Métodos Quantitativos para Ciência da Computação Experimental

Regressão Linear

Jussara Almeida
DCC-UFMG
2017

Apresentacao derivada dos slides originais de Virgilio Almeida

Modelos de Regressão Linear

- O que é um bom modelo?
- Como estimar os parâmetros do modelo?
- Como alocar variações?
- Intervalos de Confiança para Regressões
- Inspeçao Visual

Apresentacao derivada dos slides originais de Virgilio Almeida

O que é um bom modelo?

- Para dados correlacionados, um modelo deve prever uma resposta dado uma entrada.
- Modelo deve ser a equação que “se adequa” (“*fit*”) aos dados.
- Uma definição padrão de “fits” está diretamente relacionada aos mínimos quadrados (“*least-squares”)*
  - Minimizar o erro ao quadrado
  - Enquanto mantém o erro médio em zero
  - Equivalente a minimizar a variância dos erros

Erro do Mínimo Quadrado

- Se $\hat{y} = b_0 + b_1 x$ então o erro na estimativa para x_i é

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

- Minimizar a Soma dos Erros ao Quadrado (SSE)

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$$

- Sujeita as restrições

$$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i) = 0$$

Estimando os Parâmetros do Modelo

- Os melhores parâmetros da regressão (levam ao menor erro) são:

$$b_1 = \frac{\sum xy - n\bar{x}\bar{y}}{\sum x^2 - n\bar{x}^2} \qquad b_0 = \bar{y} - b_1\bar{x}$$

onde

$$\bar{x} = \frac{1}{n}\sum x_i \qquad \bar{y} = \frac{1}{n}\sum y_i$$

$$\sum xy = \sum x_i y_i \qquad \sum x^2 = \sum x_i^2$$

Estimativa dos parâmetros exemplo

- Tempo de execução de um query para várias palavras:

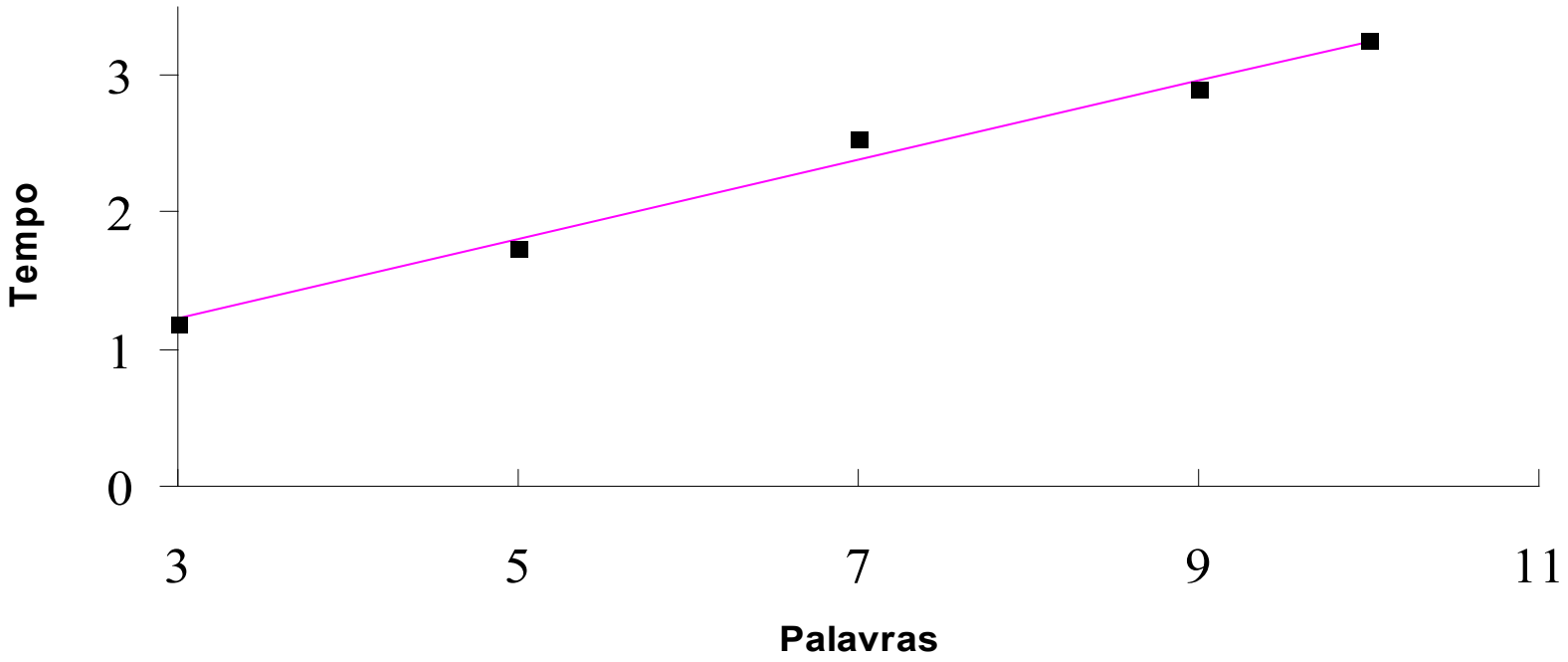
x	Palavras	3	5	7	9	10
y	Tempo	1.19	1.73	2.53	2.89	3.26

$\bar{x}$ = 6.8, $\bar{y}$ = 2.32, $\Sigma xy = 88.54$, $\Sigma x^2 = 264$

$$b_1 = \frac{88.54 - (5)(6.8)(2.32)}{264 - (5)(6.8)^2} = 0.29$$

- $b_0 = 2.32 - (0.29)(6.8) = 0.35$

Gráfico dos Parâmetros de Estimativa exemplo



Apresentacao derivada dos slides originais de Virgilio Almeida

Variantes da Regressão Linear

- Algumas relações não lineares podem ser tratadas por transformações:

  – Para $y = ae^{bx}$ pegue o logaritmo de y, faça a regressão sobre $\log(y) = b_0+b_1x$, sendo $b = b_1$, $a = e^{b_0}$

