Métodos Quantitativos
Aula 07

Regressão e Predição (Parte 2)

Roberto Massi de Oliveira
Alex Borges Vieira

Revisão: Operações e Propriedades Matriciais

- Soma de Matrizes:

$$A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{bmatrix} \; B = \begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{bmatrix}$$

$$A + B = \begin{bmatrix} 1+1 & 2+2 \\ 3+3 & 4+4 \\ 5+5 & 6+6 \end{bmatrix}$$

$$C = \begin{bmatrix} 2 & 4 \\ 6 & 8 \\ 10 & 12 \end{bmatrix}$$

```
import numpy as np
```

```
A = ([[1, 2],
      [3, 4],
      [5, 6]])

B = ([[1, 2],
      [3, 4],
      [5, 6]])

C = np.add(A,B)

print(C)
```

```
[[ 2  4]
 [ 6  8]
 [10 12]]
```

Revisão: Operações e Propriedades Matriciais

- Subtração de matrizes:

$$A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{bmatrix} \; B = \begin{bmatrix} 2 & 1 \\ 8 & 4 \\ 5 & 10 \end{bmatrix}$$

$$A - B = \begin{bmatrix} 1-2 & 2-1 \\ 3-8 & 4-4 \\ 5-5 & 6-10 \end{bmatrix}$$

$$C = \begin{bmatrix} -1 & 1 \\ -5 & 0 \\ 0 & -4 \end{bmatrix}$$

```python
import numpy as np
```

```python
A = ([[1, 2],
      [3, 4],
      [5, 6]])

B = ([[2, 1],
      [8, 4],
      [5, 10]])

C = np.subtract(A,B)

print(C)
```

```
[[-1  1]
 [-5  0]
 [ 0 -4]]
```

Revisão: Operações e Propriedades Matriciais

- Multiplicação de matrizes: $A_{m \times n} \times B_{n \times p} = C_{m \times p}$

$$A = \begin{bmatrix} 2 & 3 \\ 4 & 6 \end{bmatrix} \quad B = \begin{bmatrix} 1 & 3 & 0 \\ 2 & 1 & 1 \end{bmatrix}$$

$$A \times B = \begin{bmatrix} \boxed{2 \quad 3} \\ \boxed{4 \quad 6} \end{bmatrix} \times \begin{bmatrix} \boxed{\begin{matrix} 1 \\ 2 \end{matrix}} & \boxed{\begin{matrix} 3 \\ 1 \end{matrix}} & \boxed{\begin{matrix} 0 \\ 1 \end{matrix}} \end{bmatrix} =$$

$$\begin{bmatrix} 2 \times 1 + 3 \times 2 & 2 \times 3 + 3 \times 1 & 2 \times 0 + 3 \times 1 \\ 4 \times 1 + 6 \times 2 & 4 \times 3 + 6 \times 1 & 4 \times 0 + 6 \times 1 \end{bmatrix} = \begin{bmatrix} 8 & 9 & 3 \\ 16 & 18 & 6 \end{bmatrix}$$

Revisão: Operações e Propriedades Matriciais

- Multiplicação de matrizes: $A_{m \times n} \times B_{n \times p} = C_{m \times p}$

$$A = \begin{bmatrix} 2 & 3 \\ 4 & 6 \end{bmatrix} \quad B = \begin{bmatrix} 1 & 3 & 0 \\ 2 & 1 & 1 \end{bmatrix}$$

$$A \times B = \begin{bmatrix} 2 & 3 \\ 4 & 6 \end{bmatrix} \times \begin{bmatrix} 1 & 3 & 0 \\ 2 & 1 & 1 \end{bmatrix} =$$

$$\begin{bmatrix} 8 & 9 & 3 \\ 16 & 18 & 6 \end{bmatrix}$$

```python
import numpy as np
```

```python
a = np.array([[2, 3],
              [4, 6]])

b = np.array([[1, 3, 0],
              [2, 1, 1]])

ab = np.matmul(a,b)

print(ab)
```

```
[[ 8  9  3]
 [16 18  6]]
```

Revisão: Operações e Propriedades Matriciais

- Inversão de matrizes (apenas matrizes quadradas):

