Objetivos do Curso

- Fornecer uma ferramenta estatística unificada que permita a análise de qualquer modelo que puder ser escrito na forma linear nos parâmetros. Este estudo englobaria Análise de Variância, Análise de Covariância, Correlação e Regressão

- Estudo de alguns modelos mais específicos, que saem como casos particulares desta abordagem:

  Curvas de Crescimento (com pouco esforço, conhecida a formulação linear geral $\underset{\sim}{Y} = X\underset{\sim}{\beta} + \underset{\sim}{\epsilon}$, entende-se este tópico)

  Modelos com medidas repetidas (modelos mistos)

  Regressão Multivariada

- Fornecer um estudo das principais distribuições utilizadas na Análise de Variância e Regressão

- Apresentar resultados matriciais úteis para estatísticos

1 - Introdução e Conceitos Básicos

1.1 - Introdução - Modelo Linear Geral

Um modelo (estatístico) linear geral é dado por

$$Y = \mu(x_1, x_2 \dots x_K) + \epsilon$$

onde

Y e ϵ são variáveis aleatórias

$x_1, x_2 \dots x_K$ são variáveis não aleatórias

$\mu(x_1, x_2 \dots x_K)$ é uma função de $x_1, x_2 \dots x_K$, definida num domínio D e linear num conjunto de parâmetros desconhecidos, $\beta_0, \beta_1, \beta_2 \dots \beta_K$ pertencentes a um espaço paramétrico Ω_β

ϵ - erro não observável, são feitas suposições sobre sua distribuição de probabilidades

$\mu \rightarrow$ parte determinística

$\epsilon \rightarrow$ parte aleatória ou estocástica do modelo

Y - variável resposta ou variável dependente

$x_1, x_2 \dots x_K$ - variáveis independentes, explicativas ou preditoras

A forma de $\mu(x_1, x_2 \dots x_K)$ será admitida conhecida

Ex:

$\mu(x) = \beta_0 + \beta_1 x$

$Y = \beta_0 + \beta_1 x + \epsilon$ → modelo de regressão linear simples

$E(\epsilon) = 0$

$Var(\epsilon) = \sigma^2$ (desconhecido) $K = 1$

$\mu(x_1, x_2, x_3) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$

$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$ → modelo de regressão linear múltipla $K = 3$

$\mu(x_1, x_2, x_3) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$

$x_1 = x \quad x_2 = x^2 \quad x_3 = x^3$

$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \epsilon$ → modelo de regressão polinomial $K = 3$

$$Y = \beta_0 + \beta_1 x + \beta_2 e^x + \epsilon$$

$\epsilon \sim N(0, \sigma^2)$ σ^2 desconhecido

$x_1 = x \quad x_2 = e^x$ $\quad K=2$

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \epsilon$$

$x_3 = x_1 x_2$ $\quad K=3$

Na forma geral

$$\mu(x_1, x_2 \dots x_K) = \beta_0 + \sum_{j=1}^{K} q_j(x_1, x_2 \dots x_K)\, \beta_j$$

Como a distribuição de probabilidades de Y depende das variáveis explicativas, poderíamos escrever

$$Y_{(\underset{\sim}{x})} = \mu(\underset{\sim}{x}) + \epsilon_{(\underset{\sim}{x})} \qquad E(\epsilon_{(\underset{\sim}{x})}) = 0 \qquad \underset{\sim}{x} \in D$$

$\underset{\sim}{x} = (x_1, x_2 \dots x_K)$

Obs:
- Este é um modelo populacional que representa a situação real

Obs:

Este é um modelo populacional que representa a situação em estudo. Nenhuma menção é feita à amostra ainda.

Para inferirmos sobre os parâmetros do modelo, tomamos uma amostra de n observações.

Modelo Linear Geral (amostral)

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_K x_{Ki} + \epsilon_i$$

$$E(\epsilon_i) = 0 \qquad i = 1, 2 \ldots n$$

Consequência: $E(Y_i \mid x_{1i}, x_{2i} \ldots x_{Ki}) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_K x_{Ki}$

com

(1) Y_i v.a. observável

(2) x_{ji} v. não aleatórias pertencentes a um domínio D.

(3) β_j, $j = 0, 1, 2 \ldots K$, parâmetros definidos em um espaço paramétrico Ω_β.

