MAE5763 - Modelos Lineares Generalizados
2º semestre 2020

Prof. Gilberto A. Paula

1ª Lista de Exercícios

1. Supor o modelo $y_{1j} = \alpha + \epsilon_{1j}$, $y_{2j} = \alpha + \Delta + \epsilon_{2j}$ e $y_{3j} = \alpha - \Delta + \epsilon_{3j}$, em que $\epsilon_{ij} \overset{\text{iid}}{\sim} \text{N}(0, \sigma^2)$, para $i = 1, 2, 3$ e $j = 1, \ldots, r$. Expresse esse modelo na forma matricial $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ especificando cada quantidade. Obtenha $\widehat{\alpha}$, $\widehat{\Delta}$, $\text{Var}(\widehat{\alpha})$, $\text{Var}(\widehat{\Delta})$ e $\text{Cov}(\widehat{\alpha}, \widehat{\Delta})$.

   Mostre que a estatística F para testar $\text{H}_0 : \Delta = 0$ contra $\text{H}_1 : \Delta \neq 0$ pode ser expressa na forma

$$\text{F} = \frac{r(3r-2)}{2} \frac{(\bar{y}_2 - \bar{y}_3)^2}{\sum_{i=1}^{3}\sum_{j=1}^{r}(y_{ij} - \hat{y}_{ij})^2}.$$

2. Considere a seguinte função densidade de probabilidade:

$$f(y; \theta, \phi) = \frac{\phi \text{a}(y, \phi)}{\pi(1+y^2)^{1/2}} \exp[\phi\{y\theta + (1-\theta^2)^{1/2}\}],$$

   em que $0 < \theta < 1$, $-\infty < y < \infty$, $\phi^{-1} > 0$ é o parâmetro de dispersão e $\text{a}(\cdot, \cdot)$ é uma função normalizadora. Mostre que essa distribuição pertence á família exponencial. Encontre a função de variância. Obtenha os componentes do desvio $d^{*2}(y_i; \hat{\mu}_i)$ supondo uma amostra de n variáveis aleatórias independentes de médias μ_i e parâmetro de dispersão ϕ^{-1}, para $i = 1, \ldots, n$.

3. Seja Y o número de ensaios independentes até a ocorrência do r-ésimo sucesso, em que π é a probabilidade de sucesso em cada ensaio. Denote

$Y \sim \text{Pascal}(r, \pi)$ (distribuição de Pascal) cuja função de probabilidades é dada por

$$f(y; r, \pi) = \binom{y-1}{r-1} \pi^r (1-\pi)^{(y-r)},$$

para $y = r, r+1, \ldots$ e $0 < \pi < 1$. Mostre que $Y^* = \frac{Y}{r}$ pertence à família exponencial de distribuições. Encontre a função de variância $V(\mu)$, em que $\mu = \text{E}(Y^*)$. Supor agora que $Y_i \overset{\text{ind}}{\sim} \text{Pascal}(r, \pi_i)$ para $i = 1, \ldots, n$. Obtenha os componentes $d^{*2}(y_i; \hat{\pi}_i)$ da função desvio.

4. Supor $Y_i|(x_i, z_i) \overset{\text{ind}}{\sim} \text{Be}(\mu_i)$, em que $\text{arcsen}(\sqrt{\mu_i}) = \eta_i = \beta(x_i - \bar{x}) + \gamma(z_i - \bar{z})$, para $i = 1, \ldots, n$. Obter a matriz $\mathbf{X}$ e as variâncias assintóticas de $\hat{\beta}$ e $\hat{\gamma}$ e $\text{Cov}(\hat{\beta}, \hat{\gamma})$. Compare a correlação linear $\rho(\hat{\beta}, \hat{\gamma})$ com a correlação linear amostral $r_{xz} = S_{xz}/\sqrt{S_{xx}}\sqrt{S_{zz}}$, em que $S_{xz} = \sum(x_i - \bar{x})(z_i - \bar{z})$, $S_{xx} = \sum(x_i - \bar{x})^2$ e $S_{zz} = \sum(z_i - \bar{z})^2$. Comente. Use o resultado: $\frac{d}{dx}\text{arcsen}\{u(x)\} = \frac{1}{\sqrt{1-u^2}}\frac{du}{dx}$. Supor que $\det\{(\mathbf{X}^\top \mathbf{X})\} > 0$.

