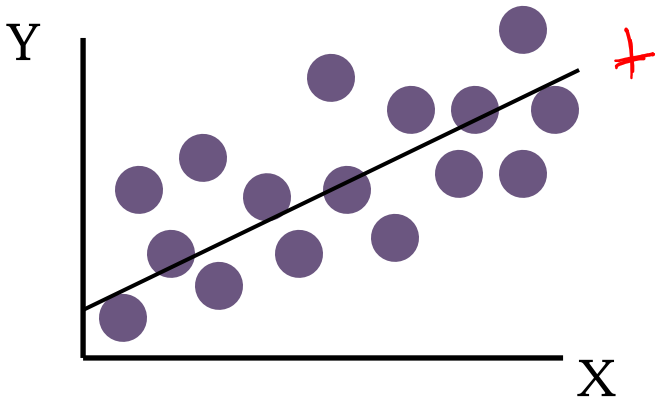
Variáveis Correlacionadas

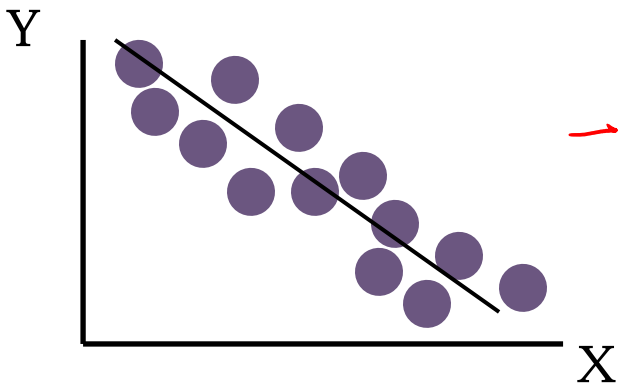


Fonte: Statistics for Managers Using Microsoft® Excel 4th Edition, 2004 Prentice-Hall