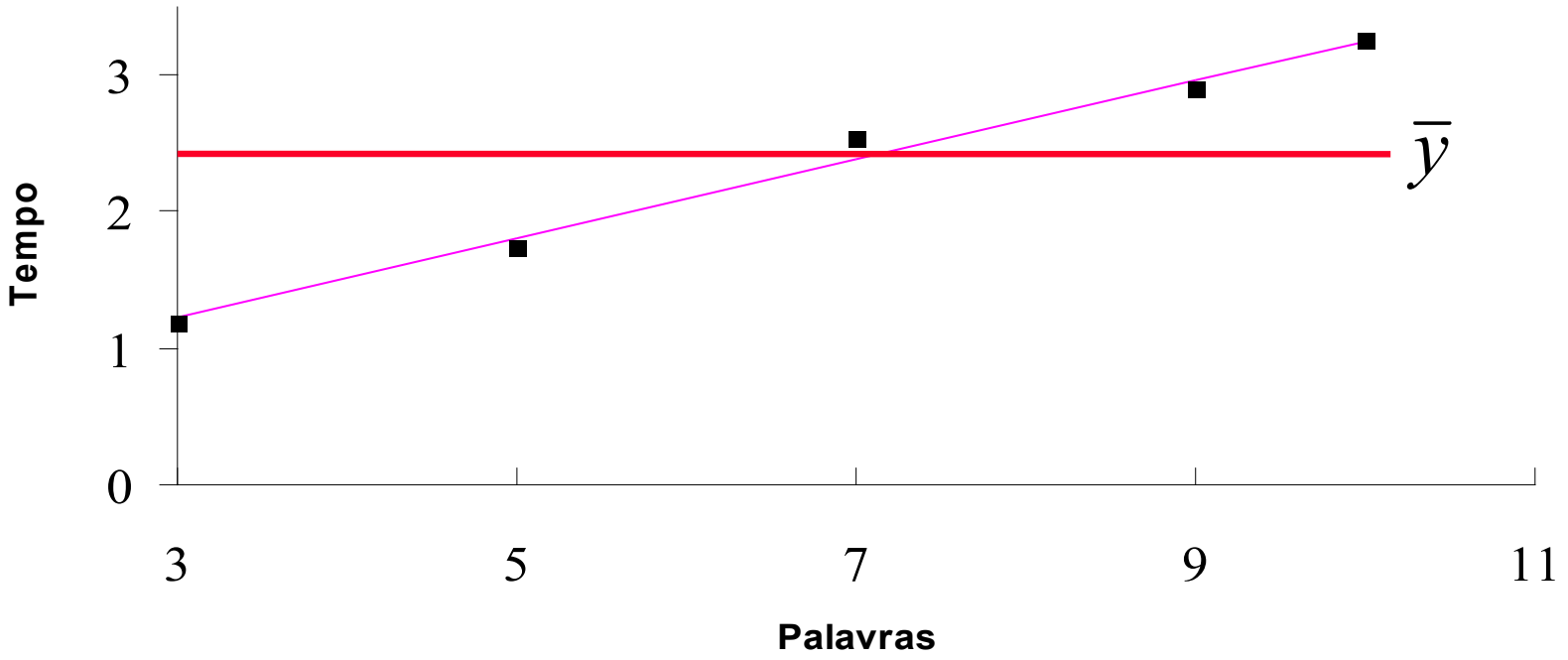
  – Para $y = a+b\log(x)$, tome o log de x antes dos parâmetros de "fitting", seja $b = b_1$, $a = b_0$

  – Para $y = ax^b$, tire o log de ambos x e y, e faça $b = b_1$, $a = e^{b_0}$

Alocando a Variação

- Sem regressão, a melhor estimativa de y é $\bar{y}$
- Valores observados de y diferem de $\bar{y}$ aumentando os erros (variação)
- Regressão provê uma melhor estimativa, mas ainda existem erros
- Nós podemos avaliar a qualidade da regressão pela alocação das fontes de erros.

Gráfico dos Parametros de Estimativa exemplo: regressão e a média



Apresentacao derivada dos slides originais de Virgilio Almeida

Notação

- SSE – Sum of Squared Errors
- SST – Total Sum of Squares

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \left(\sum_{i=1}^{n} y_i^{\,2} \right) - n\bar{y}^2 = SSY - SS0$$

- SSY – Sum of Squares of y
- SS0 – Sum of Squares of $\bar{y}$
- SSR – Sum of Squares explained by Regression

A Soma Total dos Quadrados

- Sem regressão, o erro ao quadrado é

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

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

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

$$= \left(\sum_{i=1}^{n} y_i^2\right) - n\bar{y}^2 = \text{SSY} - \text{SS0}$$

A Soma dos Quadrados da Regressão

- A soma dos erros quadrados sem regressão (=SST):

$$\sum e_i^2 = \sum \left(y_i - \bar{y}_i \right)^2$$

- SSE (com regressao):

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$$

- Assim a regressão explica SSR = SST - SSE
- Qualidade da regressão medida pelo *coeficiente de determinação:*

$$R^2 = \frac{\mathrm{SSR}}{\mathrm{SST}} = \frac{\mathrm{SST} - \mathrm{SSE}}{\mathrm{SST}}$$

- *Quanto maior o valor de R^2, melhor a regressão.*

Avaliação do Coeficiente de Determinação

- Calcule $\mathrm{SST} = (\sum y^2) - n\bar{y}^2$
- Calcule $\mathrm{SSE} = \sum y^2 - b_0 \sum y - b_1 \sum xy$
- Calcule $R^2 = \frac{\mathrm{SST} - \mathrm{SSE}}{\mathrm{SST}} = 1 - \frac{\mathrm{SSE}}{\mathrm{SST}}$

Exemplo de Coeficiente de Determinação

- Para o exemplo anterior de regressão

x	3	5	7	9	10
y	1.19	1.73	2.53	2.89	3.26

$$n\bar{y}^2 = (5)(2.32)^2 = 26.9$$

  - SSE = 29.79-(0.35)(11.60)-(0.29)(88.54) = 0.05
  - SST = 29.79-26.9 = 2.89
  - SSR = 2.89-.05 = 2.84
    - R^2 = (2.89-0.05)/2.89 = 0.98

Desvio Padrão de Erros

- Variancia de erros é SSE dividido pelos graus de liberdade (DOF):
  - DOF: n–2 porque calculamos 2 parametros de regressão dos dados.
  - Assim a variância (*mean squared error*, MSE): $\frac{SSE}{n-2}$

- Desvio padrão dos erros é a raiz quadrada:

$$s_e = \sqrt{\frac{\text{SSE}}{n-2}}$$

Coeficiente de Determinação X Correlação da Amostra

- Coeficiente de determinação

$$R^2 = \frac{\text{SSR}}{\text{SST}} = \frac{\text{SST} - \text{SSE}}{\text{SST}}$$

- Correlação da Amostra (premissa: linearidade)

$$s^2{}_{xy} = \sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})$$

$$\text{Correlação da amostra} = R_{xy} = \frac{s^2{}_{xy}}{s_x s_y}$$

Apresentacao derivada dos slides originais de Virgilio Almeida

Calculando os graus de liberdade de várias soma de quadrados

SST	n-1	Precisa computar $\bar{y}$
SSY	n	Não depende de nenhum outro parâmetro
SS0	1	Precisa computar $\bar{y}$
SSE	n-2	Precisa computar dois parâmetros da regressão
SSR	1	=SST-SSE

$$\mathrm{SST} = \mathrm{SSY} - \mathrm{SS0} = \mathrm{SSR} + \mathrm{SSE}$$

$$n - 1 = \quad n \quad - \quad 1 \quad = \quad 1 \quad + (n-2)$$

Apresentacao derivada dos slides originais de Virgilio Almeida

Exemplo de Desvio Padrão de Erros

- Para o exemplo de regressão, SSE era 0.05, então

$$MSE = 0.05/(5-2) = 0.05/3 = 0.017$$

$$s_e = \sqrt{MSE} = 0.13$$

- Observe a alta qualidade da regressão do exemplo:
  - $R^2 = 0.98$
  - $s_e = 0.13$

Intervalos de Confiança para Regressões

- Regressão é calculada de uma única amostra da população (tamanho *n*)
  - Diferentes amostras devem dar resultados diferentes.
  - Modelo verdadeiro é $y = \beta_0 + \beta_1 x$
  - Parâmetros b_0 e b_1 são na verdade médias (estimativas para parametros reais) retiradas das amostras da população.

