Regressão Linear Múltipla

- Modelos com mais de uma variável previsora
- Mas cada variável previsora tem uma relação linear com a variável de resposta
- Conceitualmente, seria equivalente a fazer um gráfico de uma linha de regressão num espaço n-dimensional, ao invés de 2-dimensões

Fórmula Básica de Regressão Linear Múltipla

- A resposta y é uma função de k variáveis previsoras $x_1, x_2, \ldots, x_k$

$$y = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_k x_k + e$$

Um Modelo de uma Regressão Linear Múltipla

Dada uma amostra de *n* observações

$$\{(x_{11}, x_{21}, \cdots, x_{k1}, y_1), \cdots, (x_{1n}, x_{2n}, \cdots, x_{kn}, y_n)\}$$

o modelo consiste de *n* equações:

$$y_1 = b_0 + b_1 x_{11} + b_2 x_{21} + \cdots + b_k x_{k1} + e_1$$
$$y_2 = b_0 + b_1 x_{12} + b_2 x_{22} + \cdots + b_k x_{k2} + e_2$$
$$\vdots$$
$$y_n = b_0 + b_1 x_{1n} + b_2 x_{2n} + \cdots + b_k x_{kn} + e_n$$

Sob a forma de aritmética matricial

y = Xb +e

$$
\begin{bmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{k1} \\ 1 & x_{12} & x_{22} & \cdots & x_{k2} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 1 & x_{2n} & x_{2n} & \cdots & x_{kn} \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ \cdot \\ \cdot \\ \cdot \\ b_k \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \cdot \\ \cdot \\ \cdot \\ e_n \end{bmatrix}
$$

Análise de Regressão Linear Múltipla

- Está descrita no *box* 15.1 do Jain.
- Não é essencialmente importante saber como foi derivada, pois nosso curso não é de estatística e nem essa é a finalidade de um curso de métodos quantitativos.
- É importante no entanto saber que existe e como usá-la.
- A maior parte do material é similar a regressão linear simples.
- Um exemplo de duas variáveis.

Exemplo de uma Regressão Linear Múltipla

- Considere uma equipe de segurança de redes desenvolveu vários esquemas alternativos para conter ataques aos servidores e rede. O grupo quer avaliar os mecanismos e definiu um índice de sucesso dos esquemas. O índice foi atribuído pela equipe.

- O índice de sucesso é baseado em dois fatores
  - Tempo do experimento (duração)
  - Número de ataques no período

- Produz uma regressão

$$\text{índice} = b_0 + b_1(\#\text{ataques}) + b_2(\text{duração})$$

Dados amostrais

Esquema	#Ataques	Duração	Índice
A	5	118	8.1
B	13	132	6.8
C	20	119	7.0
D	28	153	7.4
E	41	91	7.7
F	49	118	7.5
G	61	132	7.6
H	62	105	8.0

Aritmética Matricial

- Precisa-se calcular **X**, **X**^{T}, **X**^{T}**X**, (**X**^{T}**X**)$^{-1}$ e **X**^{t}**y**

- Por quê?
- Para obter $$\mathbf{b} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}$$

b = (8.373, .005, -.009)

- Indicando que a regressão prediz:

indice = 8.373 + 0.005*#ataques – 0.009*duração

Matriz X do Exemplo

$$\mathbf{X} = \begin{bmatrix} 1 & 5 & 118 \\ 1 & 13 & 132 \\ 1 & 20 & 119 \\ 1 & 28 & 153 \\ 1 & 41 & 91 \\ 1 & 49 & 118 \\ 1 & 61 & 132 \\ 1 & 62 & 105 \end{bmatrix}$$

Matriz Transposta X^T

$$\mathbf{X}^{\mathbf{T}} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 5 & 13 & 20 & 28 & 41 & 49 & 61 & 62 \\ 118 & 132 & 119 & 153 & 91 & 118 & 132 & 105 \end{bmatrix}$$

Multiplicação Matricial $X^T X$

$$\mathbf{X}^T\mathbf{X} = \begin{bmatrix} 8 & 279 & 968 \\ 279 & 13025 & 33045 \\ 968 & 33045 & 119572 \end{bmatrix}$$

Inversão Matricial $(X^TX)^{-1}$

$$C = \left(\mathbf{X}^T\mathbf{X}\right)^{-1} = \begin{bmatrix} 7.7134 & -0.0227 & -0.0562 \\ -0.0227 & 0.0003 & 0.0001 \\ -0.0562 & 0.0001 & 0.0004 \end{bmatrix}$$

Multiplicação para obter X^Ty

$$\mathbf{X}^T\mathbf{y} = \begin{bmatrix} 60.1 \\ 2118.9 \\ 7247.5 \end{bmatrix}$$

Multiplicação de $(X^TX)^{-1}(X^Ty)$ para obter b

$$\mathbf{b} = \begin{bmatrix} 8.37 \\ 0.005 \\ -0.009 \end{bmatrix}$$

Quão bom é este modelo de regressão?

- Qual a precisão do modelo na previsão do índice de um esquema baseado no #ataques e tempo de duração?
- A melhor forma para determinar isto analiticamente é calcular

$$SSE = \left\{ \mathbf{y}^T \mathbf{y} - \mathbf{b}^T \mathbf{X}^T \mathbf{y} \right\}$$

ou

$$SSE = \sum e_i^2$$

Cálculo dos Erros

Índice	#At.	Dur.	Indice estimado	e_i	e_i^2
8.1	5	118	7.4	-0.71	0.51
6.8	13	132	7.3	0.51	0.26
7.0	20	119	7.4	0.45	0.21
7.4	28	153	7.2	-0.20	0.04
7.7	41	91	7.8	0.10	0.01
7.5	49	118	7.6	0.11	0.01
7.6	61	132	7.5	-0.05	0.00
8.0	62	105	7.8	-0.21	0.04

Cálculo dos Erros

- Assim SSE = 1.08
- SSY = $\sum y_i^2 = 452.91$
- SS0 = $n\bar{y}^2 = 451.5$
- SST = SSY - SS0 = 452.91- 451.5 = 1.4
- SSR = SST - SSE = .33
- $$R^2 = \frac{SSR}{SST} = \frac{.33}{1.41} = .23$$
- Isto é, esta regressão está RUIM!

