Estatística Aplicada II

- Correlação e Regressão

Aula de hoje

- Tópicos
  - Correlação e Regressão
- Referência
  - Barrow, M. Estatística para economia, contabilidade e administração. São Paulo: Ática, 2007, Cap. 7

Aula de hoje

Objetivos:

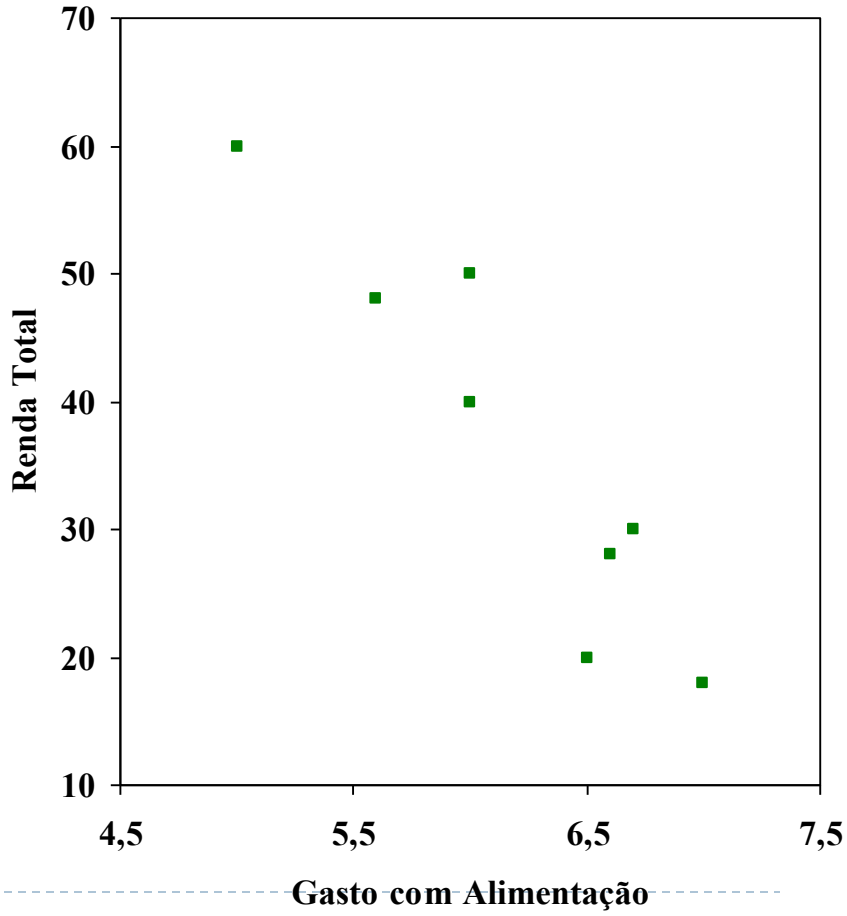
- Analisar os movimentos simultâneos de variáveis:
  - Entender o grau de relação linear entre elas através do cálculo do coeficiente de correlação
  - Entender a causalidade entre elas através da análise de regressão

Correlação

Exemplo: Renda familiar e gastos com alimentação (em % da renda)

- Como esperado, à medida em que aumenta a renda familiar, diminui o percentual da renda destinado à alimentação

Família	Renda Total	Gasto em Alimentação
A	12	7,2
B	16	7,4
C	18	7,0
D	20	6,5
E	28	6,6
F	30	6,7
G	40	6,0
H	48	5,6
I	50	6,0
L	60	5,0



Exemplo livro (Bussab-Morettin), p.81

- Consideremos as duas variáveis abaixo

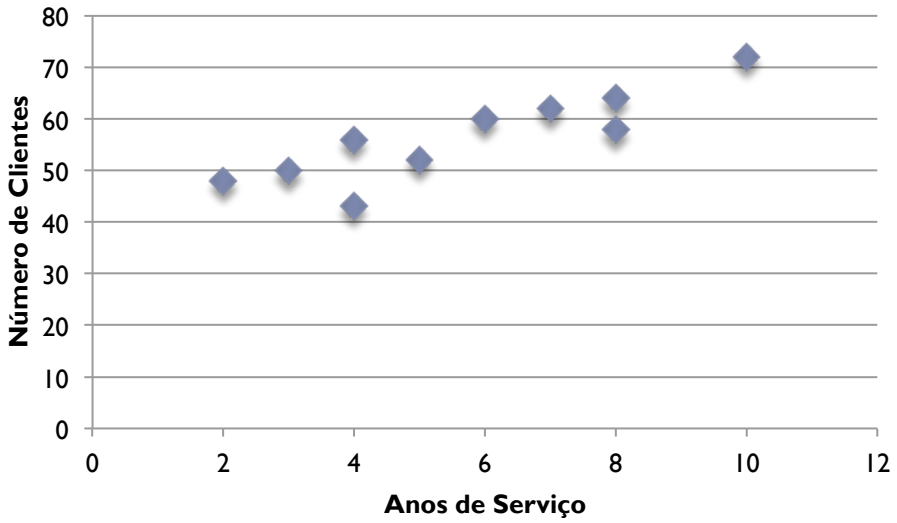
  Número de anos de serviço (X) por número de clientes de agentes de uma cia de seguros

Agente	Anos de serviço (X)	Número de clientes
A	2	48
B	3	50
C	4	56
D	5	52
E	4	43
F	6	60
G	7	62
H	8	58
I	8	64
J	10	72

- Dados hipotéticos

Exemplo livro (Bussab-Morettin), p.81

- Gráfico de Dispersão



- Dados hipotéticos

Covariância

- Dados n pares de valores $(x_1, y_1)..., (x_n, y_n)$, chamaremos de covariância entre as variáveis X e Y, na população:

