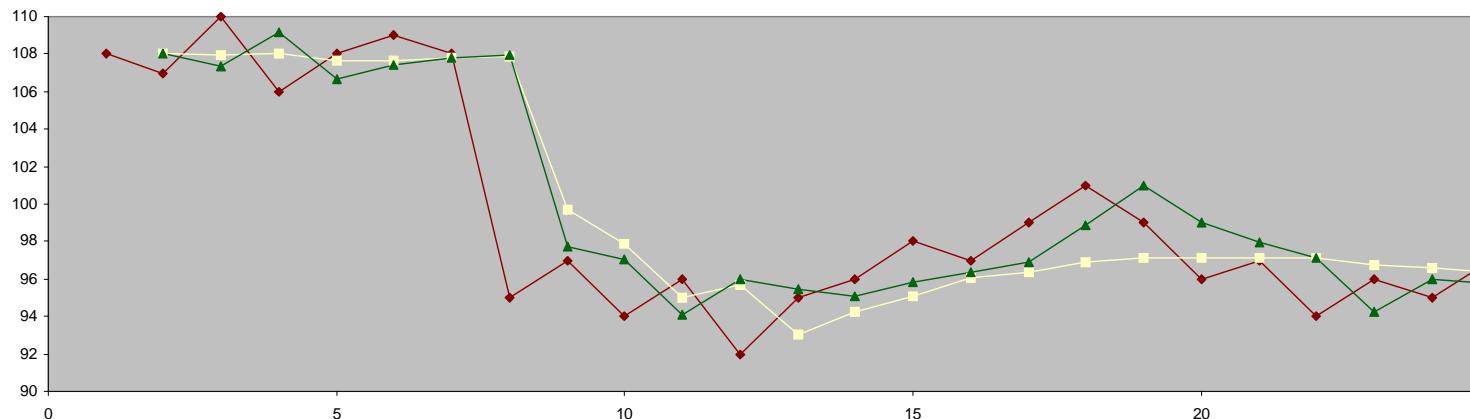


MÉTODOS ESTATÍSTICOS DE PREVISÃO



Regressão Linear

Bernardo Almada-Lobo

Regressão

A **regressão** é uma das técnicas estatísticas mais potentes e de utilização mais frequente.

É um método matemático utilizado para descrever a relação entre variáveis

- Regressão linear simples
- Regressão linear múltipla
- Regressão não linear

Regressão

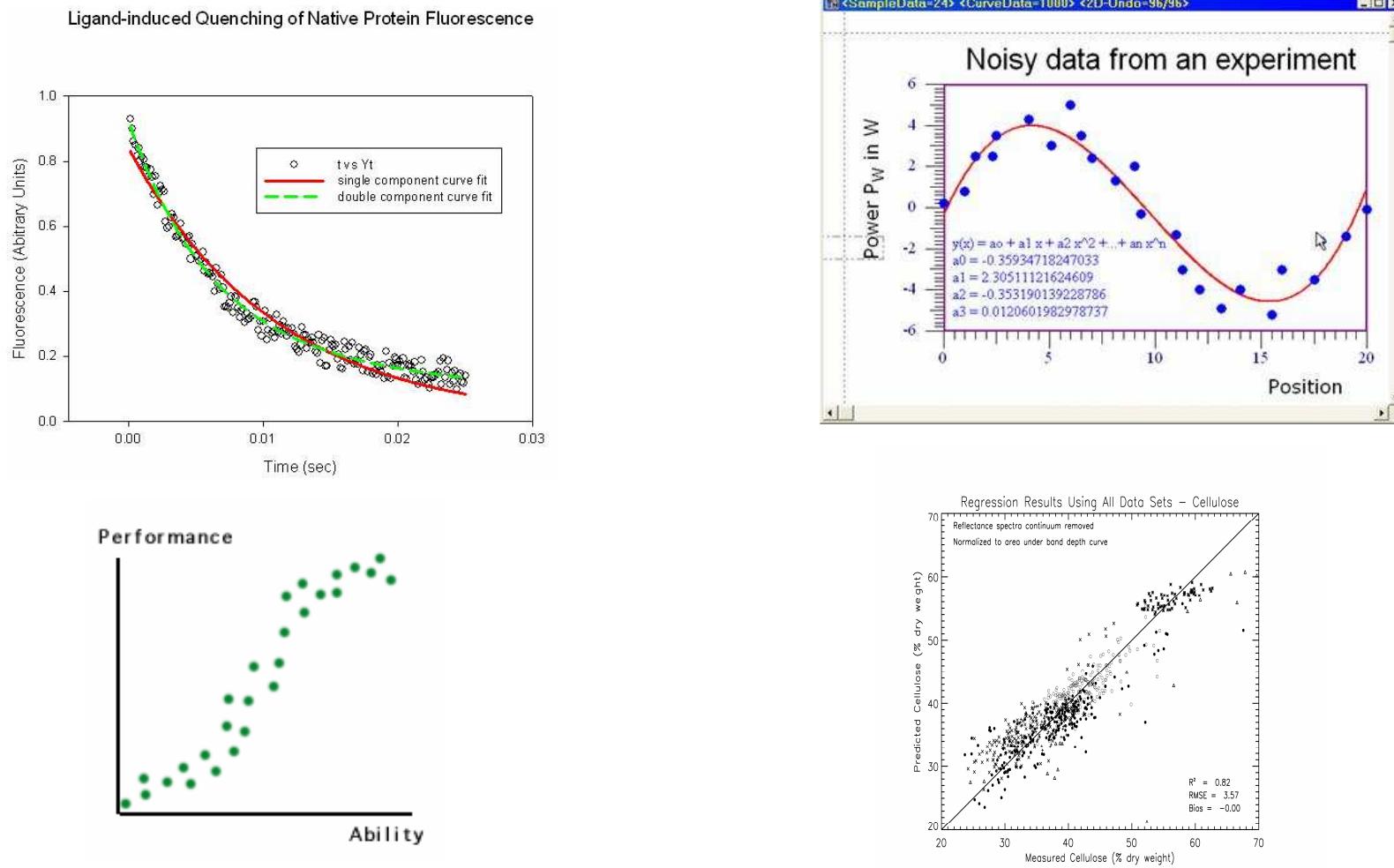
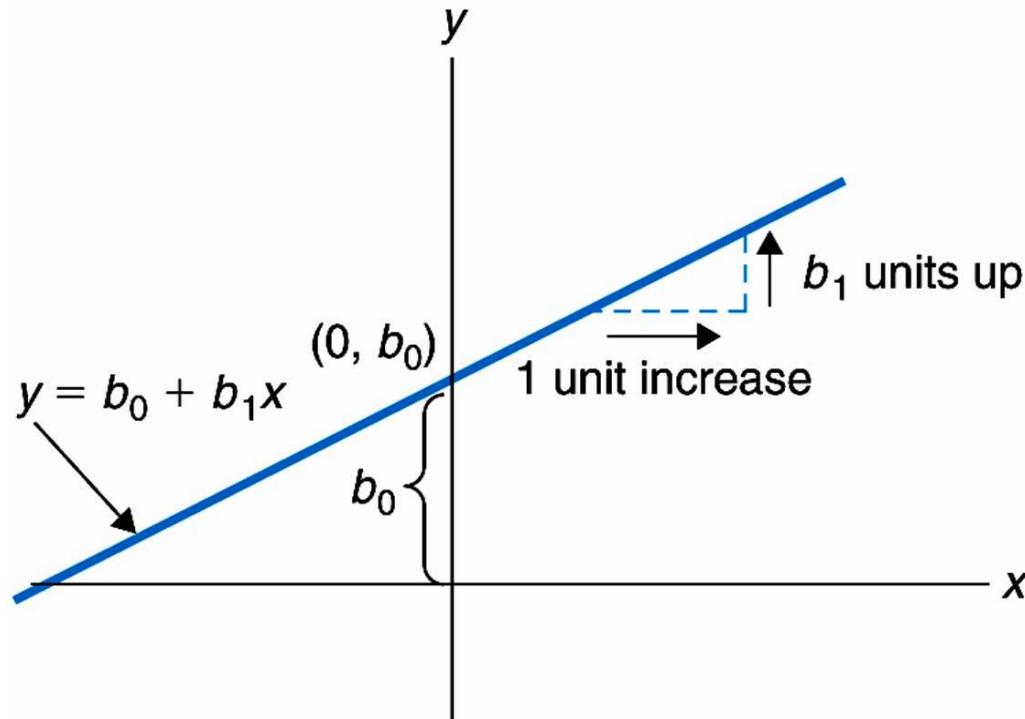


Figure 6c

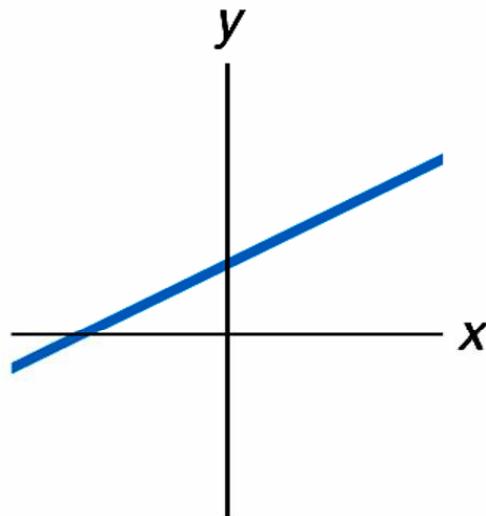
Regressão

Para uma equação de uma recta $y = b_0 + b_1x$, ao valor b_0 chama-se **ordenada na origem** e ao valor b_1 chama-se **declive**.

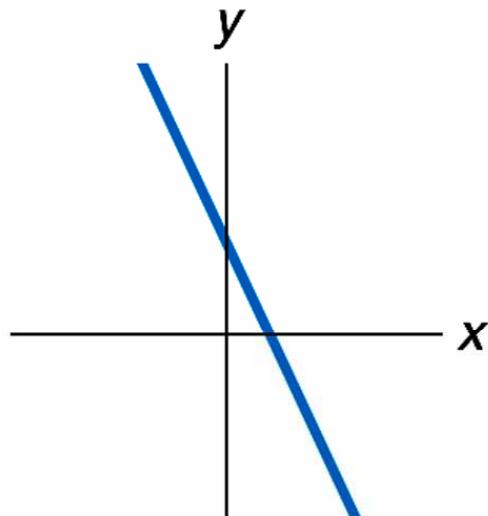


Regressão

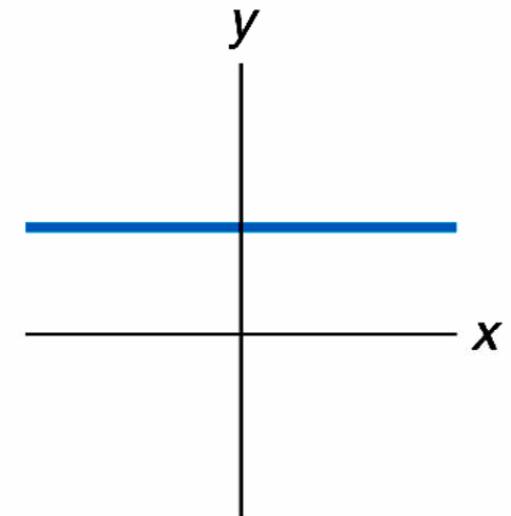
Interpretação gráfica do declive:



$$b_1 > 0$$



$$b_1 < 0$$



$$b_1 = 0$$

Regressão Linear Simples

Um modelo de regressão linear simples descreve uma relação entre duas variáveis quantitativas X, independente, e Y, dependente

$$Y_n = \alpha + \beta \cdot (X_n - \bar{X}) + E_n \Leftrightarrow Y_n = \beta_0 + \beta \cdot X_n + E_n \quad (\beta_0 = \alpha - \beta \cdot \bar{X})$$

(X_n, Y_n) – n – ésima observação de (X, Y) ($n = 1, \dots, N$)

α, β – parâmetros fixos a estimar

E_n – erro aleatório associado a Y_n

Regressão Linear Simples

Aos **valores observados X_n** não estão associados quaisquer erros, devendo ser encarados como constantes.

Os **erros considerados** no modelo de regressão linear simples incidem sobre os valores observados de **Y**.

Na teoria da regressão **admitem-se as seguintes hipóteses** sobre os **erros**:

1. valor esperado nulo e variância constante
2. são mutuamente independentes
3. são normalmente distribuídos

$$E_n \sim IN(0, \sigma^2)$$

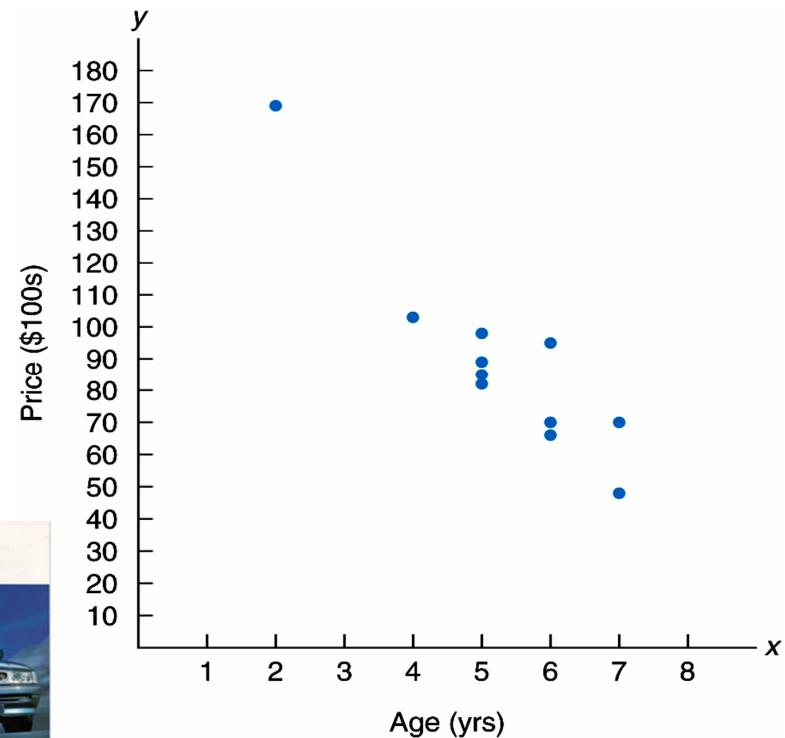
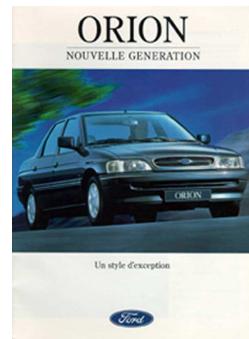
Note que, a hipótese de que existe uma recta de regressão que se ajusta aos dados está implícita em todo o processo.