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 3 & 1 \\ 1 & 2 & 0 \end{bmatrix} \qquad \mathbf{A}^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ -1/2 & 0 & 1/2 \\ 1/2 & 1 & -3/2 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 0 & 0 \\ 1 & 3 & 1 \\ 1 & 2 & 0 \end{bmatrix} \cdot \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$\begin{bmatrix} a & b & c \\ a + 3d + g & b + 3e + h & c + 3f + i \\ a + 2d & b + 2e & c + 2f \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Revisão: Operações e Propriedades Matriciais

- Inversão de matrizes (apenas matrizes quadradas):

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 3 & 1 \\ 1 & 2 & 0 \end{bmatrix}$$

$$\mathbf{A}^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ -1/2 & 0 & 1/2 \\ 1/2 & 1 & -3/2 \end{bmatrix}$$

```python
import numpy as np
```

```python
a = np.array([[1, 0, 0],
              [1, 3, 1],
              [1, 2, 0]])

ai = np.linalg.inv(a)

print(ai)
```

```
[[ 1.   0.   0. ]
 [-0.5  0.   0.5]
 [ 0.5  1.  -1.5]]
```

Revisão: Operações e Propriedades Matriciais

- Transposição de matrizes:

$$A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{bmatrix} \Rightarrow A^t = \begin{bmatrix} 1 & 3 & 5 \\ 2 & 4 & 6 \end{bmatrix}$$

```python
import numpy as np
```

```python
a = np.array([[1, 2],
              [3, 4],
              [5,6]])

at = np.transpose(a)

print(at)
```

```
[[1 3 5]
 [2 4 6]]
```