Cálculo de Intervalos para Parâmetros da Regressão

- Desvio Padrão dos Parâmetros:

$$s_{b_0} = s_e \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum x^2 - n\bar{x}^2}}$$

$$s_{b_1} = \frac{s_e}{\sqrt{\sum x^2 - n\bar{x}^2}}$$

- Intervalos de confiança são $b_i \pm t_{[1-\alpha/2,n-2]} s_{bi}$
- Onde t tem n - 2 graus de liberdade
- s_e é o desvio padrão dos erros

Exemplo do Intervalo de Confiança da Regressão

- Lembre que $s_e = 0.13$, $n = 5$, $\Sigma x^2 = 264$, $\bar{x} = 6.8$
- Assim

$$s_{b_0} = 0.13\sqrt{\frac{1}{5} + \frac{(6.8)^2}{264 - 5(6.8)^2}} = 0.16$$

$$s_{b_1} = \frac{0.13}{\sqrt{264 - 5(6.8)^2}} = 0.004$$

- Usando um intervalo de confiança de 90%:

$t_{0.95;3} = 2.353$

Exemplo do Intervalo de Confiança da Regressão

- Assim, o intervalo b_0

$$0.35 \pm 2.353(0.16) = (-0.03, 0.73)$$

- b_1 é

$$0.29 \pm 2.353(0.004) = (0.28, 0.30)$$

Intervalos de Confiança para Predições

- Intervalos de confiança vistos são para os *parâmetros*
  - Quão certo podemos estar que os parâmetros estão corretos?

- Finalidade da regressão é a *predição*
  - Quão precisas são as predições?
  - Regressão oferece APENAS uma média das respostas previstas, baseadas nas amostras usadas.

Predições baseadas em *m* amostras

- Desvio padrão para a média de futuras amostras de *m* observações em x_p é $y_p = b_0 + b_1 x_p$

$$S_{y_{mp}} = s_e \sqrt{\frac{1}{m} + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum x^2 - n\bar{x}^2}}$$

- Note que o desvio diminui qdo $m \to \infty$
- Variância mínima em $x = \bar{x}$

Exemplo de Confiança das Predições

- Usando modelo desenvolvido, qual é o tempo previsto para **uma** execução com 8 palavras?
- Tempo = 0.35 + 0.29(8) = 2.67
- Desvio padrão de erros $s_e = 0.13$

$$S_{y_p} = 0.13\sqrt{1+\frac{1}{5}+\frac{(8-6.8)^2}{264-5(6.8)^2}} = 0.14$$

- 90% do intervalo é então

$$2.67 \pm 2.353(0.14) = (2.34, 3.00)$$

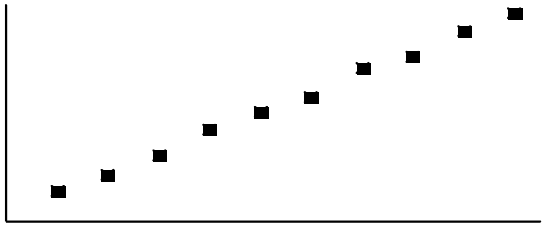
Verificando as hipóteses ("assumptions") visualmente

- Regressões são baseadas em hipóteses:
  - Relação linear entre a resposta *y* e previsor *x*
  - Previsor *x* livre de erro
  - Erros do modelo são estatisticamente independentes
    - Com distribuição normal $N(0,c)$ para desvio padrão constante c

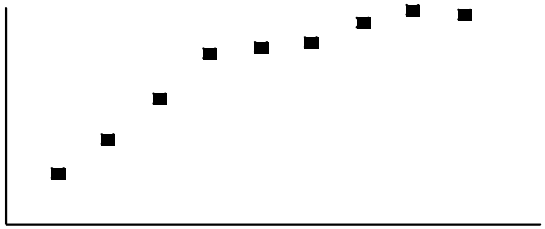
- Se as hipóteses são violadas, o modelo pode ser inadequado ou inválido.

Testando a Linearidade

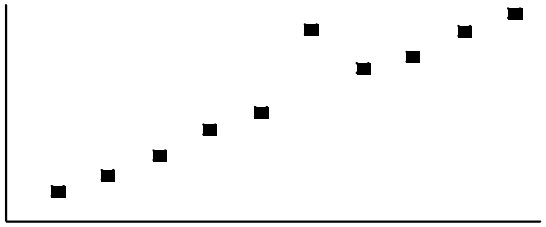
- Gráficos de pontos *x* vs. *y* para ver o tipo básico da curva