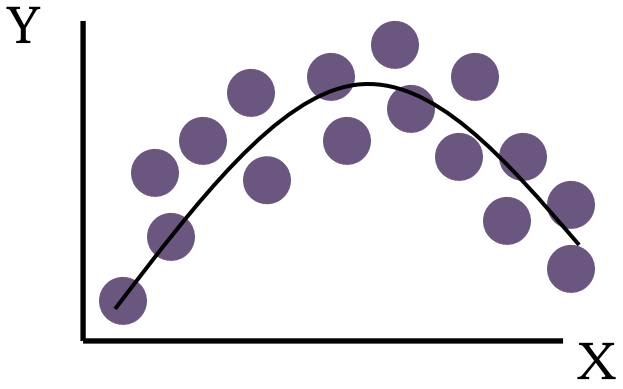
Variáveis Correlacionadas

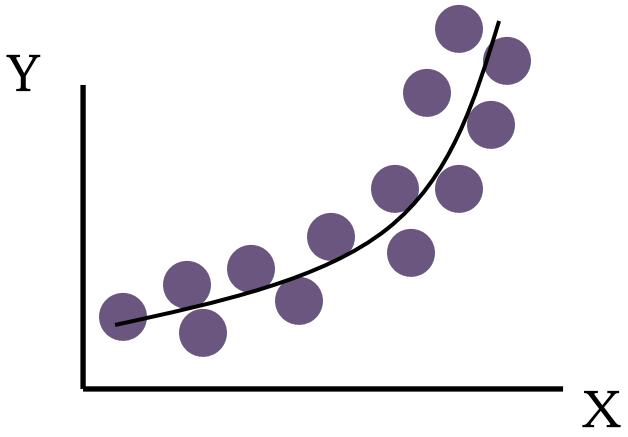
Relação Linear





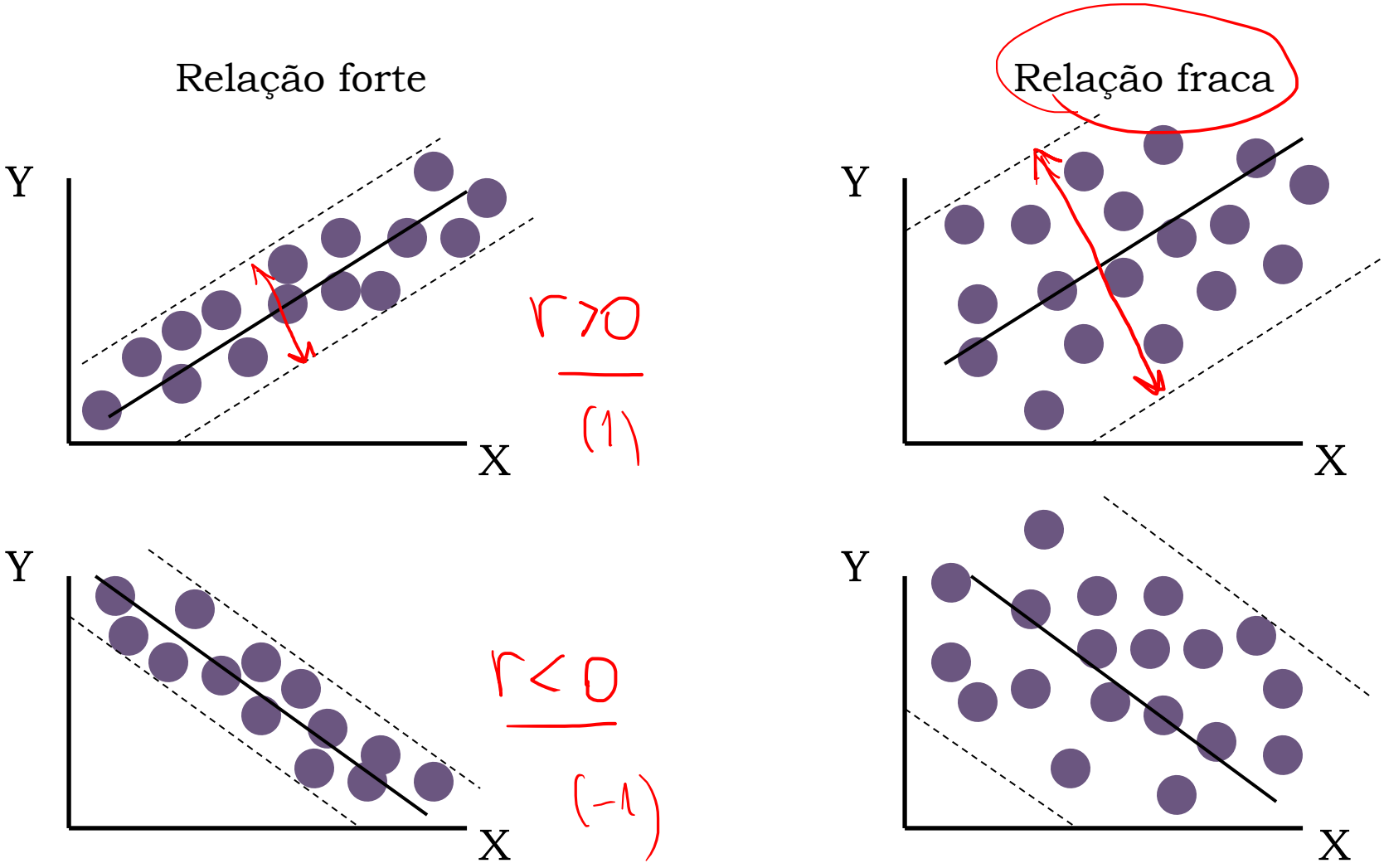
Relação não-linear





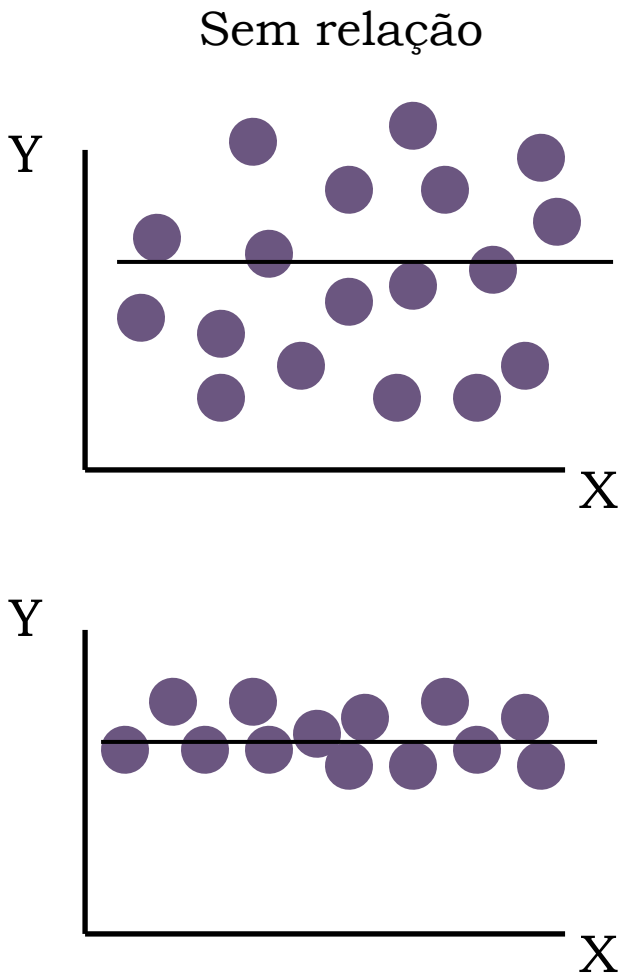
Fonte: Statistics for Managers Using Microsoft® Excel 4th Edition, 2004 Prentice-Hall

Variáveis Correlacionadas



Fonte: Statistics for Managers Using Microsoft® Excel 4th Edition, 2004 Prentice-Hall

Variáveis Correlacionadas



Fonte: Statistics for Managers Using Microsoft® Excel 4th Edition, 2004 Prentice-Hall

Regressão Linear Simples

Uma variável dependente

Uma variável independente

Regressão Linear Simples
Exemplo

- Uma rede vende peças para manutenção de motores. Foram registrados os níveis de venda (“Y”) para 10 semanas (em centenas de itens), para diferentes níveis de preços (“X”), já que os preço variam segundo o valor do câmbio.

Semana Preços Vendas

#	X	Y
1	1,3	10
2	2,0	6
3	1,7	5
4	1,5	12
5	1,6	10
6	1,2	15
7	1,6	5
8	1,4	12
9	1,0	17
10	1,1	20

Regressão Linear Simples
Coeficiente de Correlação

$$r = \frac{n\sum XY - (\sum X)(\sum Y)}{\sqrt{n\sum X^2 - (\sum X)^2}\sqrt{n\sum Y^2 - (\sum Y)^2}}$$

$$= \frac{10(149{,}3) - (14{,}4)(112{,}0)}{\sqrt{10(21{,}56) - (14{,}4)^2}\sqrt{10(1.488) - (112)^2}} = -0{,}8635$$

Há correlação linear negativa

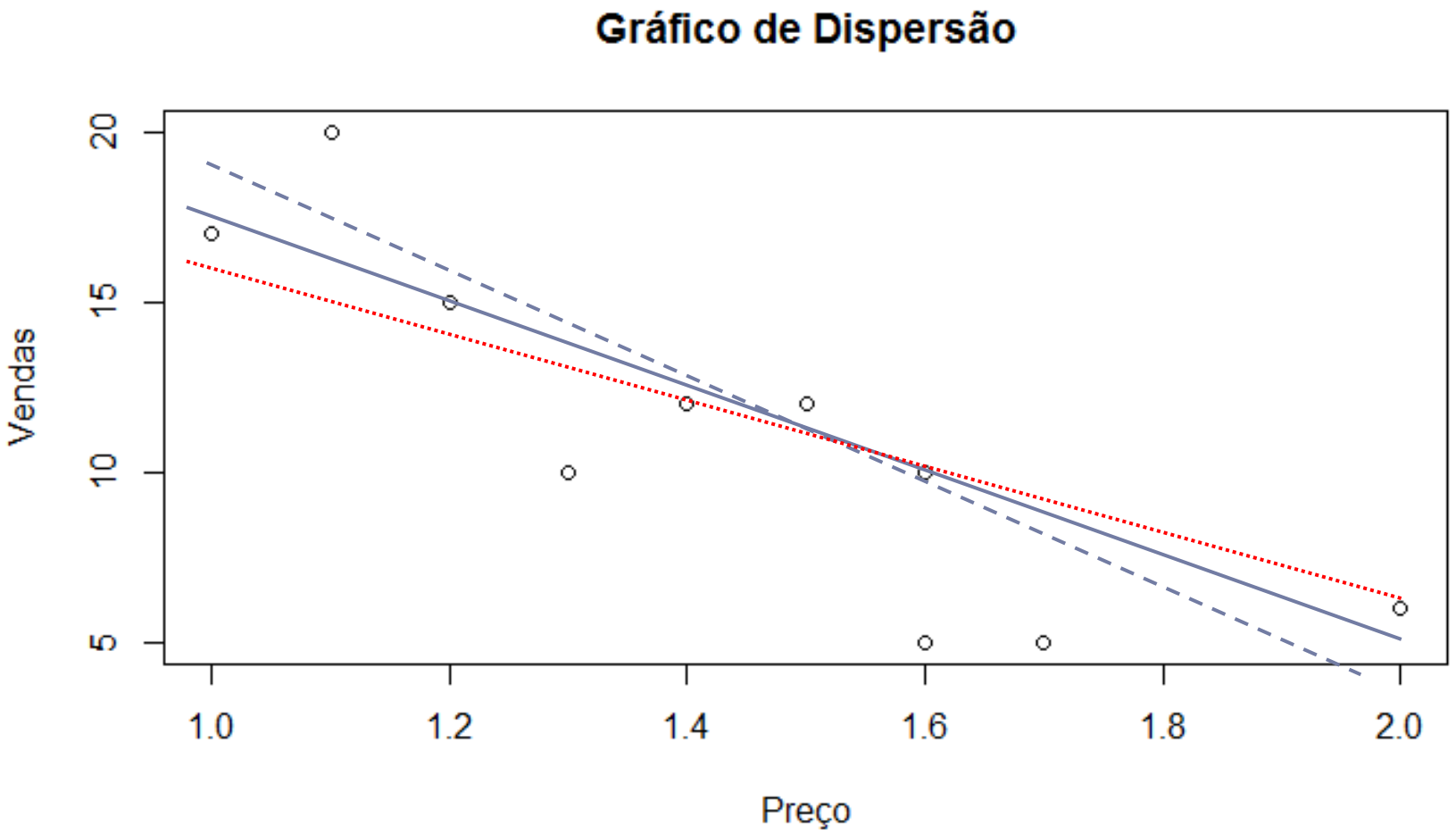
- Em R: `cor(ex0)`

```
             Preço    Vendas
Preço    1.000000 -0.863489
Vendas -0.863489  1.000000
```

Regressão Linear Simples
Coeficiente de Correlação

#	X	Y	XY	X^2	Y^2
1	1,30	10,00	13,00	1,69	100,00
2	2,00	6,00	12,00	4,00	36,00
3	1,70	5,00	8,50	2,89	25,00
4	1,50	12,00	18,00	2,25	144,00
5	1,60	10,00	16,00	2,56	100,00
6	1,20	15,00	18,00	1,44	225,00
7	1,60	5,00	8,00	2,56	25,00
8	1,40	12,00	16,80	1,96	144,00
9	1,00	17,00	17,00	1,00	289,00
10	1,10	20,00	22,00	1,21	400,00
Soma	**14,40**	**112,00**	**149,30**	**21,56**	**1.488,00**

Regressão Linear Simples
Determinando uma reta (curva) de regressão



Qual é a melhor reta de regressão?