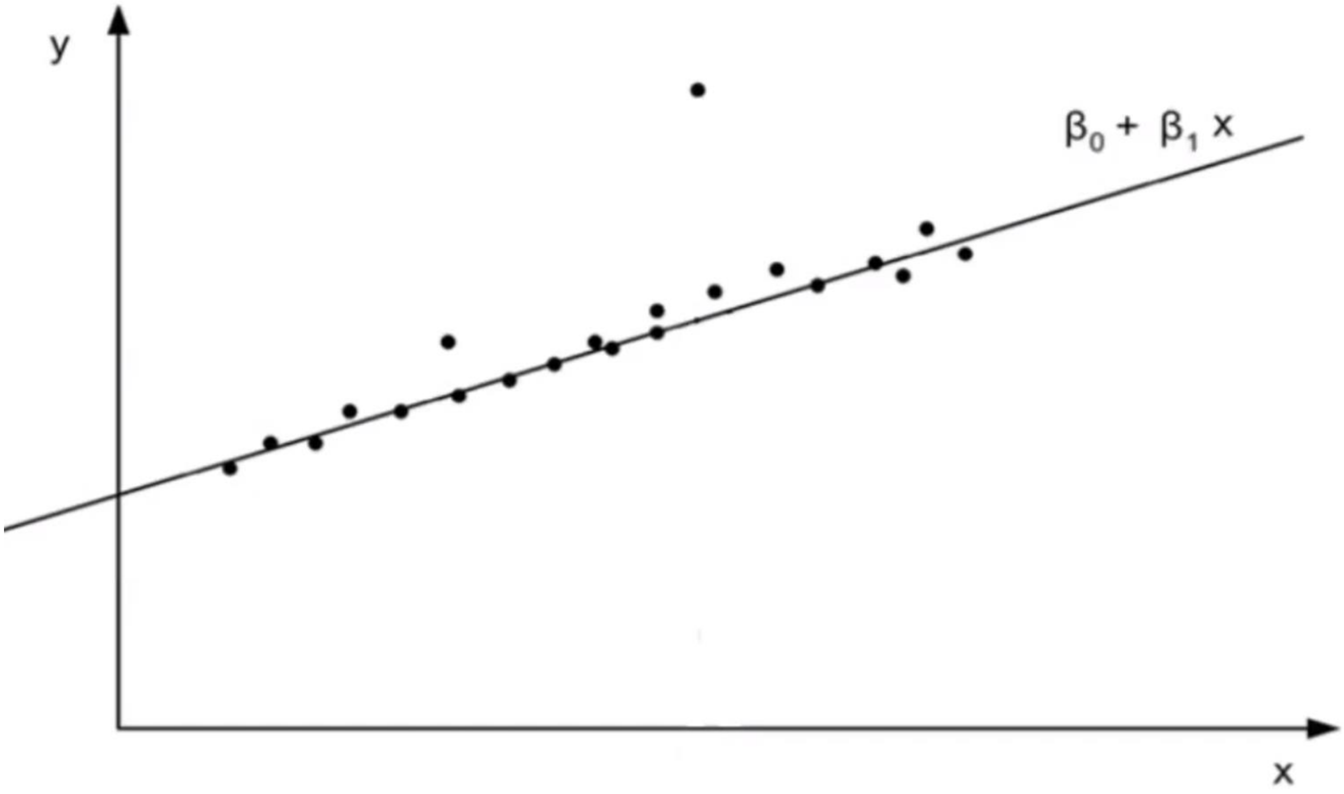
Revisão: Regressão Linear Simples

- Modela a relação entre duas variáveis
- Sendo essa relação linear, ela é matematicamente expressa por:

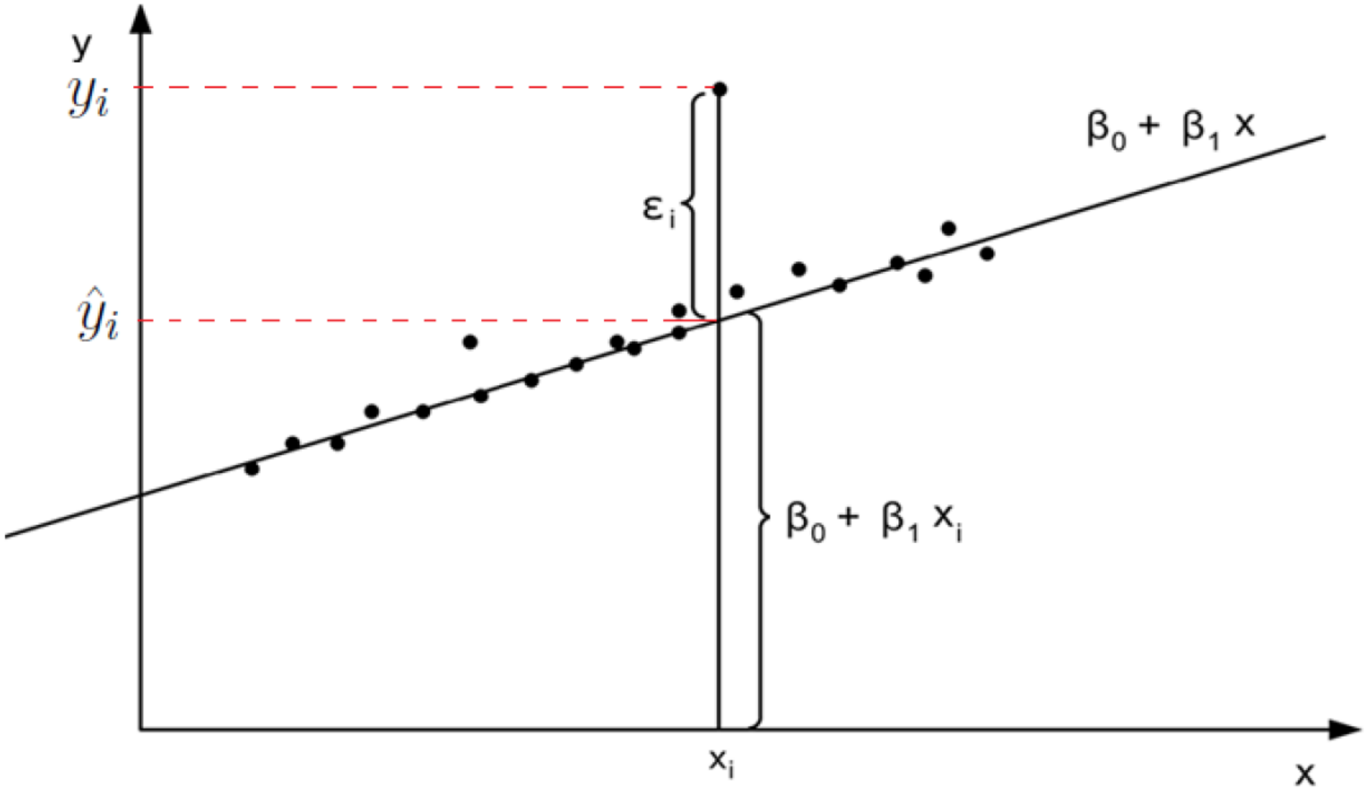
$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

  - **y** é a variável **dependente** ou variável **resposta**
  - **x** é a variável **independente**, **regressora**, **explicativa** ou **previsora**
  - **ε** é o termo de **erro**. É a flutuação aleatória que ocorre ao tentar explicar a variável **y** por **x.** Seja por imperfeições do modelo, erros de medida, ou outras variáveis fora de controle.