Por que é ruim?

- Vamos examinar as propriedades dos parâmetros da regresão

$$s_e = \sqrt{\frac{SSE}{n-3}} = \sqrt{\frac{1.08}{5}} = .46$$

Graus de liberdade: n -3 (3 parametros)

- Vamos calcular o desvio padrão dos parâmetros da regressão

Cálculo do Desvio Padrão

- São estimativas, pois estamos trabalhando com uma amostra
- Desvio padrão estimado de:

$$b_0 = s_e\sqrt{c_{00}} = .46\sqrt{7.71} = 1.2914$$

$$b_1 = s_e\sqrt{c_{11}} = .46\sqrt{.0003} = .0097$$

$$b_2 = s_e\sqrt{c_{22}} = .46\sqrt{.0004} = .0083$$

Cálculo de Intervalos de Confiança

- Em um nível de confiança de 90%, por exemplo
- Intervalos de confiança são:

$$b0 = 8.37 \quad \pm \quad (2.015)(1.29) = (5.77, 10.97)$$

$$b1 = .005 \quad \pm \quad (2.015)(.01) = (-.02, .02)$$

$$b2 = -.009 \quad \pm \quad (2.015)(.008) = (-.03, .01)$$

- Somente b_0 é significativo, neste nível

Análise da Variância

- Podemos então dizer que realmente nenhuma das variáveis previsoras é significativa?

- O teste F pode ser usado para essa finalidade
  - Por exemplo, para determinar se o SSR é significativamente maior que o SSE
  - Equivalente a testar se y não depende de qualquer das variáveis previsoras

Executando o F-Teste

- Calcule SSR e SSE e seus graus de liberdade:
  - SSR tem k graus de liberdade (# previsores)
  - SST tem n-1 graus de liberdade
  - Logo: SSE tem n-(k+1) graus de liberdade (k+1 parametros)
- Calcule o quadrado das médias da regressão (MSR) e dos erros (MSE)
  - MSR = SSR/DOF(SSR)
  - MSE = SSE/DOF(SSE)
- MSR/MSE tem uma distribuição F
- Se MSR/MSE > F-tabela, os previsores explicam uma fração significativa da variação da resposta
  - Em outras palavras: SSR e significativamente maior que SSE
  - OU: y depende de *pelo menos uma variavel* previsora
- Vide Tabela 15.3 do Jain: Tabela ANOVA

O F-Teste do Exemplo

- SSR = .33
- SSE = 1.08
- MSR = SSR/k = .33/2 = .16
- MSE = SSE/(n-k-1) = 1.08/(8 - 2 - 1) = .22
- F-calculado = MSR/MSE = .76
- $F_{[90;\ 2,5]}$ = 3.78 (em 90%)
- Assim o teste F falha em 90%

Multipla colinearidade

- Se dois previsores são linearmente dependentes, eles são co-lineares
  - Significa que são relacionados
  - E assim uma segunda variável não melhora a regressão
  - Pode inclusive piorar a regressão.
- Sintoma típico são resultados inconsistentes em vários testes de significância.
  - F-teste da que SSR e significativamente maior que SSE
  - Mas ICs para coeficientes incluem 0

Determinação de Multipla colinearidade

- Deve haver uma correlação entre as variáveis previsoras.
- Se a correlação for alta, elimine uma e repita a regressão sem ela.
- Se a significância da regressão melhorar, deve-se provavelmente à co-linearidade entre as duas variáveis.

A múltipla co-linearidade é um problema no nosso exemplo?

- Provavelmente não, pois não há testes inconsistentes.
- Como verificar?
- Calcular a correlação de #ataques e duração

- O cálculo indica: -.25
  - Não são (fortemente) correlacionados

- Ponto importante: adicionar uma variável previsora nem sempre aumenta a precisão da regressão.