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

- Para calcular a covariância na amostra, devemos dividir por n-1 e não por n
- É a média dos produtos dos valores centrados das variáveis
- Tendo esta definição, podemos escrever o coeficiente de correlação como:

$$corr(X,Y) = \frac{\mathrm{cov}(X,Y)}{dp(X).dp(Y)}$$

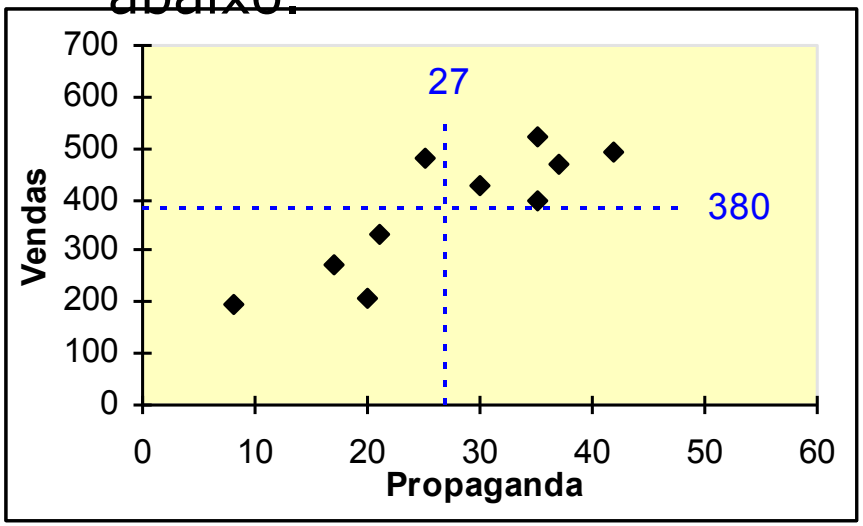
Características da covariância

- Das expressões da covariância, população e amostra:
  - As duas variáveis devem ter o mesmo número de dados.
  - Os pares de dados ocorrem ao mesmo tempo, são pares casados. Embora possa parecer redundante, é importante observar que não se pode mudar a ordem de uma única variável; a mudança de ordem deverá ser realizada nas duas amostras sem descasar os pares de dados.

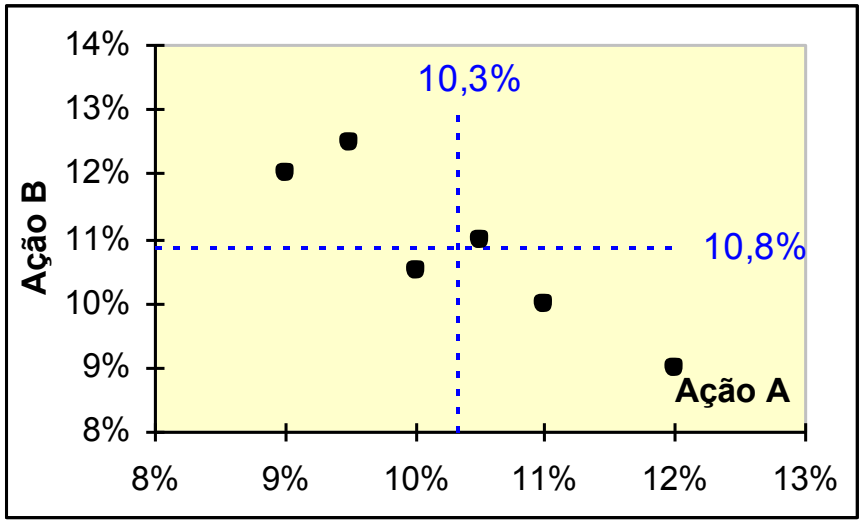
Características da covariância

- A covariância é:
  - No caso de população, a soma dos produtos dos desvios de duas variáveis dividida pela quantidade de dados das variáveis.
  - No caso de amostra, a soma dos produtos dos desvios de duas variáveis dividida pela quantidade de dados das variáveis menos um.
- Os numeradores das expressões da covariância para população e para amostra são iguais, o resultado da soma dos produtos dos desvios.

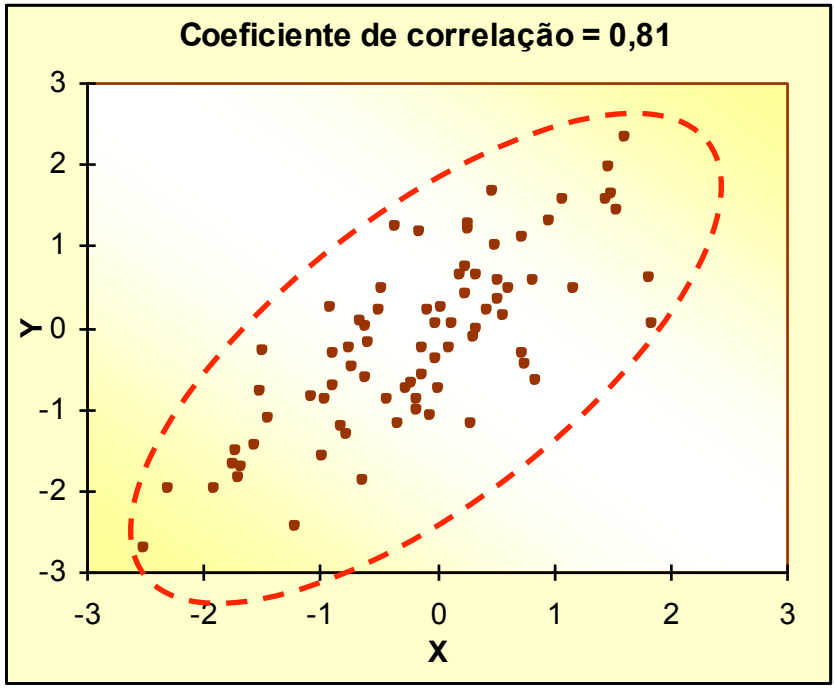
- A covariância pode ser nula, negativa ou positiva.
- A covariância é a medida do afastamento simultâneo das respectivas médias.
- Se as ambas variáveis aleatórias tendem a estar simultaneamente acima, ou abaixo, de suas respectivas médias, então a covariância tenderá a ser positiva e nos outros casos poderá ser negativa, como mostram os gráficos abaixo.

