Modelos Lineares
Regressão Linear Simples

Professora Ariane Ferreira

Instituto Politécnico
Campus Regional da UERJ
Nova Friburgo - RJ

Modelos de Regressão: Ajuste da reta

Definem-se desvios como os valores, segundo Y, das diferenças entre os valores observados e os valores sobre a linha l_i ajustada ao conjunto de pares (X, Y).

Como regra estabelece-se que a melhor linha l_i corresponde àquela cujo somatório dos desvios tende a zero (é minimizado).

A melhor linha ajustada define o modelo de regressão e pode ser obtida pela derivação de mínimos quadrados ordinários.

Modelo de Regressão Linear Simples

RLS

Na regressão linear simples, nós nos preocupamos em modelar a relação entre duas variáveis,

Exemplos:

- rendimento e número de anos de educação,
- altura e peso de pessoas, comprimento e largura de envelopes,
- altitude e temperatura de ebulição da água,
- dose de uma droga e resposta,
- quantidade de adubo e produção de gramíneas.

Para uma relação linear, nós usamos um modelo da forma:

$$\mathrm{Y_i} = \alpha + \beta x_i + \varepsilon_i \qquad \mathrm{E}(\mathrm{Y_i}) = \alpha + \beta x_i$$

onde *Y* é a variável dependente ou variável resposta e *X* é a variável independente ou variável preditora.

A variável aleatória ε_i é o termo de erro no modelo.

Modelo de Regressão Linear Simples

Nesse contexto,o erro não significa engano ou equívoco, mas sim um termo estatístico que representa flutuações aleatórias, erros de medidas ou o efeito de fatores não controlados.

A linearidade do modelo em si é uma suposição.

Geralmente, nós adicionamos outras suposições sobre a distribuição do erro, independência dos valores observados de Y, assim por diante.

Usando valores observados de X e Y, nós estimamos $\beta 0$ e $\beta 1$ e fazemos inferências tais como intervalos de confiança e testes de hipóteses sobre $\beta 0$ e $\beta 1$.

Nós também podemos usar o modelo estimado para prever ou predizer o valor de Y para um particular valor de X.

Estimação e procedimentos inferenciais para o modelo de regressão linear simples são desenvolvidos nesta disciplina.

Modelo de Regressão Linear Simples

RLS

- X e Y duas variáveis contínuas.
- Os valores de x_i de X são amostrados sem erros de medida.
- Observamos os valores correspondentes $y_1, \ldots, y_n$ de Y.

Exemplos

- X pode ser o tempo e Y um valor medido em diferentes datas.
- Y pode ser a diferença de potencial medida nas extremidades de uma resistência para diferentes valores de intensidade de corrente.

Hypótese

- X e Y não são independentes e lo conhecimento X permete adquirir o conhecimento de Y.
- O valor médio E(Y|X=x), é a esperança condicional de Y sabendo que X = x.

Função Linear

$$\mathrm{E}(\mathrm{Y}_\mathrm{i}) = \alpha + \beta x_i$$

$$\mathrm{Y}_\mathrm{i} = \alpha + \beta x_i + \varepsilon_i$$

com $\mathrm{E}(\varepsilon_i)=0$, para todo $i=1,\ldots,n$

n = n° d'observações; ε_i =resíduos das obs i

Os Dados

Y = Variável à explicar

- numérica
- (ou dependente)

X_p = Variável explicativa

- numérica ou binária
- (obs independentes)

O Quadro de Dados

	X	Y
1	x_{11}	y_1
⋮		
i	x_{1i}	y_i
⋮		
n	x_{1n}	y_n

Onde os x_{ji} são fixos e ε_i é um termo aleatório de lei N(0,σ);
Os ε_i são indépendentes entre si.