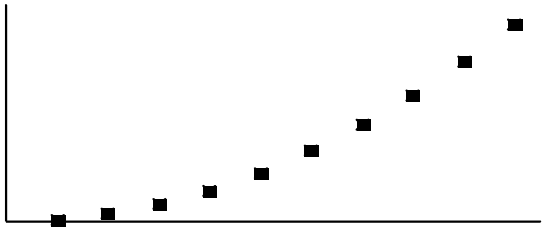
Linear



Linear por partes



Outlier/Excecão

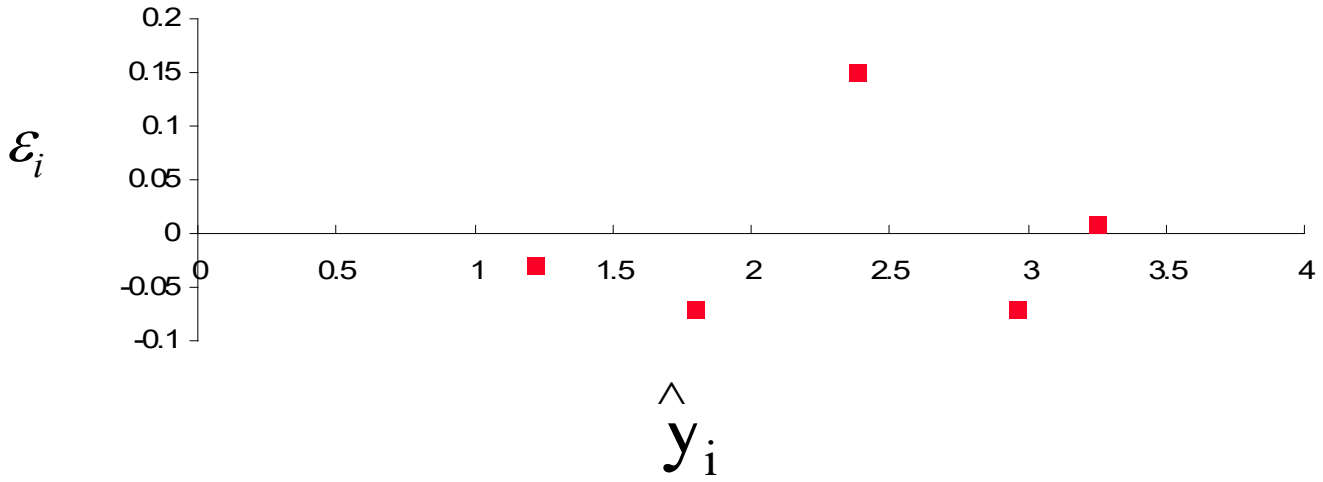


Não linear (Função de Potência)

Apresentacao derivada dos slides originais de Virgilio Almeida

Testando a Independência dos Erros

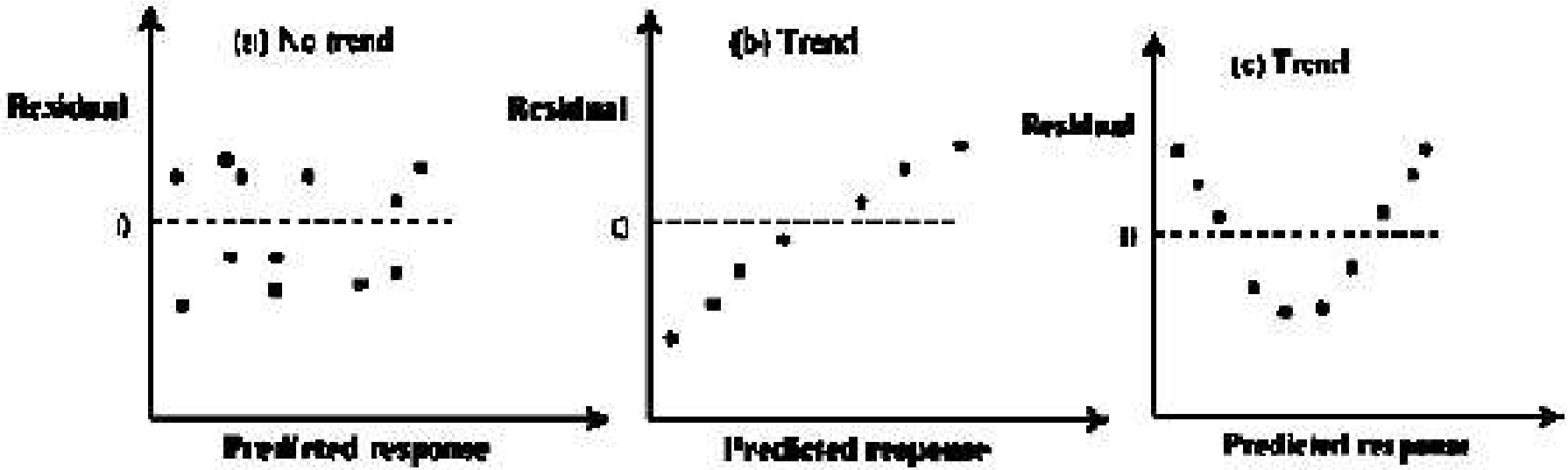
- Gráfico de pontos ε_i versus $\hat{y}_i$
- Não deve haver tendência visível
- Exemplo do ajuste de curva feito:



Apresentacao derivada dos slides originais de Virgilio Almeida

Testando a Independência

1. Scatter plot of ε_i versus the predicted response $\hat{y}_i$



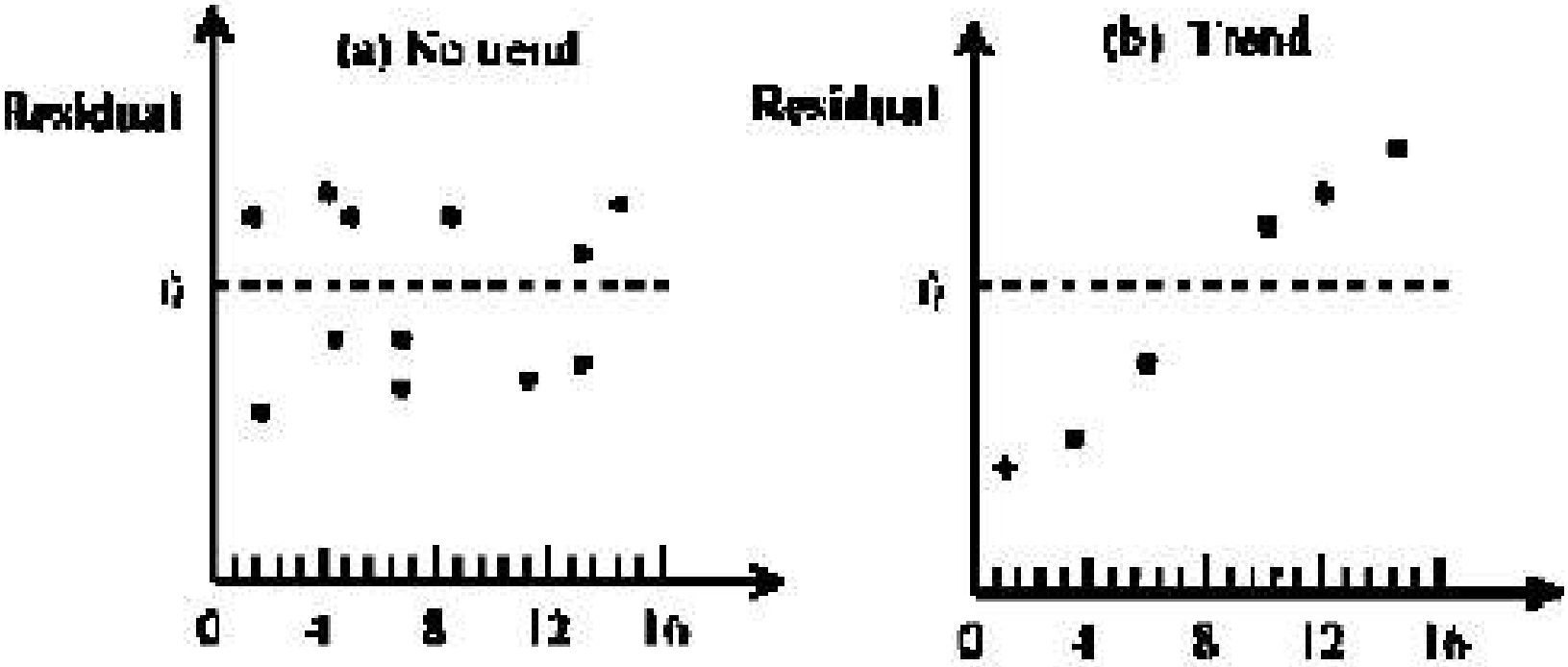
- All tests for independence simply try to find dependence.