Calculo da Correlacao

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

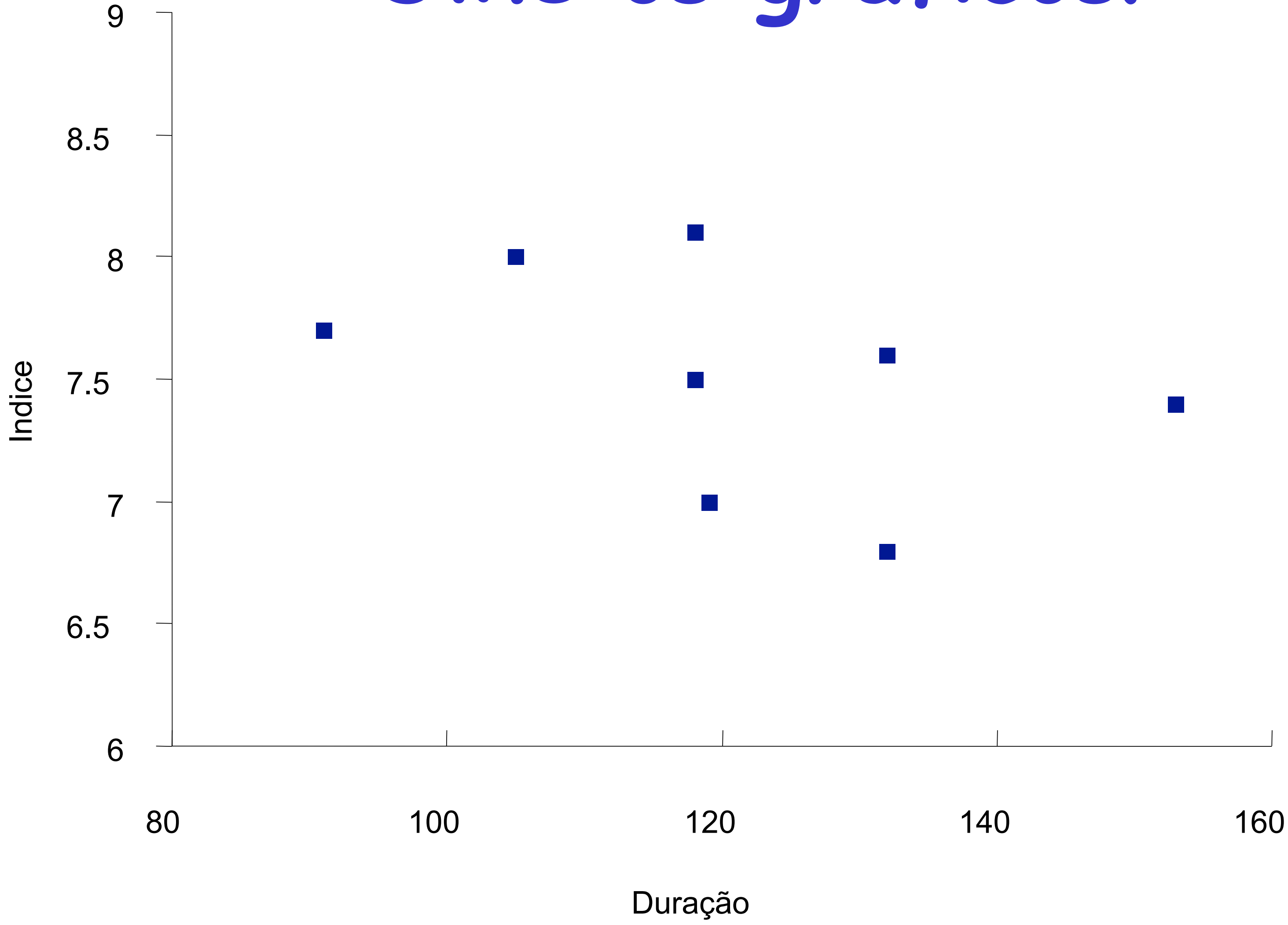
$$\text{Correlação entre } x \text{ e y} = R_{xy} = \frac{s^2{}_{xy}}{s_x s_y}$$

Por que a regressão não funcionou bem neste exemplo?

- Verifique os gráficos de pontos
  - Índice vs. #ataques
  - Índice vs. duração

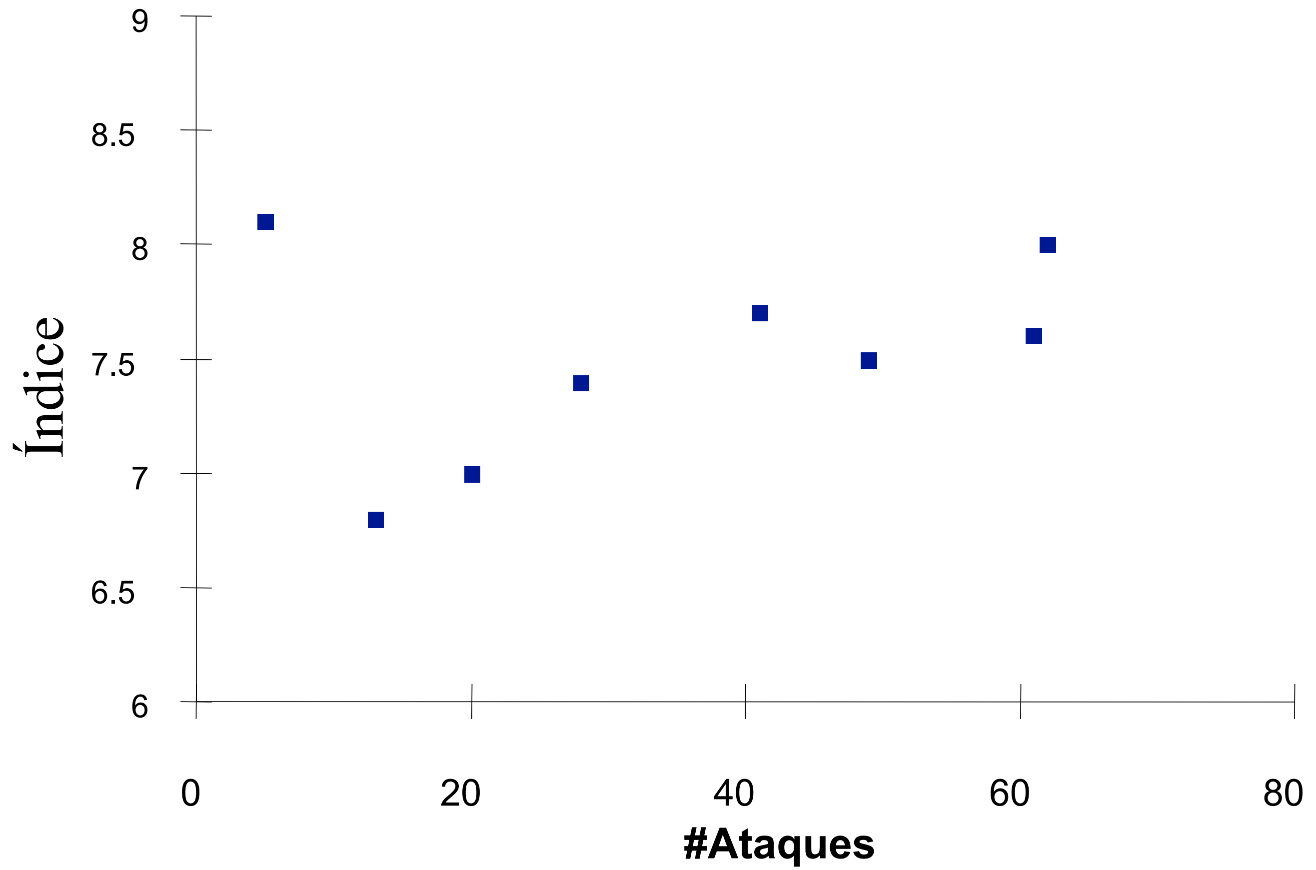
- Independente de quão boa ou ruim é a regressão (coeficiente de determinação), **sempre verifique os gráficos de pontos.**

Olhe os gráficos!