O Modelo de Regressão Multipla

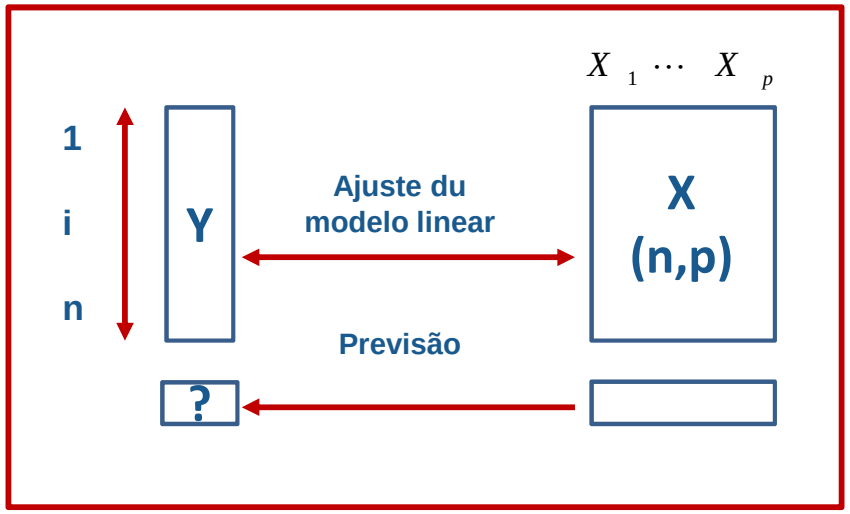
$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon$$

Suposição : independência linear des X_i.