Revisão: Regressão Linear Simples



Visão Geral: Regressão Linear Simples



Revisão: Regressão Linear Simples

- Usaremos uma amostra $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ para estimar os parâmetros do modelo β_0 e β_1
- Métodos de estimação:
  - Mínimos quadrados ordinários (MQO)
  - Máxima verossimilhança (MV)
  - Método os momentos (MM)
  - Melhor estimador não-enviesado (BLUE)
- A seguir, faremos inferências acerca dos parâmetros β_0 e β_1
  - ex.: propriedades dos estimadores, intervalos de confiança, testes de hipótese

Revisão: Regressão Linear Múltipla

- Modelo linear:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \varepsilon$$

- Matematicamente, dizemos que as derivadas parciais de y em relação aos coeficientes de regressão não dependem desses coeficientes

$$\frac{\partial y}{\partial \beta_0} = 1; \ \frac{\partial y}{\partial \beta_1} = x_1; \ldots; \frac{\partial y}{\partial \beta_p} = x_p$$

- Contra exemplo:

$$y_i = \beta_0 + e^{\beta_2 x_i} + \varepsilon_i$$

Revisão: Tipos de Regressão

- Modelo linear simples:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

- Modelo linear múltiplo:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \varepsilon$$

- Modelo não-linear:

$$y_i = \beta_0 + e^{\beta_2 x_i} + \varepsilon_i$$

Revisão: Objetivos da Regressão

- Previsão: influência das variáveis independentes **x** na variável resposta **y**

- Descrição dos dados ou explanação: usar modelos para sumarizar ou descrever dados

- Seleção de variáveis ou triagem: determinar a importância de cada variável independente **x** na determinação de **y**. Quanto menor a contribuição de determinada variável, maior a possibilidade de sua exclusão do modelo

- Controle da saída: modelo estimado pode ser usado para controlar a saída **y.** É possível encontrar um modelo ótimo para a variável de saída

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

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

$$y = b_0 + b_1x_1 + b_2x_2 + \ldots + b_kx_k + e$$

Regressão Linear Múltipla

- Dada uma amostra com ***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:

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

Regressão Linear Múltipla

- Representação matricial do modelo:

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

Regressão Linear Múltipla

- Ex. 01: Uma equipe de segurança de redes desenvolveu vários esquemas alternativos para conter ataques a servidores. O grupo quer avaliar os mecanismos e definiu um índice de sucesso dos esquemas.
- O índice de sucesso é baseado em dois fatores
  - Tempo do experimento (duração)
  - Número de ataques no período

Esse enunciado nos leva ao seguinte modelo de regressão linear múltipla:

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

Regressão Linear Múltipla

Ex. 01: $\text{índice} = b_0 + b_1(\#\text{ataques}) + b_2(\text{duração})$

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

Regressão Linear Múltipla

- Ex. 01:
  - Para a estimação do modelo, precisamos calcular:

$$\mathbf{X},\ \mathbf{X}^{T},\ \mathbf{X}^{T}\mathbf{X},\ (\mathbf{X}^{T}\mathbf{X})^{-1}\ \text{e}\ \mathbf{X}^{t}\mathbf{y}$$

  - Essas matrizes e operações entre matrizes, são usadas para estimar os parâmetros:

$$\mathbf{b} = \left(\mathbf{X}^{T}\mathbf{X}\right)^{-1}\mathbf{X}^{T}\mathbf{y}$$

  - Cada elemento da matriz **b** resultante corresponde a um parâmetro b_k

Regressão Linear Múltipla

- Ex. 01:

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

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

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

Regressão Linear Múltipla

- Ex. 01: $\text{índice} = b_0 + b_1(\#\text{ataques}) + b_2(\text{duração})$

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

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

$$C = \left(\mathbf{X}^T\mathbf{X}\right)^{-1} = \begin{bmatrix} 7.7134 & -0.0227 & -0.0562 \\ -0.0227 & 0.0003 & 0.0001 \\ -0.0562 & 0.0001 & 0.0004 \end{bmatrix} \qquad \mathbf{X}^T\mathbf{y} = \begin{bmatrix} 60.1 \\ 2118.9 \\ 7247.5 \end{bmatrix}$$