Apresentacao derivada dos slides originais de Virgilio Almeida

Olhe os gráficos!



Exemplo

Sete programas foram monitorados quanto as suas demandas por recursos, particularmente, o numero de operacoes de I/Os (disco), o consumo de memoria (em KB) e o tempo de CPU (em ms). Os dados sao mostrados a seguir

Tempo de CPU y_i	2	5	7	9	10	13	20
Disk I/Os x_{1i}	14	16	27	42	39	50	83
Tamanho da Memoria x_{2i}	70	75	144	190	210	235	400

Encontre um modelo linear para estimar o tempo de CPU em funcao dos outros dois recursos e assim quantificar o impacto do uso destes recursos no tempo de execução)

Exemplo

CPU time = $b_0 + b_1$ (# disk I/Os) + b_2 (tamanho da mem)

$$\mathbf{X} = \begin{bmatrix} 1 & 14 & 70 \\ 1 & 16 & 75 \\ 1 & 27 & 144 \\ 1 & 42 & 190 \\ 1 & 39 & 210 \\ 1 & 50 & 235 \\ 1 & 83 & 400 \end{bmatrix}$$

Exemplo

CPU time = b_0 + b_1 (# disk I/Os) + b_2 (tamanho da mem)

$$\mathbf{X}^T X = \begin{bmatrix} 7 & 271 & 1324 \\ 271 & 13855 & 67188 \\ 1324 & 67188 & 326686 \end{bmatrix}$$

$$C = (\mathbf{X}^T X)^{-1} = \begin{bmatrix} 0.6297 & 0.0223 & -0.0071 \\ 0.0223 & 0.0280 & -0.0058 \\ -0.0071 & -0.0058 & 0.0012 \end{bmatrix}$$

Exemplo

CPU time = b_0 + b_1 (# disk I/Os) + b_2 (tamanho da mem)

$$\mathbf{X}^T y = \begin{bmatrix} 66 \\ 3375 \\ 16388 \end{bmatrix}$$

$$b = (\mathbf{X}^T X)^{-1} \mathbf{X}^T y = \begin{bmatrix} -0.1614 & 0.1182 & 0.0276 \end{bmatrix}^T$$

A equacao de regressao:

Cpu time = -0.1614 + 0.1182(# disk I/Os) + 0.0276(tam. Mem)

Exemplo

Vamos fazer a analise de variancia (ANOVA) da regressao:

Calculo das previsoes, erros e erros quadrados

y_i	2	5	7	9	10	13	20
x_{1i}	14	16	27	42	39	50	83
x_{2i}	70	75	144	190	210	235	400
$\hat{y}_i$	3.3490	3.7180	6.8472	9.8400	10.0151	11.9783	20.2529
e_i	-1.3490	1.2820	0.1528	-0.8400	-0.0151	1.0217	-0.2529
$(e_i)^2$	1.8198	1.6436	0.0233	0.7053	0.0002	1.0439	0.0639

$$SSE = \sum_i e_i^2 = 5.3 = \{y^T y - b^T X^T y\}$$

Exemplo

Calculo dos SS*

$$SSY = \sum_i y_i^2 = 828 \qquad SS0 = n\bar{y}^2 = 622.29$$

$$SST = SSY - SS0 = 828 - 622.29 = 205.71$$

$$SSR = SST - SSE = 205.71 - 5.3 = 200.41$$

$$R^2 = \frac{SSR}{SST} = \frac{200.41}{205.71} = 0.97$$

A regressao explica 97% da variabilidade dos dados: BOM!

Exemplo

Calculo do desvio padrao dos erros e dos coeficientes

$$s_e = \sqrt{\frac{SSE}{n-3}} = \sqrt{5.3/4} = 1.2$$

Desvio padrao estimado para

$$\mathrm{b}_0 = s_e\sqrt{c_{00}} = 1.2\sqrt{0.6297} = 0.9131$$

$$b_1 = s_e\sqrt{c_{11}} = 1.2\sqrt{0.0280} = 0.1925$$

$$b_2 = s_e\sqrt{c_{22}} = 1.2\sqrt{0.0012} = 0.0404$$

Exemplo

Calculo dos CI de 90%:

95% da variavel t com 4 graus de liberdade $t_{0.95,4} = 2.132$

$$b_0 = -0.1614 \pm (2.132)(0.9131) = (-2.11, 1.79)$$
$$b_1 = 0.1182 \pm (2.132)(0.1925) = (-0.29, 0.53)$$
$$b_2 = 0.0265 \pm (2.132)(0.0404) = (-0.06, 0.11)$$

Nenhum parametro e significativo

Exemplo

Realizando o teste F:

SSE = 5.3

Graus de liberdade do SSE = n-(k+1) = n-3 = 4

MSE = SSE/n-(k+1) = 5.3/4 = 1.33

SSR = 200.41

Graus de liberdade do SSR = k = 2

MSR = 200.41/2 = 100.205

MSR / MSE = 75.40 Tabela F: 4.32

Ja que MSR/MSE > F -> regressao passou o teste F

Isto significa que a hipotese de que todos parametros sao 0 nao pode ser aceita.

