MAE5763 - Modelos Lineares Generalizados
2º semestre 2023

Prof. Gilberto A. Paula

1ª Lista de Exercícios

1. Considere a seguinte função densidade de probabilidade:

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

em que $0 < \theta < 1$, $-\infty < y < \infty$, $\phi^{-1} > 0$ é o parâmetro de dispersão e $\mathrm{a}(\cdot,\cdot)$ 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 $i = 1,\ldots,n$. Expresse também a medida R^2.

2. Supor $Y_1,\ldots,Y_k$ variáveis aleatórias independentes tais que Y_i tem função de probabilidade expressa na forma

$$f(y_i;\psi) = \binom{m}{y_i} \left(\frac{\psi}{1+\psi}\right)^{y_i} \left(\frac{1}{1+\psi}\right)^{(m-y_i)},$$

com $\log(\psi) = \alpha$, $y_i = 0,1,\ldots,m$ e $i = 1,\ldots,k$. Como fica a estatística do teste da razão de verossimilhanças para testar $\mathrm{H}_0 : \alpha = 1$ contra $\mathrm{H}_1 : \alpha \neq 1$? Qual a distribuição nula assintótica da estatística do teste?

3. Supor que $Z_i \overset{\text{iid}}{\sim} \mathrm{ZABI}(n,\mu,\pi)$, para $i = 1,\ldots,k$. Obter $\mathrm{E}(Z)$, $\mathrm{Var}(Z)$, as estimativas de máxima verossimilhança $\hat{\mu}$ e $\hat{\pi}$, as respectivas variâncias assintóticas $\mathrm{Var}(\hat{\mu})$ e $\mathrm{Var}(\hat{\pi})$ e a covariância assintótica $\mathrm{Cov}(\hat{\mu},\hat{\pi})$. Sugestão: supor que os r primeiros elementos são zeros e que $R \sim$

$\mathrm{B}(k, \pi)$. Lembre que a função de probabilidade para a variável Z pode ser expressa na forma

$$f_Z(z; \mu, \pi) = \begin{cases} \pi & \text{se} \quad z = 0 \\ (1-\pi)\frac{f_Y(z;\mu)}{1-P(Y=0)} & \text{se} \quad z = 1, 2, \ldots, n, \end{cases}$$

em que $f_Y(z; \mu)$ é a função de probabilidade de uma $\mathrm{B}(n, \mu)$.

4. Supor que $Z_i \overset{\text{iid}}{\sim} \text{ZANBI}(\mu, \nu, \pi)$, para $i = 1, \ldots, n$, em que a função de probabilidade 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 probabilidade de uma $\mathrm{BN}(\mu, \nu)$. Supondo ν fixo obter a estatística da razão de verossimilhanças para testar H: $\mu = 1$ contra A: $\mu \neq 1$? Supor que os primeiros r elementos são zeros.

5. Supor $Y_{ij} \overset{\text{ind}}{\sim} \mathrm{Ge}(\mu_i)$ (distribuição geométrica), em que $\mu_i = \mathrm{E}(Y_i)$ e $\mathrm{Var}(Y_i) = V(\mu_i) = \mu_i(\mu_i - 1)$, $i = 1, 2$ e $j = 1, \ldots, m$, com $\log(\mu_1) = \eta_1 = \alpha - \Delta$ e $\log(\mu_2) = \eta_2 = \alpha + \Delta$. Como ficam as matrizes $\mathbf{X}$ e $\mathbf{W}$? Obter as variâncias assintóticas $\mathrm{Var}(\hat{\alpha})$ e $\mathrm{Var}(\hat{\Delta})$ além da covariância assintótica $\mathrm{Cov}(\hat{\alpha}, \hat{\Delta})$, deixando as expressões em função dos pesos $\mathbf{W}$. Mostre que a estatística do teste de escore para testar $\mathrm{H}_0 : \Delta = 0$ contra $\mathrm{H}_1 : \Delta \neq 0$ fica expressa na forma

$$\xi_{SR} = \frac{m(\bar{y}_2 - \bar{y}_1)^2}{2\bar{y}(\bar{y} - 1)}.$$

Qual a distribuição nula assintótica da estatística do teste?

6. Supor $Y_i \overset{\text{iid}}{\sim} \text{PoissonTruncada}(\lambda)$, para $i = 1, \ldots, n$. Obter a estimativa de máxima verossimilhança $\hat{\lambda}$ e a respectiva variância assintótica $\mathrm{Var}(\hat{\lambda})$. A função de probabilidade da Poisson truncada é expressa na forma

$$f(y; \lambda) = \frac{e^{-\lambda}\lambda^y}{y!(1 - e^{-\lambda})},$$

para $y = 1, 2, \ldots$, em que $\lambda > 0$. Obter $\mathrm{E}(Y)$ e $\mathrm{Var}(Y)$.

7. 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. Apresente a matriz de correlação entre as variáveis, boxplot robusto, densidade da variável resposta e diagramas de dispersão (com tendência) entre cada variável explicativa e a variável resposta. Comente. Aplique o procedimento `stepAIC` para selecionar as variáveis explicativas. Verifique se é possível incluir alugma interação de $1^{\underline{a}}$ ordem. Com o modelo selecionado faça uma análise de diagnóstico: análise de resíduos e distância de Cook com $k = 2$. 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.