Regressão Linear Múltipla

- Ex. 01:

$$\mathbf{b} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$$

$$\mathbf{b} = \begin{bmatrix} 7.7134 & -0.0227 & -0.0562 \\ -0.0227 & 0.0003 & 0.0001 \\ -0.0562 & 0.0001 & 0.0004 \end{bmatrix} \begin{bmatrix} 60.1 \\ 2118.9 \\ 7247.5 \end{bmatrix} = \begin{bmatrix} 8.37 \\ 0.005 \\ -0.009 \end{bmatrix}$$

índice = b_0 + b_1(#ataques) +b_2(duração)

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

Qualidade da Regressão

- Algumas notações importantes:
  - SSE – Sum of Squared Errors (soma dos quadrados residuais, com regressão)

$$SSE = \{\mathbf{y}^T\mathbf{y} - \mathbf{b}^T\mathbf{X}^T\mathbf{y}\} \quad \text{ou} \quad SSE = \sum e_i^2$$

  - SST – Total Sum of Squares (soma dos quadrados residuais, sem regressão)

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

  - SSY – Sum of Squares of y
  - SS0 – Sum of Squares of $\bar{y}$
  - SSR – Sum of Squares explained by Regression (SSR = SST - SSE)

Revisão: Qualidade da Regressão

- Para avaliar a qualidade da regressão:

  1. Calcule SST

  2. Calcule SSE

  3. Calcule o coeficiente de determinação (valor entre 0 e 1):

$$R^2 = \frac{SST - SSE}{SST} = 1 - \frac{SSE}{SST}$$

- Quanto maior o coeficiente de determinação, melhor a regressão

Qualidade da Regressão

- Voltando ao Ex. 01:

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

Índice	#At.	Dur.	Índice 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

Qualidade da Regressão

- Ex. 01:

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!

índice = b_0 + b_1(#ataques) +b_2(duração)

Qualidade da Regressão

- Ex. 01: Por que a regressão encontrada 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

Qualidade da Regressão

- Ex. 01: Cálculo do desvio padrão:

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

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

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

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

$$b_i \pm t_{[1-\alpha, n-2]} s_{b_i}$$

Qualidade da Regressão

- Ex. 01:

Em um nível 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

90% já é um nível de confiança baixo e 2 dos 3 parâmetros não têm significância

Qualidade da Regressão

- Ex. 01: Análise da variância

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

  O Teste-F pode ser usado para essa finalidade

  - Verificar se y depende ou não das variáveis previsoras
  - Tabela F utilizada no teste:
    - https://drive.google.com/open?id=1trjh4htB9TgARp7XuBL097oRMDwIs7QR

Qualidade da Regressão

- Utilização do Teste-F (análise da variância):

    1. Calcule SSR e SSE e seus graus de liberdade:
        a. SSR tem **k** graus de liberdade (**k** = nº de parâmetros - 1)
        b. SSE tem **n-(k+1)** graus de liberdade (**k+1** parâmetros)
    2. Calcule o quadrado das médias da regressão (MSR) e dos erros (MSE)
        a. MSR = SSR/GL
        b. MSE = SSE/GL
        c. MSR/MSE tem uma distribuição F
    3. Se MSR/MSE > tabela-F, variáveis previsoras (**x**) explicam uma fração significativa de **y**
        a. **y** depende de pelo menos uma variável previsora

índice = $b_0 + b_1$(#ataques) $+b_2$(duração)

Qualidade da Regressão

- Voltando ao Ex. 01:
  - SSR = 0.33
  - SSE = 1.08
  - MSR = SSR/k = 0.33/2 = 0.16
  - MSE = SSE/(n-k-1) = 1.08/(8 - 2 - 1) = 0.22
  - **F-calculado = MSR/MSE = 0.76**
  - **F[90; k-1, n-k-1] = F[90; 2, 5] = 3.78 (em 90% de confiança)**
  - **F-calculado < F-tabelado**

GL	V1	
V2	1	2
1	39.864	49.500
2	8.526	9.000
3	5.538	5.462
4	4.545	4.325
5	4.060	3.780
6	3.776	3.463

Conclusão: as variáveis previsoras não contribuem significativamente para o modelo (o modelo estimado é inadequado)