Inconsistencia???

Exemplo

Vamos calcular a correlacao entre as variaveis previsoras (numeros de I/Os e tamanho de memoria)

$$n = 7 \qquad \sum x_{1i} = 271 \qquad \sum x_{2i} = 1324$$

$$\sum x_{1i}^2 = 1385 \qquad \sum x_{2i}^2 = 32668$$

$$\sum x_{1i} x_{2i} = 67188$$

$$Correlacao(x_1, x_2) = R_{x_1, x_2} =$$

$$\frac{\sum x_{1i} x_{2i} - \frac{1}{n}(\sum x_{1i})(\sum x_{2i})}{\left[\sum x_{1i}^2 - \frac{1}{n}(\sum x_{1i})(\sum x_{1i})\right]^{1/2}\left[\sum x_{2i}^2 - \frac{1}{n}(\sum x_{2i})(\sum x_{2i})\right]^{1/2}}$$

$$= 0.9947$$

Exemplo

- Alta correlacao: multicolineariedade prejudica a regressao.
- Precisa refazer regressao somente com # de I/Os e, separadamente, com tamanho de memoria, e escolher melhor previsor (isto e, aquele que resulta no maior R2)
  - Neste caso e regressao linear simples

Regressão com Previsores Categóricos

- Os métodos de regressão vistos ate aqui assumiram valores numéricos!
- O que acontece se algumas variaveis são por natureza categóricas, não numéricas? Por exemplo, o tipo de processador pode ser uma variável categórica.
- Existem técnicas se todas variáveis são categóricas.
  – Projetos fatoriais: estatisticamente mais precisos
- As tecnicas apresentadas a seguir sao para regressoes com previsores mistos (alguns categoricos e outros numericos)
- **Níveis** – número de valores que uma categoria pode assumir.

Trabalhando com Previsores Categóricos

- Se somente dois níveis são usados, defina x_i assim:
  - $x_i = 0$ para primeiro valor, $x_i = 1$ para segundo valor

  b_i representa a diferenca no efeito das duas alternativas

- Pode-se usar +1 and -1 como valores, também.

  $2b_i$ representa a diferenca entre duas alternativas

Trabalhando com Previsores Categóricos

- Precisa-se de k-1 variáveis previsoras para k níveis
  - Para evitar implicações de ordem nas categorias

$$\cancel{x_1 = \begin{cases} 1 \rightarrow tipo\,A \\ 2 \rightarrow tipo\,B \\ 3 \rightarrow tipo\,C \end{cases}}$$

Reflete B no meio entre A e C
Parametros sem significado

$$(x_1, x_2) = (1,0) \rightarrow tipo\,A$$
$$(x_1, x_2) = (0,1) \rightarrow tipo\,B$$
$$(x_1, x_2) = (0,0) \rightarrow tipo\,C$$

$$y = b_0 + b_1 x_1 + b_2 x_2 + e$$

$$\overline{y_A} = b_0 + b_1$$

$$\overline{y_B} = b_0 + b_2$$

$$\overline{y_C} = b_0$$

Exemplo de Variáveis Categóricas

- O desempenho de uma chamada de procedimento remota (RPC) foi comparada em dois sistemas operacionais UNIX e ARGUS. A metrica avaliada foi o tempo total para diferentes tamanhos de dados. A Tabela abaixo mostra os resultados das medicoes.

Unix:

Data bytes	64	64	64	64	234	590	846	1060	1082	1088	1088	1088	1088
Tempo	26.4	26.4	26.4	26.2	33.8	41.6	50.0	48.4	49.0	42.0	41.8	41.8	42.0

Argus:

Data bytes	92	92	92	92	348	604	860	1074	1074	1088	1088	1088	1088
Tempo	32.8	34.2	32.4	34.4	41.4	51.2	76.0	80.8	79.8	58.6	57.6	59.8	57.4

Qual o custo de processamento por byte para os dois sistemas? E o custo de setup?

Exemplo de Variáveis Categóricas

- $y = b_0 + b_1x_1 + b_2x_2$
  - y = tempo de processamento da RPC
  - x_1= numero de bytes
  - x_2= 1 se sistema e Unix, e 0 se sistema e Argus

- Resultado da Regressao:

Parametro	Media	Desvio Padrao	IC
b_0	36.739	3.251	(31.1676,42.3104)
b_1	0.025	0.004	(0.0192, 0.0313)
b_2	-14.927	3.165	(-20.3509,-9.5024)

R^2 = 0.765

Custo por byte em ambos sistemas e 0.025 milisegundos
Custo de setup e 36.73 ms no ARGUS e (36.739 – 14.927) no UNIX

Premissa da solucao: custo per byte independe do sistema operacional.
E se isto nao for verdade?

Regressão Curvilinear

- Regressão linear assume relações lineares entre variáveis previsoras e a resposta.
- O que acontece quando essas relações não são lineares?
  - Coeficientes de determinação R^2 pobres
- É necessário encontrar outro tipo de função para a relação entre previsores e resposta.

Quando devemos usar uma regressão curvilinear?

- A forma mais direta é fazer uma inspeção visual nos dados.
- Faça um gráfico de pontos
  - Se o gráfico não se apresenta como linear (alguma indicação de linearidade), use então uma regressão curvilinear.
- Ou então quando há outras razões para suspeitar que as relações não são lineares (ex., fenômenos claramente modelados por *power laws, Zipfs Law*, etc).
- Relações devem ser convertidas para formas lineares.

Tipos de Regressão Curvilinear

- Existem muitos tipos possíveis, baseados numa variedade de relações entre as variáveis:

$$y = bx^{a}$$

$$y = a + b/x$$

$$y = ab^{x}$$

- Existem várias outras possibilidades

Transformação para Relações Lineares

- Use qualquer transformação que leve a representar a relação através de funções de forma linear, como : logaritmos, multiplicação, divisão, etc.