8. No arquivo **heart.txt** (Hosmer et al., 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)

rqres.plot(ajuste, howmany=8, ylim.all=1)

rqres.plot(ajuste, howmany=40, plot="all")
```

Comente os gráficos de resíduos. Obter R^2 do modelo final.

9. Considere o arquivo **BigMac2003** da biblioteca `alr4` do R, em que são descritas as seguintes variáveis de 69 cidades de diversos países:

   - `BigMac`: minutos de trabalho para comprar um Big Mac
   - `Bread`: minutos de trabalho para comprar 1kg de pão
   - `Rice`: minutos de trabalho para comprar 1kg de arroz
   - `FoodIndex`: índice de preços de alimentos
   - `Bus`: valor da passagem de ônibus (em USD)
   - `Apt`: valor do aluguel (em USD) de um apartamento padrão de 3 dormitórios
   - `TeachGI`: salário bruto anual (em 1000 USD) de um professor de ensino fundamental
   - `TeachNI`: salário líquido anual (em 1000 USD) de um professor de ensino fundamental
   - `TaxRate`: imposto pago (em porcentagem) por um professor de ensino fundamental

- `TeachHours`: carga horária semanal (em horas) de um professor de ensino fundamental.

Para disponibilizar e visualizar um resumo dos dados use na sequência os seguintes comandos do R:

```
require(alr4)
require(MASS)
attach(BigMac2003)
summary(BigMac2003).
```

O objetivo principal do estudo é relacionar a variável `BigMac` com as demais variáveis explicativas. Apresente a densidade da variável resposta, as correlações lineares amostrais bem como os diagramas de dispersão (com tendência) entre a variável resposta e cada uma das variáveis explicativas. Comente. Padronize as variáveis explicativas. Por exemplo, para padronizar a variável explicativa `Bread` use o comando

```
sBread = scale(Bread, center = TRUE, scale = TRUE).
```

Ajustar inicialmente um modelo com resposta gama e ligação logarítmica no GAMLSS através do comando

```
fit1.bigmac = gamlss(BigMac ~ ·, family=GA, data=BigMac2003).
```

Através do procedimento `stepGAIC` fazer uma seleção das variáveis explicativas

```
fit2.bigmac = stepGAIC(fit1.bigmac).
```

Para o submodelo selecionado aplicar análises de resíduos através dos comandos `plot(fit2.bigmac)` e `wp(fit2.bigmac)`. Construir o gráfico da distância de Cook. Comente. Classifique as variáveis explicativas segundo o impacto na explicação da média da variável resposta. Apresente e comente o `term.plot(fit2.bigmac)`. Obter R^2 do modelo final.

10. No arquivo **visitas.txt**, extraído de Zeileis et al. (2008), são descritas as seguintes variáveis observadas numa amostra aleatória de 4380 indivíduos com mais de 65 anos e atendidos através de um programa de saúde pública durante os anos de 1987-88: (i) `nvis` (número de visitas ao médico), (ii) `hosp` (número de internações hospitalares), (iii)

`altacond` (0:não, 1:sim), (iv) `baixacond` (0:não, 1:sim), (v) `nucron` (número de condições crônicas), (vi) `gênero` (0:feminino, 1:masculino), (vii) `escol` (escolaridade em anos de estudo) e (viii) `seguro` (seguro particular, 0:não, 1:sim). Quando `altacond=0` e `baixacond=0` tem-se condição regular. Quando (altocond=0 e baixacond=0) tem-se média condição que é a casela de referência.

O objetivo do estudo é explicar a demanda por serviços médicos através de modelos de regressão em que a resposta é o número de visitas ao médico e as demais variáveis como explicativas. Compare os modelos com resposta Poisson e com resposta binomial negativa. Para ler esse arquivo no R faça o seguinte:

```
visitas = read.table("visitas.txt", header=TRUE)
```

```
attach(visitas).
```

Fazer inicialmente uma análise descritiva dos dados, por exemplo histograma de `nvis`, boxplots robustos de `nvis` segundo os níveis das variáveis categóricas e diagrama de dispersão (com tendência) entre `nvis` e `escol`.

Para ajustar o modelo de Poisson com todas as variáveis explicativas use o comando

```
require(gamlss)
```

```
fit1.visitas = gamlss(nvis ∼., family=PO).
```

Use o comando `stepGAIC` para selecionar um submodelo. As análises de resíduos podem ser realizadas através dos comandos

```
plot(fit1.visitas)
```

```
rqres.plot(fit1.visitas, howmany=8, ylim.all=1)
```

```
rqres.plot(fit1,visitas, howmany=40, plot="all").
```

Para ajustar o modelo com resposta binomial negativa use o comando

```
fit2.visitas = gamlss(nvis ∼.,family=NBI).
```

Para ajustar o modelo com resposta binomial negativa com todas as variáveis explicativas ajustando conjuntamente a média e o parâmetro de dispersão, use o comando

```
fit3.visitas = gamlss(nvis ∼., ∼., family=NBI).
```

Use o comando `stepGAIC` para selecionar um submodelo. As análises de resíduos podem ser realizadas através dos comandos

```
plot(fit3.visitas)
```

```
rqres.plot(fit3.visitas, howmany=8, ylim.all=1)
```

```
rqres.plot(fit3.visitas, howmany=40, plot="all").
```

Qual modelo ajusta melhor os dados? Interpretar as estimativas do modelo selecionado apresentando estimativas intervalares de 95%. Obter R^2 do modelo final.

Referências

Hosmer DW, Lemeshow S, Sturdivant R (2013) *Applied Logistic Regression, 3nd Edition.* Wiley.

Weisberg S (2014) *Applied Linear Regression, Fourth Edition.* Wiley.

Weisberg S (2022) Data to Accompany Applied Linear Regression, Fourth Edition: http://CRAN.R-project.org/package=alr4.

Zeileis A, Kleiber C, Jackman S (2008) Regression models for count data in R. J Stat Software 27:1-25.