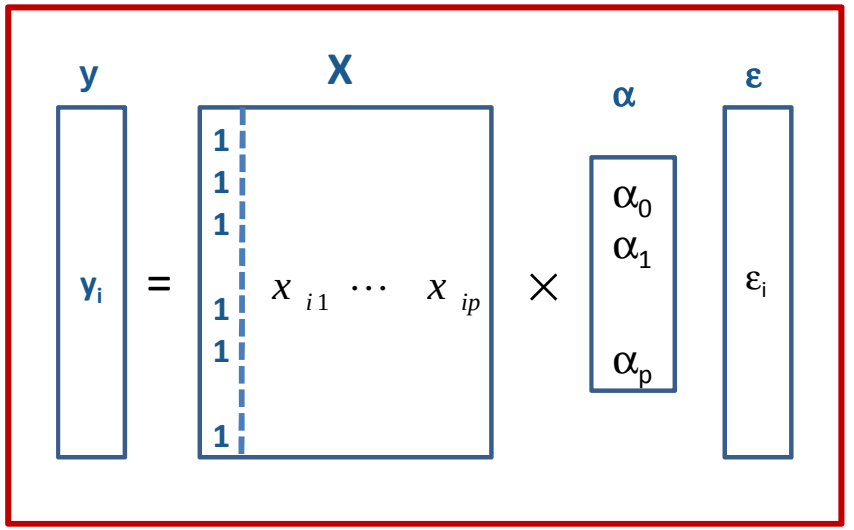
Modelo na forma matricial

$$y = \mathbf{X}\boldsymbol{\alpha} + \varepsilon$$

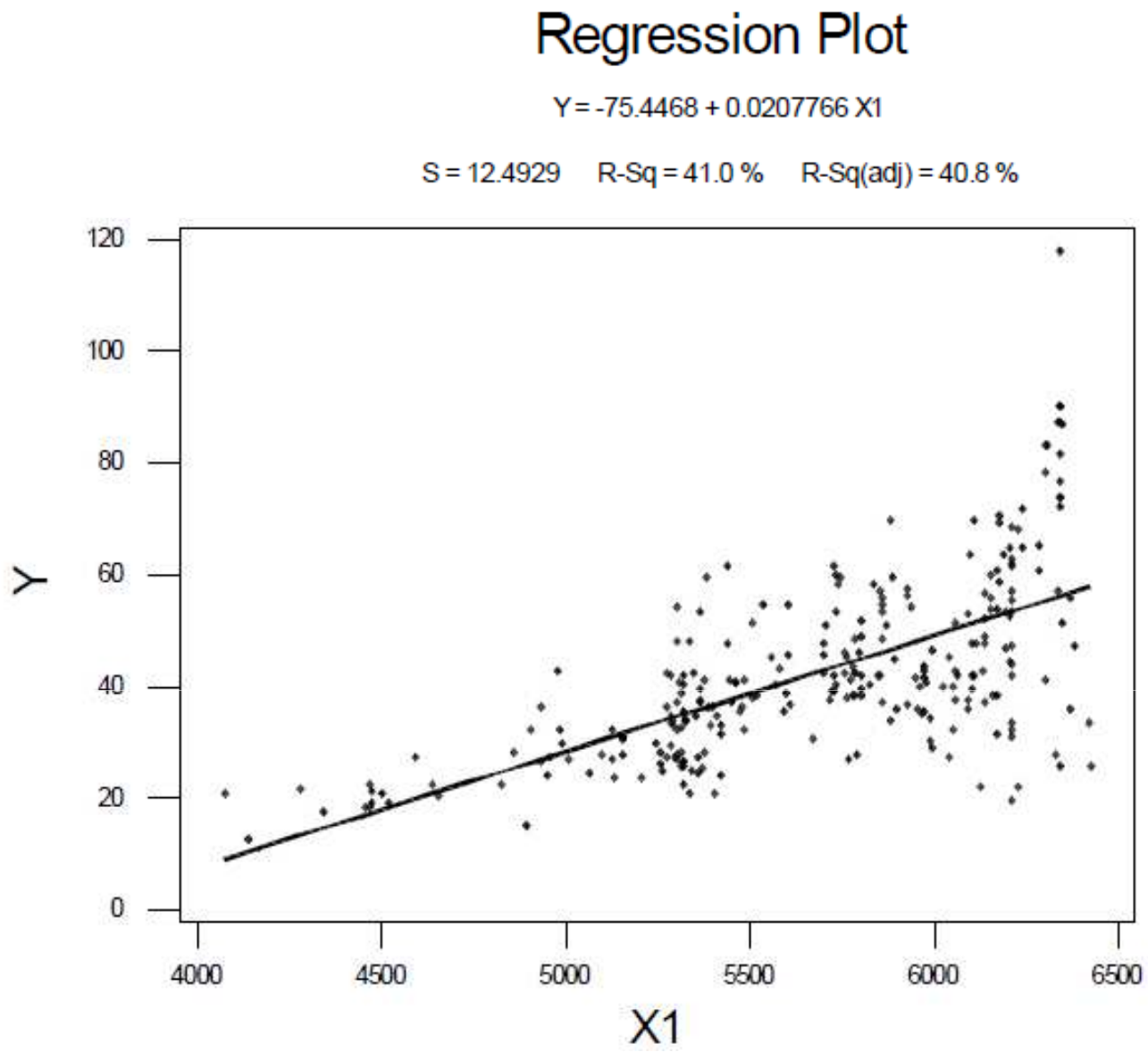
(n,1) (n,p+1) (p+1,1) (n,1)



Previsão Linear



Schema du modelo Linear


Regression Plot
Y = -75.4468 + 0.0207766 X1
S = 12.4929 R-Sq = 41.0 % R-Sq(adj) = 40.8 %
120
100
80
60
40
20
0
Y
4000
4500
5000
5500
6000
6500
X1

Hipóteses do modelo linear

Resíduos

- A variância dos resíduos é a mesma para todos os valores de X
  - **Homoscedasticidade** : $V(\varepsilon_i) = \sigma^2$
- Os resíduos são linearmente independentes: $cov(\varepsilon_i,\varepsilon_j) = 0 \ \forall \ i \neq j$
- Os resíduos são normalmente distribuídos : $\varepsilon_i \sim N(0,\sigma^2)$

A existência da componente estocástica (ε_i) corresponde ao fato de que :

- **variação sincrônica :** indivíduos com mesmo valor de x_i podem ter respostas Y diferentes;
- **Variação discrônica** : um mesmo indivíduo medido varias vezes para um mesmo valor de x_i pode ter respostas Y diferentes.

- **Temos equivalência de** $\varepsilon_i \sim N(0,\sigma^2)$ et $Y/X=x_i \sim N(\alpha + \beta x_i,\sigma^2)$

Simplificando

•Supomos um sistema onde medimos a variavel resposta Y que depende da variavel de regressão X:

$$y = ax + b + \epsilon$$

Onde *a* e *b* são os coeficientes de regressão e ε é o termo do erro aleatorio verificando:

$$E(\epsilon) = 0 \text{ et } V(\epsilon) = \sigma^2$$

Realizamos n experiências: *n* valores $x_1, x_2, ..., x_n$ da variavel de regessão x;

Medimos no sistema : n valores $y_1, y_2, ..., y_n$ da variavel de resposta y.