Régressão Linear Simples

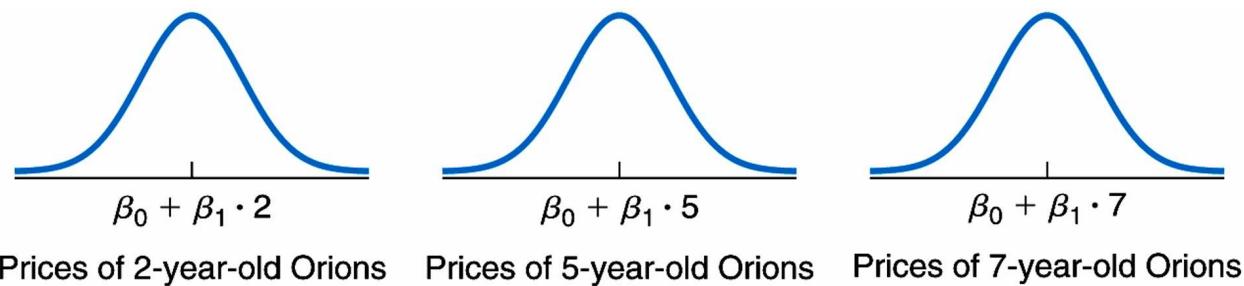
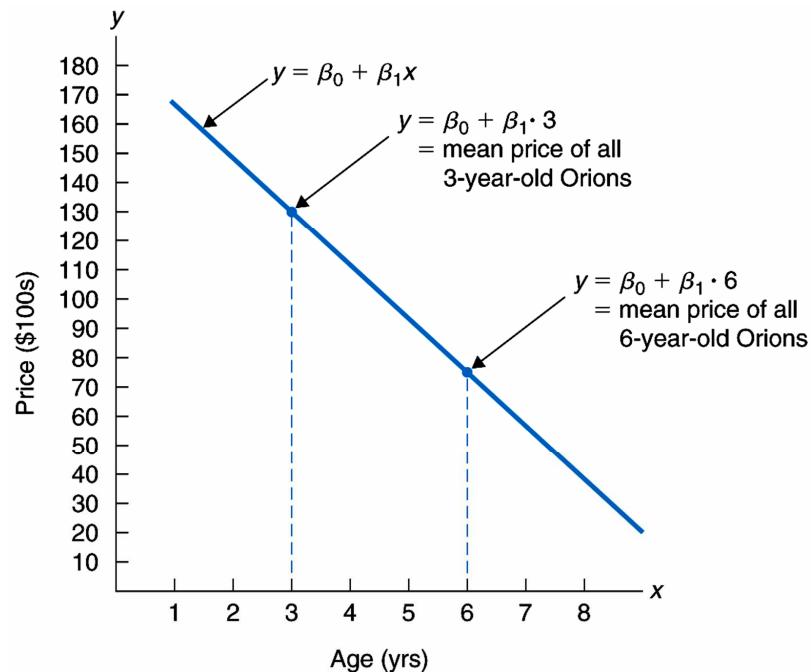
Exemplo:

Idade e preço de uma amostra 11 Orions

Car	Age (yrs) x	Price (\$100s) y
1	5	85
2	4	103
3	6	70
4	5	82
5	5	89
6	5	98
7	6	66
8	6	95
9	2	169
10	7	70
11	7	48

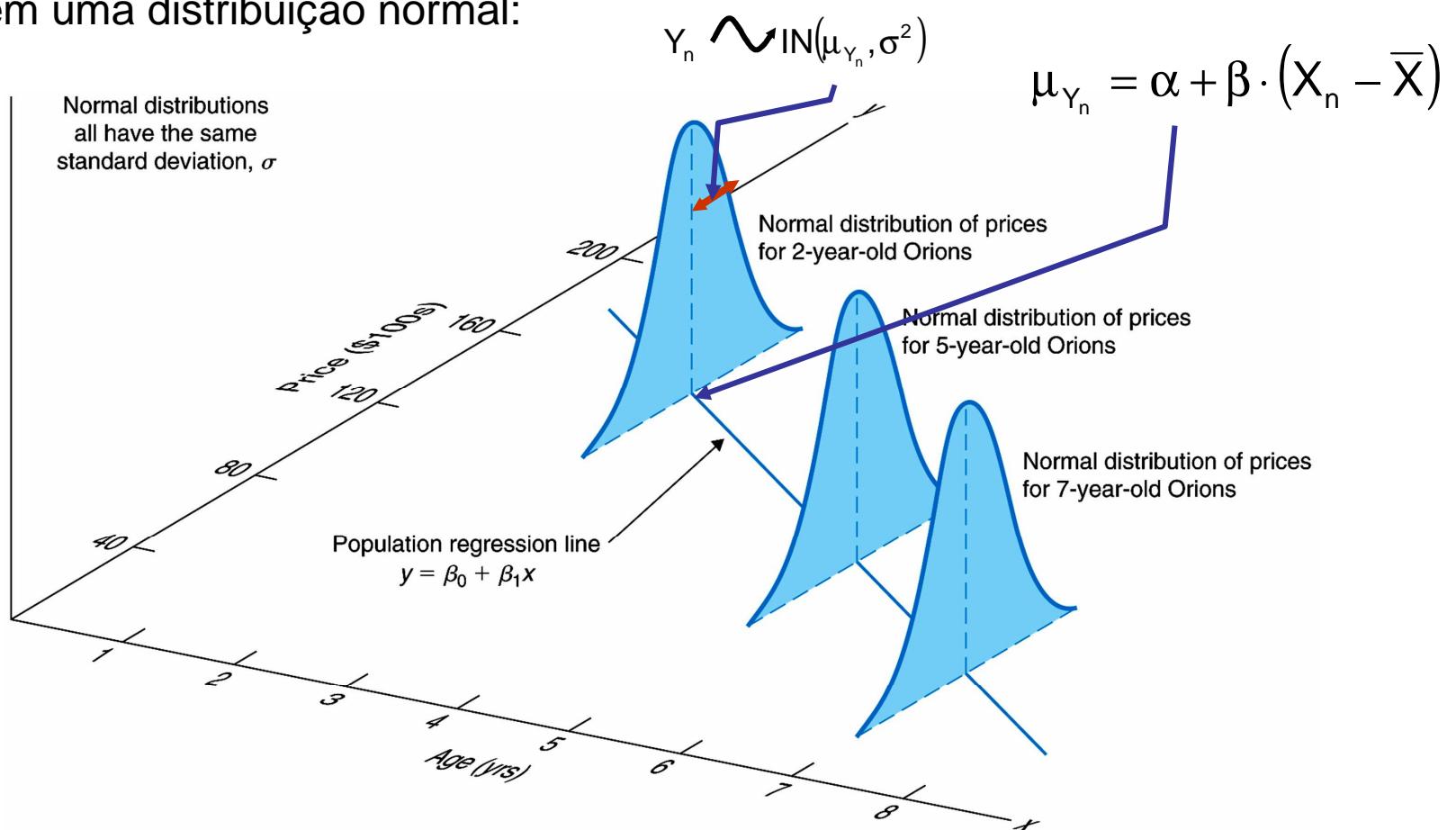


Régressão Linear Simples



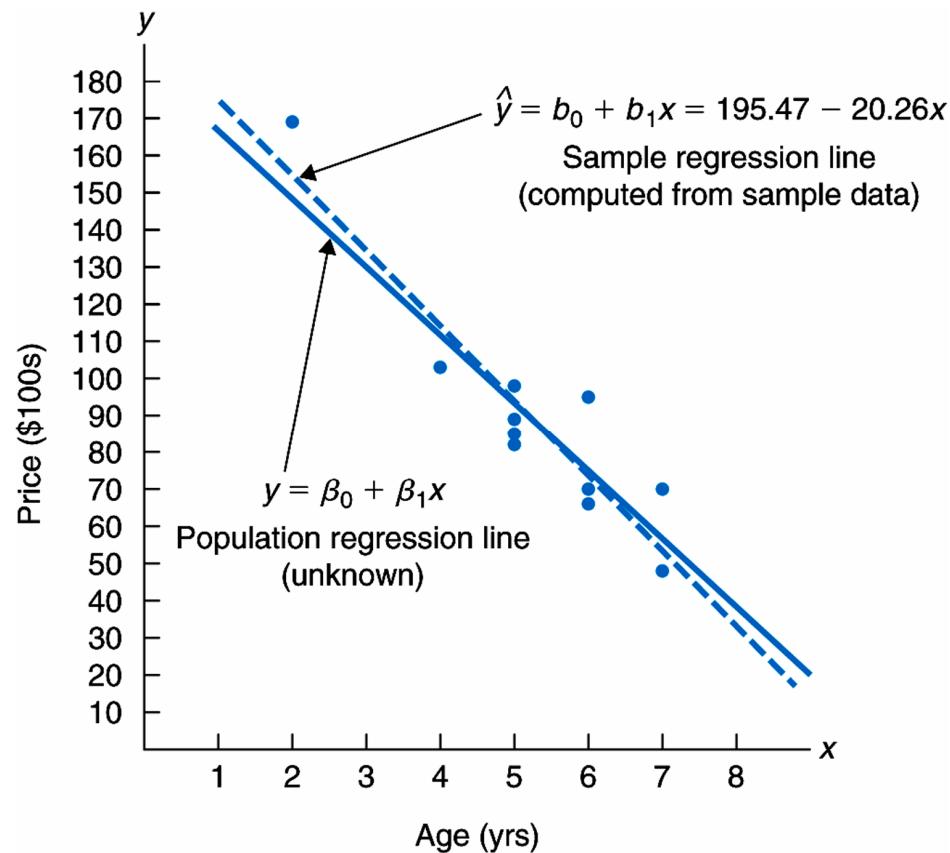
Regressão Linear Simples

Se as hipóteses referidas se verificarem, os valores de Y_n são independentes e seguem uma distribuição normal:



Regressão Linear Simples

A figura assinala a diferença entre a recta de regressão da população (que gostaríamos de conhecer mas não conhecemos) e a recta estimada a partir da amostra.

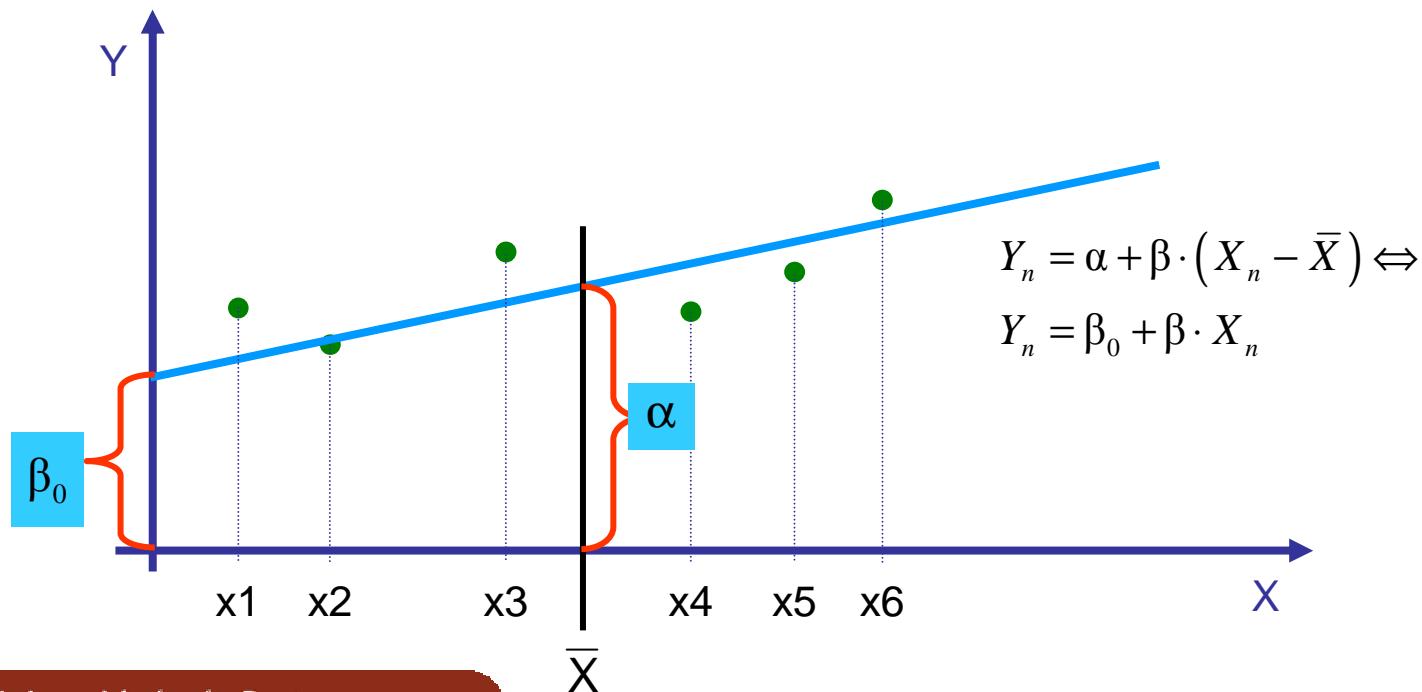


Régressão Linear Simples

Uma vez que os valores X_n são constantes, \bar{X} é também constante.

Logo, $\beta_0 = \alpha - \beta \cdot \bar{X}$ também é constante

A figura ilustra o significado de α e β_0



Regressão Linear Simples

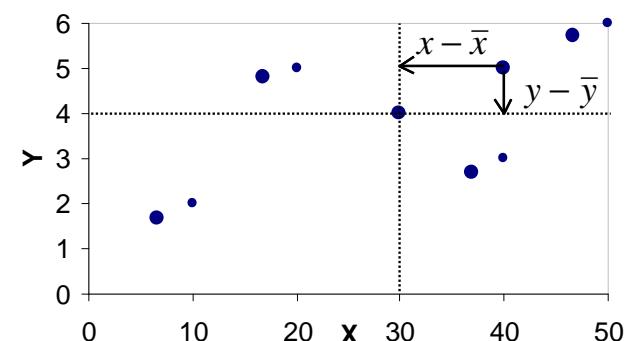
Os parâmetros da recta de regressão podem ser estimados pelo **método dos mínimos quadrados**

$$\text{MIN SEQ} = \sum_{n=1}^N E_n^2 = \sum_{n=1}^N \{Y_n - [\alpha + \beta \cdot (X_n - \bar{X})]\}^2 \Rightarrow \begin{cases} A = \frac{1}{N} \cdot \sum_{n=1}^N Y_n = \bar{Y} \\ B = \frac{\sum_{n=1}^N [(X_n - \bar{X}) \cdot (Y_n - \bar{Y})]}{\sum_{n=1}^N (X_n - \bar{X})^2} = \frac{S_{XY}}{S_{XX}} \end{cases}$$

A partir de um **conjunto particular de observações** (x_n, y_n) obtêm-se as estimativas seguintes:

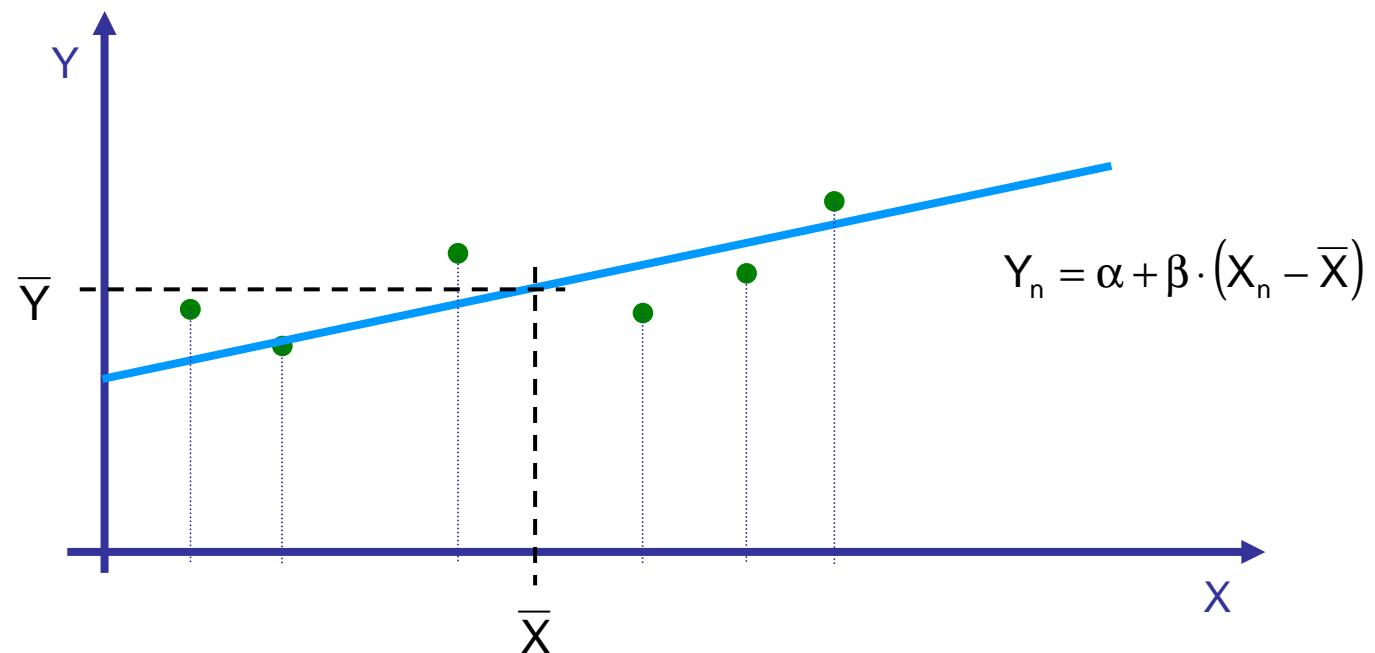
$$a = \hat{\alpha} = \frac{1}{N} \cdot \sum_{n=1}^N y_n = \bar{y}$$

$$b = \hat{\beta} = \frac{S_{XY}}{S_{XX}}$$



Régressão Linear Simples

Note que, a recta estimada passa sempre pelo ponto (\bar{X}, \bar{Y})



$$a = \hat{\alpha} = \bar{y}$$

$$Y_n = \bar{y} + \beta \cdot (\bar{x} - \bar{X}) = \bar{y}$$

Régressão Linear Simples

Se a **relação** entre X_n e μ_{Y_n} for efectivamente linear e os **erros forem independentes**, tiverem **valor esperado nulo** e **variância constante**, pode demonstrar-se que:

1. **A** e **B** são estimadores não-enviesados, eficientes e consistentes
2. A matriz de variância-covariância dos estimadores é

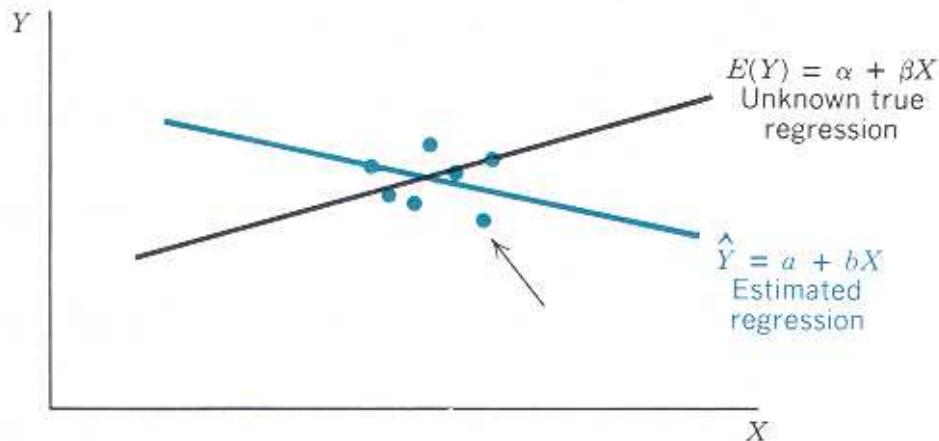
$$\begin{bmatrix} \text{VAR}(A) & \text{COV}(A,B) \\ \text{COV}(B,A) & \text{VAR}(B) \end{bmatrix} = \begin{bmatrix} \sigma^2/N & 0 \\ 0 & \sigma^2/S_{xx} \end{bmatrix}$$

Como $\text{Cov}(a,b)=0$, conclui-se que **A e B não são correlacionados**, o que não acontece com **B_0 e B_1** , uma vez que β_0 é função de β . $\beta_0 = \alpha - \beta \cdot \bar{X}$

Deste facto resulta a vantagem do modelo

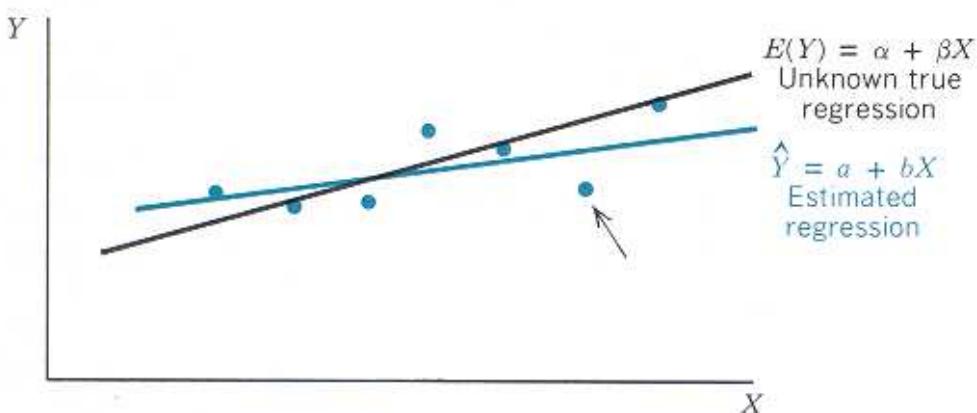
$$Y_n = \alpha + \beta \cdot (X_n - \bar{X}) + E_n \quad \text{relativamente a} \quad Y_n = \beta_0 + \beta \cdot X_n + E_n$$

Regressão Linear Simples



$$Var(B) = \frac{\sigma^2}{S_{XX}}$$

Estimativa de β é melhor quando os valores de x estão mais espalhados.



Régressão Linear Simples

3. Um estimador não-enviesado de σ^2 é

$$S^2 = \frac{1}{N-2} \cdot \sum_{n=1}^N \hat{E}_n^2 = \frac{1}{N-2} \cdot \sum_{n=1}^N [Y_n - A - B \cdot (X_n - \bar{X})]^2$$

Regressão Linear Simples

Exemplo

Admitindo que a relação entre as variáveis X e Y é linear, pretende-se **estimar os parâmetros do modelo de regressão** correspondente:

n	x _n	y _n	N = 5
1	12	33	$\bar{x} = 10$
2	8	24	$\bar{y} = 30$
3	14	39	$s_{xx} = \sum_n (x_n - \bar{x})^2 = 40$
4	10	31	$s_{xy} = \sum_n (x_n - \bar{x}) \cdot (y_n - \bar{y}) = 82$
5	6	23	

$$a = \bar{y} = 30$$

$$b = s_{xy}/s_{xx} = 2.05$$

Podemos também **estimar a variância dos erros e dos estimadores**:

n	y _n	$\mu_{y_n} = a + b \cdot (x_n - \bar{x})$	$\hat{e}_n = y_n - \hat{\mu}_{y_n}$
1	33	$30 + 2.05 \cdot 2 = 34.1$	$33 - 34.1 = -1.1$
2	24	$30 + 2.05 \cdot (-2) = 25.9$	$24 - 25.9 = -1.9$
3	39	$30 + 2.05 \cdot 4 = 38.2$	$39 - 38.2 = 0.8$
4	31	$30 + 2.05 \cdot 0 = 30.0$	$31 - 30.0 = 1.0$
5	23	$30 + 2.05 \cdot (-4) = 21.8$	$23 - 21.8 = 1.2$

$$s^2 = \hat{\sigma}^2 = \frac{1}{N-2} \cdot \sum_{n=1}^N \hat{e}_n^2 = 2.63$$

$$\hat{\sigma}_A^2 = s^2/N = 0.527$$

$$\hat{\sigma}_B^2 = s^2/s_{xx} = 0.0658$$

Régressão Linear Simples

Se $E_n \sim IN(0, \sigma^2)$, é possível especificar as distribuições dos estimadores de α , β e σ^2

$$1. \quad A \sim N(\alpha, \sigma^2/N) \Rightarrow \frac{A - \alpha}{S/\sqrt{N}} \sim t_{N-2}$$

$$\begin{aligned} Y_n &= \alpha + \beta \cdot (X_n - \bar{X}) \Leftrightarrow \\ Y_n &= \beta_0 + \beta \cdot X_n \end{aligned}$$

$$2. \quad B_0 \sim N\left(\beta_0, \left(\frac{1}{N} + \frac{\bar{X}^2}{S_{xx}}\right) \cdot \sigma^2\right) \Rightarrow \frac{B_0 - \beta_0}{S \cdot \sqrt{\frac{1}{N} + \frac{\bar{X}^2}{S_{xx}}}} \sim t_{N-2} \quad (B_0 = A - \bar{X} \cdot B)$$

$$3. \quad B \sim N(\beta, \sigma^2/S_{xx}) \Rightarrow \frac{B - \beta}{S/\sqrt{S_{xx}}} \sim t_{N-2}$$

A partir destas expressões é possível definir **I.C.** e **T.H.**

Régressão Linear Simples

Intervalos de Confiança para os parâmetros de regressão

Os intervalos de confiança bilaterais a $(1 - \gamma) \cdot 100\%$ vêm

$$1. \quad \alpha : A \pm t_{N-2}(\gamma/2) \cdot S \cdot \sqrt{1/N}$$

$$2. \quad \beta_0 : (A - \bar{X} \cdot B) \pm t_{N-2}(\gamma/2) \cdot S \cdot \sqrt{1/N + \bar{X}^2 / S_{xx}}$$

$$3. \quad \beta : B \pm t_{N-2}(\gamma/2) \cdot S \cdot \sqrt{1/S_{xx}}$$

EXEMPLO (CONT.)

$$t_3(0.05/2) = 3.18$$

$$\alpha : 30 \pm 3.18 \cdot \sqrt{2.63} \cdot \sqrt{1/5} \Rightarrow [27.69, 32.31]$$

$$\beta_0 : (30 - 10 \cdot 2.05) \pm 3.18 \cdot \sqrt{2.63} \cdot \sqrt{1/5 + 10^2 / 40} \Rightarrow [1.05, 17.95]$$

$$\beta : 2.05 \pm 3.18 \cdot \sqrt{2.63} \cdot \sqrt{1/40} \Rightarrow [1.23, 2.87]$$

Regressão Linear Simples

Testes de hipóteses para os parâmetros de regressão

Os **testes de hipóteses** relativos aos parâmetros podem ser realizados recorrendo aos intervalos de confiança. Alternativamente, os testes podem ser efectuados pelo **método clássico**:

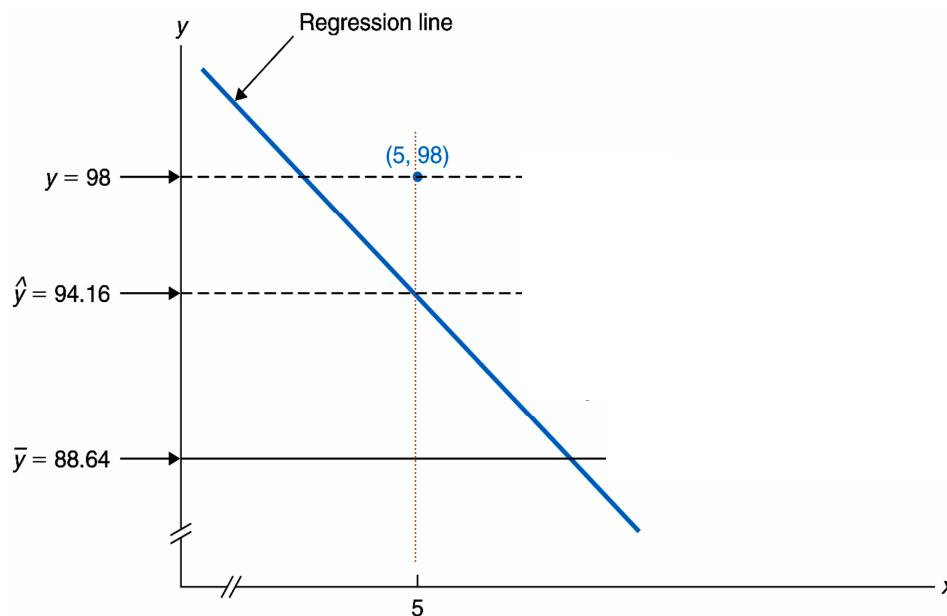
1. $\left. \begin{array}{l} H_0 : \alpha = \alpha_0 \\ H_1 : \alpha \neq \alpha_0, \alpha > \alpha_0, \alpha < \alpha_0 \end{array} \right\} \Rightarrow H_0 \text{ VERD.: } ET = \frac{A - \alpha_0}{S/\sqrt{N}} \sim t_{N-2}$
2. $\left. \begin{array}{l} H_0 : \beta_0 = (\beta_0)_0 \\ H_1 : \beta_0 \neq (\beta_0)_0, \beta_0 > (\beta_0)_0, \beta_0 < (\beta_0)_0 \end{array} \right\} \Rightarrow H_0 \text{ VERD.: } ET = \frac{A - \bar{X} \cdot B - (\beta_0)_0}{S \cdot \sqrt{\frac{1}{N} + \frac{\bar{X}^2}{S_{xx}}}} \sim t_{N-2}$
3. $\left. \begin{array}{l} H_0 : \beta = \beta_0 \\ H_1 : \beta \neq \beta_0, \beta > \beta_0, \beta < \beta_0 \end{array} \right\} \Rightarrow H_0 \text{ VERD.: } ET = \frac{B - \beta_0}{S/\sqrt{S_{xx}}} \sim t_{N-2}$

Regressão Linear Simples

Quando se pretende realizar um **teste bilateral** à hipótese nula $H_0: \beta=0$, pode recorrer-se a um procedimento baseado na **ANOVA**

$$\sum_n (Y_n - \bar{Y})^2 = \underbrace{\sum_n \{[A + B \cdot (X_n - \bar{X})] - \bar{Y}\}^2}_{\text{V. DEVIDA À REGRESSÃO } VDR=B^2 \cdot S_{XX}} + \underbrace{\sum_n \{Y_n - [A + B \cdot (X_n - \bar{X})]\}^2}_{\text{V. RESIDUAL } VR=S_{YY}-B^2 \cdot S_{XX}}$$

