Lista 01

Prof. Dr. Leandro Ferreira - profleandro.ferreira@fiap.com.br

2024-09-09

1. O arquivo **claims.csv** contém uma amostra aleatória de 996 apólices de seguros de veículos referente ao período 2004-2005. As variáveis do arquivo estão na seguinte ordem : (i) **valorv** (valor do veículo em 10000 dolares australianos), (ii) **expos** (exposição do veículo), (iii) **nsinistros** (número de sinistros no período), (iv) **csinistros** (custo total dos sinistros em dolares australianos), (v) **tipov** (tipo do veículo em 11 categorias), (vi) **idadev** (idade do veículo em 4 categorias), (vii) **sexoc** (sexo do condutor principal), (viii) **areac** (área de residência do condutor principal) e (ix) **idadec** (idade do condutor principal em 6 categorias).

a. Faça uma análise descritiva dos dados e procure agrupar em um número menor de categorias algumas variáveis categóricas. Considere como variável resposta (target) **cmsinistros** = csinistros/nsinistros.

b. Qual é a relação da variável cmsinistros com sexo? Algum sexo é maior em média? Você consegue defender seu argumento utilizando as tecnicas que estudou?

c. Estude cmsinistros em relação a idade do condutor. Sua impressão muda se adicionarmos a variável sexo?

d. Existe uma relação clara de aumento de risco cmsinistros em relação ao valor do veículo?

e. A variável cmsinistros segue normalidade? Argumente com base no histograma, gráfico qq-plot e teste de normalidade.

2. Para esse exercício será utilizado os dados referentes a um estudo de caso-controle realizado no Setor de Anatomia e Patologia do Hospital Heliópolis em São Paulo, no período de 1970 a 1982 (canc3.csv). Um total de **175** pacientes com processo infecioso pulmonar atendido no hospital no período acima foi classificado segundo as seguintes variáveis: **tipo**, tipo de tumor (1: maligno, 0: benigno); **IDADE**, idade em anos; **SEXO** (0:masculino, 1: feminino); **HL**, intensidade da célula histócitos-linfócitos (1:ausente, 2: discreta, 3: moderada, 4: intensa) e **FF**, intensidade da célula fibrose-frouxa (1:ausente, 2: discreta, 3: moderada, 4: intensa).

a. Faça uma análise descritiva das variáveis do problema.

b. Existe alguma variável que chama atenção em relação ao tipo de tumor? (Dica: Cruze as variáveis em relação ao tipo de tumor).

3. No arquivo ‘expecVidas.csv’ são descritas as seguintes variáveis referentes a 50 estados norte-americanos: (i) **estado** (nome do estado), (ii) **pop** (população estimada em julho de 1975), (iii) **percap** (renda percapita em 1974 em USD), (iv) **analf** (proporção de analfabetos em 1970), (v) **expvida** (expectativa de vida em anos 1969-70), (vi) **crime** (taxa de criminalidade por 100000 habitantes 1976), (vii) **estud** (porcentagem de estudantes que concluem o segundo grau 1970), (viii) **ndias** (número de dias do ano com temperatura abaixo de zero grau Celsus na cidade mais importante do estado) e (ix) **area** (área do estado em milhas quadradas). O objetivo do exercício é tentar explicar a **expvida** média usando um modelo de regressão linear dadas as variáveis explicativas **percap**, **analf**, **crime**, **estud**, **ndias** e **dens**, em que **dens** = pop/area

a. Faça uma análise descritiva dos dados, use também gráficos para auxiliar na análise. O mais importante é entender a relação da variável *expvida* com as demais. variáveis explicativas. Comente essa parte descritiva.

b. Ajuste um modelo de regressão linear com todas as variáveis explicativas (completo) e um outro somente usando as variáveis significativas.