Regressão Linear Simples

- Reta (curva) de Regressão: $\hat{Y} = b_0 + bX$
- A melhor reta de regressão é aquela que minimiza a soma das diferenças quadráticas (distância) entre os pontos e a reta – *método dos mínimos quadrados:* (MMQ)

$$b = \frac{n\sum XY - \sum X \sum Y}{n\sum X^2 - (\sum X)^2} \qquad b_0 = \frac{\sum Y}{n} - \frac{b\sum X}{n}$$

Regressão Linear Simples

- Para o exemplo numérico: $b_0 = 32{,}136$ e $b = -14{,}539$.
- Em R ("linear model"):

```
regressão <- lm(ex0$Vendas~ex0$Preço)
```

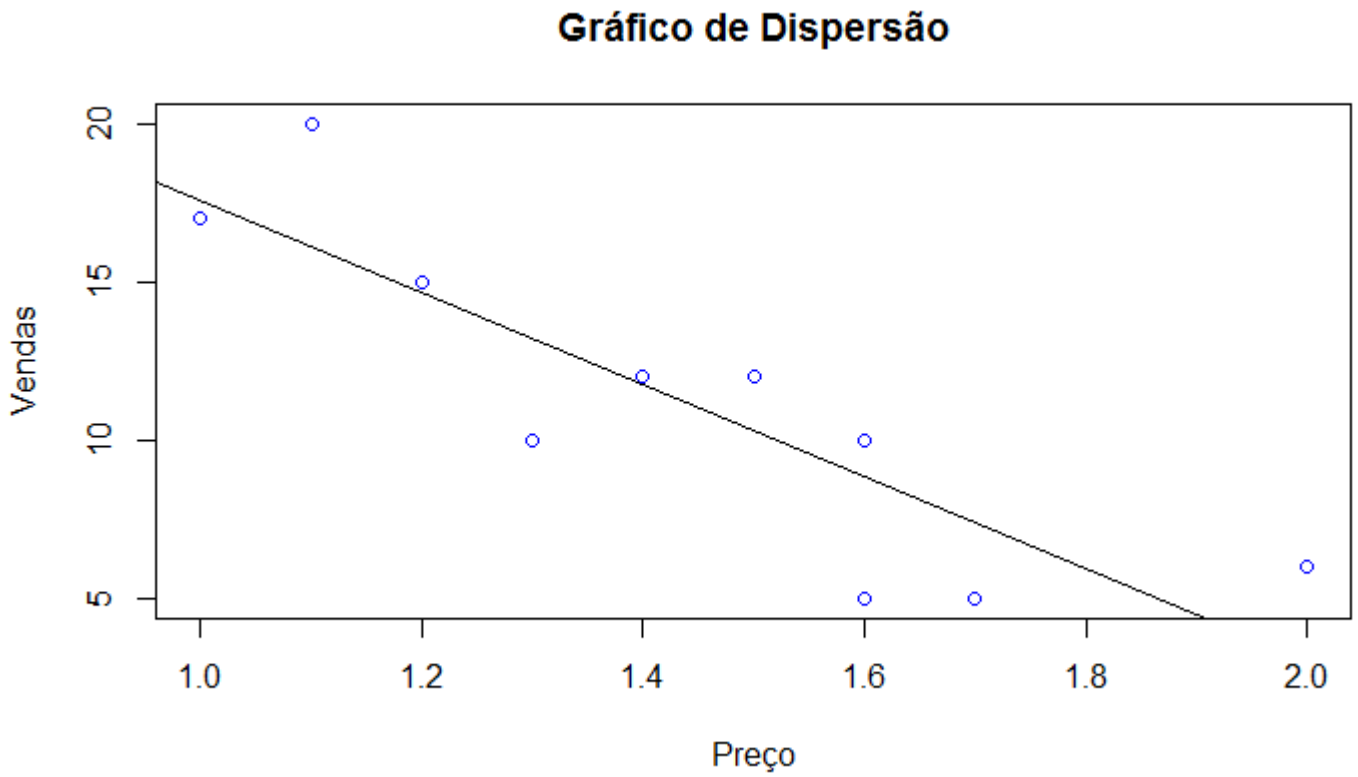
Regressão Linear Simples

- Em R: summary(regressão)

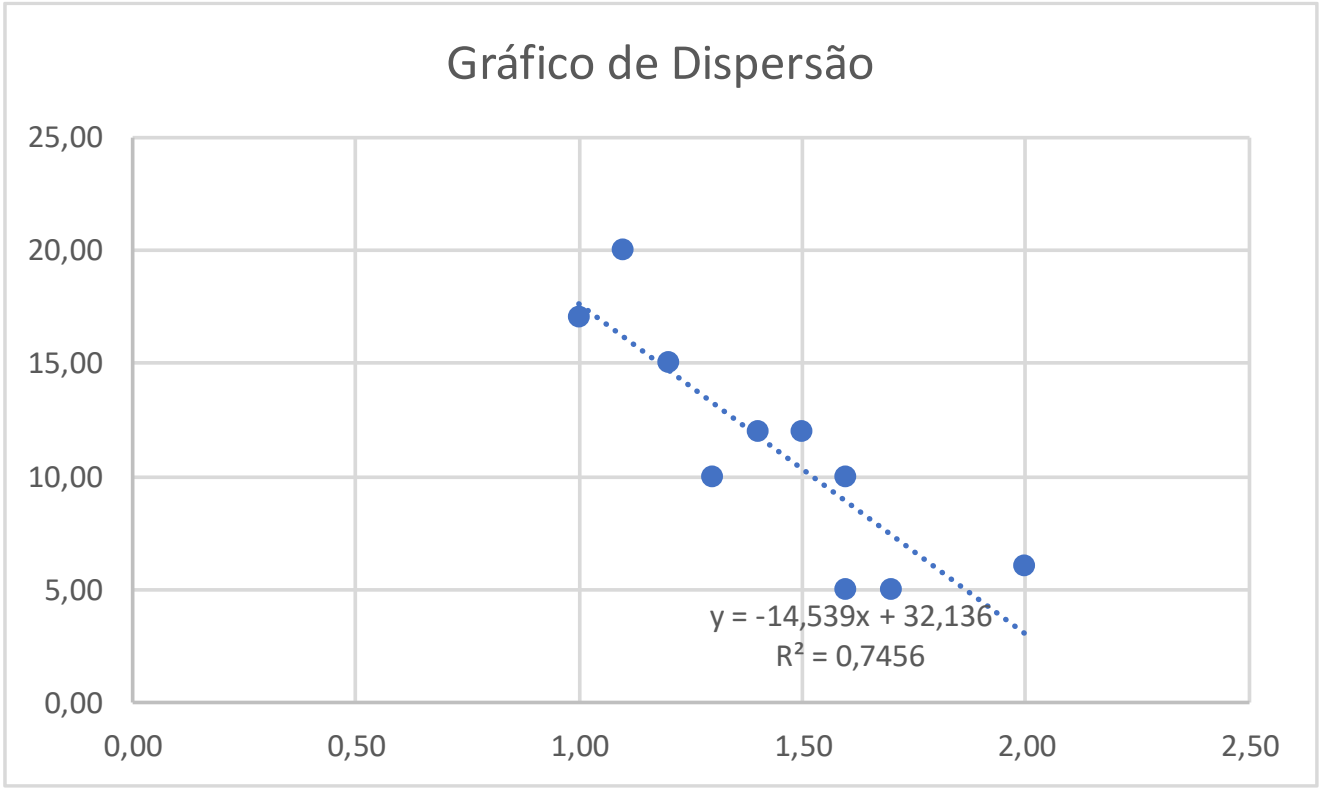
```
Call: lm(formula = ex0$Vendas ~ ex0$Preço)

Residuals:
     Min       1Q   Median       3Q      Max
 -3.8738 -1.9642   0.2646  1.5358  3.8568

Coefficients:
             Estimate  Std. Error   t value  Pr(>|t|)
(Intercept)   32.136        4.409    7.289    8.48e-05 ***
ex0$Preço    -14.539        3.002   -4.842     0.00128 **
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.725 on 8 degrees of freedom
Multiple R-squared: 0.7456,     Adjusted R-squared: 0.7138
F-statistic: 23.45 on 1 and 8 DF, p-value: 0.001284
```