(4) ϵ_i v.a. não observáveis, $Cov(\epsilon_i, \epsilon_l) = \delta_{il}$

$i, l = 1, 2 \ldots n$

Na forma matricial, este modelo fica

$$\underset{\sim}{Y} = X\underset{\sim}{\beta} + \underset{\sim}{\epsilon}$$

onde

(1) $\underset{\sim}{Y}$ é um vetor aleatório observável $n \times 1$

(2) X é uma matriz $n \times (K+1)$ de valores observáveis fixos (os elementos de X não são variáveis aleatórias) $n > K+1$

(3) $\underset{\sim}{\beta}$ é um vetor $(K+1) \times 1$ de parâmetros desconhecidos, definido num espaço paramétrico ℓ_β

(4) $\underset{\sim}{\epsilon}$ é um vetor aleatório não observável com

$$E(\underset{\sim}{\epsilon}) = \underset{\sim}{0} \quad \text{e} \quad Var(\underset{\sim}{\epsilon}) = \begin{bmatrix} Var(\epsilon_1) & Cov(\epsilon_1,\epsilon_2) & \cdots & Cov(\epsilon_1,\epsilon_n) \\ & Var(\epsilon_2) & & Cov(\epsilon_2,\epsilon_n) \\ & & & \\ & & & Var(\epsilon_n) \end{bmatrix}$$

$$= \Sigma$$

$$\underset{\sim}{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \qquad X = \begin{bmatrix} 1 & x_{11} & \cdots & x_{K1} \\ 1 & x_{12} & & x_{K2} \\ & & & \\ & & & \\ 1 & x_{1n} & & x_{Kn} \end{bmatrix} \qquad \underset{\sim}{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_K \end{bmatrix} \qquad \underset{\sim}{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

Obs:

Este modelo admite vários casos especiais dependendo de

i) A distribuição de probabilidades de $\underset{\sim}{\epsilon}$.

ii) A estrutura de covariância definida na matriz Σ.

iii) O posto e a estrutura da matriz X.

Casos Particulares: Modelos de Regressão
Modelos de Análise de Variância
Modelos Mistos

Exemplo 1

Y - Pressão Sanguínea

x - idade

$Y = \beta_0 + \beta_1 x + \epsilon$

$E(\epsilon) = 0 \quad Var(\epsilon) = \sigma^2$

para $50 \leq x \leq 60$

$D = \{x \mid 50 \leq x \leq 60\}$

Tomada uma amostra de 6 observações independentes

$x_1 = 50 \quad x_2 = 52 \quad x_3 = 55 \quad x_4 = 56 \quad x_5 = 58 \quad x_6 = 60$

fixados, observa-se $y_1, y_2 \dots y_6$.

$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad i = 1, 2 \dots 6$

ϵ_i variáveis aleatórias independentes

$\epsilon_i \sim N(0, \sigma^2)$

↓ sup. adicional

a partir da amostra → inferência sobre β_0, β_1 e σ^2.

$\forall x \in D$, admite-se que $E(Y|x) = \beta_0 + \beta_1 x$.

$Y|x \sim N(\beta_0 + \beta_1 x, \sigma^2)$

Por ex., $x = 54 \quad Y|x=54 \sim N(\beta_0 + 54\beta_1, \sigma^2)$

$E(Y|x=54) = \beta_0 + 54\beta_1$

Podemos estimar a média de $Y|x=54$ sem amostrar esta distribuição

$\hat{E}(Y|x=54) = \hat{\beta}_0 + \hat{\beta}_1 . 54 \quad \hat{\beta}_0$ e $\hat{\beta}_1$ estimadores de β_0 e β_1.

Esta análise não é possível para $x = 62$ porque $x = 62 \notin D$.

No caso, $X = \begin{bmatrix} 1 & 50 \\ 1 & 52 \\ 1 & 55 \\ 1 & 56 \\ 1 & 58 \\ 1 & 60 \end{bmatrix}$ $\quad \Sigma = \sigma^2 I_6$

Modelo de Regressão Linear Simples

É uma suposição forte $E(Y|x) = \beta_0 + \beta_1 x$

$\forall x, \ 50 \leq x \leq 60$

$X_1, X_2 \ldots X_k$ → quantitativas : modelo de regressão

→ qualitativas : modelo de planejamento de experimentos
(atributos : cor, tipo de máquina, sexo, religião)

Exemplo 2

Y - Pressão sanguínea

x - sexo $n = 6$ 3 homens, 3 mulheres

$$Y_{ij} = \mu_i + \epsilon_{ij} \qquad E(\epsilon_{ij}) = 0$$

$i = 1, 2$ (para cada sexo) μ_1 - Pressão média dos Homens

$j = 1, 2, 3$ (para indivíduo) μ_2 - Pressão média das mulheres

$$\underset{\sim}{Y} = X \underset{\sim}{\beta} + \underset{\sim}{\epsilon} \qquad \beta = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$$

$$\begin{bmatrix} Y_{11} \\ Y_{12} \\ Y_{13} \\ Y_{21} \\ Y_{22} \\ Y_{23} \end{bmatrix} = \underbrace{\begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}}_{6 \times 2} \underbrace{\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}}_{2 \times 1} + \begin{bmatrix} \epsilon_{11} \\ \epsilon_{12} \\ \epsilon_{13} \\ \epsilon_{21} \\ \epsilon_{22} \\ \epsilon_{23} \end{bmatrix}$$