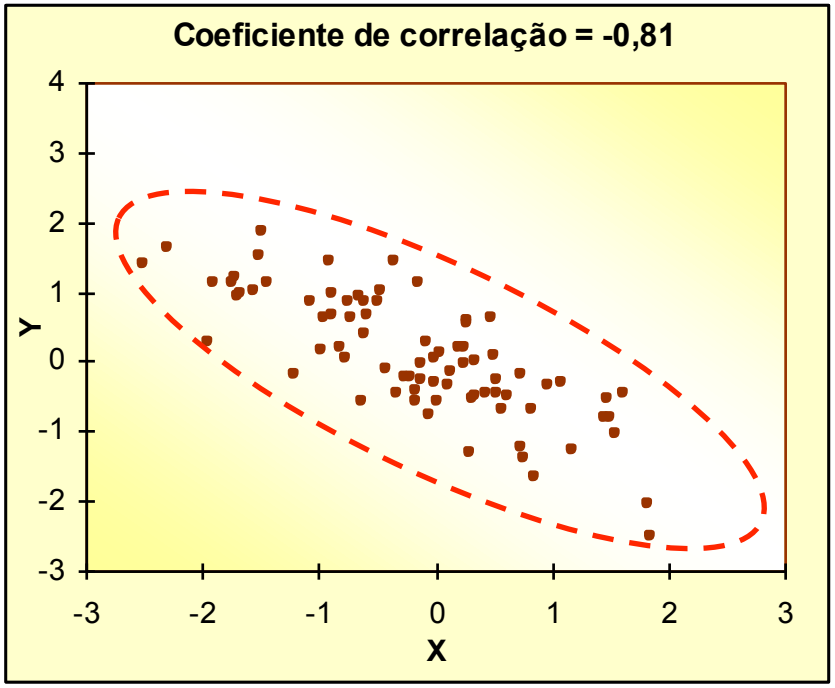


A maioria dos pares de valores tem os dois valores acima de sua média correspondente, provocando covariância positiva.



A maioria dos pares de valores tem um valor acima de sua média e outro abaixo da média correspondente, provocando covariância negativa.





O gráfico de dispersão da esquerda mostra uma relação direta ou positiva entre as variáveis *X* e *Y*, tendência destacada pela declividade positiva da elipse tracejada. Enquanto o gráfico de dispersão da direita mostra uma relação inversa ou negativa, tendência também destacada pela declividade negativa da elipse tracejada.

Características da covariância

- A covariância de uma variável e ela mesma é a própria variância da variável, seja no caso de população ou amostra. Como *Y* = *X*,

$$\sigma_{XX} = \frac{\sum_{i=1}^{N}(X_i - \mu_X) \times (X_i - \mu_X)}{N} = \frac{\sum_{i=1}^{N}(X_i - \mu_X)^2}{N} = \sigma_X^2$$

- A permutação das variáveis não altera o resultado da covariância, se os pares de valores não forem alterados

$$\sigma_{XY} = \sigma_{YX}$$

Características da covariância

- Da mesma forma que a variância, a covariância é afetada pelos valores extremos da variável, ela não é uma medida resistente.
- A unidade de medida é o resultado do produto das unidades dos valores das variáveis.

Coeficiente de correlação

- Para facilitar o entendimento da relação entre duas variáveis e evitar a influência da unidade de medida, foi definido o coeficiente de correlação $r_{XY.}$
- Os valores de r_{XY} estão limitados entre os valores -1 e +1, e sem nenhuma unidade de medida

Coeficiente de correlação

- O coeficiente de correlação busca auferir a direção da relação entre as variáveis, dentro de um intervalo determinado entre -1 e 1
- O objetivo do intervalo é discriminar a direção e a intensidade da relação:
  - valores próximos de zero indicam ausência de relação entre as variáveis
  - valores próximos de 1 indicam forte relação positiva
  - valores próximos de -1 indicam forte relação negativa

Coeficiente de correlação

- O coeficiente de correlação é a medida do grau de associação linear entre duas variáveis
- Fórmula do coeficiente de correlação:

$$corr(X,Y) = \frac{1}{n}\sum\left(\frac{x_i - \bar{x}}{dp(X)}\right)\left(\frac{y_i - \bar{y}}{dp(Y)}\right)$$

Cálculo do coeficiente de correlação

Agente	Anos de serviço (X)	Número de clientes	$x - \bar{x}$	$y - \bar{y}$	$\frac{x - \bar{x}}{dp(X)} = z_x$	$\frac{y - \bar{y}}{dp(Y)} = z_y$	zx.zy
A	2	48	-3,7	-8,5	-1,54	-1,05	1,608
B	3	50	-2,7	-6,5	-1,12	-0,80	0,897
C	4	56	-1,7	-0,5	-0,71	-0,06	0,043
D	5	52	-0,7	-4,5	-0,29	-0,55	0,161
E	4	43	-1,7	-13,5	-0,71	-1,66	1,173
F	6	60	0,3	3,5	0,12	0,43	0,054
G	7	62	1,3	5,5	0,54	0,68	0,366
H	8	58	2,3	1,5	0,95	0,18	0,176
I	8	64	2,3	7,5	0,95	0,92	0,882
J	10	72	4,3	15,5	1,78	1,91	3,407

