CIÊNCIA DE DADOS (BIG DATA)

ANÁLISE ESTATÍSTICA

Professor curador: Mário Olímpio de Menezes

EaD Mackenzie

TRILHA 4

PARTE A – MODELAGEM ESTATÍSTICA

PARTE A – MODELAGEM ESTATÍSTICA

MODELAGEM ESTATÍSTICA

ASPECTOS CHAVES DA MODELAGEM ESTATÍSTICA

- Entender que tipo de variável resposta se tem.
  - Aquela cuja variação queremos entender.
  - Aquela que vai no eixo y do gráfico.
- Natureza das variáveis explicativas.
  - Queremos entender como estas variáveis explicam a variação da variável resposta.

VARIÁVEL RESPOSTA E VARIÁVEIS EXPLICATIVAS

Precisamos considerar como as variáveis medem os atributos.

- **Variáveis numéricas** – altura, peso, idade... – números reais e números inteiros.
- **Variáveis categóricas** – sexo, cor, fabricante... – representam classes, categorias.

VARIÁVEL RESPOSTA E VARIÁVEIS EXPLICATIVAS

- Qual das variáveis é a variável resposta?
- Quais são as variáveis explicativas?
- As variáveis explicativas são numéricas ou categóricas, ou misturadas?
- Que tipo de variáveis resposta temos:
    - Uma medida contínua? Contagem? Proporção? Categoria?

MÉTODO ESTATÍSTICO APROPRIADO

QUANDO AS VARIÁVEIS EXPLICATIVAS:

1. São contínuas → Regressão.
2. São categóricas → Análise de Variância (ANOVA).
3. São contínuas ou categóricas → Análise de Covariância (ANCOVA).

MÉTODO ESTATÍSTICO APROPRIADO

QUANDO A VARIÁVEL RESPOSTA É:

1. Contínua → **Regressão Normal, ANOVA ou ANCOVA**
2. Proporção → **Regressão Logística**
3. Contagem → **Modelos log-linear (Poisson)**
4. Binária → **Regressão Logística binária**
5. Tempo na morte → **Análise de Sobrevivência**

MODELAGEM ESTATÍSTICA

O objetivo da modelagem estatística é determinar os valores dos parâmetros em um modelo específico que levam ao **melhor ajuste do modelo aos dados**.

- Buscamos um modelo **mínimo** – princípio da parcimônia –, mas também um modelo **adequado**.
- Não há **um** modelo que seja o correto; dentre os modelos diferentes, buscamos um **adequado**.

REGRESSÃO LINEAR

- Análise de Regressão: uma variável resposta Y e uma ou mais variáveis preditoras, ou explicativas, X_1, X_2, ..., X_p.
  - Quando p = 1, temos regressão simples.
  - Quando p > 1, temos regressão múltipla
- A variável resposta deve ser contínua.
- As variáveis explicativas podem ser contínuas, discretas ou categóricas.

REGRESSÃO LINEAR

Um modelo linear entre duas variáveis X e Y é definido matematicamente como uma equação com dois parâmetros desconhecidos.

$$Y = \beta_0 + \beta_1 X$$

REGRESSÃO POR MÍNIMOS QUADRADOS

HIPÓTESES REQUERIDAS

- Normalidade da variável dependente.
- Independência dos valores da variável dependente.
- Linearidade.
- Homocedasticidade – variância de Y é constante.

MODELO DE REGRESSÃO LINEAR

O modelo linear ajustado conterá estimativas dos valores verdadeiros da população.

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

Para cada valor de Y_i, temos um valor estimado pela equação de regressão:

$$Y_i = \hat{Y}_i + e_i$$

MODELO DE REGRESSÃO LINEAR

Ou seja, o modelo pode ser expresso como:

DADOS = AJUSTE + RESÍDUO

$$Y_i = \hat{Y}_i + e_i$$

Da expressão anterior, temos:

$$\begin{aligned} e_i &= Y_i - \hat{Y}_i \\ e_i &= Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i) \end{aligned}$$


M
Mackenzie