Em cada experimento temos: $y_i = ax_i + b + \epsilon_i,$

Onde ε_i são v.a. idenpendentes

Calculos iniciais

•Para os calculos definimos:

$$y_i = ax_i + b + \epsilon_i,$$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

Esperança matematica de y:

$$E(\bar{y}) = E\left\{\frac{1}{n}\sum_{i=1}^{n}(ax_i + b + \epsilon_i)\right\} = \frac{a}{n}\sum_{i=1}^{n} x_i + b + \frac{1}{n}\sum_{i=1}^{n}\underbrace{E(\epsilon_i)}_{0} = a\bar{x} + b$$

Então:

$$E(y_i - \bar{y}) = E(ax_i + b + \epsilon_i - a\bar{x} - b) = a(x_i - \bar{x})$$

Se $V(\epsilon_i) = \sigma^2$, temos que $V(\bar{y}) = \frac{\sigma^2}{n}$

Estimação de a e b

•Método dos minimos quadrados: consiste em procurar os valores de $\hat{a}$ e $\hat{b}$

•Que minimisam a seguinte função:

$$L = \sum_{i=1}^{n} (y_i - ax_i - b)^2$$

Derivando em relação à *a* e *b* :

$$\sum_{i=1}^{n} x_i(y_i - \hat{a}x_i - \hat{b}) = 0$$

$$\sum_{i=1}^{n} (y_i - \hat{a}x_i - \hat{b}) = 0$$

Então:

$$\hat{a}\sum_{i=1}^{n} x_i^2 + \hat{b}n\bar{x} = \sum_{i=1}^{n} x_i y_i$$

$$\hat{a}\bar{x} + \hat{b} = \bar{y}$$

Estimação de a e b

•Solução do sistema: valores de $\hat{a}$ e $\hat{b}$

$$\hat{a} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$$

$$\hat{b} = \bar{y} - \hat{a}\bar{x}$$

Ou ainda:

$$\hat{a} = \frac{\sum_{i=1}^{n} y_i(x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

Dessa forma $\hat{a}$ é uma combinação linar de y_i

$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

$$S_{xy} = \sum_{i=1}^{n} y_i(x_i - \bar{x})$$

Então:

$$\hat{a} = \frac{S_{xy}}{S_{xx}}$$

Residuos

Para cada valor de x_i : $\hat{y}_i = \hat{a}x_i + \hat{b}$

Correspondendo ao valor da resposta calculado pelo modelo linear.

Para cada observação da resposta definimos o residuo:

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

O qual reflete o desvio entre a observação da resposta y_i e o valor calculado pelo modelo linear $\hat{y}_i$

Os residuos tem as propriedades de ortogonalidades:

$$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} e_i x_i = \sum_{i=1}^{n} e_i \hat{y}_i = 0$$

Residuos: prova da ortogonalidade

Calculamos : $\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} (y_i - \hat{y}_i)$

Usando na expressão acima: $\hat{y}_i = \hat{a}x_i + \hat{b}$ obtemos:

$$\sum_{i=1}^{n} e_i = n \underbrace{(\bar{y} - \hat{a}\bar{x} - \hat{b})}_{0} = 0$$

Da mesma maneira:

$$\sum_{i=1}^{n} e_i x_i = \sum_{i=1}^{n} (y_i - \hat{y}_i) x_i = \sum_{i=1}^{n} (y_i - \hat{a}x_i - \hat{b}) x_i$$

Desenvolvendo a soma:

$$\sum_{i=1}^{n} e_i x_i = \sum_{i=1}^{n} x_i y_i - \hat{a} \sum_{i=1}^{n} x_i^2 - \hat{b} n \bar{x} = 0$$

Residuos: prova da ortogonalidade

Desenvolvendo a soma:

$$\sum_{i=1}^{n} e_i x_i = \sum_{i=1}^{n} x_i y_i - \hat{a} \sum_{i=1}^{n} x_i^2 - \hat{b} n \bar{x} = 0$$

