MI 406 - Regressão
Primeiro semestre de 2021
Lista de Exercícios III.

1. Resolva TODOS os exercícios deixados em sala.
2. Para todas as questões de análises de dados, faça uma análise residual apropriada (não somente para o modelo inicial para os eventuais modelos reduzidos). Se houver dois ou mais modelos competidores (p.e., uma reta e uma parábola) compare os ajustes através dessa metodologia.
3. Considere o conjunto de dados disponível no arquivo Sef1999.xls (veja a descrição dos dados no próprio arquivo). Considere que o objetivo é comparar os quatro grupos formados pelas combinações dos fatores escova e dentrifício (assuma que as observações são não correlacionadas e homocedásticas) e que a variável resposta é $\frac{\text{IPB depois}}{\text{IPB antes}}$ (neste caso quanto maior, melhor o desempenho). Proponha um modelo que possa responder às perguntas de interesse e reduza-o até obter o modelo mais simples compatível com os dados (realizando análise de resíduos apropriadas para cada modelo). No final, conclua qual das escovas é melhor e em que circunstância (com ou sem dentrifício).
4. Repita a questão anterior considerando como resposta $\frac{\text{IPB antes–IPB depois}}{\text{IPB antes}}$.
5. Repita a questão 3 considerando como resposta o IPB depois e o IPB antes como mais uma variável explicativa, comparando os resultados e os modelos (Questões 3, 4 e 5). Qual das três variáveis resposta/modelo você considera mais apropriada(o) e em que sentido? Justifique, adequadamente, sua resposta.
6. Considere os estimadores de mínimos quadrados de β_0 e β_1 do modelo de regressão linear simples, apresentados aqui. Suponha as três seguintes situações, em relação aos erros:
   (a) $\mathcal{V}(\xi_i) = \sigma_i^2$ (conhecidos), $i = 1, 2, ..., n$.
   (b) $\xi \overset{i.i.d.}{\sim} t_{(\nu)}$, ν (conhecido).
   (c) Corre$(\xi_i, \xi_j) = \rho$ (conhecido), $i \neq j$, $\rho \in (-1, 1)$.

   Encontre, para cada um das três situações acima, a distribuição dos referidos estimadores, das estatísticas do teste (sob H_0 e H_1) bem como da quantidade pivotal, propondo IC's (apropriados).
7. Pesquise sobre testes de hipótese e construção de intervalos de confiança para o coeficiente de correlação linear de Pearson.
8. Pesquise sobre o coeficiente de correlação linear múltipla.

9. Pesquisa sobre a relação entre o teste Anova (para o modelo de regressão linear simples) e o teste de nulidade acerca do coeficiente de correlação linear de Pearson.

10. Pesquise sobre o Teorema de Gauss-Markov.

11. Seja o modelo $Y_i = \beta_0 + \beta_1(x_i - \overline{x}) + \xi_i$ com as seguintes suposições:

    (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.

12. Considere o modelo de regressão normal linear visto em , 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 conhecido. 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 $\nu = 4$, $\sigma^2 = 10$ e $Y_i = 2 - 1x_i + \xi_i$, n =50, estime os parâmetros e compare as estimativas (pontual e intervalarmente) com os verdadeiros valores. Compare os resultados com aquele obtidos sob normalidade via MQO (função lm).

13. 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.

14. 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.