$\underbrace{\sum_n (Y_n - \bar{Y})^2}_{\text{V. TOTAL } VT=S_{YY}}$



Regressão Linear Simples

Tabela ANOVA para o modelo de regressão linear simples

FONTES DE VARIAÇÃO	VARIAÇÕES (Somas de quadrados)	GRAUS DE LIBERDADE (Número de termos independentes)	DESVIOS QUADRÁTICOS MÉDIOS	VALORES ESPERADOS
DEVIDA À REGRESSÃO (DR)	$VDR = B^2 \cdot S_{XX}$ $= B \cdot S_{XY}$ <i>(Variação explicada pela regressão)</i>	$GL_1 = 1$	$DQMDR = VDR$	$E [DQMDR] = \sigma^2 + \beta^2 \cdot S_{XX}$
RESIDUAL (R)	$VR = S_{YY} - B \cdot S_{XY}$ <i>(Variação residual, não explicada)</i>	$GL_2 = N - 2$	$DQMR = VR/GL_2$	$E [DQMR] = \sigma^2$
TOTAL (T)	$VT = S_{YY}$ <i>(Variação total)</i>	$GL = GL_1 + GL_2 = N - 1$		

O procedimento de teste ANOVA tem a seguinte estrutura:

$$\left. \begin{array}{l} H_0 : \beta = 0 \\ H_1 : \beta \neq 0 \end{array} \right\} \Rightarrow H_0 \quad \text{VERD.: } ET = \frac{DQMDR}{DQMR} \sim F_{1,N-2}$$

Regressão Linear Simples

Exemplo (anterior)

Admitindo que a relação entre as variáveis X e Y é linear, pretende-se estimar os parâmetros do modelo de regressão correspondente

n	x _n	y _n
1	12	33
2	8	24
3	14	39
4	10	31
5	6	23

$$a = \bar{y} = 30$$

$$b = s_{XY}/s_{XX} = 2.05$$

Régressão Linear Simples

Exemplo (cont.)

$$H_0 : \beta = 0 \quad H_1 : \beta \neq 0 \quad \alpha = 5\%$$

I.C.: $\beta \in 2.05 \pm 3.18 \cdot \sqrt{2.63} \cdot \sqrt{1/40} \Rightarrow [1.23, 2.87]$

T.H.: $ET = \frac{2.05}{1.62/\sqrt{40}} = 7.99 > t_3(0.05/2) = 3.18$ (Valor de prova $P = 0.41\%$)

ANOVA:

FONTES	VARIACÕES	G.L.	DQM
DR	168.1	1	168.1
R	7.9	3	2.633
T	176.0	4	

$ET = \frac{168.1}{2.633} = 63.84 > F_{1,3}(0.05) = 10.13$ (Valor de prova $P = 0.41\%$)

Rejeitar H_0

Regressão

Previsões com base no modelo de regressão linear simples

Para cada valor de X , a **melhor previsão de Y** é dada por

$$\hat{Y} = \hat{\mu}_Y = A + B \cdot (X - \bar{X})$$

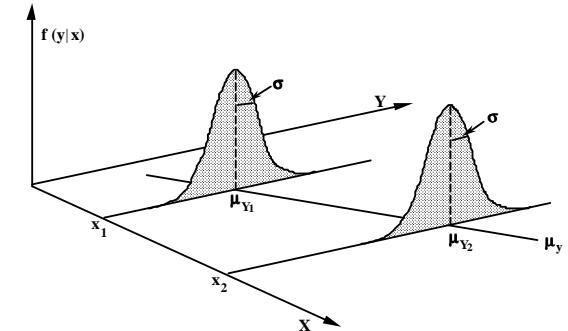
O **erro** que se comete na **previsão** vem

$$\delta = Y - \hat{Y} = [\alpha + \beta \cdot (X - \bar{X}) + E] - [A + B \cdot (X - \bar{X})]$$

Dado que cada novo valor Y se admite independente dos anteriores, α , β , X e \bar{X} são constantes e A e B são independentes, pode-se mostrar que a variância do erro de previsão vem:

$$\text{VAR}(\delta) = \text{VAR}(Y) + \text{VAR}(\hat{Y}) = \left[1 + \frac{1}{N} + \frac{(X - \bar{X})^2}{S_{xx}} \right] \cdot \sigma^2$$

$$\begin{bmatrix} \text{VAR}(A) & \text{COV}(A,B) \\ \text{COV}(B,A) & \text{VAR}(B) \end{bmatrix} = \begin{bmatrix} \sigma^2/N & 0 \\ 0 & \sigma^2/S_{xx} \end{bmatrix}$$



Régressão

Admitindo a normalidade dos erros E_n , temos que Y , \hat{Y} e δ seguem também distribuições Normais.

Assim, o **intervalo de previsão a $(1-\gamma) \cdot 100\%$** será:

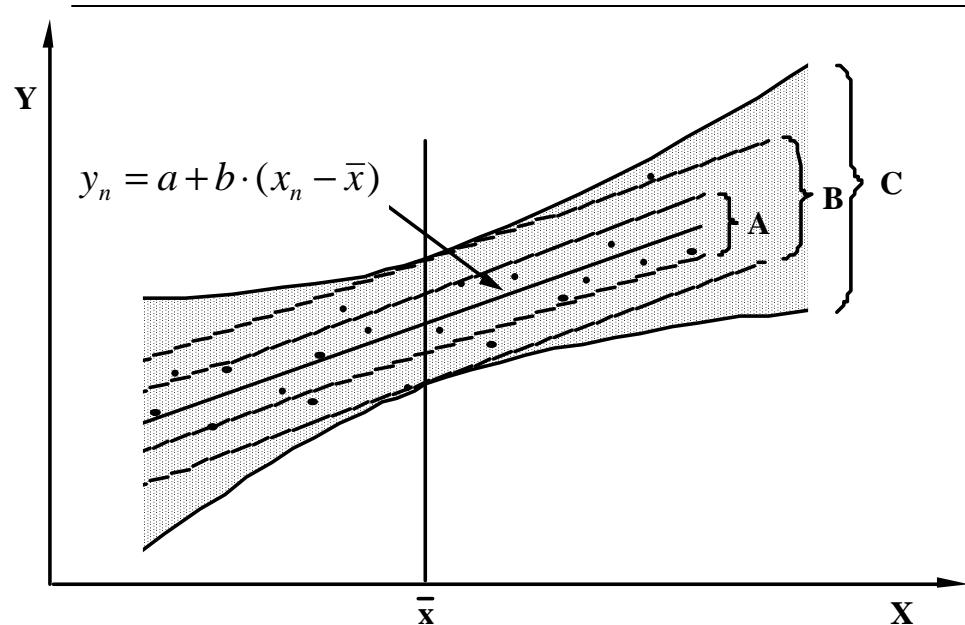
$$\hat{Y} \pm t_{N-2}(\gamma/2) \cdot S \cdot \sqrt{1 + \frac{1}{N} + \frac{(x - \bar{x})^2}{S_{xx}}}$$

Quando se considera todos os valores possíveis de X , os intervalos formam uma banda de previsão.

Se $t_{n-2}(\gamma/2)=1$ a banda de previsão a $(1-\gamma) \cdot 100\%$ é definida por:

$$\hat{Y} \pm S \cdot \sqrt{1 + \frac{1}{N} + \frac{(x - \bar{x})^2}{S_{xx}}}$$

Regressão



BANDA A: $\hat{Y} \pm S \cdot \sqrt{1}$

BANDA B: $\hat{Y} \pm S \cdot \sqrt{1 + \frac{1}{N}}$

BANDA C: $\hat{Y} \pm S \cdot \sqrt{1 + \frac{1}{N} + \frac{(x - \bar{x})^2}{S_{xx}}}$

Banda A - tem apenas a ver com a estimativa do **desvio-padrão do erro E**

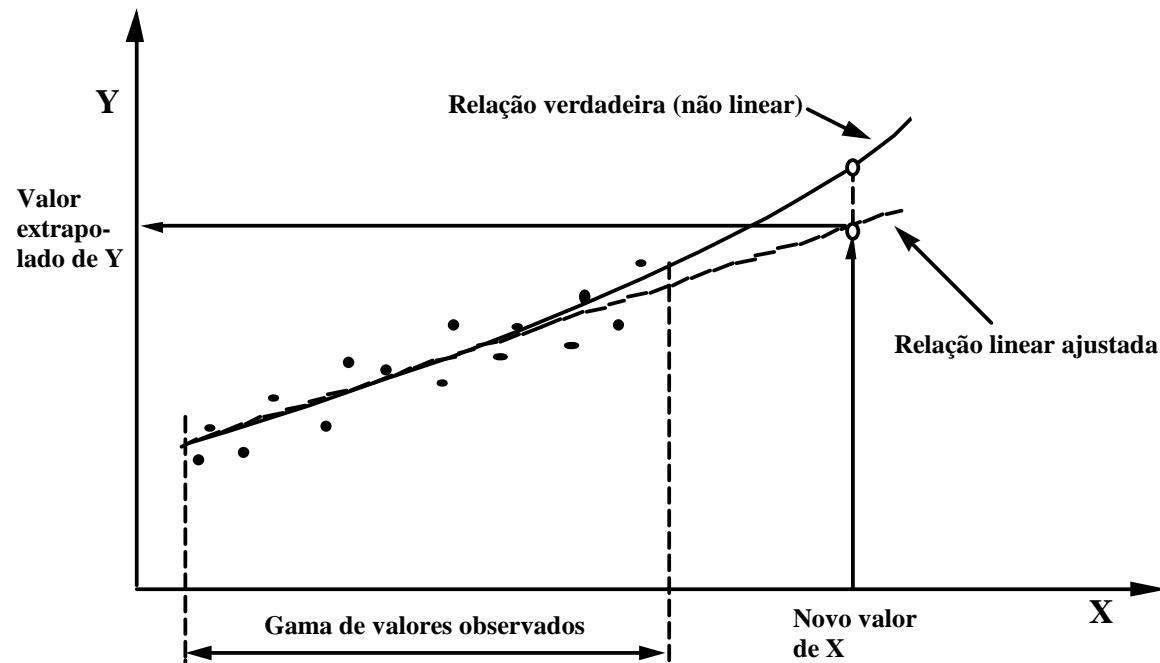
Banda B - acrescenta, relativamente à banda A, o termo correspondente ao **erro de estimação de α**

Banda C - acrescenta, relativamente à banda B, o termo correspondente ao **erro de estimação de β**

Régressão

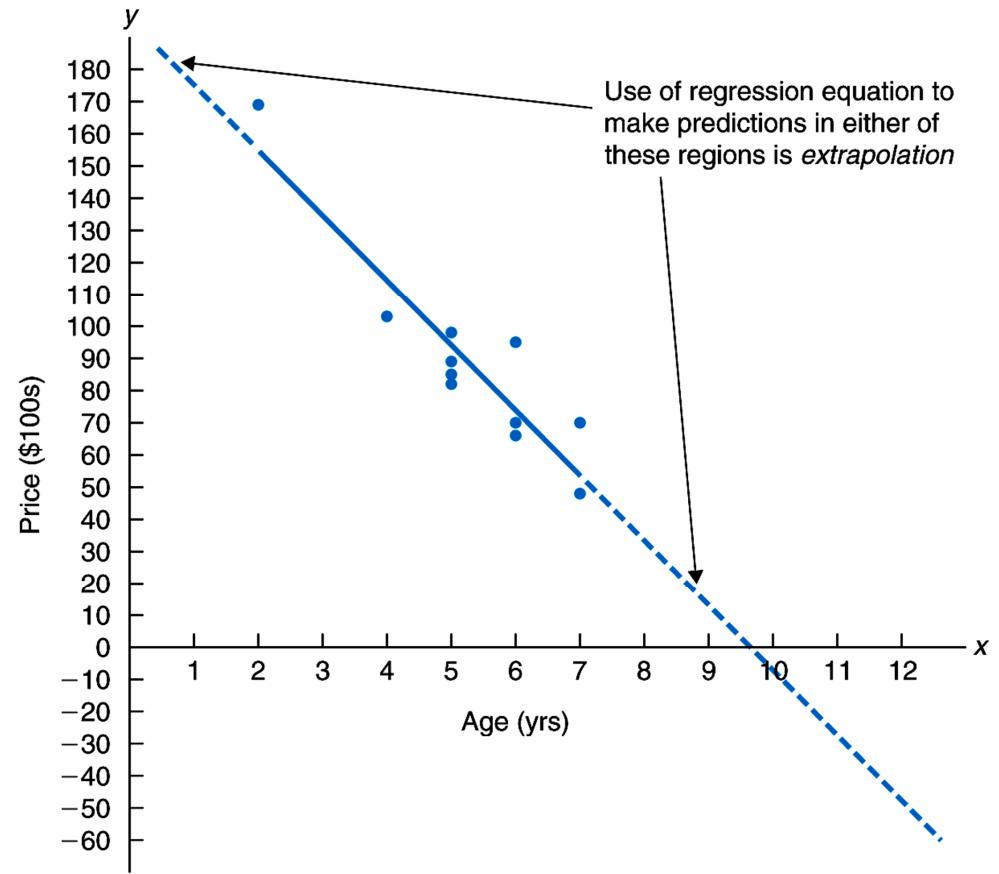
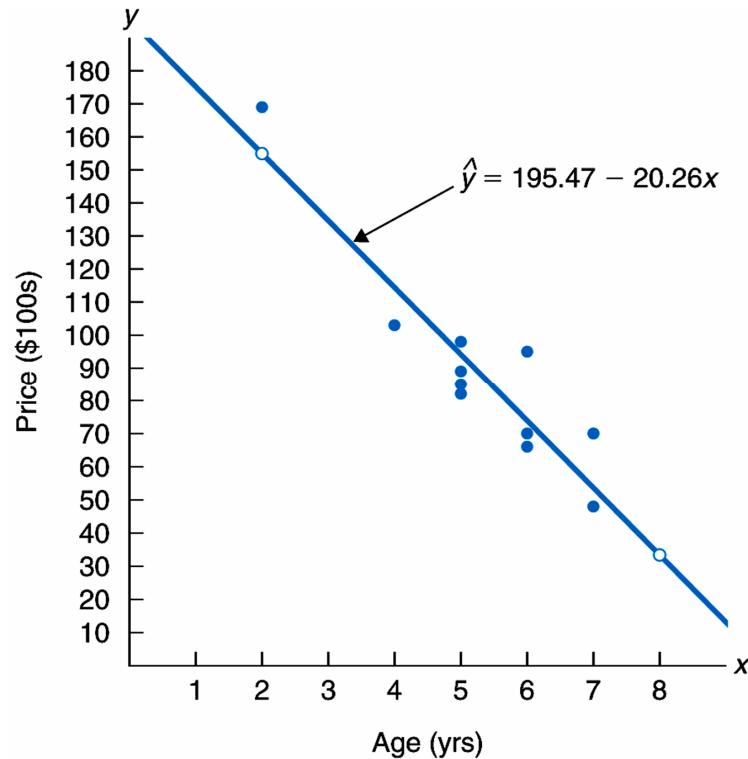
Uma relação que pareça linear dentro da gama de valores observados X_n , pode ter um comportamento não-linear numa gama mais alargada.