Para calcular o coeficiente de correlação, devemos dividir o somatório dos valores da última coluna (8,77) pelo número de observações (n=10)

Então: Corr(X,Y) = 8,77/10=0,877

Coeficiente de correlação

- O coeficiente de correlação pode ser escrito da seguinte forma:

$$corr(X,Y) = \frac{1}{n}\sum\left(\frac{x_i - \bar{x}}{dp(X)}\right)\left(\frac{y_i - \bar{y}}{dp(Y)}\right)$$

$$corr(X,Y) = \frac{\sum x_i y_i - n\overline{xy}}{\sqrt{\left(x_i^2 - n\bar{x}^2\right)\left(y_i^2 - n\bar{y}^2\right)}}$$

Sendo que -1≤ corr(X,Y) ≤1

- Lembremos da **variância**, que usamos para observar a dispersão de uma só variável

$$\text{var}(X) = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}{n}$$

Voltando ao coeficiente de correlação

- Da fórmula do coeficiente de correlação pode-se obter também a covariância das mesmas variáveis quando conhecidos os desvios padrões correspondentes:

$$\sigma_{XY} = r_{XY} \times \sigma_X \times \sigma_Y$$

Características de r

- Se a variável *Y* é a mesma variável *X*, então o coeficiente de correlação é igual a 1:

$$r_{XX} = \frac{\sigma_{XX}}{\sigma_X \times \sigma_X} = \frac{\sigma_X^2}{\sigma_X^2} = 1$$

- A permutação das variáveis não altera o resultado do coeficiente de correlação, se os mesmos pares de valores forem mantidos.

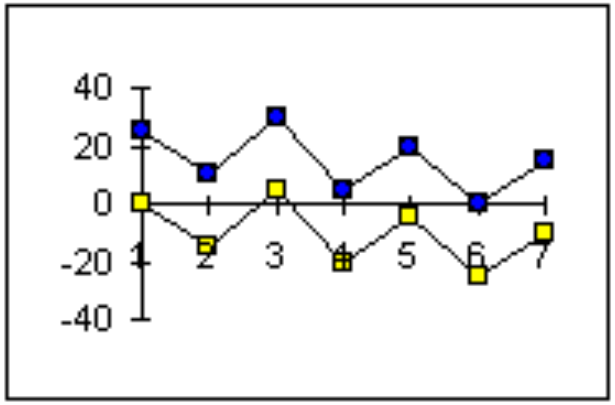
$$r_{XY} = r_{YX}$$

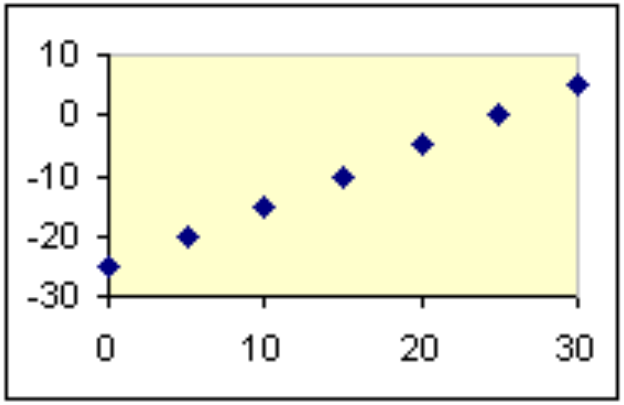
$r = +1$

Amostras perfeitamente correlacionadas em sentido positivo

X	Y
25	0
10	-15
30	5
5	-20
20	-5
0	-25
15	-10

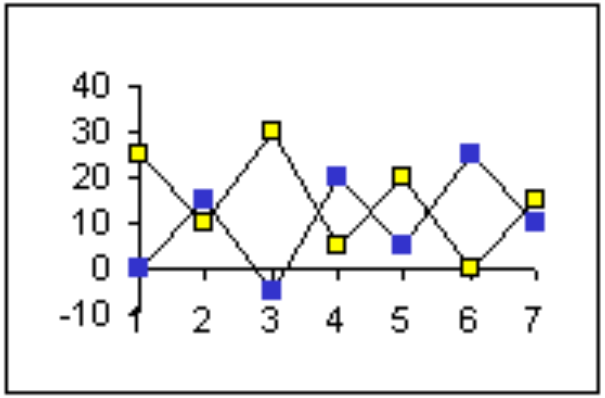
r	1,00

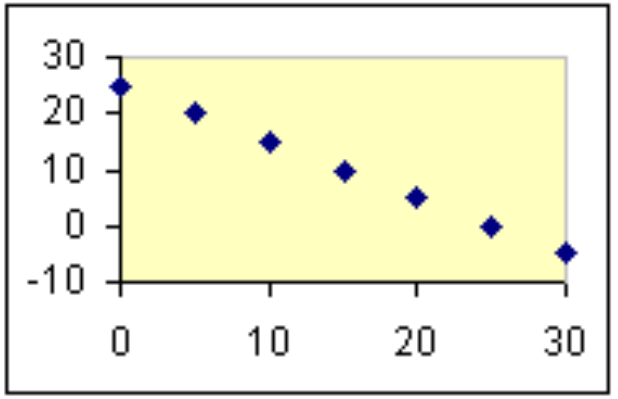




$r = -1$

	A	B	C	D	E	F	G	H	I
13	**Amostras perfeitamente correlacionadas em sentido negativo**								
14		**X**	**Y**			***r***	**-1,00**		
15		25	0						
16		10	15						
17		30	-5						
18		5	20						
19		20	5						
20		0	25						
21		15	10						
22									