Múltipla Colinearidade

- Se dois previsores são linearmente dependentes, eles são colineares
  - Significa que são relacionados
  - Uma segunda variável (**x**) não melhora a regressão, pode inclusive piorar a regressão
- Sintomas típicos:
  - Resultados inconsistentes em vários testes de significância
  - F-calculado > F-tabelado, mas ICs para coeficientes incluem 0 (inconsistência nos testes)
- Detecção de múltipla colinearidade:
  - Se a correlação entre variáveis previsoras for alta, elimine uma e repita a regressão sem ela
  - Se a significância da regressão melhorar, provavelmente havia múltipla colinearidade

Múltipla Colinearidade

- Cálculo da correlação:

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

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

```python
import numpy as np

x = [1.6, 1.7, 1.8, 1.9]
y = [60, 70, 80, 90]
xy = [x, y]

r = np.corrcoef(xy)
```

- Breve revisão:
  - Coeficiente de correlação varia de -1 a 1

Múltipla Colinearidade

- Ex. 02: Sete programas foram monitorados quanto às suas demandas por recursos: número de operações de I/Os (disco), consumo de memória (em KB) e tempo de CPU (em ms). Os dados são 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 outros função dos dois recursos

Múltipla Colinearidade

- Ex. 02:

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

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

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

Múltipla Colinearidade

- Ex. 02: CPU time = b_0 + b_1 (# disk I/Os) + b_2 (tamanho da mem)

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

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

Múltipla Colinearidade

- Ex. 02: CPU time = $b_0 + b_1$ (# disk I/Os) + b_2 (tamanho da mem)

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

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

Cpu time = -0.1614 + 0.1182(# disk I/Os) + 0.0265(tam. Mem)

Múltipla Colinearidade

- Ex. 02: 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

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

Múltipla Colinearidade

- Ex. 02:

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

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

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

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

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

Múltipla Colinearidade

- Ex. 02: 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

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

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

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

Múltipla Colinearidade

- Ex. 02: Cálculo de IC para nível de confiança de 90% (t-student):

4 graus de liberdade $t_{0.90,4} = 2.132$

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

Nenhum parâmetro significativo.

Múltipla Colinearidade

- Ex. 02: 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
  MSR/MSE > F

Regressão passou no teste-F. Hipótese de que todos os parâmetros são 0 não pode ser aceita. Inconsistência?

Múltipla Colinearidade

- Ex. 02: Vamos calcular a correlação entre as variáveis previsoras:

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

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

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

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

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

$$= 0.9947$$

Múltipla Colinearidade

- Ex. 02:
  Conclusões:
  - Alta correlação (0,9947): multicolinearidade prejudica a regressão
  - Precisa refazer regressão somente com # de I/Os e, separadamente, com tamanho de memória, e escolher melhor previsor (isto é, aquele que resulta no maior R^2)
  - Neste caso, o modelo indicado é o de regressão linear simples

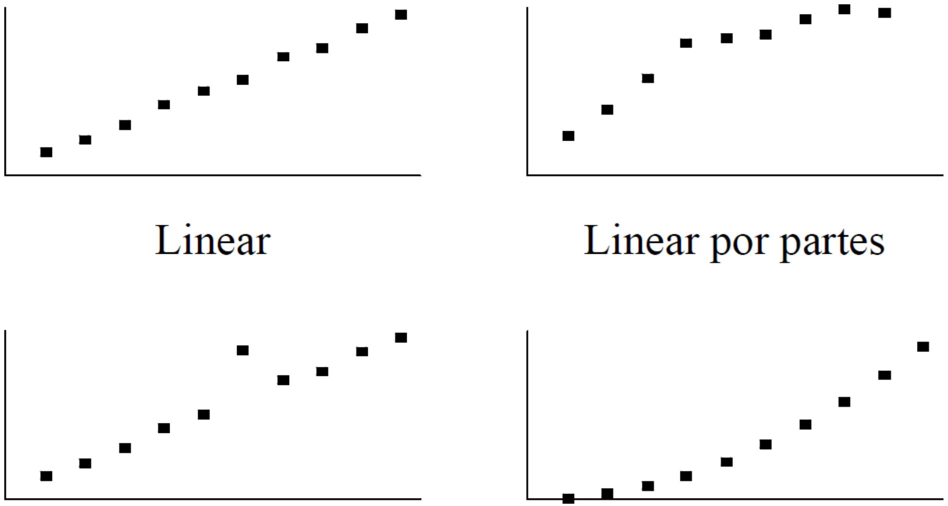
Regressão Curvilinear

- Regressão linear assume relações lineares entre previsoras e a resposta
- O que acontece quando essas relações não são lineares?
  - Coeficientes de determinação R^2 com baixos valores
- Possível solução: modelar o problema com regressão curvilinear
- Inspeção visual (dispersão) pode revelar que o modelo deve ser curvilinear
- Deve-se tentar transformar modelos curvilineares para lineares