Regressão Linear Simples
Plotando a Curva de Regressão

- `abline(regressão)`



Regressão Linear Simples
Plotando a Curva de Regressão (Excel)



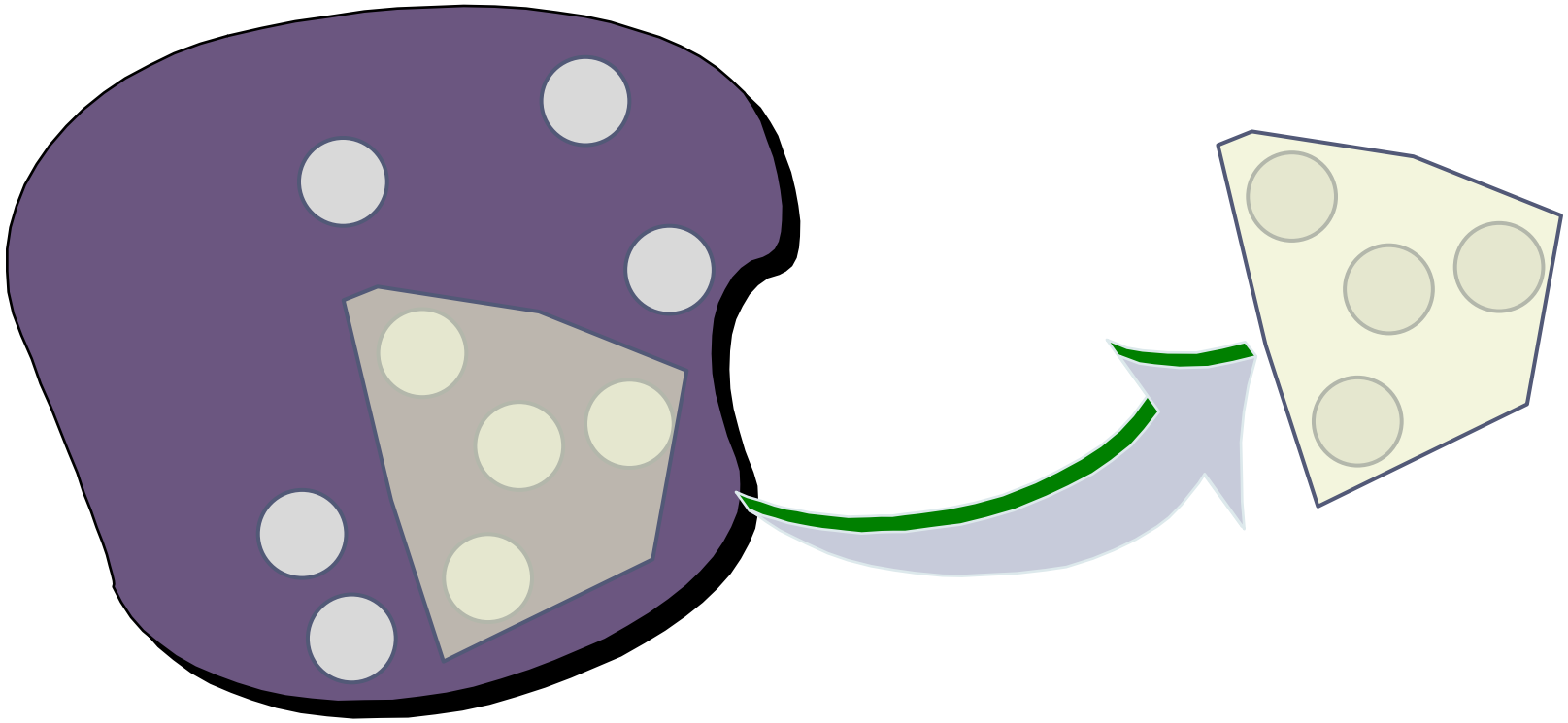
Avaliação do Modelo

Regressão Linear Simples

- É importante observar que a regressão linear elaborada baseia-se em **dados amostrais**.
- Se a amostra for diferente (ou maior), a equação de regressão poderá ser diferente.
- Assim, para a regressão realizada existe um erro associado, que passará ser analisado.