Apresentacao derivada dos slides originais de Virgilio Almeida

Testando a Independência

- Pode ser útil "plotar" os resíduos de erro versus o número do experimento
  - No exemplo anterior dá o mesmo gráfico, exceto para a escala de *x*

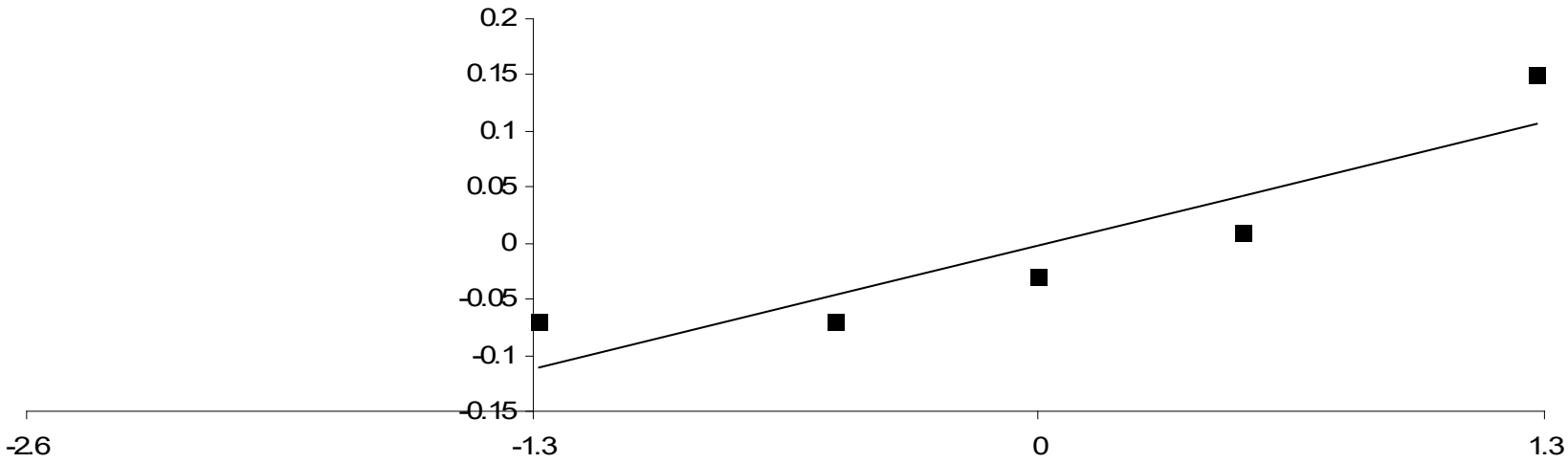
Testando a Independência



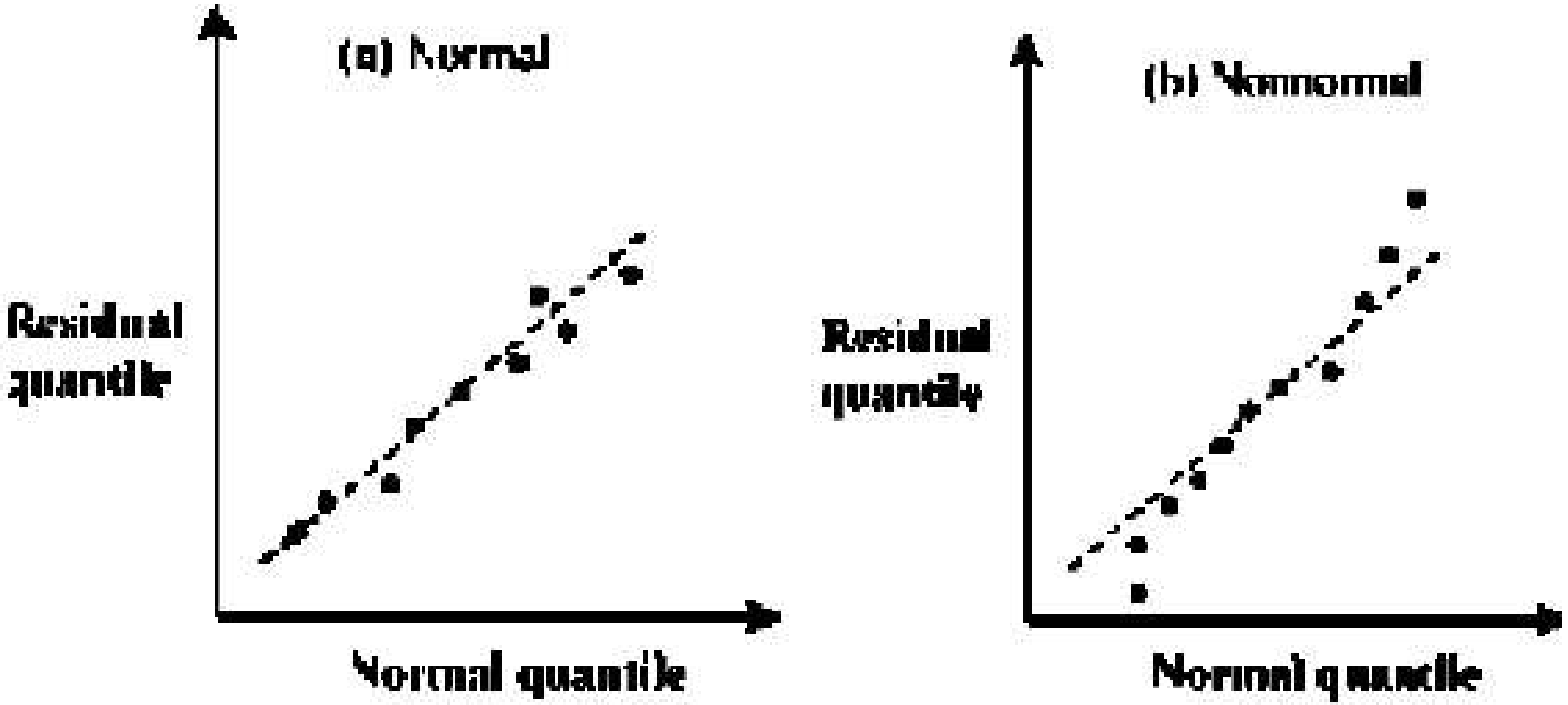
Apresentacao derivada dos slides originais de Virgilio Almeida