Apesar de a banda de previsão definida alargar à medida que as observações se afastam da média de X, esta não contempla esse tipo de situações, assentando no pressuposto de que a relação é efectivamente linear.



Regressão

Extrapolação no exemplo do Orion



Regressão Linear Simples

Regressão linear simples e correlação entre variáveis

Embora a **análise de correlação** seja uma técnica menos potente do que a **regressão linear simples**, pois apenas revela o grau de relacionamento linear entre variáveis sem especificar a forma que ele assume, ambas estão intimamente ligadas.

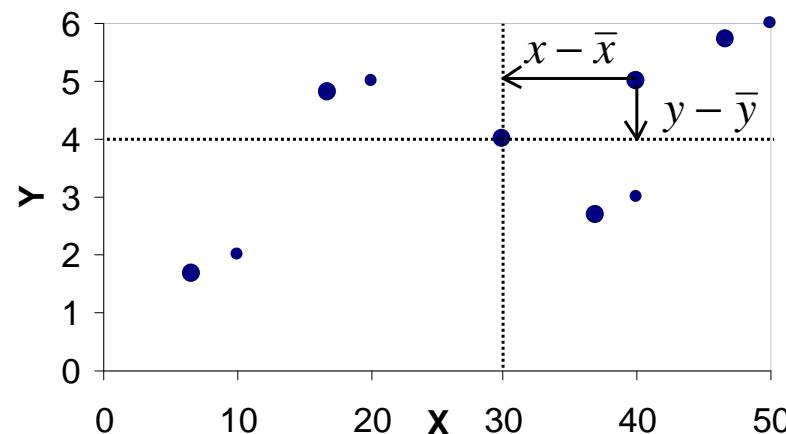
Na **regressão linear simples**, os valores observados de **X** são encarados como **constantes**, podendo não ser representativos de uma qualquer distribuição populacional

Na **análise de correlação**, para que a partir do coeficiente de correlação amostral se possam fazer inferências relativas ao coeficiente de correlação populacional, é preciso que as observações (X_n, Y_n) sejam representativas da população conjunta de X e Y

Régressão Linear Simples

Covariância amostral (permite inferir acerca da população)

$$c_{XY} = \frac{1}{N-1} \cdot \sum_{n=1}^N (x_n - \bar{x}) \cdot (y_n - \bar{y})$$

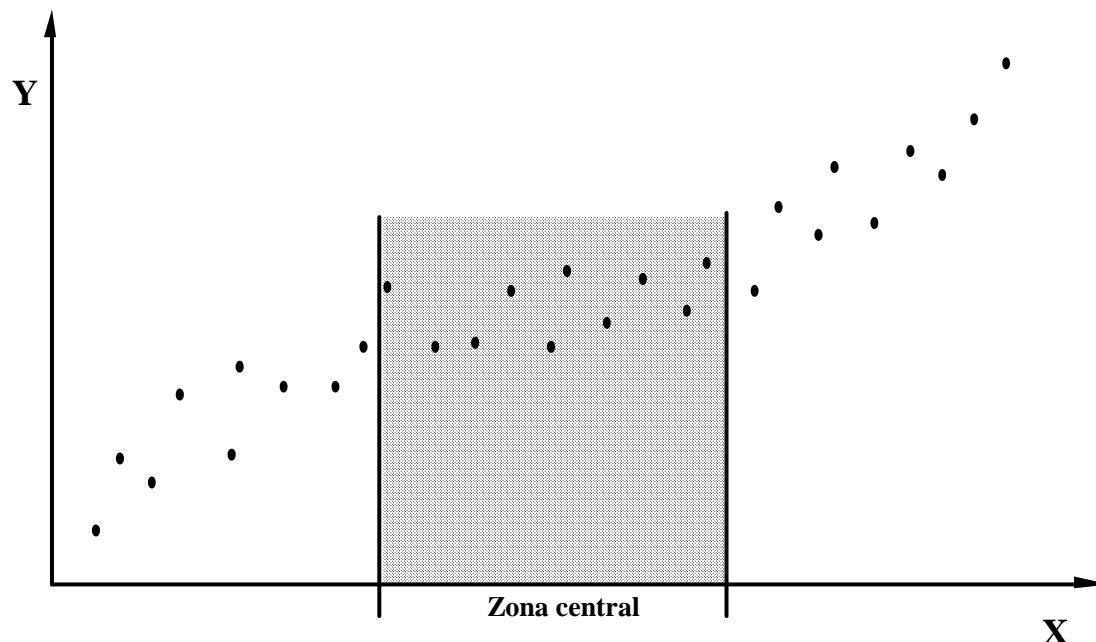


Coeficiente de correlação amostral (medida adimensional)

$$r_{XY} = \frac{\frac{1}{N-1} \cdot \sum_{n=1}^N (x_n - \bar{x}) \cdot (y_n - \bar{y})}{\sqrt{\frac{1}{N-1} \cdot \sum_{n=1}^N (x_n - \bar{x})^2} \cdot \sqrt{\frac{1}{N-1} \cdot \sum_{n=1}^N (y_n - \bar{y})^2}} = \frac{c_{XY}}{s_X \cdot s_Y} \quad (-1 \leq r_{XY} \leq 1)$$

Régressão Linear Simples

Se os valores X_n não forem obrigatoriamente representativos da população de X , podem situar-se numa zona restrita dessa população, provocando o enviesamento do coeficiente de correlação amostral



Enviesamento do coeficiente de correlação amostral calculado a partir de observações pertencentes à zona central da população de x

Régressão Linear Simples

Na análise de correlação não se faz qualquer distinção entre variável dependente e variável independente

A existência de correlação implica que

- X é causa de Y, ou
- Y é causa de X, ou ainda
- Uma outra variável é causa simultânea de X e Y

Que sentido fará, no contexto da regressão, calcular o coeficiente de correlação amostral R_{XY} ?

Regressão Linear Simples

Se X for claramente assumida como variável predeterminada, o cálculo do quadrado do coeficiente de correlação amostral, o **coeficiente de determinação**, representa a proporção da variação de y que é explicada pela regressão

$$R_{XY}^2 = \frac{B^2 \cdot S_{XX}}{S_{YY}} = \frac{B^2 \cdot \sum_n (X_n - \bar{X})^2}{\sum_n (Y_n - \bar{Y})^2}$$

$$\begin{aligned} S_{YY} &= \sum_{n=1}^N (y_n - \bar{y})^2 = \sum_{n=1}^N [(y_n - \hat{y}_n) + (\hat{y}_n - \bar{y})]^2 \\ &= \dots = \underbrace{\sum_{n=1}^N e_n^2}_{\text{Não-}} + \underbrace{b^2 \cdot \sum_{n=1}^N (x_n - \bar{x})^2}_{\text{-explicada}} \end{aligned}$$

Régressão Linear Simples

O montante global dos seguros de vida efectuados pelas famílias de um determinado país depende do rendimento anual do agregado familiar. Na tabela seguinte apresentam-se os valores destas variáveis, expressas em unidades monetárias do país em causa, para um conjunto de 12 famílias considerado representativo da população.

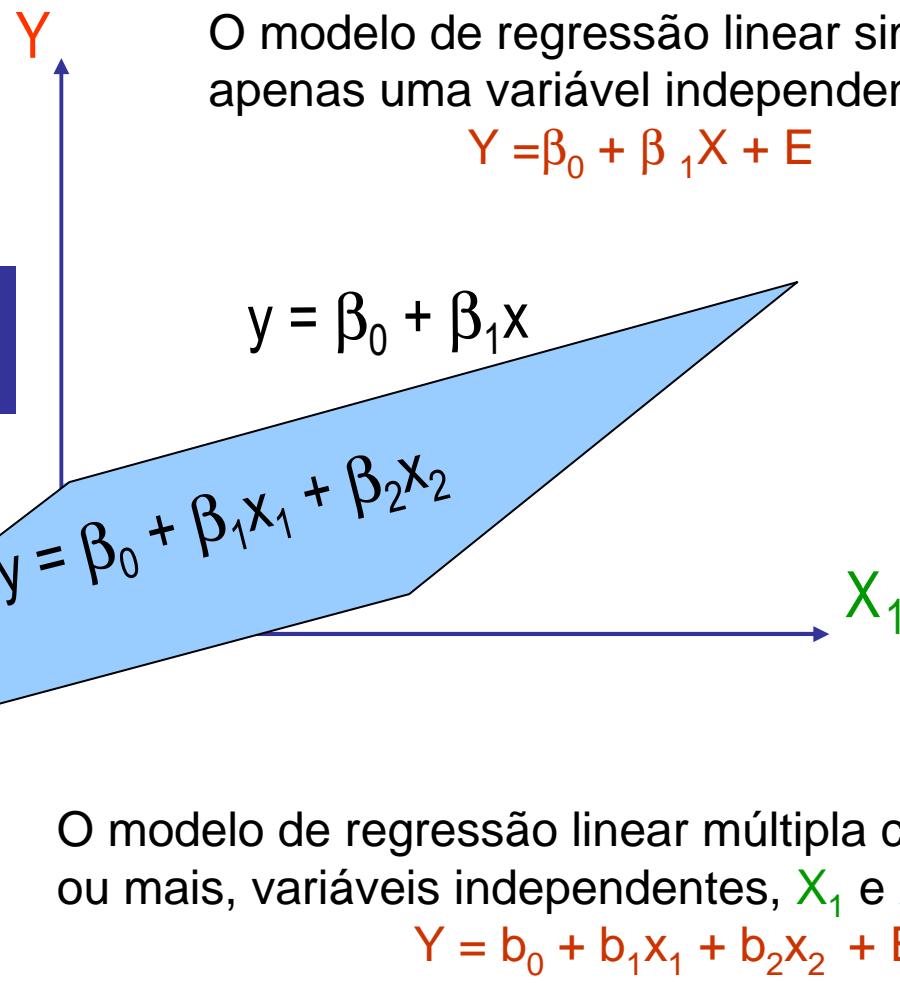
Rendimento anual [1000 u.m.]	Capital seguro [1000 u.m.]
14	31
19	40
23	49
12	20
9	21
15	34
22	54
25	52
15	28
10	21
12	24
16	34

Estime a relação entre as duas variáveis e o intervalo de previsão do montante global de seguros de vida efectuados por uma família com um rendimento anual agregado de 20 000 [u.m]

Regressão Linear Múltipla

- O modelo de regressão linear múltipla é uma extensão do modelo de regressão linear simples.
- Permite descrever uma relação entre um **conjunto** de variáveis quantitativas independentes, X_j ($j=1,2,\dots,J$), e uma variável, Y , quantitativa dependente.
- O objectivo será o de construir um modelo que se ajuste melhor aos dados do que o modelo de regressão linear simples.

Regressão Linear Múltipla



O modelo de regressão linear múltipla considera duas, ou mais, variáveis independentes, X_1 e X_2 .

$$Y = b_0 + b_1x_1 + b_2x_2 + E$$

Regressão Linear Múltipla

O modelo de regressão linear múltipla :

$$Y_n = \alpha + \beta_1 \cdot (X_{1n} - \bar{X}_1) + \cdots + \beta_J \cdot (X_{Jn} - \bar{X}_J) + E_n$$

$(X_{1n}, \dots, X_{Jn}, Y_n)$ n-ésima observação das variáveis X_1, \dots, X_J e Y

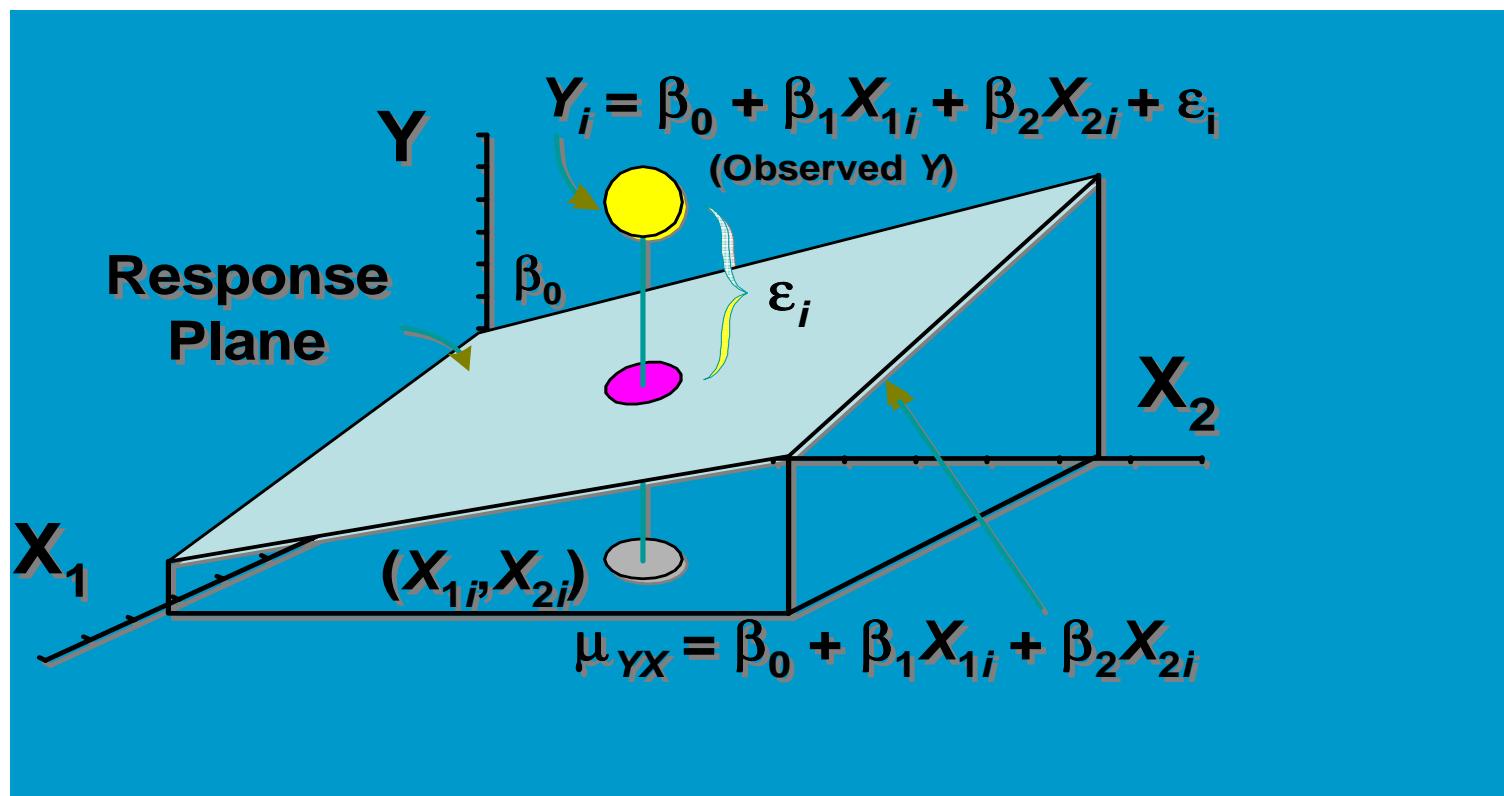
\bar{X}_j média das observações da variável X_j