Regressão Linear Simples

População **Amostra Aleatória**

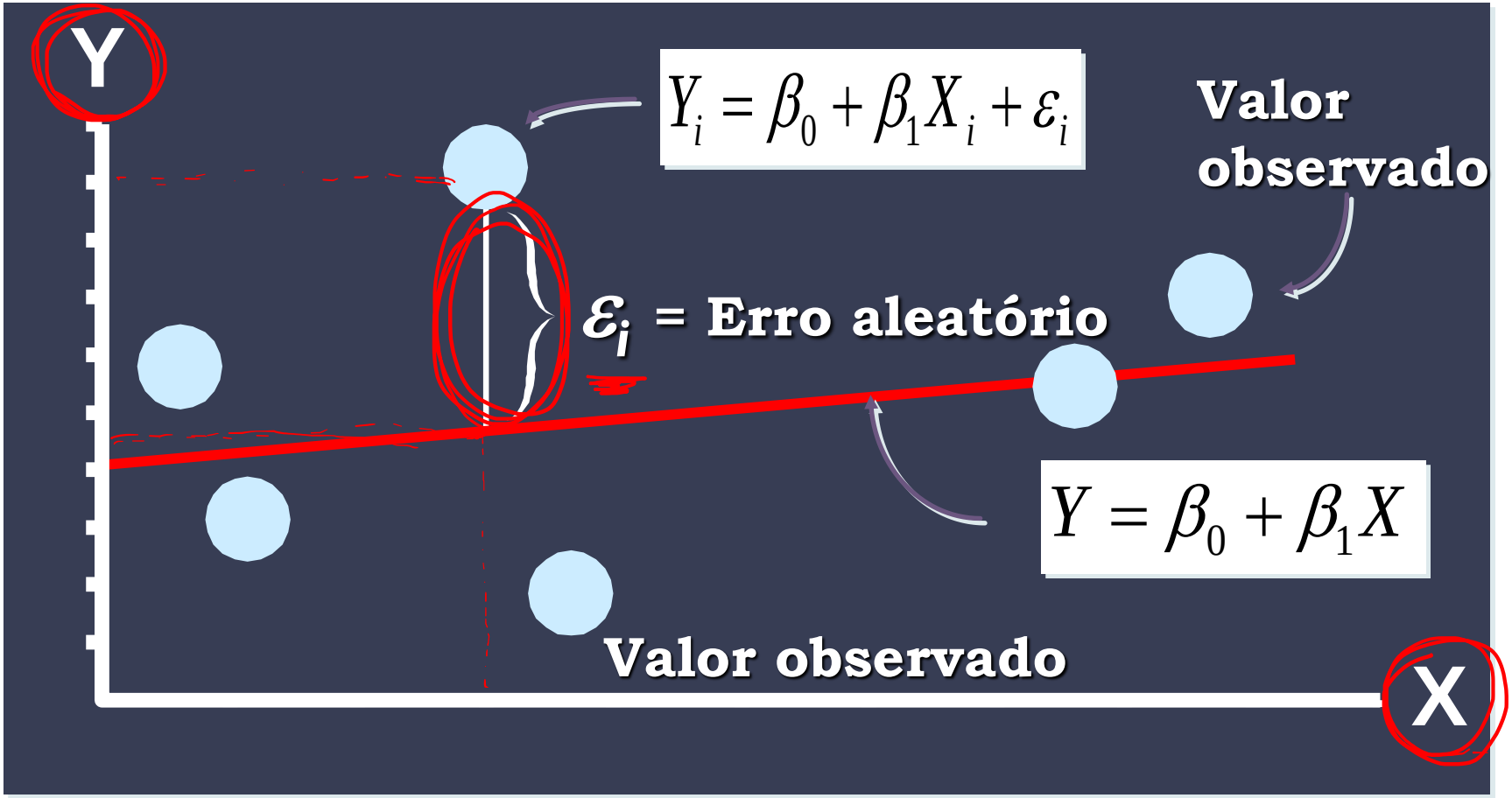


$$Y = \beta_0 + \beta X \qquad \hat{Y} = b_0 + bX$$

$$Y_i = \beta_0 + \beta X_i + \varepsilon_i \qquad Y_i = b_0 + bX_i + e_i$$

Adaptado de https://msu.edu/~fuw/teaching/Fu_Ch11_linear_regression.ppt

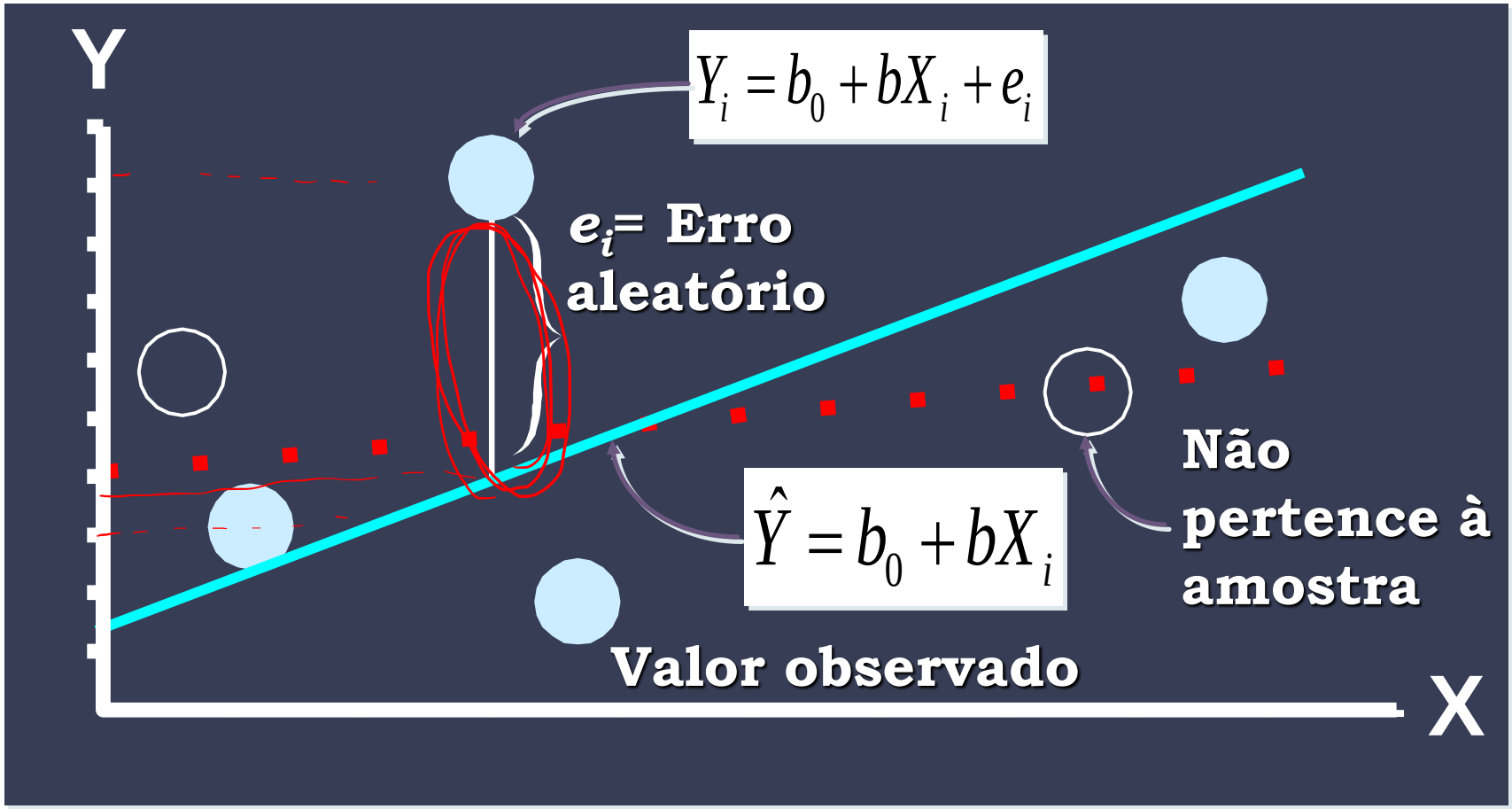
Regressão Linear Simples - População



valor observado - valor previsto

Adaptado de https://msu.edu/~fuw/teaching/Fu_Ch11_linear_regression.ppt

Regressão Linear Simples - Amostra



Adaptado de https://msu.edu/~fuw/teaching/Fu_Ch11_linear_regression.ppt

1. Soma dos Erros Quadráticos

- Soma dos erros quadráticos (SSE - *sum of squares for errors*), é a diferença entre os pontos e a curva de regressão.
- Permite aferir quanto que a curva se adere aos dados.

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

y_i → dado observado

$\hat{y}_i$ → dado obtido reta

ou

$$SSE = \sum_{i=1}^{n} {y_i}^2 - b_0 \sum_{i=1}^{n} y_i - b \sum_{i=1}^{n} x_i y_i$$

1. Soma dos Erros Quadráticos

#	X	Y	XY	X^2	Y^2	$Y - \hat{Y}$	$(Y - \hat{Y})^2$
1	1,30	10,00	13,00	1,69	100,00	-3,24	10,47
2	2,00	6,00	12,00	4,00	36,00	2,94	8,65
3	1,70	5,00	8,50	2,89	25,00	-2,42	5,86
4	1,50	12,00	18,00	2,25	144,00	1,67	2,80
5	1,60	10,00	16,00	2,56	100,00	1,13	1,27
6	1,20	15,00	18,00	1,44	225,00	0,31	0,10
7	1,60	5,00	8,00	2,56	25,00	-3,87	15,01
8	1,40	12,00	16,80	1,96	144,00	0,22	0,05
9	1,00	17,00	17,00	1,00	289,00	-0,60	0,36
10	1,10	20,00	22,00	1,21	400,00	3,86	14,87
Soma	**14,40**	**112,00**	**149,30**	**21,56**	**1.488,00**		**59,425**

SSE

2. Erro Padrão da Estimativa

- O erro padrão da estimativa consiste no valor padrão pelo qual o valor real difere do valor estimado pela regressão.
- O erro médio é zero.
- Se o erro padrão σ_ε for baixo, os erros tenderão a ficar próximos de zero, indicando que o modelo de regressão se adere aos dados. Também indicará que o uso de um modelo linear é válido.
- Um estimador de σ_ε pode ser dado por s_ε.

