UNIVERSIDADE FEDERAL DA BAHIA
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA
DEPARTAMENTO DE ESTATÍSTICA
MAT229 - Análise de Regressão
Professora: Edleide de Brito **2023.1**

LABORATÓRIO 3
Regressão Linear Simples - Diagnóstico do modelo

Atividade 1: Pesquisa sobre Cerejeiras *Black cherry*

O conjunto de dados `trees`, disponível no pacote `datasets`, contém informações de 31 cerejeiras (*Black cherry*) da Floresta Nacional de Allegheny, relativas a três variáveis/características: volume de madeira útil (`Volume`), em pés cúbicos; altura (`Height`), em pés, e circunferência (`Girth`) a 4.5 pés (1,37 metros) de altura. Para esta atividade considere apenas as informações referentes ao volume e altura das árvores. Com base nestes dados:

(a) Ajuste um modelo linear simples para volume como função da altura da árvore;

(b) Avalie o gráfico de resíduos Jacknife para diagnóstico do modelo ajustado. Há algum problema?

(c) Considere as seguintes transformações para Y: c.1) $T_1 = \sqrt{Y}$, c.2) $T_2 = \log(Y)$; c.3) $T_3 = Y^2$. Para cada uma das transformações, ajuste um modelo linear simples e compare os respectivos gráficos de resíduos Jacknife.

(d) Verifique que transformação seria mais apropriada dentro da família proposta por Box e Cox. Defina graficamente o λ, a ser considerado. Compare o gráfico dos resíduos do modelo ajustado usando a transformação Box-Cox com os anteriores. Houve alguma alteração nos resultados? *Sugestão:* Utilize a função `boxcox` do pacote `MASS`.

(e) Qual das transformações anteriores você indicaria para o investigador deste estudo?

LABORATÓRIO 3
Regressão Linear Múltipla - Estimação pontual

Atividade 2: Pesquisa sobre prevalência de obesidade, diabetes e outros fatores de risco cardiovasculares

O conjunto de dados "`Dados2.csv`", contém informações de 403 afro-americanos residentes no Estado da Virginia (EUA), entrevistados em um estudo referente à prevalência de obesidade, diabetes e outros fatores de risco cardiovasculares. As variáveis/características apresentadas são descritas a seguir: número de identificação do sujeito (`id`), colesterol total (`chol`), glicose estabilizada (`stab.glu`), lipoproteína de alta densidade (colesterol bom) (`hdl`), razão colesterol total e colesterol bom (`ratio = chol/hdl`), hemoglobina glicada (`glyhb`), município de residência (Buckingham ou Louisa) (`location`), idade em anos (`age`), sexo (masculino – *male* ou feminino – *female*) (`gender`), altura (em polegadas) (`height`), peso (em libras) (`weight`), pressão sanguínea sistólica (1ª medida) (`bp.1s`), pressão sanguínea diastólica (1ª medida) (`bp.1d`), pressão sanguínea sistólica (2ª medida) (`bp.2s`), pressão sanguínea diastólica (2ª medida) (`bp.2d`), cintura (em polegadas) (`waist`) e quadril (em polegadas) (`hip`). Pede-se:

(a) Checar a base de dados quanto a possíveis inconsistências, dados ausentes e para identificar a escala de cada variável;

(b) Converter os valores de variáveis expressas em escalas não usuais ao padrão brasileiro;

(c) Tentar identificar uma variável que poderia ser considerada a variável resposta em uma análise de regressão. Na medida do possível, explore a relação dessa variável com as demais;

(d) Se acharem pertinente, propor novas variáveis, baseadas nas variáveis disponíveis. A título de exemplo, uma nova variável poderia ser o IMC.

(e) Ajustar um modelo de regressão linear múltiplo, interpretar e verificar significância dos parâmetros do modelo ajustado. Usar apenas variáveis quantitativas.

Observações:

- Apresente as conclusões sobre os resultados em forma de relatório (arquivo PDF). O arquivo deve ter texto corrido, sem inclusão de *outputs*.
- O *script* utilizado também deverá ser enviado pelo AVA Moodle.
- Lembrete: toda modelagem estatística deve ser precedida por uma análise descritiva/ exploratória, composta por gráficos e medidas descritivas pertinentes.

Algumas sugestões para a redação dos relatórios:

- Sejam parcimoniosos quanto aos resultados incluídos no relatório. Obviamente, nem todos os resultados produzidos na análise precisam ser relatados. Algumas representações fundamentais:
  - Gráficos e ou tabelas de análise descritiva/exploratória;
  - Os resultados referentes ao(s) modelo(s) ajustado(s) na forma de quadros, gráficos ou tabelas;
  - Figuras (que podem ser compostas por múltiplos gráficos) referentes ao diagnóstico do ajuste;
- A depender da análise, figuras, quadros ou tabelas para outros tipos de resultados podem ser necessários. Alguns resultados (como medidas e testes de qualidade de ajuste) podem ser inseridos no próprio texto;
- Todos os quadros, tabelas e figuras deverão ter títulos e numeração. Todos eles deverão ser mencionados em algum momento no texto, com a discussão dos respectivos resultados;
- Os resultados deverão ser devidamente editados. Saídas cruas (*outputs*) do *software* `R` serão desconsideradas;
- Não incluir códigos de programação. O *script* utilizado deverá ser enviado separadamente pelo AVA Moodle ou como apêndice do relatório.
- As páginas do relatório deverão ser numeradas.