$\alpha, \beta_1, \dots, \beta_J$ parâmetros fixos a estimar

E_n erro aleatório associado a Y_n

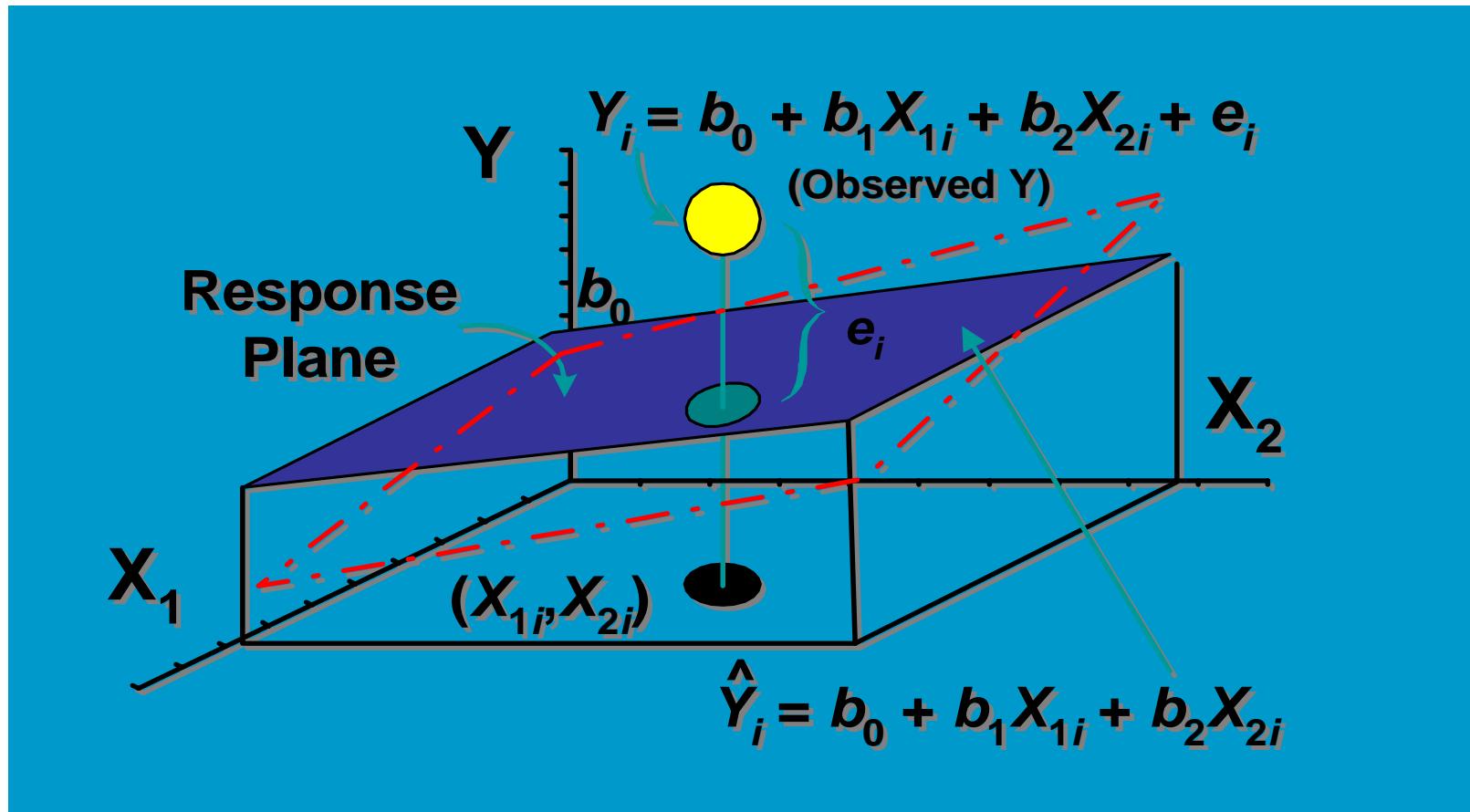
Regressão Linear Múltipla

Modelo de regressão linear **populacional** bivariado



Regressão Linear Múltipla

Modelo de regressão linear **amostral** bivariado



Régressão Linear Múltipla

A este modelo **estão subjacentes as seguintes hipóteses:**

1. Os valores X_{jn} , e portanto, os seus valores esperados são encarados como constantes predeterminadas, sem erro
2. Os erros E_n são mutuamente independentes, têm valor esperado nulo, variância constante e são normalmente distribuídos, isto é,

$$E_n \sim IN(0, \sigma^2)$$

Régressão Linear Múltipla

Os parâmetros $\alpha, \beta_1, \dots, \beta_J$ podem ser estimados recorrendo ao método dos mínimos quadrados minimizando a seguinte função:

$$SEQ = \sum_n E_n^2 = \sum_n \{Y_n - [\alpha + \beta_1 \cdot (X_{1n} - \bar{X}_1) + \dots + \beta_J \cdot (X_{Jn} - \bar{X}_J)]\}^2$$

Sendo o mínimo atingido para

$$\frac{\partial SEQ}{\partial \alpha} = (-2) \cdot \sum_n [Y_n - \alpha - \beta_1 \cdot (X_{1n} - \bar{X}_1) - \dots - \beta_J \cdot (X_{Jn} - \bar{X}_J)] = 0$$

$$\frac{\partial SEQ}{\partial \beta_1} = (-2) \cdot \sum_n \{(X_{1n} - \bar{X}_1) \cdot [Y_n - \alpha - \beta_1 \cdot (X_{1n} - \bar{X}_1) - \dots - \beta_J \cdot (X_{Jn} - \bar{X}_J)]\} = 0$$

(...)

$$\frac{\partial SEQ}{\partial \beta_J} = (-2) \cdot \sum_n \{(X_{Jn} - \bar{X}_J) \cdot [Y_n - \alpha - \beta_1 \cdot (X_{1n} - \bar{X}_1) - \dots - \beta_J \cdot (X_{Jn} - \bar{X}_J)]\} = 0$$

Régressão Linear Múltipla

A primeira equação permite obter o **estimador de α** , que é idêntico ao que foi definido para o modelo de regressão linear simples:

$$\alpha = \frac{1}{N} \cdot \sum_n Y_n = \bar{Y}$$

Desenvolvendo as restantes equações, obtém-se o seguinte sistema cuja resolução permite obter **os estimadores de β_1, \dots, β_J**

$$B_1 \cdot S_{X_1 X_1} + B_2 \cdot S_{X_1 X_2} + \dots + B_J \cdot S_{X_1 X_J} = S_{X_1 Y}$$

$$B_1 \cdot S_{X_2 X_1} + B_2 \cdot S_{X_2 X_2} + \dots + B_J \cdot S_{X_2 X_J} = S_{X_2 Y}$$

(...)

$$B_1 \cdot S_{X_J X_1} + B_2 \cdot S_{X_J X_2} + \dots + B_J \cdot S_{X_J X_J} = S_{X_J Y}$$

onde $S_{X_{j_1} X_{j_2}} = \sum_n (X_{j_1 n} - \bar{X}_{j_1}) \cdot (X_{j_2 n} - \bar{X}_{j_2})$ e $S_{X_j Y} = \sum_n (X_{j n} - \bar{X}_j) \cdot (Y_n - \bar{Y})$

Régressão Linear Múltipla

Se a relação entre as variáveis X_j e μ_y for linear e se os erros E_n forem independentes, tiverem valor esperado nulo e variância constante pode demonstrar-se que:

1. Os estimadores A e B_1, \dots, B_J são **não-enviesados, eficientes e consistentes**.
2. A **matriz de variância-covariância** dos estimadores A e B_1, \dots, B_J é:

$$\begin{bmatrix} \text{Var}(A) & \text{Cov}(A, B_1) & \cdots & \text{Cov}(A, B_J) \\ \text{Cov}(B_1, A) & \text{Var}(B_1) & \cdots & \text{Cov}(B_1, B_J) \\ \cdots & \cdots & \cdots & \cdots \\ \text{Cov}(B_J, A) & \text{Cov}(B_J, B_1) & \cdots & \text{Var}(B_J) \end{bmatrix} = \sigma^2 \cdot \begin{bmatrix} N & 0 & \cdots & 0 \\ 0 & S_{x_1 x_1} & \cdots & S_{x_1 x_J} \\ \cdots & \cdots & \cdots & \cdots \\ 0 & S_{x_J x_1} & \cdots & S_{x_J x_J} \end{bmatrix}^{-1}$$

Régressão Linear Múltipla

3. Um estimador não-enviesado de σ^2 é definido pela expressão:

$$\begin{aligned} S^2 &= \frac{1}{N-J-1} \cdot \sum_n \hat{\epsilon}^2 = \\ &= \frac{1}{N-J-1} \cdot \sum_n [Y_n - A - B_1 \cdot (X_{1n} - \bar{X}_1) - \dots - B_J \cdot (X_{Jn} - \bar{X}_J)]^2 \end{aligned}$$

Regressão Linear Múltipla

Exemplo

Considerem-se as observações das variáveis X_1 , X_2 e Y que constam da tabela.

Admitindo que o valor esperado de Y é uma função linear de X_1 e X_2 , estimem-se os parâmetros do modelo de regressão correspondente.

n	x_{1n}	x_{2n}	y_n	$N=10$
1	5.0	7.2	51.7	$\bar{x}_1 = \frac{1}{10} \cdot (5.0 + 5.8 + \dots + 4.6) = 5.10$
2	5.8	7.8	56.4	
3	4.2	8.1	49.3	$\bar{x}_2 = \frac{1}{10} \cdot (7.2 + 7.8 + \dots + 6.9) = 7.65$
4	6.0	8.7	60.7	
5	4.8	6.6	48.9	
6	5.6	7.5	54.1	$\bar{y} = \frac{1}{10} \cdot (51.7 + 56.4 + \dots + 50.4) = 53.41$
7	4.4	9.0	54.9	
8	5.2	6.3	49.8	$S_{x_1 x_1} = \sum_i (x_{1i} - \bar{x}_1)^2 = (5.0 - 5.1)^2 + \dots + (4.6 - 5.1)^2 = 3.3$
9	5.4	8.4	57.9	
10	4.6	6.9	50.4	$S_{x_2 x_2} = \sum_i (x_{2i} - \bar{x}_2)^2 = (7.2 - 7.65)^2 + \dots + (6.9 - 7.65)^2 = 7.42$

$$S_{x_1 x_2} = S_{x_2 x_1} = \sum_i (x_{1i} - \bar{x}_1) \cdot (x_{2i} - \bar{x}_2) = (5.0 - 5.1) \cdot (7.2 - 7.65) + \dots + (4.6 - 5.1) \cdot (6.9 - 7.65) = 0.45$$

$$S_{x_1 y} = \sum_i (x_{1i} - \bar{x}_1) \cdot (y_i - \bar{y}) = (5.0 - 5.1) \cdot (51.7 - 53.41) + \dots + (4.6 - 5.1) \cdot (50.4 - 53.41) = 15.67$$

$$S_{x_2 y} = \sum_i (x_{2i} - \bar{x}_2) \cdot (y_i - \bar{y}) = (7.2 - 7.65) \cdot (51.7 - 53.41) + \dots + (6.9 - 7.65) \cdot (50.4 - 53.41) = 24.16$$

$$S_{yy} = \sum_i (y_i - \bar{y})^2 = (51.7 - 53.41)^2 + \dots + (50.4 - 53.41)^2 = 147.19$$

Regressão Linear Múltipla

Exemplo (cont.)

As estimativas dos parâmetros de regressão podem então obter-se nos seguintes termos:

$$a = \hat{\alpha} = \bar{y} = 53.41$$

$$\begin{cases} b_1 \cdot S_{x_1 x_1} + b_2 \cdot S_{x_1 x_2} = S_{x_1 y} \\ b_1 \cdot S_{x_2 x_1} + b_2 \cdot S_{x_2 x_2} = S_{x_2 y} \end{cases} \Rightarrow \begin{cases} 3.3 \cdot b_1 + 0.45 \cdot b_2 = 15.67 \\ 0.45 \cdot b_1 + 7.42 \cdot b_2 = 24.16 \end{cases} \Rightarrow \begin{cases} b_1 = \hat{\beta}_1 = 4.34 \\ b_2 = \hat{\beta}_2 = 2.99 \end{cases}$$

Régressão Linear Múltipla

A partir de $\hat{\mu}_{Y_i} = 53.41 + 4.34 \cdot (x_{1i} - 5.10) + 2.99 \cdot (x_{2i} - 7.65)$ pode-se calcular a estimativa de σ^2

n	y_n	$\hat{\mu}_{Y_n}$	$\hat{e}_i = y_i - \hat{\mu}_{Y_n}$
1	51.7	51.630	0.070
2	56.4	56.897	-0.497
3	49.3	50.850	-1.550
4	60.7	60.458	0.242
5	48.9	48.967	-0.067
6	54.1	55.132	-1.032
7	54.9	54.410	0.490
8	49.8	49.306	-0.006
9	57.9	56.956	0.944
10	50.4	48.996	1.404

$$s^2 = \hat{\sigma}^2 = \frac{1}{N-J-1} \cdot \sum_n \hat{e}_n^2 = \frac{1}{10-2-1} \cdot [0.07^2 + \dots + 1.404^2] = 0.983$$

e estimar a variância dos estimadores:

$$\begin{bmatrix} \hat{\text{Var}}(B_1) & \hat{\text{Cov}}(B_1, B_2) \\ \hat{\text{Cov}}(B_2, B_1) & \hat{\text{Var}}(B_2) \end{bmatrix} = \hat{\sigma}^2 \cdot \begin{bmatrix} S_{x_1 x_1} & S_{x_1 x_2} \\ S_{x_2 x_1} & S_{x_2 x_2} \end{bmatrix}^{-1} = 0.983 \cdot \begin{bmatrix} 3.30 & 0.45 \\ 0.45 & 7.42 \end{bmatrix}^{-1} = \begin{bmatrix} 0.300 & -0.018 \\ -0.018 & 0.133 \end{bmatrix}$$

$$\hat{\text{Var}}(B_1) = 0.300 \Rightarrow \hat{\sigma}_{B_1} = 0.548$$

$$\hat{\text{Var}}(B_2) = 0.133 \Rightarrow \hat{\sigma}_{B_2} = 0.365$$

$$\hat{\text{Cov}}(B_1, B_2) = -0.018 \Rightarrow \hat{\rho}_{B_1, B_2} = -\frac{0.018}{0.548 \cdot 0.365} = -0.090$$

Regressão Linear Múltipla

Se $E_n \sim IN(0, \sigma^2)$, é possível especificar as distribuições dos estimadores A e B_1, \dots, B_J

$$A \sim N(\alpha, \sigma^2/N)$$

$$B_1 \sim N[\beta_1, \text{Var}(B_1)]$$

.....

$$B_J \sim N[\beta_J, \text{Var}(B_J)]$$

Onde $\text{var}(B_1), \dots, \text{Var}(B_J)$ são definidos a partir da matriz variância-covariância.

Nestas condições, temos as distribuições seguintes:

$$\frac{A - \alpha}{S/\sqrt{N}} \sim t_{N-J-1}$$

$$\frac{B_1 - \beta_1}{\sqrt{\text{Var}(B_1)}} \sim t_{N-J-1}$$

$$\frac{B_J - \beta_J}{\sqrt{\text{Var}(B_J)}} \sim t_{N-J-1}$$

Regressão Linear Múltipla

A partir das expressões anteriores é possível definir **intervalos de confiança e testes de hipóteses** envolvendo os parâmetros de regressão.

