

## Resumo para Exame de Análise Exploratória de Dados e Regressão Múltipla

### Condições do Exame:

- Sem consulta; duração: 1h00
- 20 valores totais; todas as questões têm a mesma cotação
- Formatos: escolha múltipla, verdadeiro/falso (erro desconta 50%), resposta aberta

### 1. Análise Exploratória de Dados (AED)

Definição: etapa inicial para detectar erros, entender variáveis e relações antes de modelar.

#### Variáveis:

- Qualitativas: nominais (sem ordem), ordinais (com ordem)
- Quantitativas: discretas, contínuas; escalas: intervalo (zero não real), razão (zero natural)

#### Análise Univariada:

- Medidas de localização: média, mediana, moda
- Medidas de dispersão: variância, desvio padrão, amplitude, IQR, coeficiente de variação
- Quantis, percentis, quartis; interpretação de assimetria pela posição de Q1/Q3 em relação à mediana

#### Assimetria e Curtose:

- Skewness (g1):  $>0$  (positiva),  $<0$  (negativa),  $=0$  (simétrica).
- Kurtosis (g2):  $>3$  leptocúrtica (caudas pesadas),  $<3$  platicúrtica (caudas leves),  $=3$  mesocúrtica.
- Transformações (log, raiz) para simetrizar dados.

#### Tabelas e Gráficos:

- Tabelas de frequência (absoluta, relativa, acumuladas) para categóricas e discretas.
- Histogramas para variáveis contínuas; definição de classes (Regra de Sturges).
- Gráficos de barras e setores para variáveis discretas/categóricas.
- Boxplot para visualizar dispersão e outliers.

### 2. Regressão Linear Múltipla (MLR)

Modelo:  $Y = \alpha_0 + \alpha_1 X_1 + \dots + \alpha_p X_p + \varepsilon$ , com  $\varepsilon \sim N(0, \sigma^2)$ .

Dummies para categóricas: K categorias  $\rightarrow$  K-1 dummies; referência codificada como 0.

#### Interpretação de Coeficientes:

- Numérico:  $\alpha_j$  = variação em Y para +1 em  $X_j$ , mantidos outros.
- Binário: diferença em Y entre categoria e referência.

#### Avaliação do Modelo:

- Somas de quadrados:  $SST = SSR + SSE$ .
- $R^2 = SSR/SST$ ;  $R^2$  ajustado penaliza p para evitar overfitting.

#### Pressupostos:

1. Linearidade; 2. Normalidade dos resíduos; 3. Ausência de multicolinearidade (VIF);

4. Ausência de autocorrelação (Durbin–Watson); 5. Homoscedasticidade.

Inferência:

- Teste F global:  $H_0: \alpha_j=0$ .
- Testes t individuais:  $H_0: \alpha_j=0$ ;  $p\text{-valor} < \alpha \rightarrow$  rejeita.
- Intervalos de confiança:  $\alpha_j \pm t_{\{\alpha/2, n-p-1\}} \cdot SE(\alpha_j)$ .

Seleção de Variáveis:

- Métodos forward, backward, stepwise; parcimônia no modelo.

### 3. Regressão Logística Múltipla

Resposta binária  $Y \in \{0,1\}$ ;  $\text{logit}(p) = \ln(p/(1-p)) = \alpha_0 + \sum \alpha_j X_j$ .

Odds =  $p/(1-p)$ ; OR =  $\exp(\alpha_j)$  para variável binária (razão de chances).

Avaliação:

- Likelihood, pseudo- $R^2$ , teste de razão de verossimilhança.
- Matriz de confusão, sensibilidade, especificidade, AUC.

Dicas para o Exame:

- Em T/F, evite respostas precipitadas (erro = -50%).
- Relembre fórmulas-chave e interpretação de gráficos.
- Em perguntas abertas, seja conciso e claro, cite definições e exemplos.