5. Supor que $Z_i \overset{\text{iid}}{\sim} \text{ZANBI}(\mu, \nu, \pi)$, para $i = 1, \ldots, n$, em que a função de probabilidades de z_i fica dada por

$$f_z(z_i; \mu, \nu, \pi) = \begin{cases} \pi & \text{se} \quad z_i = 0 \\ (1-\pi)\frac{f_y(z_i;\mu,\nu)}{1-f_y(0;\mu,\nu)} & \text{se} \quad z_i = 1, 2, \ldots, \end{cases}$$

em que $f_y(y_i; \mu, \nu)$ denota a função de probabilidades de uma $\text{BN}(\mu, \nu)$. Supondo ν fixo obter a estatística da razão de verossimilhanças para testar H: $\mu = 1$ contra A: $\mu \neq 1$?

6. No arquivo **fuel2001.txt** (Weisberg, 2014, Cap.3) são descritas as seguintes variáveis referentes aos 50 estados norte-americanos mais o Distrito de Columbia no ano de 2001: (i) `UF`, unidade da federação, (ii) `Drivers`, número de motoristas licenciados, (iii) `FuelC`, total de gasolina vendida (em mil galões), (iv) `Income`, renda per capita em 2000 (em mil USD), (v) `Miles`, total de milhas em estradas federais, (vi) `MPC`, milhas per capita percorridas, (vii) `Pop`, população ≥ 16 anos e (viii) `Tax`, taxa da gasolina (em cents por galão). A fim de possibilitar uma comparação entre as UFs duas novas variáveis são consideradas `Fuel = 1000*FuelC/Pop` e `Dlic = 1000*Drivers/Pop`, além da variável `Miles` ser substituída por `log(Miles)`. Para ler o arquivo no R use os comandos

`fuel2001 = read.table(“fuel2001.txt”, header=TRUE).`

Considere como resposta a variável Fuel e como variáveis explicativas Dlic, log(Miles), Income e Tax. Faça inicialmente uma análise descritiva dos dados. Por exemplo, boxplots robustos para cada variável e diagrama de dispersão de cada variável explicativa e a variável resposta Fuel. Comente. Aplique o procedimento `stepAIC` para selecionar as variáveis explicativas. Verifique se é possível incluir alugma interação. Com o modelo selecionado faça uma análise de diagnóstico: análise de resíduos e distância de Cook. Avalie o impacto dos pontos destacados. Interprete os coeficientes estimados. Apenas de forma ilustrativa ajustar o modelo final no GAMLSS. Apresentar os gráficos de resíduos e comentar.

7. No arquivo **heart.txt** (Hosmer, Lemeshow e Sturdivant, 2013, Cap.1) são descritos os dados de $n = 100$ pacientes com ausência (HD=0) e evidência (HD=1) de doença arterial coronariana, além da idade (Age) do paciente e a faixa etária (FE). Para ler os dados use o comando

   `heart = read.table(“heart.txt”, header=TRUE)`

   Fazer uma análise descritiva dos dados, por exemplo boxplots robustos da idade para cada um dos grupos, comente. Construa uma tabela de contigência com as frequências relativas de pacientes com evidência e ausência da doença segundo as faixas etárias, comente. Ajustar um modelo logístico para explicar a probabilidade Pr(HD=1) dado Age. Comente as estimativas. Fazer uma análise de diagnóstico como gráfico de resíduos e distância de Cook. Avalie o impacto das observações destacadas como possivelmente influentes. Construa uma banda de confiança de 95% para Pr(HD=1) dado Age. Encontre uma estimativa intervalar de 95% para a razão de chances entre um paciente com Age+1 e um paciente com Age ter presença da doença. Construa a curva ROC e estabeleça um critério para classificar pacientes como suspeitos de terem presença da doença. Para esse critério obter as taxas de positivo positivo e de falso positivo. Ajustar o modelo pelo GAMLSS através dos comandos

   `y.heart = cbind(HD, 1-HD)`

   `ajuste = gamlss(y.heart ~ Age, family=BI)`

```
plot(ajuste)
```

```
wp(ajuste)
```

Comente os gráficos de resíduos.