Intervalos de Confiança:

$$\alpha : A \pm t_{N-J-1}(\gamma/2) \cdot S \cdot \sqrt{1/N}$$

$$\beta_j : B_j \pm t_{N-J-1}(\gamma/2) \cdot \sqrt{\text{Var}(B_j)}$$

Note que, os intervalos assim definidos estão correctamente especificados quando considerados individualmente. No entanto, o nível de confiança para o conjunto dos intervalos definidos para A e B_j ($j=1,\dots,J$) é, de facto, diferente do considerado.

Teste de Hipóteses:

O teste relativo ao parâmetro α será:

$$H_0: \alpha = \alpha_0$$

$$H_1: \alpha \neq \alpha_0, \alpha < \alpha_0 \text{ ou } \alpha > \alpha_0$$

$$ET = \frac{A - \alpha_0}{S / \sqrt{N}}$$

$$H_0 \text{ verdadeira} \Rightarrow ET \sim t_{N-J-1}$$

Régressão Linear Múltipla

Teste de Hipóteses (cont.):

Relativamente aos parâmetros β_j deverá primeiro ser testada a hipótese de que todos eles são nulos contra a hipótese de que pelo menos um deles é diferente de zero.

Tal teste será realizado recorrendo à técnica de **Análise de Variância** que se fundamenta na seguinte decomposição:

$$\sum_n (Y_n - \bar{Y})^2 = \underbrace{\sum_n [A + B_1 \cdot (X_{1n} - \bar{X}_1) + \dots + B_J \cdot (X_{Jn} - \bar{X}_J) - \bar{Y}]^2}_{VT=S_{YY}} +$$
$$+ \underbrace{\sum_n [Y_n - A - B_1 \cdot (X_{1n} - \bar{X}_1) - \dots - B_J \cdot (X_{Jn} - \bar{X}_J)]^2}_{VR=S_{YY} - (B_1 \cdot S_{X_1Y} + \dots + B_J \cdot S_{X_JY})}$$

Regressão Linear Múltipla

Na **ANOVA** referente à regressão linear múltipla adopta-se esta decomposição.

FONTES DE VARIAÇÃO	VARIACÕES (Somas de quadrados)	GRAUS DE LIBERDADE (Número de termos independentes)	DESVIOS QUADRÁTICOS MÉDIOS	VALORES ESPERADOS
DEVIDA À REGRESSÃO (DR)	$VDR = B_1 \cdot S_{X_1Y} + \dots + B_J \cdot S_{X_JY}$	$GL_1 = J$	$DQMDR = VDR/GL_1$	$E [DQMDR] = \sigma^2 + [f(B_1, \dots, B_J)]^2$
RESIDUAL (R)	$VR = S_{YY} - VDR$	$GL_2 = N - J - 1$	$DQMR = VR/GL_2$	$E [DQMR] = \sigma^2$
TOTAL (T)	$VT = S_{YY}$	$GL = GL_1 + GL_2 = N - 1$		

Donde decorre a estrutura do **teste ANOVA**

$$H_0: \beta_1 = \beta_2 = \dots = \beta_J = 0$$

$$H_1: \text{algum } \beta_j \neq 0$$

$$ET = \frac{DQMDR}{DQMR}$$

H_0 verdadeira \Rightarrow

$$ET \sim F_{J, N-J-1}$$

Régressão Linear Múltipla

Exemplo (cont.)

Utilizando os dados do [exemplo](#) anterior, vamos testar as hipóteses

$$H_0: \beta_1 = \beta_2 = 0$$

$$H_1: \beta_1 \text{ ou } \beta_2 \neq 0.$$

A **tabela ANOVA** correspondente vem:

FONTES	VARIACÕES	G.L.	DQM
DR	140.3	2	70.15
R	6.9	7	0.983
T	147.2	9	

$$ET = \frac{70.15}{0.983} = 71.34 > F_{2,7}(\alpha = 0.05) = 4.74$$

H_0 é rejeitada ao nível de significância de 5% (**valor de prova quase nulo**)

Régressão Linear Múltipla

Quando a hipótese nula é rejeitada é necessário verificar quais os β_j que são diferentes de zero. Uma via possível consiste na realização dos seguintes testes individuais aos parâmetros.

Das expressões [anteriores](#) relativas aos estimadores de B_j decorre a seguinte estrutura para os testes:

$$H_0: \beta_j = 0$$

$$H_1: \beta_j \neq 0, \beta_j > 0 \text{ ou } \beta_j < 0,$$

$$ET = \frac{\beta_j}{\sqrt{Var(B_j)}} \quad H_0 \text{ verdadeira} \Rightarrow \quad ET \sim t_{N-J-1}$$

Régressão Linear Múltipla

Exemplo (cont.)

$$\begin{aligned} H_0: \beta_1 &= 0 \\ H_1: \beta_1 &\neq 0 \end{aligned}$$

$$\begin{aligned} H_0: \beta_2 &= 0 \\ H_1: \beta_2 &\neq 0 \end{aligned}$$

Sabendo que $b_1 = 4.34$ $b_2 = 2.99$ $\hat{\sigma}_{B_1} = 0.548$ $\hat{\sigma}_{B_2} = 0.365$

$$ET = \frac{b_1}{\sqrt{\hat{Var}(B_1)}} = \frac{4.34}{0.548} = 7.92 > t_7(0.025) = 2.365$$

$H_0: \beta_1 = 0$ é rejeitada ao nível de significância de 5% (valor de prova quase nulo)

$$ET = \frac{b_2}{\sqrt{\hat{Var}(B_2)}} = \frac{2.99}{0.365} = 8.19 > t_7(0.025) = 2.365$$

$H_0: \beta_2 = 0$ também é rejeitada ao nível de significância de 5% (igualmente com valor de prova praticamente nulo).

Regressão Linear Múltipla

Exemplo.

Estime e teste o modelo de regressão para as seguintes variáveis:

Y	X1	X2
69	60	19
49	90	9
56	95	11
62	100	13
117	180	22
45	85	9
38	80	9
83	230	18

Régressão Linear Múltipla

Y	X1	X2	Syy	Sx1x1	Sx2x2	Sx1x2	Sx1y	Sx2y
69	60	19	17.0	3025.0	27.6	-288.8	-226.9	21.7
49	90	9	252.0	625.0	22.6	118.8	396.9	75.4
56	95	11	78.8	400.0	7.6	55.0	177.5	24.4
62	100	13	8.3	225.0	0.6	11.3	43.1	2.2
117	180	22	2717.0	4225.0	68.1	536.3	3388.1	430.0
45	85	9	395.0	900.0	22.6	142.5	596.3	94.4
38	80	9	722.3	1225.0	22.6	166.3	940.6	127.7
83	230	18	328.5	13225.0	18.1	488.8	2084.4	77.0
64.9	115.0	13.8	4518.9	23850.0	189.5	1230.0	7400.0	852.8
			a= 64.88					
23850	b1	+	1230	b2	=	7400		
1230	b1	+	189.5	b2	=	852.75		
	b1= 0.12							
	b2= 3.74							
ANOVA								
FV	Var	GL	DQM		ET= 21.93798			
DR	4056.595	2	2028.298		F2,5= 5.786135			
R	462.2799	5	92.45598					
T	4518.875	7			R2xy= 89.77%			

Régressão Linear Múltipla

O problema associado à realização destes testes reside no facto de o nível de significância do conjunto dos testes ser diferente daquele que foi especificado.

Esta dificuldade pode ser contornada com o método de **selecção de regressores**.

Regressão Linear Múltipla

Selecção de Regressores

- No modelo de regressão múltipla admitiu-se que as variáveis independentes (os regressores) eram designadas à partida.
- Na maioria das situações práticas não é possível especificar à partida, com segurança, o conjunto ideal de regressores.
- Num cenário real podem existir uma multiplicidade de regressores potencialmente úteis na explicação do comportamento da variável dependente, havendo que seleccionar de entre eles aqueles que devem figurar no modelo.
- De seguida serão discutidos diferentes métodos de selecção de regressores.

Regressão Linear Múltipla

Método Exaustivo

1. Construir os modelos de regressão que combinem de todas as maneiras possíveis os regressores potenciais.
2. Ordenar os modelos de regressão de acordo com um critério de qualidade (por exemplo, minimizar os DQMR).
3. Avaliar em detalhe um número restrito de modelos considerados melhores, de acordo com o critério fixado em (2).

O ponto (3) está associado à incapacidade de definir um critério único que, em todas as circunstâncias, permita comparar objectivamente a qualidade dos modelos.

Se o número de regressores potenciais for J , o número de modelos alternativos a construir é $2^J - 1$.

Regressão Linear Múltipla

Método Progressivo

1. Ajustar tantos modelos de regressão linear simples quantos os regressores potenciais. Incluir no modelo aquele que explica a maior proporção da variação da variável dependente. Se nenhum regressor explicar uma proporção significativa da variação o método termina.
2. Construir modelos de regressão dupla que associem o regressor seleccionado em (1) e cada um dos restantes regressores potenciais.
De entre os novos regressores que explicam uma proporção adicional significativa da variação total, incluir no modelo aquele que explica a maior proporção.
3. Prosseguir a tentativa de construção de modelos de ordem superior adoptando um procedimento idêntico ao descrito.
4. O método termina quando nenhum dos regressores potenciais explica um proporção adicional significativa da variação total ou quando todos os regressores forem incluídos no modelo.
5. Este método ***não garante*** a selecção do melhor conjunto de regressores.

Regressão Linear Múltipla

Exemplo

Considere-se o problema da selecção de regressores admitindo que se dispõe de 20 observações de **uma variável dependente** (Y) e de **três variáveis candidatas** a figurarem como regressores num modelo de regressão múltipla (X_1 , X_2 e X_3).

Passo (1): construir três modelos de **regressão linear simples**

Modelo $Y = Y(X_1)$:			
Fontes	Variações	G.L.	DQM
DR (X_1)	60	1	60
R	140	18	140/18
T	200	19	

Modelo $Y = Y(X_2)$:			
Fontes	Variações	G.L.	DQM
DR (X_2)	110	1	110
R	90	18	90/18
T	200	19	

Modelo $Y = Y(X_3)$:			
Fontes	Variações	G.L.	DQM
DR (X_3)	20	1	20
R	180	18	180/18
T	200	19	

O regressor que **explica a maior proporção** da variação total é X_2 .

$$H_0 : \beta_2 = 0 \rightarrow \left[ET = \frac{110}{90/18} = 22.0 > F_{1,18}(0.05) = 4.41 \right] \Rightarrow H_0 \text{ rejeitada}$$

O teste ANOVA permite **verificar que a proporção da variação explicada é significativa**.

Regressão Linear Múltipla

Passo (2) construir os dois modelos de regressão linear dupla $Y = Y(X_2, X_1)$ e $Y = Y(X_2, X_3)$

Modelo $Y = Y(X_2, X_1)$:

Fontes	Variações	G.L.	DQM
DR (X_2, X_1)	120	2	120/2
R	80	17	80/17
T	200	19	

Modelo $Y = Y(X_2, X_3)$:

Fontes	Variações	G.L.	DQM
DR (X_2, X_3)	135	2	135/2
R	65	17	65/17
T	200	19	

O regressor que explica uma proporção adicional maior da variação total é X_3 .

Vamos agora de testar se o contributo adicional de X_3 para a explicação da variação de Y é significativo. Para tal tem-se de alterar a tabela ANOVA correspondente decompondo:

$$VDR(X_2, X_3) = VDR(X_2) + VDR(X_3|X_2)$$

Fontes	Variações	G.L.	DQM
DR (X_2, X_3)	135	2	135/2
DR (X_2)	(110)	(1)	
DR ($X_3 X_2$)	(25)	(1)	25
R	65	17	65/17
T	200	19	

Regressão Linear Múltipla

O teste ANOVA a realizar tem a seguinte estrutura:

$$H_0: \beta_3 = 0$$

$$H_1: \beta_3 \neq 0$$

$$ET = \frac{DQMDR(X_3 | X_2)}{DQMR} \quad H_0 \text{ verdadeira} \Rightarrow ET \sim F_{1,N-3}.$$

$$ET = \frac{DQMDR(X_3 | X_2)}{DQMR} = \frac{25}{65/17} = 6.54 > F_{1,17}(0.05) = 4.45$$

Nestas condições, é incluído no modelo o regressor X_3 , que assim se junta ao regressor X_2

Passo (3) construir o modelo de regressão linear tripla $Y = Y(X_2, X_3, X_1)$

Fontes	Variações	G.L.	DQM
DR (X_2, X_3, X_1)	140	3	135/2
DR (X_2, X_3)	(135)	(2)	
DR ($X_1 X_2, X_3$)	(5)	(1)	5
R	60	16	60/16
T	200	19	

Neste caso, o teste ANOVA permite verificar que, a proporção adicional da variação total que é explicada por X_1 não é significativa

$$\left[ET = \frac{DQMDR(X_1 | X_2, X_3)}{DQMR} = \frac{5}{60/16} = 1.33 < F_{1,16}(0.05) = 4.50 \right] \Rightarrow H_0 \text{ não rejeitada}$$

Regressão Linear Múltipla

Método Regressivo

1. Incluir no modelo todos os regressores potenciais.
2. Retirar do modelo, um a um, regressores cuja presença não contribua para explicar uma proporção significativa da variação total
3. Prosseguir a tentativa de construção de modelos de ordem inferior adoptando um procedimento idêntico ao descrito.

Método Regressão Passo a Passo

Consistem em versões dos métodos progressivo e regressivo nas quais os regressores que tenham sido incorporados no modelo ou dele excluídos em passos anteriores são reexaminados

Regressão Linear Múltipla

2. Utilize o método de selecção de regressores para o exemplo:

Y	69	49	56	62	117	45	38	83
X1	60	90	95	100	180	85	80	230
X2	19	9	11	13	22	9	9	18

Régressão Linear Múltipla

ANOVA $y=f(X_1)$						ANOVA $y=f(X_1, X_2)$					
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>V.P.</i>		<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>V.P.</i>
DR(X_1)	1	2296.0	2296.0	6.2	0.047	DR(X_1, X_2)	2	4056.6	2028.3	21.9	0.003
R	6	2222.9	370.5			R	5	462.3	92.5		
Total	7	4518.9				Total	7	4518.9			

ANOVA $y=f(X_2)$						ANOVA $y=f(X_1, X_2)$					
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>V.P.</i>		<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>V.P.</i>
DR(X_2)	1	3837.4	3837.4	33.8	0.001	DR(X_1, X_2)	2	4056.6	2028.3		
R	6	681.5	113.6			DR(X_2)	1	3837.4			
Total	7	4518.9				DR($X_1 X_2$)	1	219.2	219.2	2.4	0.184
						Residual	5	462.3	92.5		
						Total	7	4518.9			

Regressão Linear Múltipla

Incorporação de regressores qualitativos: variáveis mudas

As **variáveis mudas** são incorporadas nos modelos de regressão com o objectivo de representar o **efeito de factores qualitativos**.