2. Erro Padrão da Estimativa

$$s_{y.x} = s_{\varepsilon} = \sqrt{\frac{\sum\left(Y - \hat{Y}\right)^2}{n-2}} = \sqrt{\frac{SSE}{n-2}}$$

- $s_{\varepsilon} = \sqrt{\frac{59{,}425}{8}} = 2{,}725$
- Em R: `summary(regressão)`

```
Residual standard error: 2.725 on 8 degrees
of freedom
```

3. Teste de Hipótese Coeficiente Angular

- Quando não houver relação linear entre duas variáveis, a curva de regressão deve ser uma reta horizontal (coeficiente angular = 0).
- Inferência sobre β por meio de um teste de hipótese em b:
- $H_0: \beta = 0$
- $H_1: \beta \neq 0$
- A estatística de teste é:

$$t = \frac{b - \beta}{s_b}, \qquad s_b = \frac{s_\varepsilon}{\sqrt{\sum (X - \bar{X})^2}}$$

3. Teste de Hipótese Coeficiente Angular

- Sendo o erro normalmente distribuído, a estatística segue a distribuição t de Student, com n-2 graus de liberdade.
- $s_b = \frac{s_\varepsilon}{\sqrt{\sum(X-\bar{X})^2}} = \frac{2{,}725}{\sqrt{0{,}824}} = 3{,}00$
- $t = \frac{b-\beta}{s_b} = \frac{-14{,}539-0}{3{,}00} = -4{,}8$
- Estatística de referência, para nível de significância α=5% e α=1%, com 8 graus de liberdade: -2.306004 e -3.355387

  Comando no R: `qt(α/2,GL)`
- Rejeito H_0 !

4. Análise de Variação dos Resíduos

- É importante conhecer qual é a variação na variável dependente Y que está associada à variação na variável dependente X.
- Vamos considerar a variação dos valores de Y em torno de sua média $\bar{Y}$ (chamaremos de SSTO - *total sum of squares*).

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

4. Análise de Variação dos Resíduos

- Vamos separar o valor de SSTO em duas componentes: uma será a variação dos valores previstos pelo modelo de regressão $\hat{Y}$ em relação à média $\bar{Y}$ (SSR – *sum of squares due regression*); a outra medirá a variação dos valores em relação aos valores previstos (SSE – *sum of squares due to error*), que é uma variação não explicada pelo modelo.

4. Análise de Variação dos Resíduos

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

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

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

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

- OBS: relação válida já que a soma dos demais termos é nula.

4. Análise de Variação dos Resíduos

- **Coeficiente de determinação**

$$r^2 = 1 - \frac{SSE}{SSTO} = \frac{SSR}{SSTO}$$

- r^2 indica a proporção da variação em Y que pode ser explicada pela variação em X.
- $r^2 = (-0{,}8635)^2 = 0{,}75$
- Aproximadamente 75% das vendas podem ser explicadas pela variação de preço, enquanto que 25% das vendas são atribuídas a outros fatores.

5. Usando o Modelo para Previsão

- Para o preço X=1,63, prever a quantidade a ser vendida.
- $\hat{Y} = b_0 + bX = 32{,}136 - 14{,}539(1{,}63) = 8{,}440$

6. Erro da Previsão

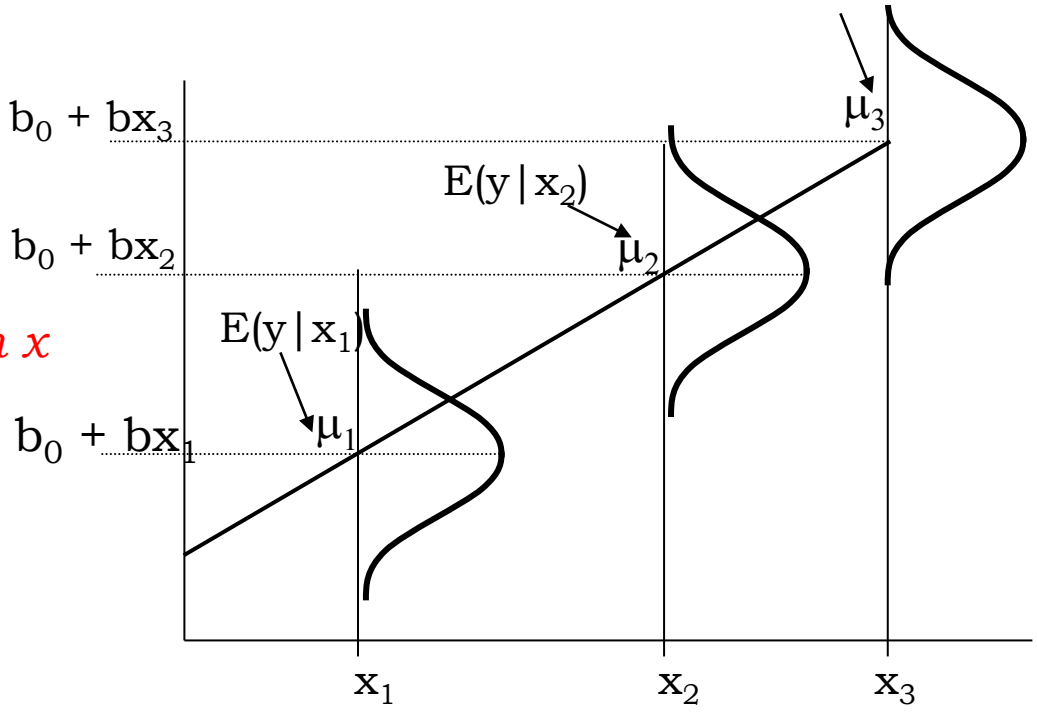
$$s_f = s_\varepsilon \sqrt{1 + \frac{1}{n} + \frac{(X - \bar{X})^2}{\sum (X - \bar{X})^2}}$$

- Intervalo de previsão: $\hat{Y} \pm z s_f$
- Para usar a estatística z as seguintes hipóteses devem ser feitas:
  - Amostra $n \geq 30$.
  - Os valores de Y são normalmente distribuídos em relação à linha de regressão.
  - A dispersão (variância) dos valores de Y em relação à linha de regressão é constante.
  - Os erros são independentes uns dos outros.
  - A regressão é linear.

6. Erro da Previsão

Desvio padrão permanece constante!