Testando Erros Normais

- Preparar um gráfico quantil-quantil
- Exemplo da regressão anterior:



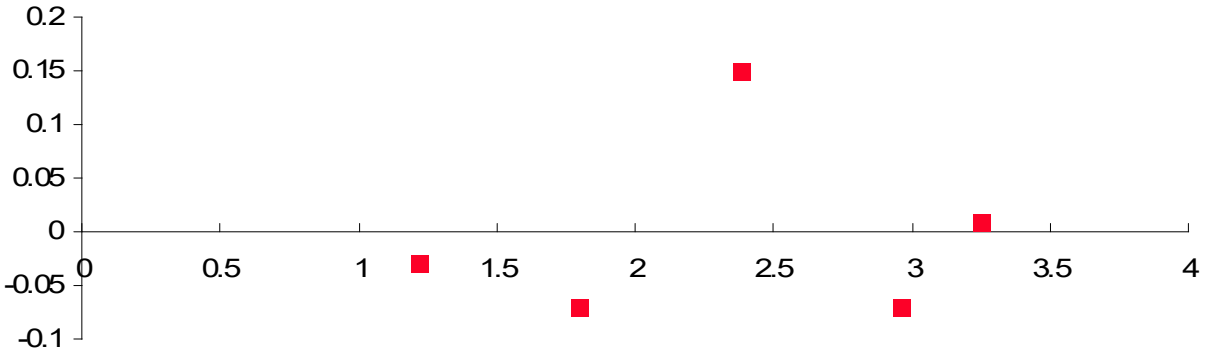
Testando Erros Normais



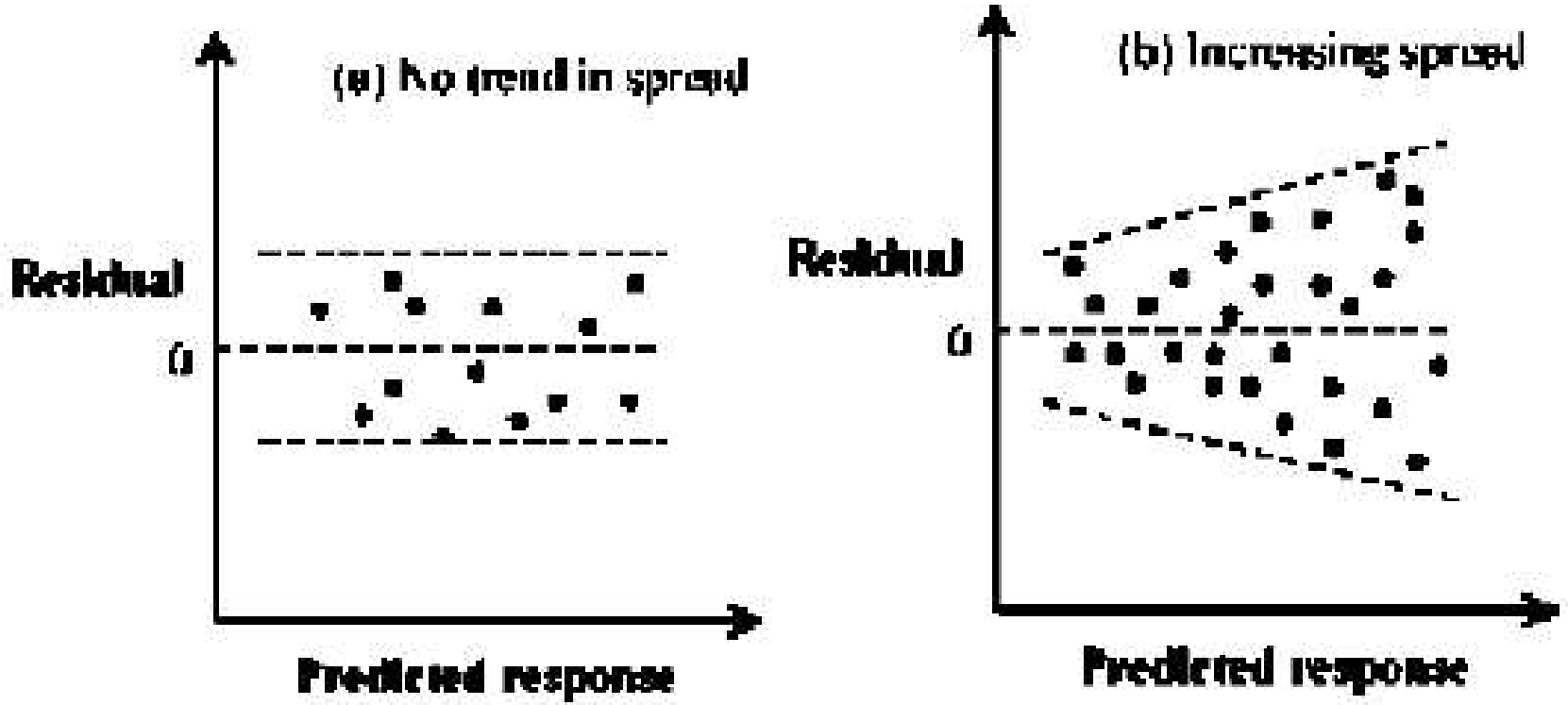
Apresentacao derivada dos slides originais de Virgilio Almeida

Testando para Desvio-Padrão Constante

- *Homoscedasticity (esta hipótese assume que a variância ao longo da linha de regressão é a mesma para todos previsores x)*
- Retorno ao gráfico de independência
- Verificar tendência no espalhamento
- Exemplo:



Testando para Desvio-Padrão Constante



- Trend ⇒ Try curvilinear regression or transformation

Apresentacao derivada dos slides originais de Virgilio Almeida

Regressão linear pode ser "enganadora" (misleading)

- Regressão despreza alguma informação sobre os dados
  - Para permitir uma sumarização compacta

- Algumas vezes características vitais são perdidas
  - No geral, examinando os gráficos de dados pode-se determinar se ha um problema ou não