Revisão: Teste Visual de Pressupostos

- Ex. (linearidade):

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



Linear

Linear por partes

Outlier/Exceção

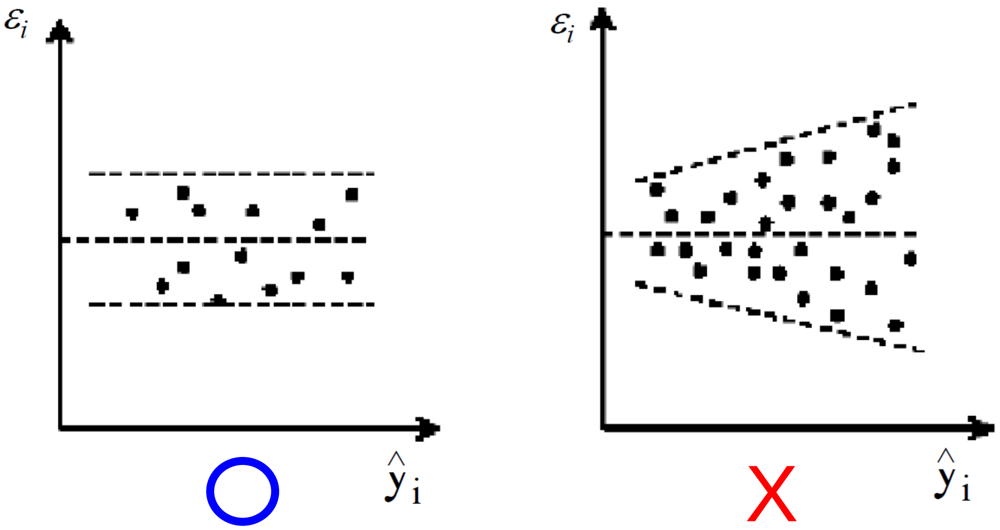
Não linear (Função de Potência)

Revisão: Teste Visual de Pressupostos

- Ex. (homocedasticidade):

Gráfico de pontos ε_i versus $\hat{y}_i$

Verificar tendência no espalhamento



Caso haja tendência ao espalhamento, usar regressão não-linear ou linearização

Linearização de Modelos

- Tipos comuns de modelos de regressão curvilinear:

$$y = bx^a$$

$$y = a + {}^{b}\!/\!_{x}$$

$$y = ab^x$$

- Para transformar esses equações em modelos lineares, costuma-se usar logaritmos, multiplicações, divisões, etc., sobre os modelos curvilineares
- Quer se obter algo como: $y' = a + bx'$
  - y' e x' obtidos através da transformação

Linearização de Modelos

- Alguns exemplos de transformação de curvilinear para linear:

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

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

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

$$y = x/(a + bx) \quad \Rightarrow (x/y) = a + bx$$

$$y = a \times b^x \quad \Rightarrow \ln y = \ln a + x \ln b \qquad y' = A + Bx'$$

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

Linearização de Modelos

- Ex. 03: A Lei de Amdahl para operações de I/Os em sistemas de computação diz que a taxa de I/O e proporcional a velocidade do processador. Para cada instrução executada, há um bit de I/O em média.

  Para validar a lei, os números de I/Os e as utilizações de CPU de um número de computadores foram medidos. Usando a taxa MIPS nominal para o sistema e a sua utilização, a taxa de processamento de instruções (em MIPS) e a taxa de I/O (em KB/s) foram computados para um período. Os dados foram mostrados abaixo. Valide a Lei de Amdahl.

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

Linearização de Modelos

- Ex. 03:
  - 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 parâmetros $b_0 = \log \alpha$ e $b_1 = b$ podem ser estimados via regressão 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

  - Coeficientes são significativos com confiança de 90%. Como o IC para b_1 contém 1, podemos aceitar a hipótese de que o relacionamento entre I/O rate e MIPS rate é linear.