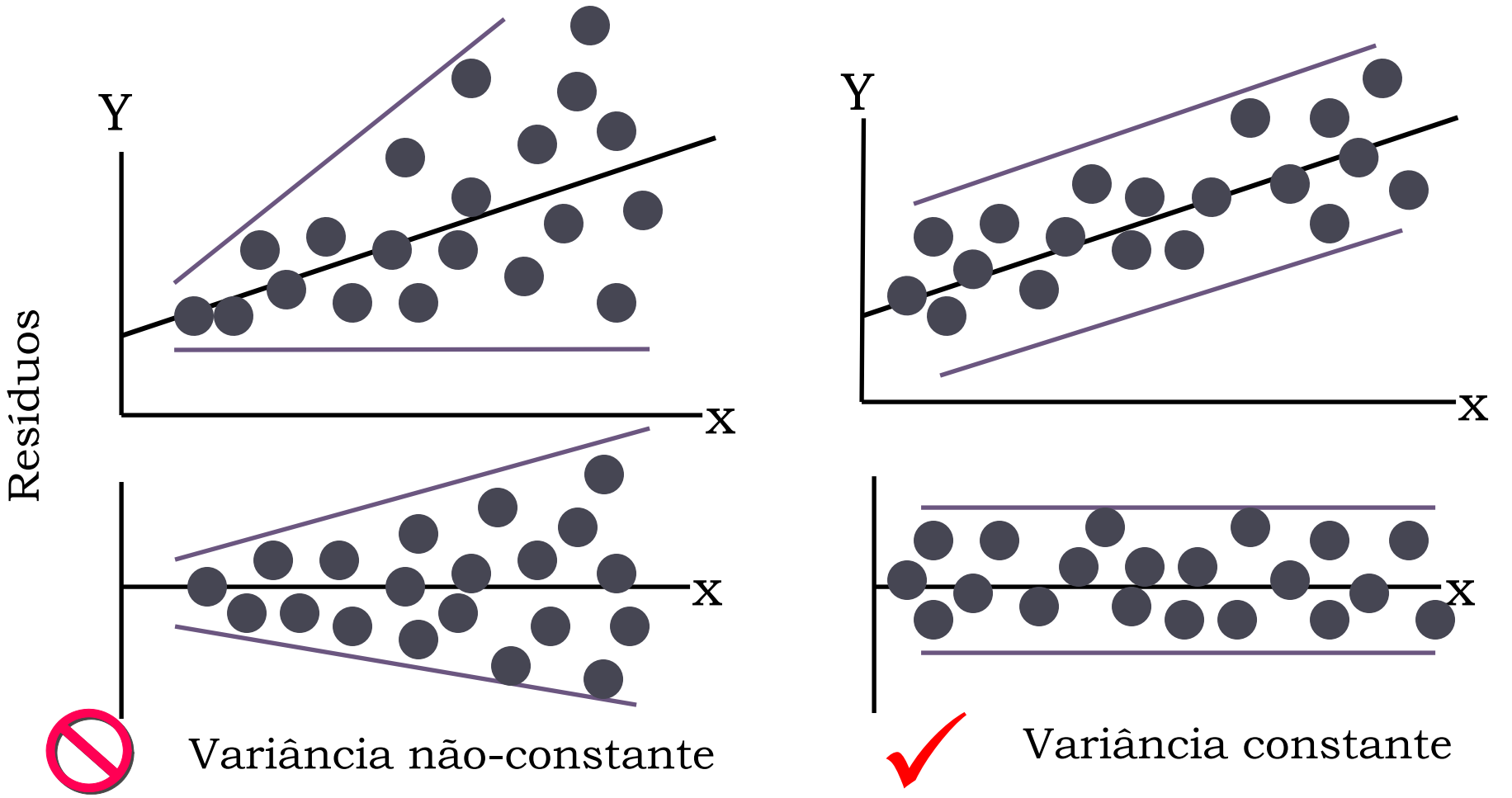
A média varia com x



Premissa:
Y segue uma distribuição normal com media $b_0 + bX$, e desvio padrão constante dado por σ_ε

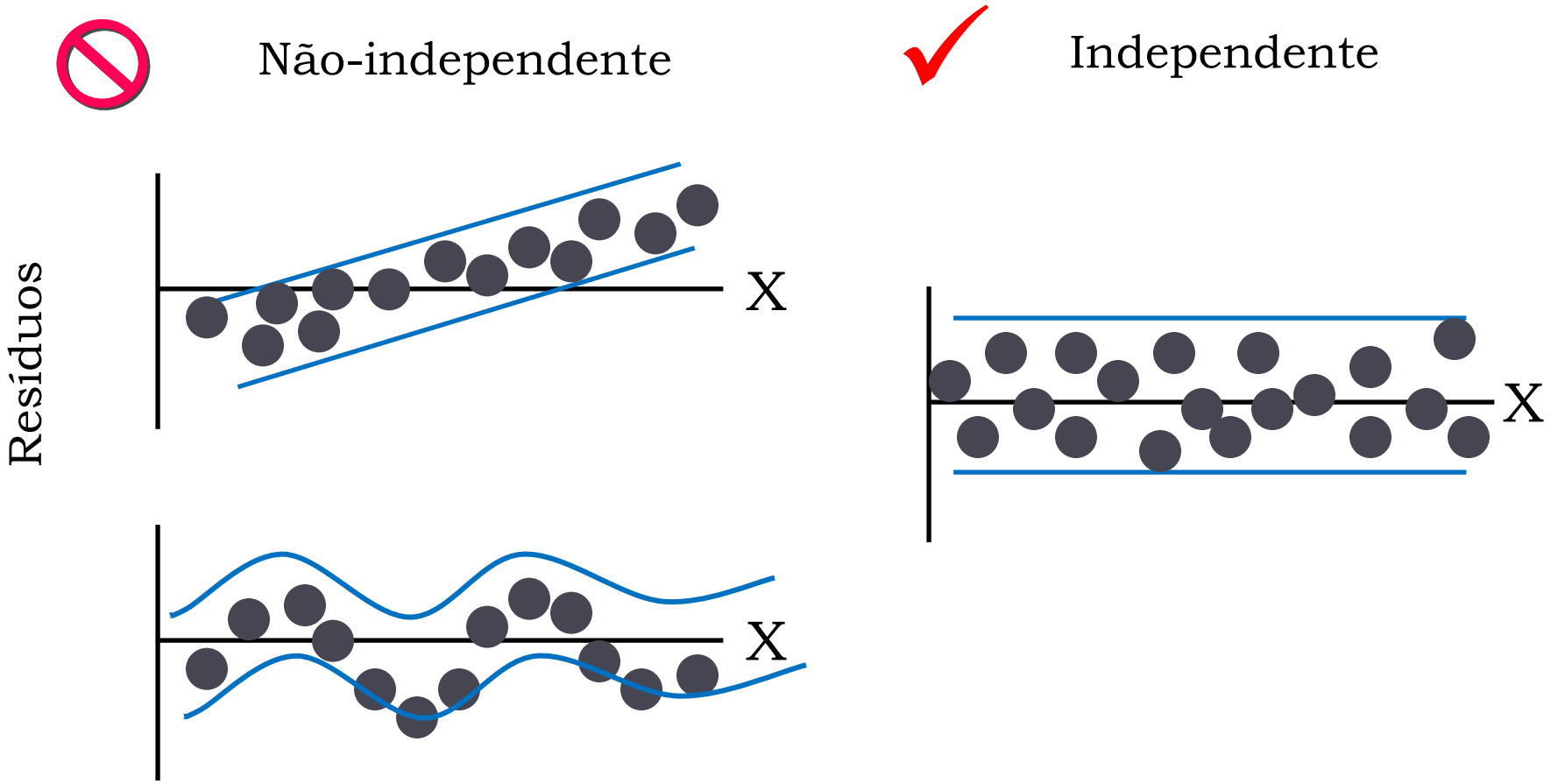
Fonte: http://www.stat.ncsu.edu/people/reiland/courses/st302/simple_lin_regress_inference.ppt

6. Erro da Previsão: Variância dos resíduos



Fonte: Statistics for Managers Using Microsoft® Excel 4th Edition, 2004 Prentice-Hall

6. Erro da Previsão: Independência dos resíduos



Fonte: Statistics for Managers Using Microsoft® Excel 4th Edition, 2004 Prentice-Hall

6. Erro da Previsão

$$s_f = s_{y.x}\sqrt{1 + \frac{1}{n} + \frac{(X - \bar{X})^2}{\sum(X - \bar{X})^2}}$$

- Para o exemplo dado, em que n =10, aproximaremos pela distribuição t de Student: $\hat{Y} \pm ts_f$
- A estatística t pode ser calculada como (Excel), para nível de significância $\alpha = 5\%$, teste bi-caudal, com $n = 10 - 2 = 8$ graus de liberdade: =INV.T(2,5%;8)
- Em R: `qt(c(.025, .975), df=8)`

```
[1] -2.306004 2.306004
```

6. Erro da Previsão

- $\hat{Y} \pm ts_f = 8{,}44 \pm 2{,}306(2{,}92) = 8{,}44 \pm 6{,}71$
- Com 95% de certeza a previsão de vendas irá variar entre 1,73 a 15,15.

