UNIVERSIDADE FEDERAL DA BAHIA
INSTITUTO DE MATEMÁTICA - DEPARTAMENTO DE ESTATÍSTICA
DISCIPLINA: MAT 229 - ANÁLISE DE REGRESSÃO
PROF: LEILA AMORIM

TÓPICO: ANÁLISE EXPLORATÓRIA DE DADOS
LABORATÓRIO 1

1. Os dados disponibilizados pelo California Standardized Testing and Reporting (STAR) contêm informações sobre performance de exames, características da escola e informações demográficas dos alunos. Os dados são provenientes de 428 escolas de ensino fundamental de distritos da Califórnia em 1998 e 1999. Os escores dos testes são médias dos escores de leitura e matemática em testes de Stanford padronizados que foram administrados em estudantes da 5^a série. Para as análises a serem feitas neste exercício, consideraremos apenas as variáveis MATH-SCR (média do escore de matemática) e STR (razão do número de estudantes por professores:(ESTUD/PROF).

   a) Descreva as variáveis através de medidas de tendência central e dispersão.

   b) Construa diagramas de dispersão e discuta a existência de pontos atípicos. Se houverem valores atípicos, retire-os e reanalise os dados.

   c) Escreva uma função no R para calcular o desvio-padrão das variáveis MATH-SCR e STR.

   d) Escreva uma função no R para calcular o coeficiente de correlação de Pearson entre MATH-SCR e STR. Interprete esses resultados.

2. O Instituto Nacional de Diabetes e de Doenças Digestivas e Renais dos EUA conduziram um estudo com 768 mulheres da tribo Pima, que residem próximo a Phoenix. As seguintes variáveis foram coletadas: número de gestações, concentração de glicose

no plasma (obtido 2 horas depois da realização de um teste de tolerância a glicose), pressão sanguínea diastólica (mmHg), largura do tríceps (mm), nível de insulina (μ U/ml), índice de massa corpórea (kg/m^2), nível de função diabética, idade (em anos) e um teste para avaliação de sinais de diabetes (0=negativo e 1=positivo). No banco de dados as variáveis encontram-se apresentadas nessa mesma ordem, mas com seus nomes em inglês.

Com base nos dados disponíveis para este estudo:

a) Descreva sumariamente os dados do estudo (amplitude, medidas de tendência central, variabilidade). Sumarize a distribuição dos dados através da construção de histogramas.

b) Verifique que algumas das variáveis apresentam valores iguais a zero em situações em que o valor "zero"não poderia existir. Neste estudo, os investigadores estão representando os valores faltantes(*"missing"*) pelo código "zero". Exclua estas observações da base de dados e refaça as análises descritivas. Você observa alguma mudança em relação aos resultados encontrados em (a)?

c) Construa diagramas de dispersão para avaliação das relações entre as variáveis neste banco de dados. Avalie as relacões entre a variável resposta (diabetes) e as demais variáveis.

d) Construa boxplots para comparar os níveis de glucose e de insulina, a pressão diastólica, o tríceps, o bmi, a idade e o nível de função diabética entre aqueles que apresentaram resultados do teste positivo e negativo. O que você pode concluir a partir da avaliação destes gráficos?

ROTEIRO EM R PARA EXECUCÃO DO LABORATÓRIO 1

QUESTÃO 1

- Cálculo de medidas descritivas: *summary(dados1)*
- Diagrama de dispersão: *plot(X1,Y)*

QUESTÃO 2

- Leitura de base de dados:

  Procurar a seguinte opção no menu do R:

  File ← Change dir... (incluir diretório onde a base de dados se encontra)

  Digitar no console do R:

  pima = read.table("pima.ascii", head=T)

  attach(pima)

- Recodificação dos "zeros"como valor faltante (NA no R): [Exemplo com uma das variáveis]: *pima$diastolic[pima$diastolic == 0] = NA*

- Informar que a variável resultado do teste é categórica: *pima$test = factor(pima$test)*

- Quando houverem variáveis com dados missing, calcular variância usando *var(pima$diastolic, na.rm=TRUE)*.

- Construção de histogramas:

  par(mfrow=c(2,4))

  hist(pima$pregnant)

  hist(pima$diastolic)

  hist(pima$triceps)

  hist(pima$glucose)

  hist(pima$insulin)

  hist(pima$bmi)

  hist(pima$age)

  hist(pima$diabetes)

- Construção de diagramas de dispersão entre todas as variáveis da base de dados: *plot(pima)*

- Construção de boxplots comparando a distribuição das variáveis por grupo de resultado do teste de diabetes: *plot(diabetes ∼ test,pima)*

  [Fazer o mesmo para as demais variáveis da base de dados]