Exemplo de Regressões Inadequadas

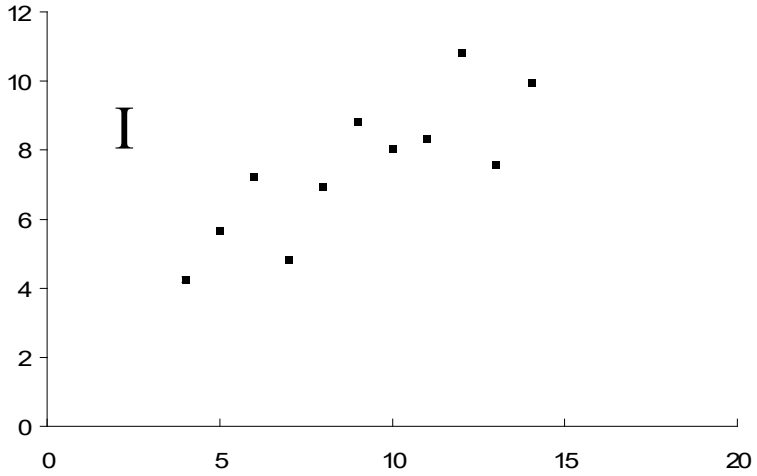
I		II		III		IV	
x	y	x	y	x	y	x	y
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.10	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.10	4	5.39	19	12.50
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

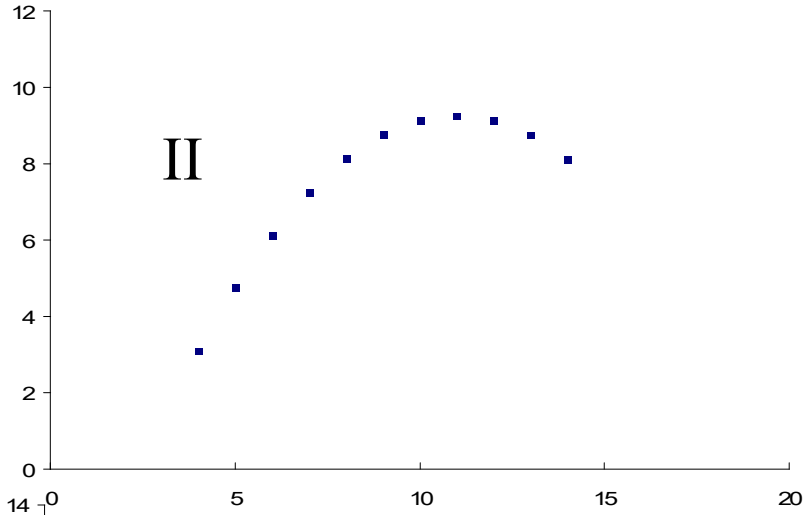
Apresentacao derivada dos slides originais de Virgilio Almeida

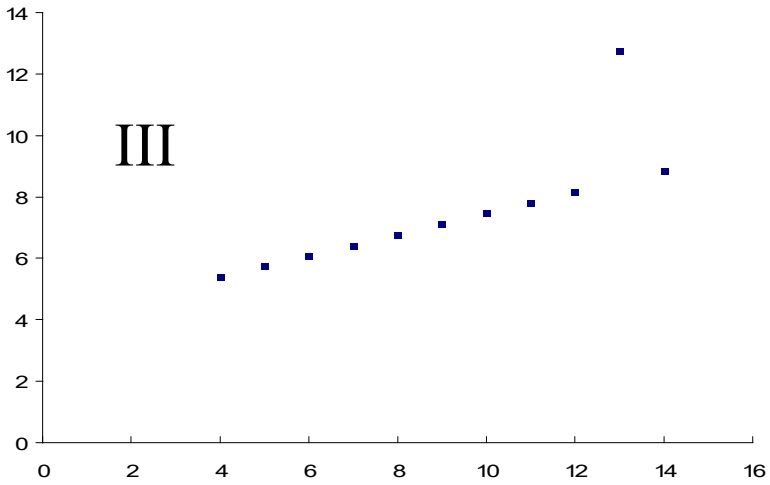
O que a regressão nos diz sobre esses conjuntos de dados?

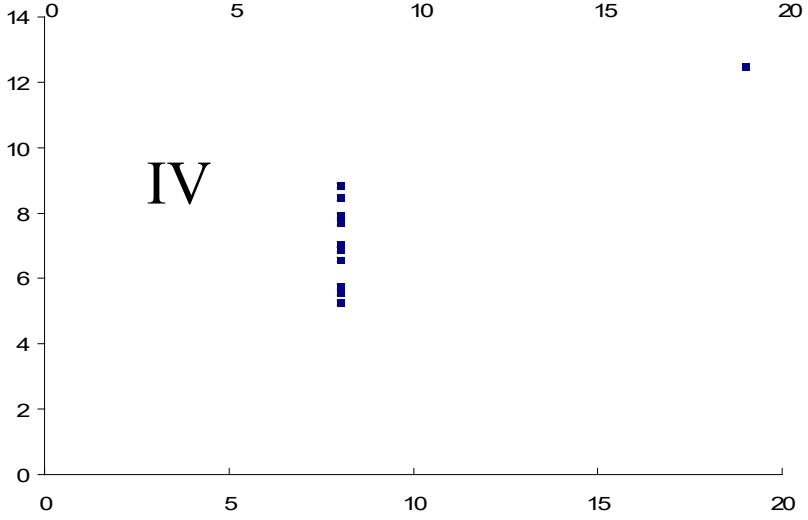
- **Exatamente a mesma coisa para cada um deles!**
- N = 11
- Média de y = 7.5
- Y = 3 + .5 X
- Erro padrão da regressão é 0.118
- Todas as somas de quadrados são as mesmas
- Coeficiente de correlação = .82
- R^2 = .67

Agora, observe estes gráficos ...









Apresentacao derivada dos slides originais de Virgilio Almeida

Sobre os gráficos anteriores

- Importância da inspeção visual dos dados experimentais...

Exemplo

- The number of disk I/O's and processor times of seven programs were measured as: (14, 2), (16, 5), (27, 7), (42, 9), (39, 10), (50, 13), (83, 20)
- For this data:

  n=7, Σ xy=3375, Σ x=271, Σ x²=13,855,
- *Σ y=66, Σ y²=828,* $\bar{x}$ *= 38.71,* $\bar{y}$ *= 9.43. Therefore,*

$$b_1 = \frac{\sum xy - n\bar{x}\bar{y}}{\sum x^2 - n(\bar{x})^2} = 0.2438 \qquad b_0 = \bar{y} - b_1\bar{x} = -0.0083$$

$$\textit{Modelo linear}: \quad \text{CPU time} = -0.0083 + 0.2438\,(\#\text{Disk I/Os})$$

Apresentacao derivada dos slides originais de Virgilio Almeida

Exemplo
Computacao do Erro

	Disk I/O's x_i	CPU Time y_i	Estimate $\hat{y}_i=b_0+b_1\ x_i$	Error $e_i=y_i-\hat{y}_i$	Error^2 e_i^2
	14	2	3.4043	-1.4043	1.9721
	16	5	3.8918	1.1082	1.2281
	27	7	6.5731	0.4269	0.1822
	42	9	10.2295	-1.2295	1.5116
	39	10	9.4982	0.5018	0.2518
	50	13	12.1795	0.8205	0.6732
	83	20	20.2235	-0.2235	0.0500
Σ	271	66	66.0000	0.00	5.8690

Apresentacao derivada dos slides originais de Virgilio Almeida

Exemplo
Alocacao da Variacao

$$\begin{aligned}
\text{SSE} &= \Sigma y^2 - b_0 \Sigma y - b_1 \Sigma xy \\
&= 828 + 0.0083 \times 66 - 0.2438 \times 3375 = 5.87
\end{aligned}$$

$$\begin{aligned}
\text{SST} &= \text{SSY} - \text{SS0} = \Sigma y^2 - n(\bar{y})^2 \\
&= 828 - 7 \times (9.43)^2 = 205.71
\end{aligned}$$

$$\text{SSR} = \text{SST} - \text{SSE} = 205.71 - 5.87 = 199.84$$

$$R^2 = \frac{\text{SSR}}{\text{SST}} = \frac{199.84}{205.71} = 0.9715$$

Modelo explica 97% da variacao: MUITO BOM!!!

