ME 613 - Regressão
Primeiro semestre de 2019
Lista de Exercícios IV

1. Resolva TODOS os exercícios deixados em sala.

2. Para cada um dos conjuntos de dados analisados (nas listas e em sala de aula) faça uma análise de influência, de forma apropriada, reajustando o modelo sem as observações que você considera como potencialmente discrepantes, comparando esses resultados com aqueles obtidos com todas as observações.

3. Para cada um dos conjuntos de dados analisados (nas listas e em sala de aula), quando pertinente e se já não tiver sido feito, selecione o melhor modelo usando as técnicas vistas em sala de aula (resíduos, testes de hipótese e estatísticas de comparação de modelos).

4. Para cada um dos conjuntos de dados analisado (nas listas e em sala de aula), quando pertinente faça uma análise da multicolinearidade dos dados, analisando o problema de forma apropriada.

5. Considere a questão 2 da Lista III e cada uma das quatro situação situações a seguir, sob o modelo $Y_i = \beta_0 + \beta_1(x_i - \overline{x}) + \xi_i$:

   (a) $\xi_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$.
   (b) $\mathcal{V}(\xi_i) = \sigma^2 x_i$.
   (c) $\xi_i \overset{i.i.d.}{\sim} t_{(\nu)}(\sigma^2), \nu = 6$, $\mathcal{V}(\xi) = \sigma^2 \frac{\nu}{\nu-2}$.
   (d) $\mathcal{V}(\xi_i) = \sigma^2$ e $\text{Corre}(\xi_i, \xi_j) = \rho$, $i \neq j, \rho = 0,90$.

   Considere, para cada situação, três tamanhos amostrais, n = 30, 50, 100. Então, tem-se 12 cenários. Gere, para cada um dos 12 cenários, R = 100 réplicas (conjuntos de valores de "Y", para um mesmo conjunto de valores de "x"), assumindo $\beta_0 = 1$, $\beta_1 = 1,5$ e $\sigma^2 = 4$. Em cada um dos cenários, simule as variáveis explicativas segundo uma $U(5, 20)$, (ou seja, um único conjunto de covariáveis para as 100 réplicas). Em cada um das situações: estime os parâmetros por mínimos quadrados, calcule intervalos de confiança (de 95%) para cada um deles e testes as hipóteses $H_0 : \beta_0 = 1$ vs $H_1 : \beta_0 \neq 1$ e $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$, ao nível de significância $\alpha = 0,05$. Usando os resultados das 100 réplicas, estude a distribuição dos estimadores, calcule a probabilidade de cobertura dos intervalos de confiança, bem como os níveis descritivos empíricos dos referidos testes, comparando-os com os verdadeiros valores. Discuta, de forma apropriada, os resultados obtidos, para cada uma das 12 situações.

6. Considere o modelo de regressão normal linear visto em `www.ime.unicamp.br/~cnaber/aula_Intro_MRLM_REG_1S_2019_parte_1.pdf`, slides 2, asssumindo que $\xi_i \overset{i.i.d.}{\sim} t_{(\nu)}(\sigma^2)$, tal que $\mathcal{V}(\xi) = \sigma^2 \frac{\nu}{\nu-2}$, considerando ν fixado. Estime $\boldsymbol{\beta}$ e σ^2 por máxima verossimilhança e implemente tal método no R (use funções já prontas como o optim, por exemplo). Simule um único conjunto de dados a partir desse modelo, considerando $\nu = 8$, $\sigma^2 = 10$ e $Y_i = 2 - 1x_i + \xi_i$, n =50, estime os parâmetros e compare as estimativas (pontual e intervalarmente) com os verdadeiros valores.

7. Considere os modelo de regressão $Y_i = 1 + 1,5x_i + \xi_i$, $i = 1, ..., n$, n = 30, 50, 100 e as duas seguintes situações:

a) $\xi \overset{i.i.d.}{\sim} N(0, 25)$

b) $\xi \overset{i.i.d.}{\sim} t_{(\nu=7)}(\sigma^2 = 25)$, $\mathcal{V}(\xi) = \sigma^2 \dfrac{\nu}{\nu - 2}$.

Para cada um dos 6 cenários, situações gere R = 100 réplicas do modelo em questão, ajuste o MRNLH e o modelo t de Student, utilize as estatísticas de comparação de modelos vistas em sala de aula e calcule a proporção de vezes em que cada modelo é selecionado por cada estatística. Sugestão, para o modelo de regressão t, considere o pacote “heavy”. Discuta, de forma apropriada, os resultados.

8. No arquivo salary.dat consta, da esquerda para a direita, informações sobre: salário anual (em mil USD), sexo, posição na empresa (escore de 1 a 9, considere-o quantitativo discreto) e experiência (em anos). O interesse é estudar como o salário é afetado pelas três outras variáveis. Proponha um modelo de regressão linear múltipla homocedástico que leve em consideração cada uma das três covariáveis com um coeficiente para cada sexo (relativo às duas covariáveis quantitativas), como também uma parte relativa ao sexo (semelhante ao modelo apresentado em classe relativo ao consumo de oxigênio, carga e etiologia cardíaca). Utilize as duas covariáveis quantitativas centradas. Reduza-o até obter o modelo mais simples compatível com os dados (realizando análise de resíduos apropriadas para cada modelo). No final, apresente as conclusões devidas.

9. O conjunto de dados coelho.dat corresponde a uma amostra de 71 observações relartivas a coelhos europeus da espécie Oryctolagus cuniculuse, onde se mediu o peso da lente ocular seca (em mg - resposta) e a idade do animal (em dias - covariável). O objetivo é explicar o peso da lente em função da idade, através de um modelo normal não-linear apropriado. Proponha um modelo adequado e realize uma análisemais completa possível, para responder as perguntas de interesse (não se esqueça de verificar a qualidade do ajuste do modelo).

10. Analise o conjunto de dados 9 (dados sobre automóveis) no que diz respeito a multicolinearidade, utilizando, em caso deste problema ser detectado, alguma técnica que você julgar apropriada, para tal. Não se esqueça de realizar uma análise de resíduos para o modelo escolhido. Compare o resultado com aqueles apresentados em sala.