ME 613 - Análise de regressão
Segundo semestre de 2016
Lista de Exercícios II.

1. Resolva TODOS os exercícios deixados em sala.

2. Considere o modelo de regressão linear normal homocedástico em sua forma matricial, visto em sala. Defina $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$, $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$, $\widehat{\boldsymbol{Y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$, $\boldsymbol{R} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}}$ e $\widehat{\boldsymbol{\theta}}_i = \boldsymbol{C}_i\widehat{\boldsymbol{\beta}} - \boldsymbol{M}_i$ em que $\boldsymbol{C}_{i(a_i \times p)}$ são matrizes não-aleatórias tais que $r(\boldsymbol{C}_i) = a_i < p$ e $\boldsymbol{M}_{i(a_i \times 1)}$ são vetores não aleatórios, $i = 1, 2...$. Obtenha as distribuições conjuntas de i) $\widehat{\boldsymbol{\beta}}$ ii) $\widehat{\boldsymbol{\theta}}_i$ iii) $(\widehat{\boldsymbol{Y}}, \boldsymbol{R})'$, iv) $(\widehat{\boldsymbol{\theta}}_1, \widehat{\boldsymbol{\theta}}_2)$ e v) $(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\theta}}_1)$.

3. Seja $X = (X_1, ..., X_n)'$ um vetor de variáveis aleatórias independentes e identicamente distribuídas com $E(X_i) = \mu$ e $Var(X_i) = \sigma^2$, i = 1, . . . , n e considere a forma quadrática $Q = \sum_{i=1}^{n}(X_i - \overline{X})^2$. Utilize as propriedades das formas quadráticas para responder o exercício.

   a) Mostre que a matriz da forma quadrática Q é idempotente.
   b) Mostre que $Q/[n(n-1)]$ é um estimador não viciado de Var($\overline{X}$).
   c) Calcule a variância de $Q/[n(n-1)]$ sob a suposição de que X tem distribuição Normal. Sugestão, se $\boldsymbol{Y} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ então $\mathcal{V}(\boldsymbol{Y}'\boldsymbol{A}\boldsymbol{Y}) = 2\mathrm{tr}\,(\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}\boldsymbol{\Sigma}) + 4\boldsymbol{\mu}'\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}\boldsymbol{\mu}$ se $\boldsymbol{A}$ for uma matriz não aleatória simétrica.

4. Considere o modelo linear $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\xi}$, cujas componentes $\boldsymbol{Y}_{(n\times 1)}, \boldsymbol{X} = [\boldsymbol{1}, \boldsymbol{Z}]_{(n\times p)}$, com $\boldsymbol{1} = (1, ..., 1)'_{(n\times 1)}$ e $r(X) = p$, $\boldsymbol{\beta} = [\alpha, \boldsymbol{\theta}^t]^t_{(p\times 1)}$ e $\boldsymbol{\xi}_{(n\times 1)}$ com $\mathcal{E}(\boldsymbol{\xi}) = \boldsymbol{0}$ e $Cov(\boldsymbol{\xi}) = \sigma^2\boldsymbol{I}_n$, têm a interpretação usual.

   a) Obtenha $\widehat{\boldsymbol{\theta}}$, o estimador de mínimos quadrados de $\boldsymbol{\theta}$.
   b) Mostre que $Cov(\widehat{\boldsymbol{\theta}}) = \sigma^2(\boldsymbol{Z}^t\boldsymbol{A}\boldsymbol{Z})^{-1}$ com $\boldsymbol{A} = \boldsymbol{I}_n - n^{-1}\boldsymbol{1}\boldsymbol{1}^t$.