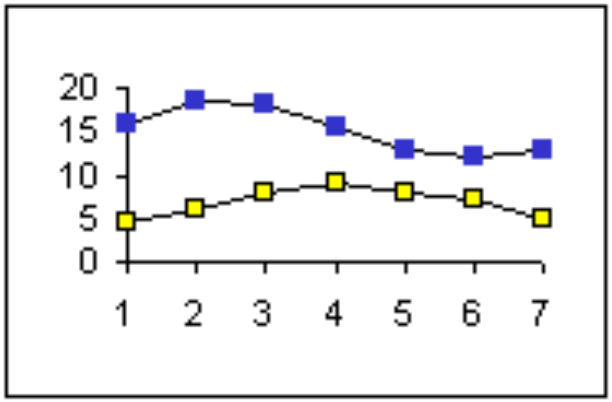


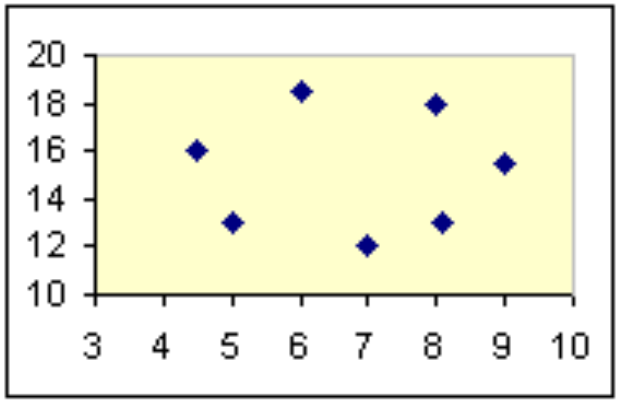
$r = 0$

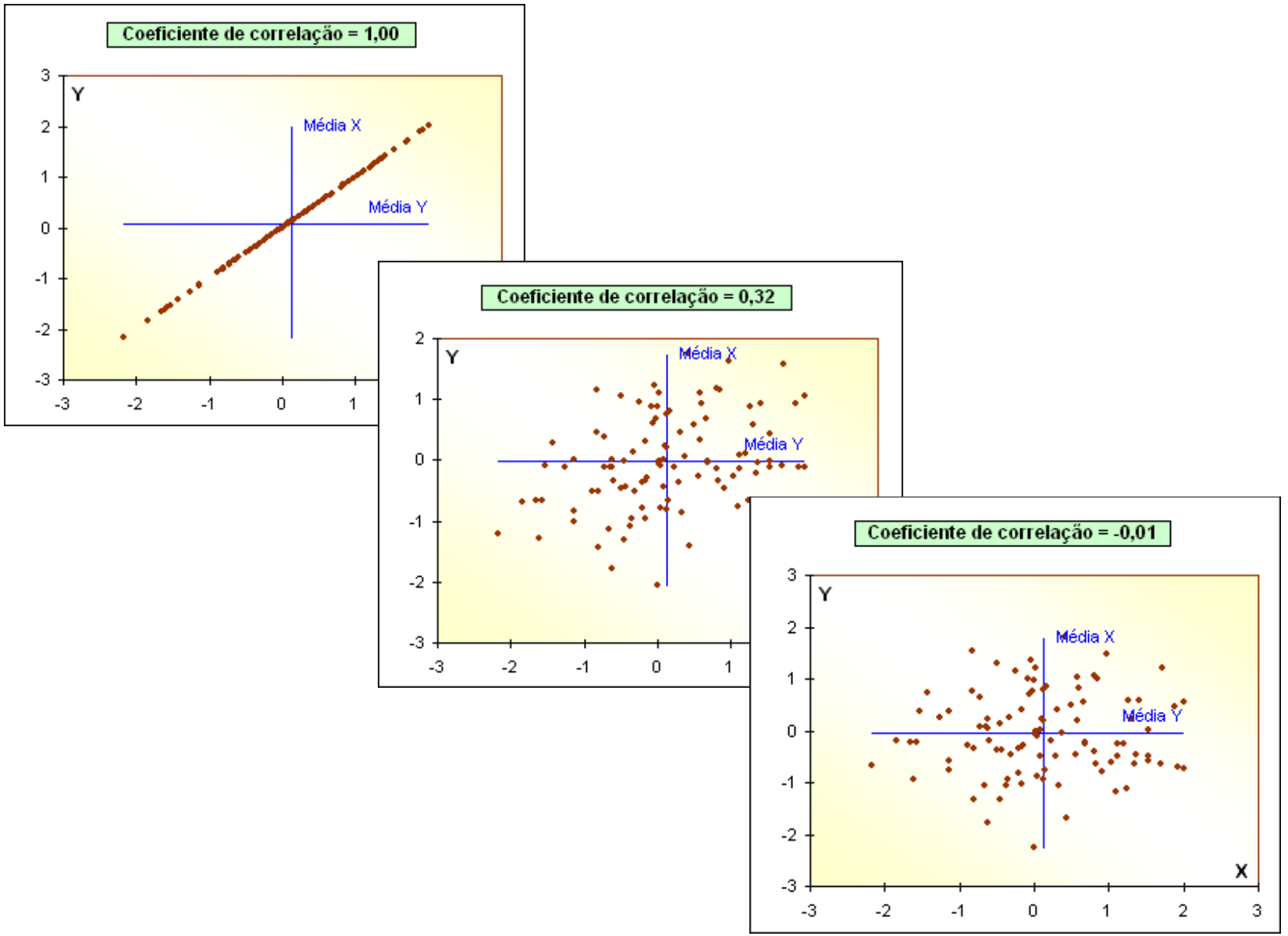
Amostras não correlacionadas

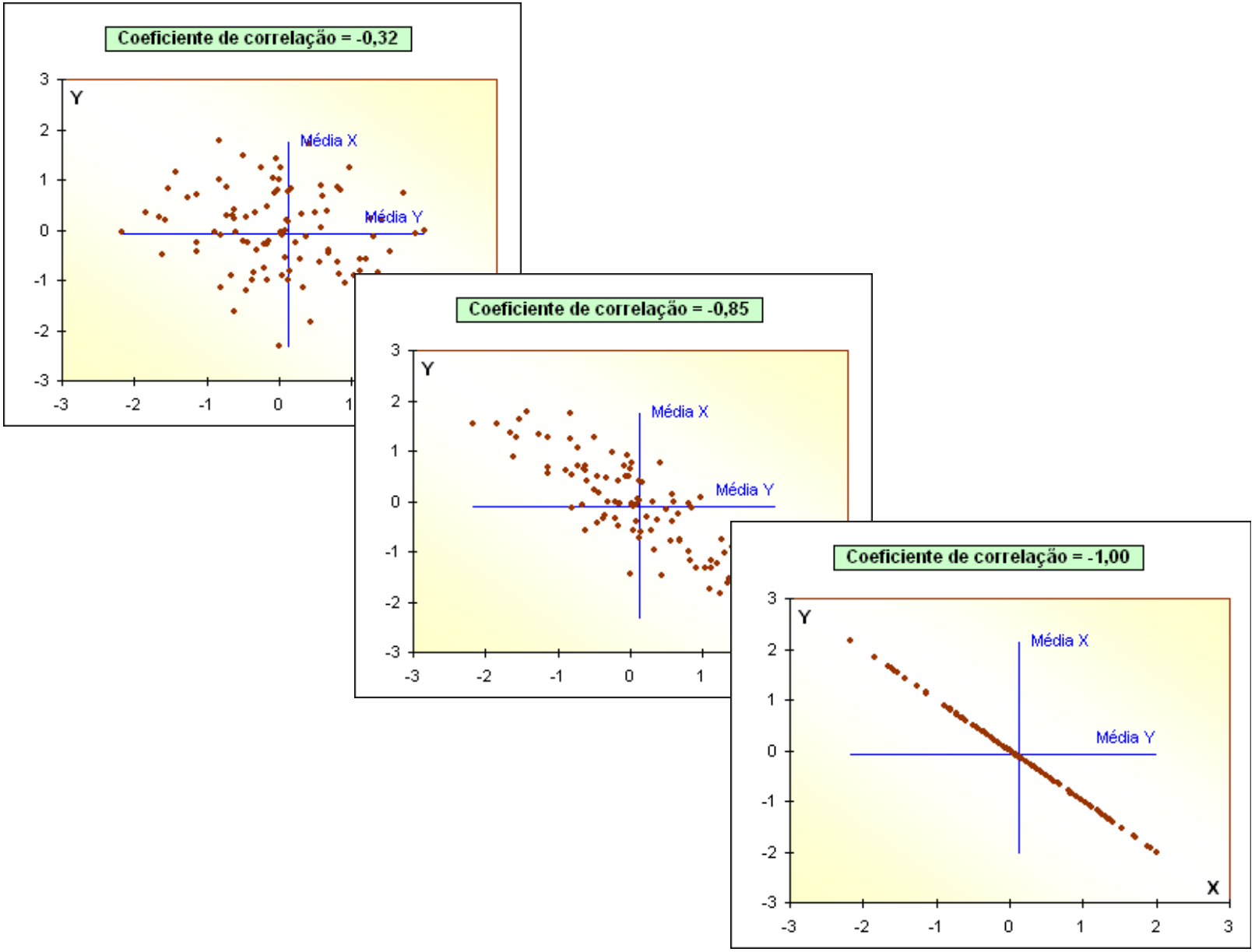
X	Y
4,5	16
6	18,5
8	18
9	15,5
8,1	13
7	12
5	13

r	0,00






Coeficiente de correlação = 1,00
Y
Média X
Média Y
Coeficiente de correlação = 0,32
Y
Média X
Média Y
Coeficiente de correlação = -0,01
Y
Média X
Média Y
X


Coeficiente de correlação = -0,32
Y
Média X
Média Y
Coeficiente de correlação = -0,85
Y
Média X
Média Y
Coeficiente de correlação = -1,00
Y
Média X
Média Y
X

Os resultados são significantes?