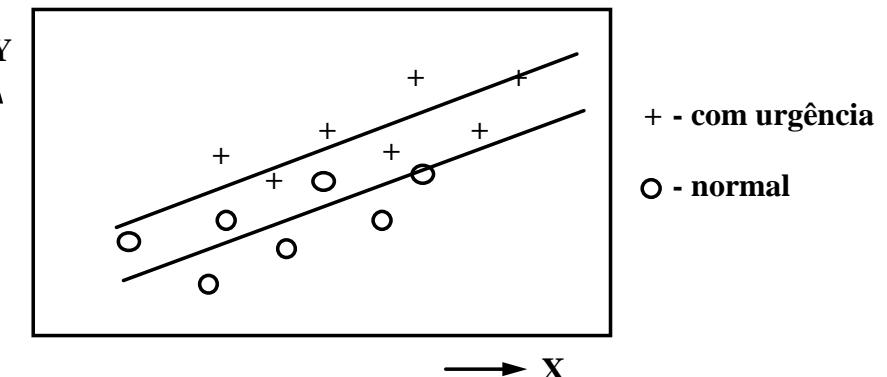
Exemplo

Numa determinada empresa de artes gráficas existe uma secção dedicada ao fabrico de um tipo de cartões. Na figura representam-se, para essa secção, observações das variáveis

X: dimensões de diferentes encomendas de cartões

Y: custos de produção associados à satisfação das encomendas.

As observações foram classificadas em encomendas satisfeitas em regime normal e encomendas satisfeitas com urgência.



Régressão Linear Múltipla

Com base na figura parece razoável adoptar o seguinte modelo

$$Y_n = \alpha + \beta \cdot (X_n - \bar{X}) + \gamma \cdot (Z_n - \bar{Z}) + E_n = \alpha' + \beta \cdot X_n + \gamma \cdot Z_n + E_n$$

Em que Z_n represente um variável muda que toma os seguintes valores

$$Z_n = \begin{cases} 0, & \text{para o regime normal} \\ 1, & \text{para o regime urgente} \end{cases}$$

No exemplo γ representa o valor esperado do custo adicional associado ao regime urgente.

Para representar adequadamente um factor com k níveis devem ser incluídas no modelo de regressão $k - 1$ variáveis mudas, por exemplo, para considerar três regimes de satisfação de encomendas

$$Z_1 = \begin{cases} 0, & \text{se o regime for normal} \\ 1, & \text{se o regime não for normal} \end{cases} \quad Z_2 = \begin{cases} 0, & \text{se o regime for urgente} \\ 1, & \text{se o regime não for urgente} \end{cases}$$

regime normal: $Z_1 = 0, Z_2 = 1$

regime urgente: $Z_1 = 1, Z_2 = 0$

regime muito urgente: $Z_1 = 1, Z_2 = 1$.

$$Y_n = \alpha' + \beta \cdot X_n + \gamma_1 \cdot Z_{1n} + \gamma_2 \cdot Z_{2n} + E_n$$

Regressão

Regressão Não-Linear

- A técnica de regressão linear simples pode ser utilizada em modelos não-lineares desde que os modelos sejam convertíveis em modelos lineares por aplicações de transformações às variáveis.
- Vamos considerar alguns exemplos de aplicação frequente.

Método:

1. Transformar a variável X
2. Transformar a variável Y
3. Aplicar método de regressão linear às variáveis transformadas

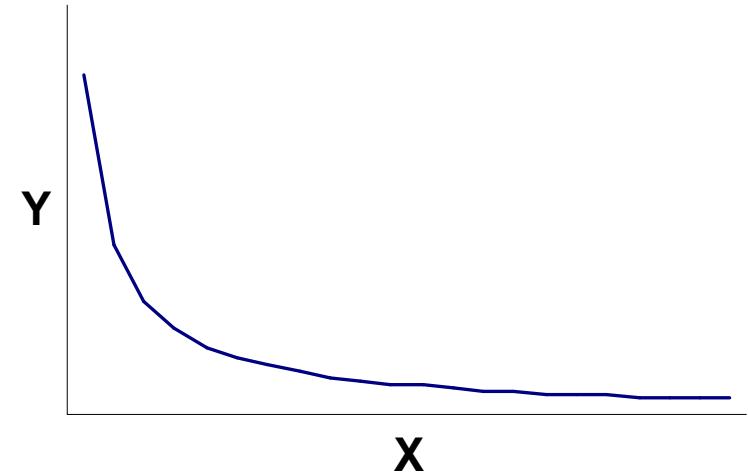
Régressão Não Linear

Exemplo 1

$$\text{Modelo : } Y_n = \alpha' + \frac{\beta}{X_n} + E_n$$

Uma relação deste tipo pode ser linearizada recorrendo à seguinte transformação da variável independente

$$U_n = \frac{1}{X_n}$$



$$\text{Mod. linearizado : } Y_n = \alpha' + \beta \cdot U_n + E_n$$

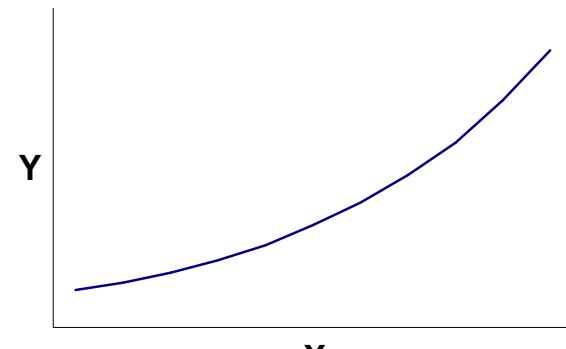
Régressão Não Linear

Exemplo 2

Modelo exponencial: $Y_n = e^{\alpha' + \beta \cdot X_n + E_n}$

Linearização através de uma
transformação logarítmica da variável
dependente $Z_n = \ln(Y_n)$

Mod. linearizado: $Z_n = \alpha' + \beta \cdot X_n + E_n$



Régressão Não Linear

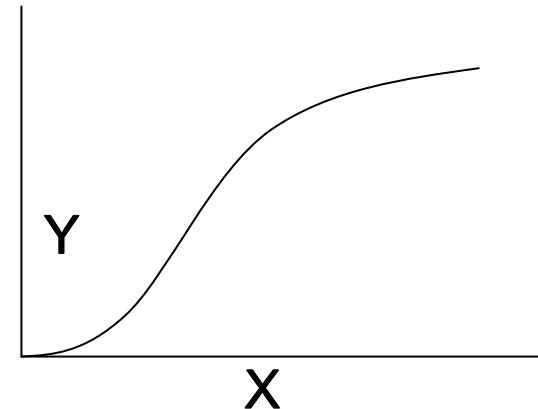
Exemplo 3

Modelo curva S: $Y_n = e^{\alpha' + \beta/X_n + E_n}$ (com $\alpha' > 0$ e $\beta < 0$)

linearização: $U_n = \frac{1}{X_n}$

$$Z_n = \ln(Y_n)$$

Mod. linearizado: $Z_n = \alpha' + \beta \cdot U_n + E_n$



Regressão

Regressão Polinomial

Entre uma variável dependente Y e uma variável independente X pode existir uma relação polinomial de grau J, que pode ser representada por um modelo do tipo:

$$Y_n = \alpha + \beta_1 \cdot (X_n - \bar{X}) + \beta_2 \cdot (X_n^2 - \bar{X}^2) + \cdots + \beta_J \cdot (X_n^J - \bar{X}^J) + E_n$$

Onde $\bar{X} = \frac{1}{N} \cdot \sum_n X_n$ $\bar{X}^2 = \frac{1}{N} \cdot \sum_n X_n^2$... $\bar{X}^J = \frac{1}{N} \cdot \sum_n X_n^J$

Este modelo designa-se por **modelo de regressão polinomial simples** e pode ser convertido num **modelo de regressão linear múltipla** fazendo corresponder a cada potência uma nova variável substituindo:

$$\text{PARA } j = 1, \dots, J \quad X_n^j = X_{jn}$$

Regressão Polinomial

Obtendo-se o modelo linearizado:

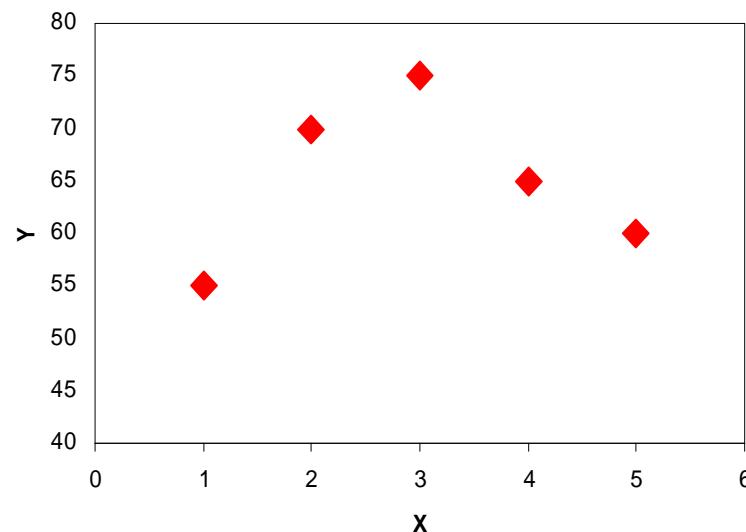
$$Y_n = \alpha + \beta_1 \cdot (X_{1n} - \bar{X}_1) + \cdots + \beta_J \cdot (X_{Jn} - \bar{X}_J) + E_n$$

O problema da escolha do grau do polinómio é equivalente ao problema da selecção de regressores anteriormente abordado.

Exemplo

X	Y
1	55
2	70
3	75
4	65
5	60

$$\text{MODELO: } Y_n = \alpha + \beta_1 \cdot X_n + \beta_2 \cdot X_n^2$$



Régressão Polinomial

Vamos definir as seguintes variáveis: $\begin{cases} X_1 \equiv X \\ X_2 \equiv X^2 \end{cases}$

obtendo-se o seguinte modelo: $Y_n = \alpha + \beta_1 \cdot X_{1n} + \beta_2 \cdot X_{2n}$

$X_1=X$	$X_2=X^2$	Y
1	1	55
2	4	70
3	9	75
4	16	65
5	25	60

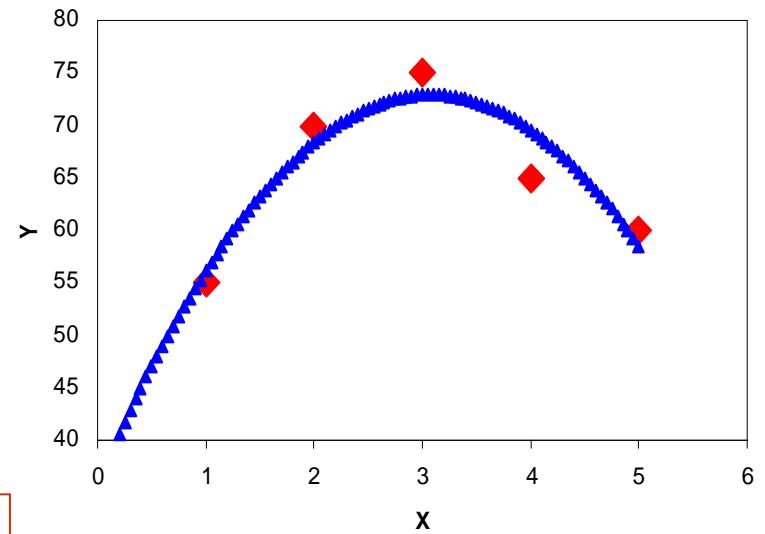
$$N=5 \quad S_{x_1 x_1} = 10 \quad S_{x_1 y} = 5$$

$$\bar{x}_1 = 3 \quad S_{x_2 x_2} = 374 \quad S_{x_2 y} = -25$$

$$\bar{x}_2 = 11 \quad S_{x_1 x_2} = 60 \quad \bar{y} = 65$$

$$\begin{cases} b_1 \cdot S_{x_1 x_1} + b_2 \cdot S_{x_1 x_2} = S_{x_1 y} \\ b_1 \cdot S_{x_2 x_1} + b_2 \cdot S_{x_2 x_2} = S_{x_2 y} \end{cases} \Leftrightarrow \begin{cases} b_1 \cdot 10 + b_2 \cdot 60 = 5 \\ b_1 \cdot 60 + b_2 \cdot 374 = -25 \end{cases} \Leftrightarrow \begin{cases} b_1 = 24.1 \\ b_2 = -3.9 \end{cases}$$

$$a = \hat{\alpha}' = \alpha - \beta_1 \cdot \bar{X}_1 - \beta_2 \cdot \bar{X}_2 = 65 - 24.1 \cdot 3 + 3.9 \cdot 11 = 36.0$$



$$\hat{\mu}_{Y_n} = 36.0 + 24.1 \cdot x_1 + 3.9 \cdot x_2$$

Regressão Polinomial

- O coeficiente de determinação (R_{xy}^2), que traduz a proporção da variação total de Y explicada pela regressão ajustada, corresponde ao coeficiente de correlação r elevado ao quadrado;
- Este coeficiente apresenta uma limitação: o denominador da expressão que lhe está subjacente tem um valor fixo, enquanto que o numerador só pode aumentar. Assim, ao adicionar-se uma nova variável na equação da regressão, o numerador aumentará, no mínimo, ligeiramente, resultando num aumento do coeficiente de determinação, mesmo que a introdução da nova variável resulte numa equação menos eficiente;
- Em teoria, usando um número infinito de variáveis independentes para explicar a variação da variável dependente, resulta num R_{xy}^2 igual a 1. Por outras palavras, o coeficiente de determinação pode ser manipulado, logo deve ser suspeitado;

Regressão Polinomial

Coeficiente de Determinação Ajustado (\bar{R}_{XY}^2)

- Dado que a introdução de um regressor irrelevante aumentará ligeiramente o R_{XY}^2 , é desejável tentar corrigi-lo, reduzindo-o de uma forma apropriada;
- O coeficiente de determinação ajustado, \bar{R}_{XY}^2 , é uma tentativa de tentar corrigir o R_{XY}^2 , ajustando o numerador e o denominador da expressão através dos respectivos graus de liberdade;

$$\bar{R}_{XY}^2 = 1 - (1 - R^2) \cdot \left(\frac{N-1}{N-K-1} \right)$$

N : n.º de observações

K : grau da regressão

N-1 : n.º total de graus de liberdade da VT

N-K-1 : n.º de graus de liberdade da VR

- Contrariamente ao coeficiente de determinação, o coeficiente de determinação ajustado pode diminuir em valor se a contribuição da variável adicional na explicação da VT, for inferior ao impacto que essa adição acarreta nos graus de liberdade.

Régressão Não Linear

3. Considere a seguinte série temporal, com 10 observações:

t	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006
Z_t	100	120	200	307	351	456	690	796	955	1195

Com base nestes dados, efectue previsões dos valores de Z_t (para $t = 2007, 2008, 2009$ e 2010), recorrendo aos seguintes métodos:

- i) régressão polinomial;
- ii) régressão linear de variáveis transformadas.

Entre as previsões efectuadas pelos dois métodos quais adoptaria?