5. (Referência: Paula, 2013, pag. 80) Os dados do arquivo censo.txt foram extraídos do censo do IBGE de 2000 e apresentam para cada unidade da federação o número médio de anos de estudo e a renda média mensal (em reais) do chefe ou chefes do domicílio. O objetivo, em nosso caso, é estudar o relacionamento da renda média mensal em função do número médio de anos de estudo. Responda aos itens:

   a) Faça uma análise descritiva dos dados.
   b) Proponha (interpretando adequadamente cada parâmetro) e ajuste um modelo de regressão linear simples (MRLS), mais especificamente uma reta, em que todos os parâmetros tenham interpretação compatível com a natureza dos dados. Apresente as estimativas pontuais, erros-padrão, intervalos de confiança (95%) e testes de hipótese de nulidade individuais. O que você pode concluir, com base no modelo, sobre a influência do número médio de anos de estudo na renda média mensal? Se for o caso, ajuste um modelo reduzido. Ou seja, inicie a análise com um modelo completo e reduza-o (se for o caso) até encontrar o modelo mais simples compatível com os dados. Em cada passo defina claramente os parâmetros e as hipóteses de interesse, estime os parâmetros e teste as hipóteses, e apresente os resultados de forma acessível para os pesquisadores (que não têm conhecimento de Estatística).

c) Apresente um gráfico de dispersão (entre renda e anos de escolaridade) com a curva ajustada (modelo final) bem como um gráfico de dispersão com os valores observados e preditos pelo modelo (com a reta identidade). O que você pode dizer sobre o poder preditivo do modelo? Você acha que uma outra estrutura de regressão linear para a média poderia ser mais apropriada do que a que foi utilizada? Qual seria? Justifique, adequadamente, sua resposta.

   OBS: nos gráficos identifique as unidades de federação, quando for possível.

6. (Referência: Narula e Stangenhaus, 1988, pgs. 31-33; Paula, 2013). No arquivo imoveis.txt são apresentados dados relativos a uma amostra de 27 imóveis. Na ordem são apresentados os valores das seguintes variáveis: (i) imposto do imóvel (em 100 USD), (ii) área do terreno (em 1000 pés quadrados), (iii) área construída (em 1000 pés quadrados), (iv) idade da residência (em anos) e (v) preço de venda do imóvel (em 1000 USD). O objetivo é avaliar a influência sobre o preço de venda por parte das demais variáveis. Responda aos itens:

   a) Faça uma análise descritiva dos dados.

   b) Proponha (interpretando adequadamente cada parâmetro) e ajuste um modelo de regressão linear múltipla (MRLM), mais especificamente uma reta, em que todos os parâmetros tenham interpretação compatível com a natureza dos dados. Apresente as estimativas pontuais, erros-padrão, intervalos de confiança (95%) e testes de hipótese de nulidade individuais. O que você pode concluir, com base no modelo, sobre a influência de cada uma das variáveis explicativas no preço de venda do imóvel? Se for o caso, ajuste um modelo reduzido. Ou seja, inicie a análise com um modelo completo e reduza-o (se for o caso) até encontrar o modelo mais simples compatível com os dados. Em cada passo defina claramente os parâmetros e as hipóteses de interesse, estime os parâmetros e teste as hipóteses, e apresente os resultados de forma acessível para os pesquisadores (que não têm conhecimento de Estatística).

   c) Apresente um gráfico de dispersão com os valores observados e preditos pelo modelo (com a reta identidade). O que você pode dizer sobre o poder preditivo do modelo? Você acha que uma outra estrutura de regressão linear para a média poderia ser mais apropriada do que a que foi utilizada? Qual seria? Justifique, adequadamente, sua resposta.

7. Em relação ao problema da Questão 13 da Lista I, apresente um modelo apropriado para responder as perguntas de interesse levando em consideração as etiologias cardíacas. Interprete cada um dos parâmetros. Se for possível, reduza o modelo através dos testes de hipótese individuais de nulidade. Em cada passo defina claramente os parâmetros e as hipóteses de interesse, estime os parâmetros e teste as hipóteses, e apresente os resultados de forma acessível para os pesquisadores (que não têm conhecimento de Estatística).