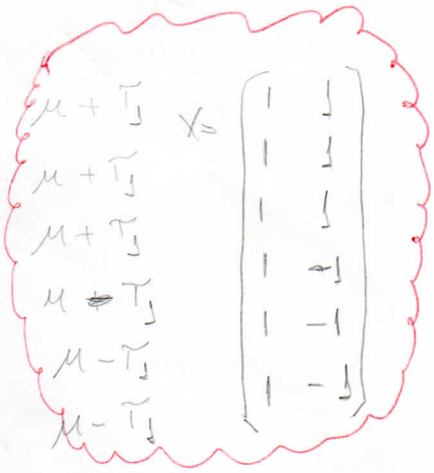
2 parâmetros

Objetivos : Estimar μ_1, μ_2 Testar $\mu_1 = \mu_2$

$\mu_i = \mu + \tau_i \qquad i = 1,2$ Forma alternativa de escrever o modelo

μ - média geral

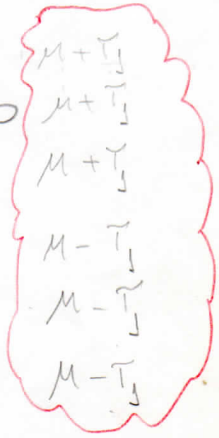
τ_i - efeito de sexo



$$Y_{ij} = \mu + \tau_i + \epsilon_{ij} \qquad \beta$$

$$\begin{bmatrix} Y_{11} \\ Y_{12} \\ Y_{13} \\ Y_{21} \\ Y_{22} \\ Y_{23} \end{bmatrix} = \underset{6\times3}{\begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix}} \underset{3\times1}{\begin{bmatrix} \mu \\ \tau_1 \\ \tau_2 \end{bmatrix}} + \begin{bmatrix} \epsilon_{11} \\ \epsilon_{12} \\ \epsilon_{13} \\ \epsilon_{21} \\ \epsilon_{22} \\ \epsilon_{23} \end{bmatrix}$$

3 parâmetros



Com relação à forma anterior

mais parâmetros

$\mu + \tau_i$ → diferente forma de escrever μ_i (desnecessária neste caso pode ser útil em planejamentos mais complicados)

$X \quad 6\times3$ posto 2 = nº de colunas linearmente independentes de X

< 3 → posto incompleto

No outro caso: $X_{6\times2}$ posto $(X) = 2$ → posto completo

Esta é a distinção entre os modelos dos tópicos 3 e 4

Nos dois casos, x é qualitativa: Modelo de planejamento de experimentos, a matriz X contem somente zeros e uns

Neste caso, a matriz de planejamento apresenta ainda padrões específicos, de acordo com o particular planejamento.

Exemplo 3 (Graybill, pag 165)

Y - Produção de Milho
em função de dois diferentes compostos e quatro diferentes métodos de aplicação.

$$Y_{ij} = \mu + \alpha_i + \tau_j + \epsilon_{ij} \quad i = 1,2 \quad j = 1,2,3,4.$$

Y_{ij} - produção de milho observada no canteiro que recebeu o i-ésimo composto aplicado pelo j-ésimo método.

μ - produção média quando não é aplicado nenhum composto e ∴ nenhum método

α_i - efeito do i-ésimo composto $\quad i = 1,2$

τ_j - efeito do j-ésimo método de aplicação
$j = 1,2,3,4.$

$$Y_{11} = \mu + \alpha_1 + \tilde{\tau}_1 + \epsilon_{11}$$
$$Y_{12} = \mu + \alpha_1 + \tau_2 + \epsilon_{12}$$
$$Y_{13} = \mu + \alpha_1 + \tilde{\tau}_3 + \epsilon_{13}$$
$$Y_{14} = \mu + \alpha_1 + \tilde{\tau}_4 + \epsilon_{14}$$
$$Y_{21} = \mu + \alpha_2 + \tau_1 + \epsilon_{21}$$
$$Y_{22} = \mu + \alpha_2 + \tau_2 + \epsilon_{22}$$
$$Y_{23} = \mu + \alpha_2 + \tilde{\tau}_3 + \epsilon_{23}$$
$$Y_{24} = \mu + \alpha_2 + \tilde{\tau}_4 + \epsilon_{24}$$

Na forma matricial $Y = X\beta + \epsilon$

$$\underset{\sim}{Y} = \begin{bmatrix} Y_{11} \\ Y_{12} \\ Y_{13} \\ Y_{14} \\ Y_{21} \\ Y_{22} \\ Y_{23} \\ Y_{24} \end{bmatrix}_{8\times 1} \qquad X = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 1 \end{bmatrix}_{8\times 7} \qquad \beta = \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \tau_1 \\ \tau_2 \\ \tilde{\tau}_3 \\ \tilde{\tau}_4 \end{bmatrix}_{7\times 1}$$

Modelo de análise de Variância com dois fatores, X de posto incompleto posto $(X) = 5$

Definido o Modelo Linear Geral estudaremos Inferência sobre $\underset{\sim}{\beta}$ para modelos com matriz de posto completo e incompleto.

Em modelos com matriz de planejamento de posto incompleto, nem todos os elementos de $\underset{\sim}{\beta}$ podem ser estimados. Veremos então nestes modelos quais funções lineares de $\underset{\sim}{\beta}$ poderão ser estimadas.