- Quer se obter algo como:

$$y' = a + bx'$$

- y' e x' obtidos com a transformacao

Funções de Regressão CurviLineares

$$\text{NaoLinear} \quad \Rightarrow \text{Linear}$$

$$y = a + b/x \quad \Rightarrow y = a + b(1/x) \qquad x' = 1/x$$

$$y = 1/(a + bx) \quad \Rightarrow 1/y = a + bx \qquad y' = 1/y$$

$$y = a \times b^x \quad \Rightarrow \ln y = \ln a + x \ln b$$

$$y = a + bx^n \quad \Rightarrow y = a + b(x^n)$$

Transformações

- O termo transformação é usado quando uma função da variável de resposta medida é usada no lugar da própria variável.
- Usar alguma função da variável resposta y (w = h(y)) em lugar do próprio y.
- Regressão curvilinear é um exemplo dessa transformação.

$$y = \frac{a}{b} x^{\alpha + 1}$$

$$\ln y = \ln(\frac{a}{b}) + (\alpha + 1) \ln x$$

$$y' = A + Bx'$$

- As técnicas tem aplicação mais geral

Quando transformar?

1. Quando as propriedades físicas conhecidas do sistema medido sugerem que a função da resposta, ao invés da própria resposta, é uma variável melhor para o modelo. Exemplo: mediu-se tempos entre chegadas mas sabe-se que relacao linear e valida para taxa de chegadas.

$$1/y \text{ melhor que } y$$

2. Quando o intervalo dos dados medidos cobre várias ordens de grandeza e a amostra e pequena. Deve-se buscar uma transformacao que reduza a variabilidade.

   Exemplo:

$$y_{\max}/y_{\min} \text{ é grande}$$

3. Quando a hipótese de uma variância homogênea dos resíduos é violada (i.e. *Homoscedasticity)*.

Transformação Devida a *Homoscedasticity*

- Se num gráfico de pontos dos resíduos (erros) versus a resposta prevista, o espalhamento não é homogêneo.
- Então os resíduos são ainda uma função das variáveis previsoras.
- A transformação da resposta pode resolver o problema.

Qual transformação deve-se usar?

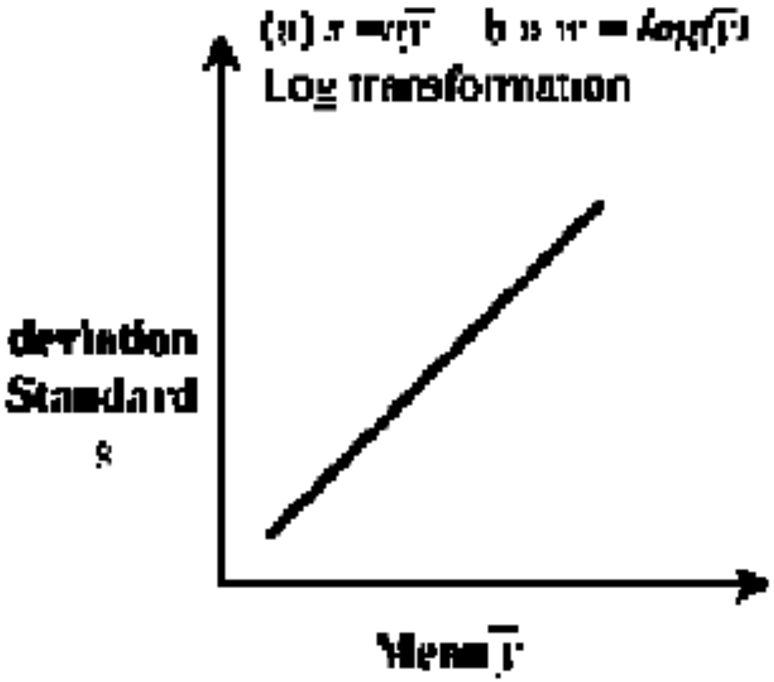
- Calcule o desvio padrão dos resíduos para cada estimativa $\hat{y}_i$.
  - Deve haver mais de um residuo para cada valor estimado para x_i.
  - Considere múltiplos experimentos para um conjunto de valores previsores.

Qual transformação deve-se usar?

- Coloque num gráfico de pontos esses desvios como função da média das observações para $\hat{y}_i$.
  - se for linear então use a transformacão logaritmica.