- H_0: r = 0

  H_1: r ≠ 0

- A estatística do teste é:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

- A qual tem distribuição t com n-2 graus de liberdade

Teste de Hipótese

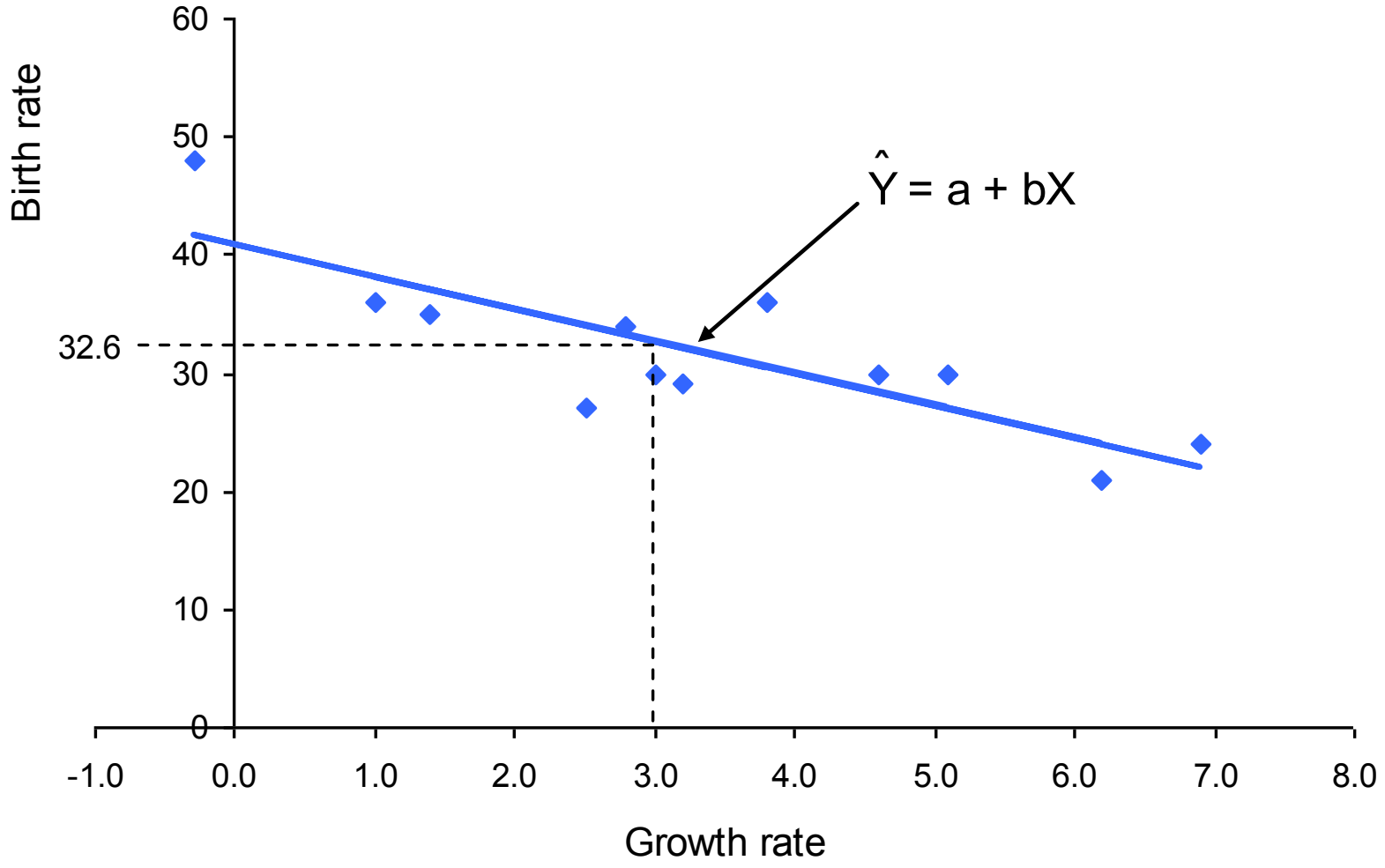
- As etapas do teste são:

1. Escrever as hipóteses alternativas e nulas
2. Escolher o nível de significância do teste α
3. Calcular a estatística t, conhecida como a **estatística do teste**
4. Calcular o **valor crítico** do teste t^*,
5. Decidir: **Comparar a estatística do teste t com o valor crítico do teste t^*,**