7. Análise de Resíduos

- `residuos <-residuals(regressão)`

  ou

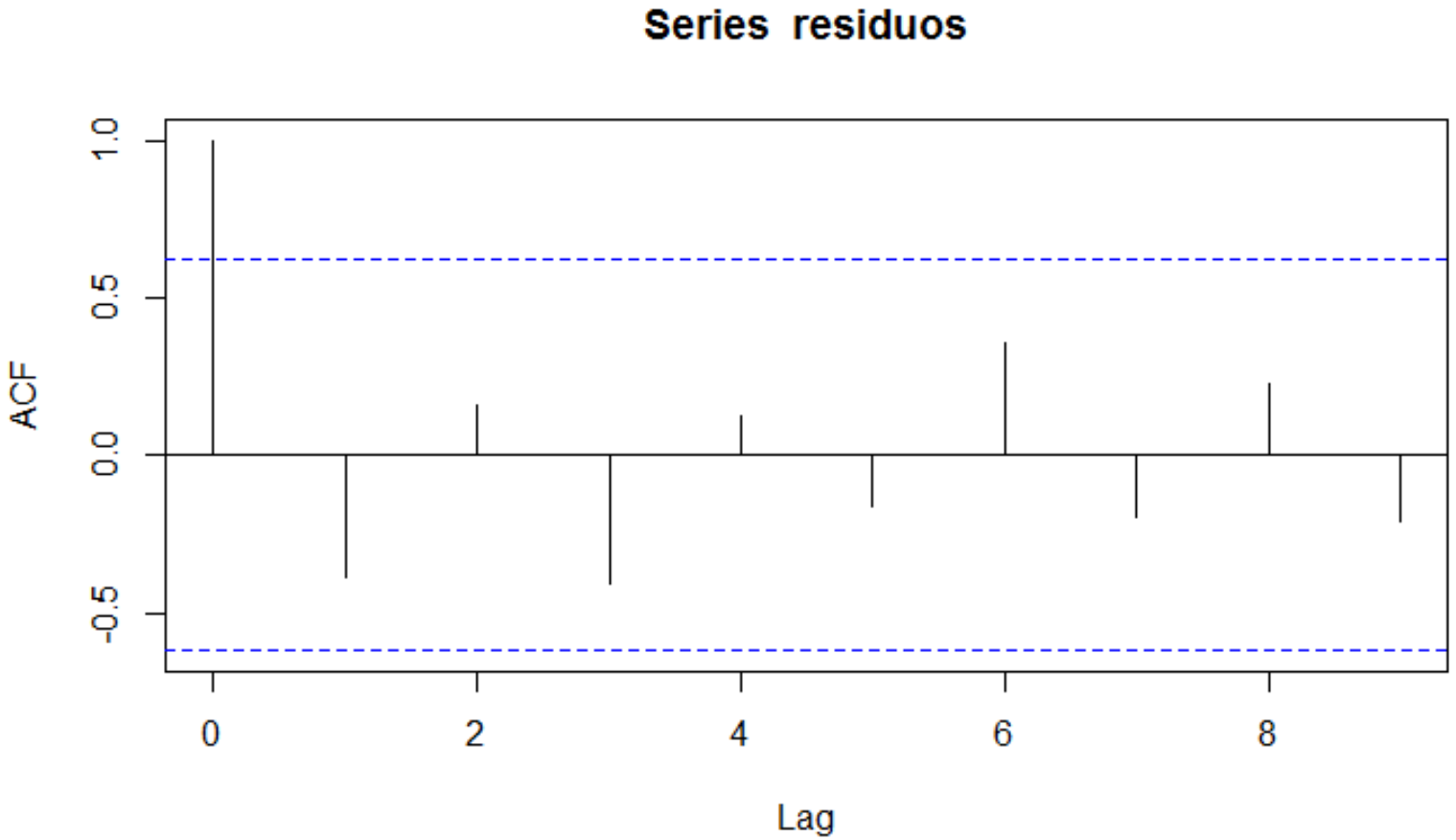
- `residuos <- c(-3.24, 2.94, -2.42, 1.67, 1.13, 0.31, -3.87, 0.22, -0.60, 3.86)`

- `summary(residuos)`

```
    Min.  1st Qu.  Median    Mean  3rd Qu.    Max.
-3.870  -1.965   0.265 0.000 1.535   3.860
```

7. Análise de Resíduos

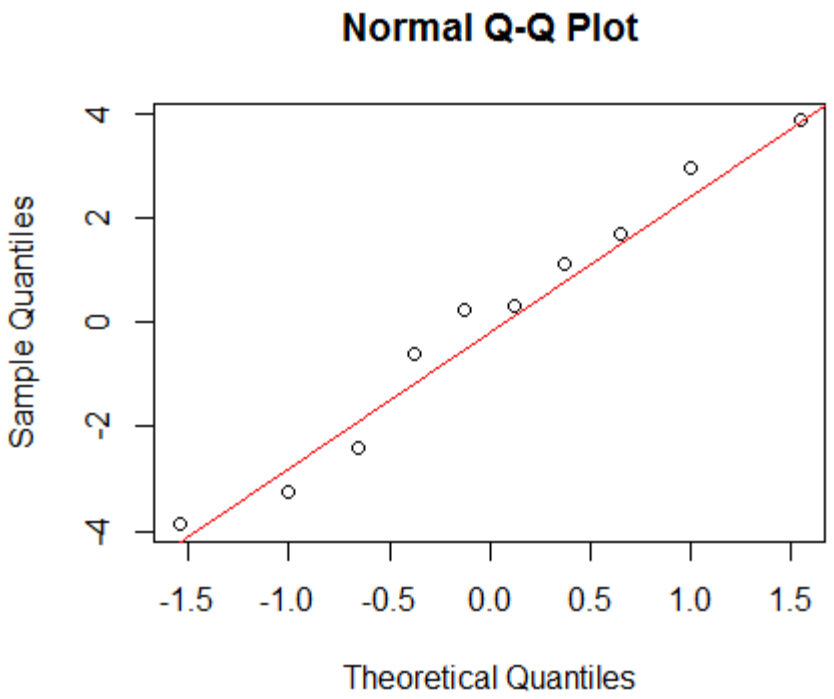
- Análise de autocorrelação: `acf(residuos)`



7. Análise de Resíduos

- Checagem de normalidade dos resíduos

```
qqnorm(residuos)
qqline(x, col = "red")
```



Exercícios

Regressão Linear Simples
Exercício 1

- Uma empresa de e-commerce registrou o nível de vendas, em função do número de propagandas veiculadas em horário nobre, nas semanas que antecediam uma certa data comemorativa.

Cidade	Vendas (x1000)	# Aparições
A	26	6
B	16	2
C	23	5
D	15	1
E	32	10
F	25	7
G	18	15
H	18	3
I	35	11
J	34	13
K	15	2
L	32	12

Regressão Linear Simples
Exercício 1

- a) Verifique se há correlação linear entre as 2 variáveis (a um nível de significância de 5%).
- b) Determine a curva de regressão.
- c) Calcule o erro padrão da estimativa.
- d) Qual o percentual de vendas que é explicado pela variação na quantidade de aparições em comerciais na televisão?
- e) Verifique se o coeficiente angular é significativamente diferente de zero, para um nível de significância de 1%.
- f) Faça uma previsão das vendas para 10 aparições na televisão.

Regressão Linear Simples
Exercício 2

- Gere diferentes amostras de tamanho=20, 40 e 60, e compare o intervalo de previsão, ao prever Y para X=305.
- Para carregar os dados no R:
  - Abrir o arquivo Ex2.txt.
  - Selecionar tudo.
  - Copiar para a área de transferência.
  - No R, digitar a linha de comando: `ex1<-read.delim("clipboard")`