$$s = a\hat{y}_i + b$$

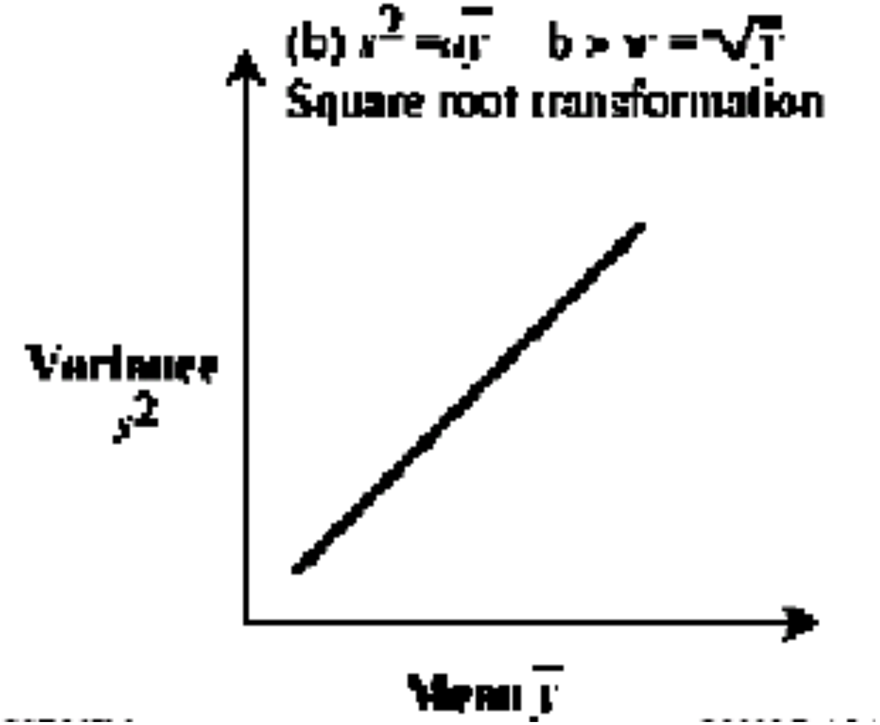
$$w = h(y) = \ln(y)$$



Apresentacao derivada dos slides originais de Virgilio Almeida

Outros testes para transformações

- Se a variância versus a média das observações medidas é linear, use uma transformação de raíz quadrada:
  w = sqrt(y)



Outros testes para transformações

- Se o desvio padrão versus o quadrado da média é linear, use uma transformação inversa: w = 1/sqrt(y)

- Se o desvio padrão versus a média elevada a uma potência a é linear use uma transformação de potência: $w = y^{1-a}$

- Outras transformações estão descritas no livro do Jain.

- Ao final basta fazer a regressao para

$$w = b_0 + b_1 x_1 + \ldots + b_k x_k + e$$

Outliers

- Medidas observadas em experimentos tipicamente contem outliers (i.e., valores muito fora do corpo da curva)
  - Medidas que não são uma característica verdadeira do sistema.
  - Erros podem ter ocorrido no processo experimental de medição.
  - Comportamentos atípicos de usuários do sistema podem existir (ex: um *nerd* que joga um game 15 horas consecutivas, quando se está analisando tempos de conexão a um provedor de serviços)
- Isso resulta no seguinte problema:
  - Devemos ou não incluir os outliers nas análises que estamos fazendo?

Como tratar os outliers?

1. Determine os outliers, analisando por exemplo os gráficos de pontos.
2. Verifique cuidadosamente os erros experimentais
3. Repita os experimentos com valores previsores para os outliers e valores proximos a eles.
4. Decida se deve ou não incluir os outliers:
   - Verifique se os outliers são parte do sistema ou se são exceções que podem ser desprezadas.
   - Analise os dados com e sem os outliers e veja o que faz mais sentido.
   - Todas as análises dependem da natureza do sistema em estudo.

Erros mais comuns nas análises usando regressões

- Geralmente baseadas em “atalhos” ou simplificação excessiva dos dados.
- Realizada sem cuidados e técnicas fundamentadas.
- Falta de entendimento dos princípios fundamentais de estatística.
- Falta de entendimento dos princípios fundamentais do método científico.

Não verificação da linearidade

- Desenhe o gráfico de pontos
- Se não for linear, verifique as possibilidades curvilineares e suas transformações.
- O uso de uma regressão linear quando as relações entre resposta e previsores não são lineares é um ERRO!

Basear em resultados sem uma inspeção visual

- Sempre verifique o gráfico de pontos, como parte das análises usando regressões.
  - Examine a linha de regressão prevista versus os pontos reais obtidos pelo experimento.
- Isso é particularmente importante no caso de uso de pacotes que fazem regressões automaticamente.

Atribuição de importância aos valores dos parâmetros

- Valores numéricos da regressão dependem da escala das variáveis previsoras.
- Não é devido ao fato de um valor ser pequeno ou grande que é necessariamente uma indicação de importância.
- Exemplo:
  - Converter segundos para microsegundos não muda nada fundamental no problema
  - Mas muda a magnitude dos valores dos parâmetros associados.

Exemplo

- Tempo de CPU em segundos = 0.01*(# oper. E/S) +
  0.001*(tamanho da memória em Mbytes)

- Tempo de CPU em milisegundos = 10*(# oper. E/S) +
  1*(tamanho da memória em Mbytes)

- Valores absolutos dos parâmetros podem ser enganadores!

- A forma correta de comparar a significância de um parâmetro da regressão é através de seu intervalo de confiança.

Ausência de cálculo de Intervalos de Confiança

- As amostras das observações medidas são aleatórias.
- Assim, a regressão executada nessas amostras gera parâmetros com propriedades aleatórias também.
- Sem intervalos de confiança, é impossível entender o significado e a confiança que se tem nos valores dos parâmetros.

Ausência de cálculo do Coeficiente de Determinação (R^2)

- Sem o cálculo de R^2, é difícil determinar quanto da variação é explicada pela regressão.

Uso Inadequado do Coeficiente de Correlação

- Coeficiente de determinação é R^2
- Coeficiente de correlação é R
- R^2 dá o percentual da variacao que é explicada pela regressão, e isso é diferente de R
- Exemplo
  - se R é 0.6, então $R^2 = 0.36$
  - a regressão explica apenas 36% da variação nos dados
  - não 60%!!

Uso de variáveis previsoras altamente correlacionadas

- Se duas variáveis previsoras são correlacionadas, o uso de ambas variáveis pode degradar a regressão.
- Exemplo:
  - num servidor Web é provável haver correlação entre tamanho de um arquivo e sua popularidade
  - assim, não use os dois num modelo de previsão de *cache hit ratio*
- O exemplo mostra que é necessário conhecer bem as variáveis previsoras e suas possíveis relações

Uso de muitas variáveis previsoras

- O acréscimo de mais variáveis previsoras não necessariamente melhora a qualidade do modelo.
- Pode-se criar problemas como o de multi-colinearidade
- Quais variáveis devem então ser usadas?
  - É o que estamos tentando aprender neste curso

Medindo um intervalo pequeno de valores ou medindo intervalos não significativos

- Uma regressão somente prevê bem valores próximos do intervalo observado de mediçoes.
- Se não forem feitas medições dos intervalos mais comuns de operação do sistema, a regressão não irá prever muita coisa.
- Exemplos
  – Se muitos programas são maiores que a memória real disponível, então medir aqueles que são menores, pode ser um erro, pois fatores como overhead estariam sendo ignorados quando fosse feita uma previsão de programas maiores.
  – Se o experimento mede os tempos de execução de queries de um conjunto de palavras pouco frequentes, então prever os tempos de palavras muito frequentes, pode ser um erro, pois há efeitos como *caching* que não estariam sendo considerados.