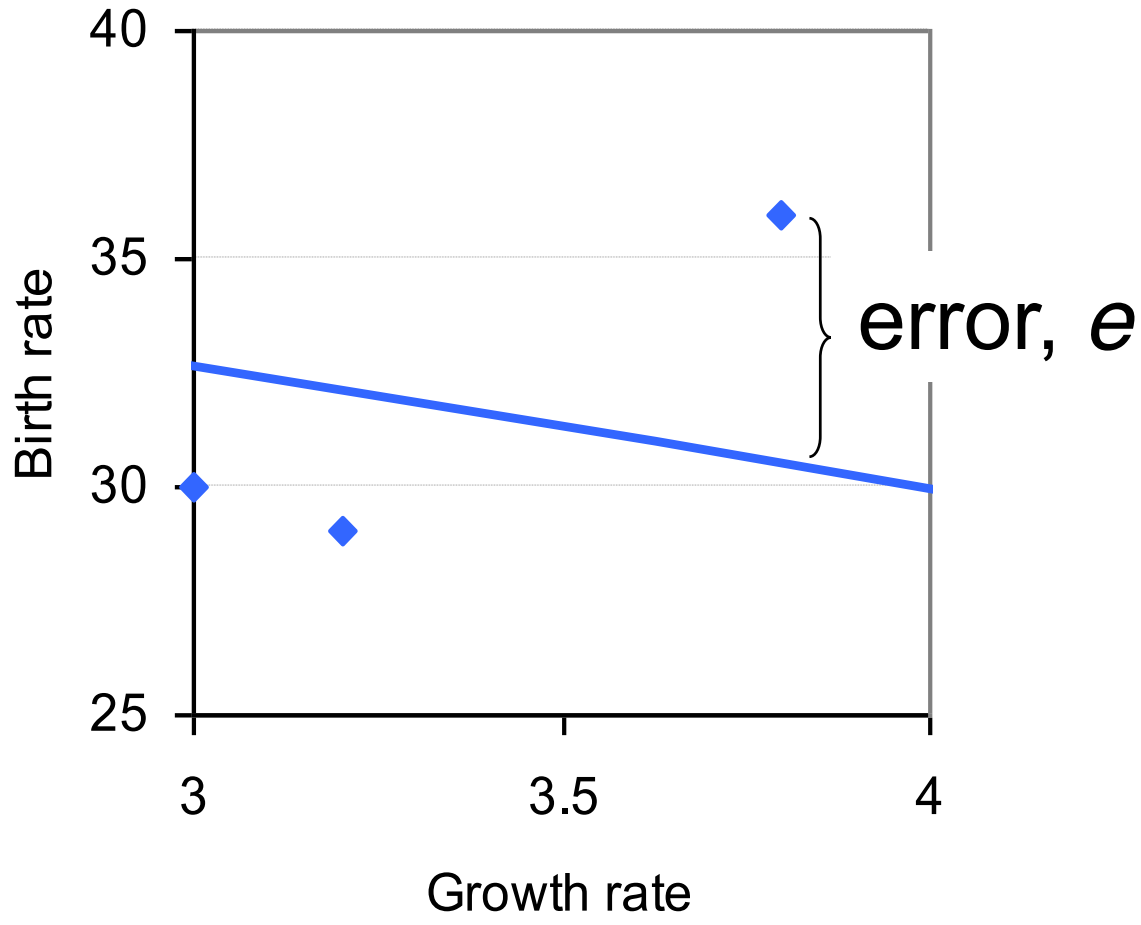
Regressão Linear Simples

- Modelo linear para explicar a variável Y, denominada variável dependente, explicada ou endógena como função da variável X, denominada variável independente, explicativa ou exógena

Regressão Linear Simples



Regressão Linear Simples



Regressão Linear Simples

- A relação entre valor observado de Y e valor previsto de Y pelo modelo é dada por:

$$Y = \hat{Y} + e$$

$$Y = a + bX + e$$

Regressão Linear Simples

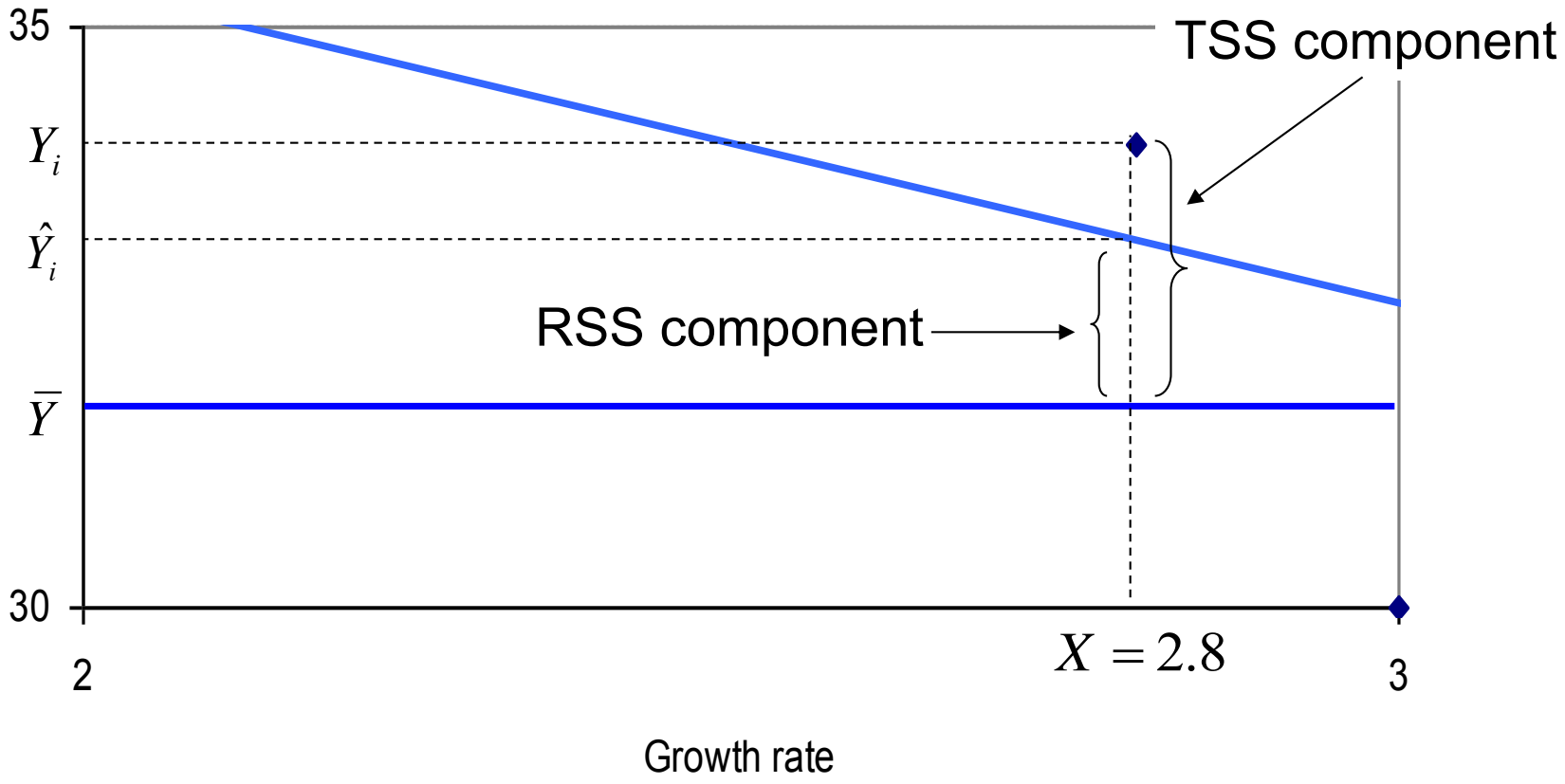
- Os valores de a e b são dados pela minimização da soma do quadrado dos erros. Tem-se:

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

e

$$a = \bar{Y} - b\bar{X}$$

Regressão Linear Simples



Regressão Linear Simples

- Mensuração da qualidade do ajuste:

$$R^2 = \frac{RSS}{TSS}$$

Em que:

$$TSS = \sum \left(Y - \bar{Y} \right)^2 = \sum Y^2 - n\bar{Y}^2$$

$$ESS = \sum \left(Y - \hat{Y} \right)^2$$