

MI 406 - Regressão

Segundo semestre de 2014

Lista de Exercícios II

Data da entrega: até o dia 08/10/2014, no começo da aula.

Exercícios selecionados para a entrega: 2, 4, 5, 6 e 7.

OBS: Para as questões relativas à análise de dados (5, 6 e 7) veja o arquivo contido no link “modelo de relatório” para ver uma sugestão de relatório. Não é necessário seguir integralmente o modelo de relatório, o mais importante é responder as perguntas de interesse de modo acessível para leitores que não têm conhecimento de Estatística, interpretar os resultados em termos do problema, descrever claramente os modelos considerados, apresentar os resultados em forma de tabelas, gráficos ou no próprio texto (quando for o caso) e seguir uma sequência lógica. Conclusões finais devem ser apresentadas. Lembrando que, nesse caso, as resoluções devem ser digitadas e apresentar, no máximo, para cada questão, 5 páginas. Note que o conjunto de dados analisado no modelo de relatório apresenta uma estrutura diferente dos dados a serem analisados nesta lista. Contudo, nas aulas, apresentou-se e está apresentando-se exemplos de análise de dados com estruturas similares.

1. Resolva TODOS os exercícios deixados em sala.
2. Seja $X = (X_1, \dots, X_n)'$ um vetor de variáveis aleatórias independentes e identicamente distribuídas com $E(X_i) = \mu$ e $Var(X_i) = \sigma^2$, $i = 1, \dots, n$ e considere a forma quadrática $Q = \sum_{i=1}^n (X_i - \bar{X})^2$. 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 $Var(\bar{X})$.
 - c) Calcule a variância de $Q/[n(n-1)]$ sob a suposição de que X tem distribuição Normal. Sugestão, se $\mathbf{Y} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ então $\mathcal{V}(\mathbf{Y}'\mathbf{A}\mathbf{Y}) = 2\text{tr}(\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}\boldsymbol{\Sigma}) + 4\boldsymbol{\mu}'\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}\boldsymbol{\mu}$ se \mathbf{A} for uma matriz não aleatória simétrica.
3. Considere o modelo linear $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\xi}$, cujas componentes $\mathbf{Y}_{(n \times 1)}$, $\mathbf{X} = [\mathbf{1}, \mathbf{Z}]_{(n \times p)}$, com $\mathbf{1} = (1, \dots, 1)_{(n \times 1)}$ e $r(X) = p$, $\boldsymbol{\beta} = [\alpha, \boldsymbol{\theta}^t]_{(p \times 1)}$ e $\boldsymbol{\xi}_{(n \times 1)}$ com $\mathcal{E}(\boldsymbol{\xi}) = \mathbf{0}$ e $Cov(\boldsymbol{\xi}) = \sigma^2 \mathbf{I}_n$, têm a interpretação usual.
 - a) Obtenha $\hat{\boldsymbol{\theta}}$, o estimador de mínimos quadrados de $\boldsymbol{\theta}$.
 - b) Mostre que $Cov(\hat{\boldsymbol{\theta}}) = \sigma^2(\mathbf{Z}'\mathbf{A}\mathbf{Z})^{-1}$ com $\mathbf{A} = \mathbf{I}_n - n^{-1}\mathbf{1}\mathbf{1}'$.
4. Considere o modelo linear $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\xi}$ com as suposições usuais. Mostre que o teste da razão de verossimilhanças para a hipótese $\mathbf{C}\boldsymbol{\beta} = \mathbf{M}$ pode ser baseado numa estatística, com distribuição F sob H_0 , definida como o quociente entre os quadrados médios devido à hipótese e ao resíduo.
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 Stangenhau, 1988, pgs. 31-33; Paula, 2013,). No arquivo imoveis.dat 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 9 da Lista I, utilize o modelo proposto naquela lista para responder às perguntas de interesse. 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).