Finalmente teremos:

$$\sum_{i=1}^{n} e_i \hat{y}_i = \sum_{i=1}^{n} e_i (\hat{a} x_i + \hat{b}) = \hat{a} \underbrace{\sum_{i=1}^{n} e_i x_i}_{0} + \hat{b} \underbrace{\sum_{i=1}^{n} e_i}_{0} = 0$$

Propriedades Estatisticas

Sabendo que: $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$

temos: $\hat{a} = \frac{1}{S_{xx}} \sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})$

Uma vez que: $E(y_i - \bar{y}) = a(x_i - \bar{x})$

Teremos: $E(\hat{a}) = \frac{a}{S_{xx}} \sum_{i=1}^{n}(x_i - \bar{x})^2 = a\frac{S_{xx}}{S_{xx}} = a$

O que prova que $\hat{a}$ é um bom estimador de a

Propriedades Estatisticas

Sabendo que os y_i são v.a. Independentes: $V(y_i) = \sigma^2$

Então Variância de $\hat{a}$: $V(\hat{a}) = \frac{\sigma^2}{S_{xx}^2} \sum_{i=1}^{n} (x_i - \bar{x})^2 = \frac{\sigma^2}{S_{xx}}$

A variância de sera menor quando os valores de x_i são dispersos em torno da média.

Como $\hat{b} = \bar{y} - \hat{a}\bar{x}$

Temos : $E(\hat{b}) = \bar{y} - a\bar{x}$

$\Big\} \; E(\hat{b}) = b$

Propriedades Estatisticas

Calculando a covariância entre y_i e $\bar{y}$:

$$C(y_i,\bar{y}) = \frac{1}{n} C\left(y_i, \sum_{j=1}^{n} y_j\right) = \frac{1}{n} \sum_{j=1}^{n} C(y_i,y_j)$$

Por definição $C(y_i,y_j) = 0$ se $i \neq j$ e $C(y_i,y_j) = \sigma^2$ se $i = j$

Temos : $$C(y_i,\bar{y}) = \frac{\sigma^2}{n}$$

Propriedades Estatisticas

Calculando a covariância entre $\hat{a}$ e $\bar{y}$:

$$C(\hat{a},\bar{y}) = C\left(\frac{1}{S_{xx}}\sum_{i=1}^{n} y_i(x_i - \bar{x}),\bar{y}\right) = \frac{1}{S_{xx}}\sum_{i=1}^{n} C(y_i,\bar{y})(x_i - \bar{x})$$

Usando $C(y_i,\bar{y}) = \frac{\sigma^2}{n}$

Temos :

$$C(\hat{a},\bar{y}) = \frac{\sigma^2}{nS_{xx}}\overbrace{\sum_{i=1}^{n}(x_i - \bar{x})}^{0} = 0$$

Propriedades Estatisticas

Finalmente sabendo que: $C(\hat{a},\bar{y}) = 0$

A variância de $\hat{b}$

$$V(\hat{b}) = V(\bar{y} - \hat{a}\bar{x}) = V(\bar{y}) + \bar{x}^2 V(\hat{a})$$

Simplificando:

$$V(\hat{b}) = \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)\sigma^2$$

Propriedades Estatisticas

Demonstração de que $\hat{a}$ e $\hat{b}$ são v.a. correlacionadas:

Calculamos: $C(\hat{a},\hat{b}) = C(\hat{a},\bar{y} - \hat{a}\bar{x}) = C(\hat{a},\bar{y}) - \bar{x}C(\hat{a},\hat{a})$

Sabendo que : $C(\hat{a},\bar{y}) = 0$ e $C(\hat{a},\hat{a}) = V(\hat{a}) = \sigma^2/S_{xx},$

Temos: $$C(\hat{a},\hat{b}) = -\frac{\bar{x}\sigma^2}{S_{xx}}$$

Uma vez que $\hat{a}$ e $\hat{b}$ são estimadores sem biais de a e b.

Então $\hat{a}x_i + \hat{b}$ e o estimador sem biais de $ax_i + b,$

Como consequência: $E(e_i) = 0$