Apresentacao derivada dos slides originais de Virgilio Almeida

Exemplo
Desvio Padrao dos Erros

$$\begin{array}{lccccccccc} \text{SS}: & \text{SST} & = & \text{SSY} & - & \text{SS0} & = & \text{SSR} & + & \text{SSE} \\ & 205.71 & = & 828 & - & 622.29 & = & 199.84 & + & 5.87 \\ \text{DF}: & 6 & = & 7 & - & 1 & = & 1 & + & 5 \end{array}$$

- The mean squared error is:

$$\text{MSE} = \frac{\text{SSE}}{\text{DF for Errors}} = \frac{5.87}{5} = 1.17$$

- The standard deviation of errors is:

$$s_e = \sqrt{\text{MSE}} = \sqrt{1.17} = 1.08$$

Apresentacao derivada dos slides originais de Virgilio Almeida

Exemplo
Desvio Padrao dos Parametros

$$s_{b_0} = s_e \left[\frac{1}{n} + \frac{\bar{x}^2}{\Sigma x^2 - n\bar{x}^2} \right]^{1/2}$$

$$= 1.0834 \left[\frac{1}{7} + \frac{(38.71)^2}{13,855 - 7 \times 38.71 \times 38.71} \right]^{1/2} = 0.8311$$

$$s_{b_1} = \frac{s_e}{[\Sigma x^2 - n\bar{x}^2]^{1/2}}$$

$$= \frac{1.0834}{[13,855 - 7 \times 38.71 \times 38.71]^{1/2}} = 0.0187$$

Apresentacao derivada dos slides originais de Virgilio Almeida

Exemplo
IC de 90% dos Parametros

0.95 quantil of t variate with 5 degrees of freedom = 2.015

$\Rightarrow$ 90% confidence interval for b_0 is:

$$-0.0083 \mp (2.015)(0.8311) = -0.0083 \mp 1.6747$$
$$= (-1.6830,\ 1.6663)$$

Since, the confidence interval includes zero, the hypothesis that this parameter is zero cannot be rejected at 0.10 significance level. $\Rightarrow$ b_0 is essentially zero.

90% Confidence Interval for b_1 is:

$$0.2438 \mp (2.015)(0.0187) = 0.2438 \mp 0.0376$$
$$= (0.2061,\ 0.2814)$$

Since the confidence interval does not include zero, the slope b_1 is significantly different from zero at this confidence level.

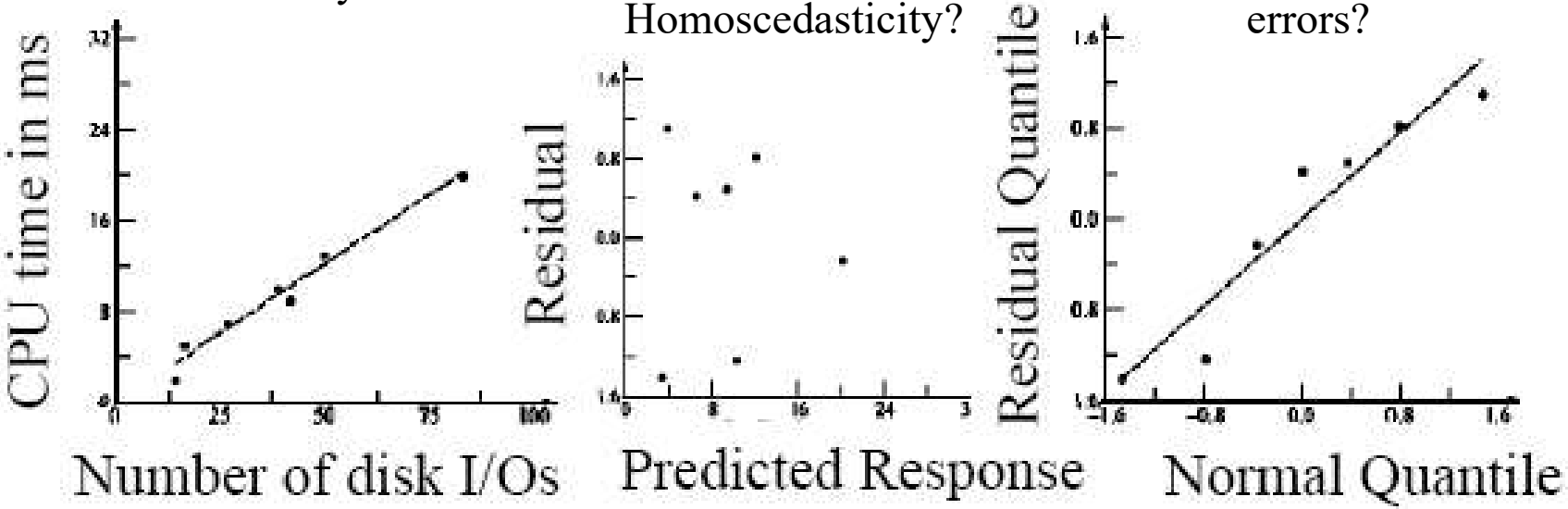
Apresentacao derivada dos slides originais de Virgilio Almeida

Exemplo
Testes Visuais

Linearity?

Independence?
Homoscedasticity?

Normality of errors?



1. Relationship is linear
2. No trend in residuals ⇒ Seem independent
3. Linear normal quantile-quantile plot ⇒ Larger deviations at lower values but all values are small

Apresentacao derivada dos slides originais de Virgilio Almeida

Outros Métodos de Regressão

- Regressão Linear Múltipla
  - mais de uma variável previsora
- Previsores Categóricos
  - **alguns** dos previsores não são quantitativos, mas representam categorias
- Regressão Curvilinear
  - relações não lineares
- Transformações
  - quando erros não são normalmente distribuídos ou variância não é constante
- Tratamento de “outliers”
  - pontos fora do corpo principal
- Erros mais comuns na análise de regressão

Apresentacao derivada dos slides originais de Virgilio Almeida