Uso de regressão muito além do intervalo de observação

- A regressão é baseada no comportamento observado de uma amostra em particular (ou conjunto de amostras). Refere-se ao comportamento do sistema numa certa faixa de valores
- É mais seguro prever dentro de uma faixa compatível com o intervalo de valores observados na medição
  - Valores muito além podem ser previstos?
- Exemplos
  - Uma regressão do tempo de execução de módulos de código que são menores que o tamanho de memória disponível, pode não ser capaz de prever o tempo de módulos que fazem muito uso de memória virtual.
  - A previsão do número de queries que chega numa máquina de busca baseada numa regressão sobre valores de um log de vários dias pode não ser capaz de prever o que acontecerá meses a frente.

Exemplo 2

- A Lei de Amdahl para operacoes de I/Os em sistemas de computacao diz que a taxa de I/O e linearmente proporcional a velocidade do processador.

  Para validar a lei, os numeros de I/Os e as utilizacoes de CPU de um numero de computadores foram medidos. Usando a taxa MIPS nominal para o sistema e a sua utilizacao, a taxa de processamento de instrucoes (em MIPS) e a taxa de I/O (em KB/s) foram computados para um periodo. Os dados foram mostrados abaixo. Voce consegue validar/refutar a Lei de Amdahl com os dados abaixo?

Sistema	1	2	3	4	5	6	7	8	9	10
MIPS Usado	19.63	5.45	2.63	8.24	14	9.87	11.27	10.13	1.01	1.26
Taxa de I/O	288.6	117.3	64.6	356.4	373.2	281.1	149.6	120.6	31.1	23.7

Exemplo 2

- Vamos assumir, por hora, o seguinte modelo curvilinear:

  $$\text{I/O rate} = \alpha\,(\text{MIPS rate})^b$$

  $$\log(\text{I/O rate}) = \log \alpha + b \log(\text{MIPS rate})$$

Os parametros $b_0 = \log \alpha$ e $b_1 = b$ podem ser estimados via regressao linear simples

Parametro	Media	Desvio Padrao	CI 90%
b_0	1.423	0.119	(1.20, 1.64)
b_1	0.888	0.135	(0.64,1.14)

R2 = 0.84 -> boa regressao

Os dois coeficientes sao significativos com a confianca de 90%.

Alem disto, como o CI para b1 contem 1, podemos aceitar a hipotese de que o relacionamento entre I/O rate e MIPS rate e linear.

Exemplo 3

Os resultados de uma regressao linear multipla baseada em nove observacoes estao mostrados na tabela abaixo. Baseado nestes resultados responda as perguntas a seguir.

j	1	2	3	4
b_j	1.3	2.7	0.5	5.0
s_{b_j}	3.6	1.8	0.6	0.3

Ponto de Intersecao = 75.3

Coeficiente de correlacao multipla = 0.95

Desvio padrao dos erros = 12.0

Exemplo 3

Qual porcentagem da variacao e explicada pela regressao?

A regressao e significativa, com uma confianca de 90%?

Quais parâmetros sao significativos com uma confianca de 90%?

Exemplo 3

Qual porcentagem da variacao e explicada pela regressao?

$R = 0.95 \quad \Rightarrow \quad R^2 = 0.95*0.95 = 0.9025$

90.25% da variacao e explicada pela regressao

A regressao e significativa, com uma confianca de 90%?

Desvio padrao dos erros $s_e = \text{sqrt}\ (SSE/n\text{-}k\text{-}1)$

$SSE = (n\text{-}k\text{-}1)*\ (s_e)^2 = (9 - 5)*12*12 = 576$

$R^2 = SSR\ /\ SST \quad = SSR\ /\ (SSR + SSE)$

$SSR/(SSR + 576) = 0.9025 \Rightarrow SSR = 519.84/0.0975 = 5331.69$

$MSR = SSR/k = 5331.69/4 = 1332.92$

$MSE = SSE/(n\text{-}k\text{-}1) = 576/4 = 144$

$MSR/MSE = 9.256$

F-value $(0.9,4,4) = 4.11 \quad \Rightarrow$ sim, a regressao e significativa

Exemplo 3

Quais parametros sao significativos com uma confianca de 90%?

Calcular IC : $b_j \pm t^* s_{bj}$

0.95 quantil da variavel t com n-k-1 (= 4) graus de liberdade = 2.132

CI para $b_1 = 1.3 \pm 2.132*3.6 = (-6.38, 8.98)$: nao e significativo pois inclui zero.

CI para $b_2 = 2.7 \pm 2.132*1.8 = (-1.14, 6.54)$: nao e significativo

CI para $b_3 = 0.5 \pm 2.132*0.6 = (-0.78, 1.7792)$: nao e significativo

CI para $b_4 = 5.0 \pm 2.132*8.3 = (-12.70,22.70)$: nao e significativo

Nenhum parametro e significativo com confianca de 90%

Exemplo 3

Qual o problema com a regressao e qual seria o seu proximo passo?

Pode ser um problema de multicolinearidade.
Testar correlacao entre varios pares de previsores.
Dentre os pares que tiverem alta correlacao, testar a regressao com cada previsor separadamente e escolher aquele que resulta no melhor R^2