MAE5763 - Modelos Lineares Generalizados
2º semestre 2023

Prof. Gilberto A. Paula

2ª Lista de Exercícios

1. Supor $Y_i \overset{\text{iid}}{\sim} \text{NI}(\mu, \phi)$, para $i = 1, \ldots, n$. Mostre que a estatística do teste da razão de verossimilhanças para testar $\text{H}_0 : \phi = 1$ contra $\text{H}_1 : \phi \neq 1$ pode ser expressa na forma

$$\xi_{RV} = n(\hat{\phi}^{-1} - 1) + n\log(\hat{\phi}),$$

e mostre que $\hat{\phi} = n/\text{D}(\mathbf{y}; \hat{\boldsymbol{\mu}})$ é a estimativa de máxima verossimilhança de ϕ. Qual a distribuição nula assintótica da estatística do teste?

2. Supor $Y_{ij} \overset{\text{ind}}{\sim} \text{FE}(\mu, \phi_i)$ (normal, gama e normal inversa), para $i = 1, 2, 3$ e $j = 1, \ldots, m$ e com parte sistemática dada por $\sqrt{\phi_1} = \lambda_1 = \alpha - \Delta$, $\sqrt{\phi_2} = \lambda_2 = \alpha$ e $\sqrt{\phi_3} = \lambda_3 = \alpha + \Delta$. Como ficam as matrizes $\mathbf{Z}$ e $\mathbf{P}$? Obter a função escore U_Δ e a variância assintótica $\text{Var}(\hat{\Delta})$. Mostre que a estatística do teste de escore para testar $\text{H}_0 : \Delta = 0$ contra $\text{H}_1 : \Delta \neq 0$ pode ser expressa na forma

$$\xi_{SR} = \frac{m(\hat{t}_3^0 - \hat{t}_1^0)^2}{-2d''(\hat{\phi}^0)},$$

em que $\hat{t}_1^0$ e $\hat{t}_3^0$ são avaliados na hipótese nula. Qual a distribuição nula assintótica da estatística do teste?

3. Supor $Y_{ij} \sim \text{Q}(y_{ij}; \mu)$, em que $\text{E}(Y_{ij}) = \mu$ e $\text{Var}(Y_{ij}) = \sigma^2\mu(1 + \mu)$ com parte sistemática dada por $\log(\mu) = \eta = \beta$, $\mu > 0$ e $\text{R}_i(\alpha)$ simétrica para $i = 1, \ldots, m$ e $j = 1, 2$. Obter $\hat{\beta}_G$ e $\text{Var}(\hat{\beta}_G) = H_1^{-1}(\beta)$. Mostre

que a estatística do teste de Wald para testar $\mathrm{H}_0 : \beta = 1$ contra $\mathrm{H}_1 : \beta \neq 1$ pode ser expressa na forma

$$\xi_W = \frac{2n\bar{y}\{\log(\bar{y}) - 1\}^2}{\hat{\sigma}^2(1+\bar{y})(1+\hat{\alpha})}.$$

Descreva $\hat{\sigma}^2$ e $\hat{\alpha}$. Qual a distribuição nula assintótica da estatística do teste?

4. Considere o modelo aditivo $Y_i \overset{\text{ind}}{\sim} \mathrm{BN}(\mu_i, \nu)$ com $g(\mu_i) = f(t_i)$, sendo $f(t)$ um spline cúbico, para $i = 1, \ldots, n$. Desenvolva um processo iterativo com penalização para estimar $\mathbf{f} = (f_1, \ldots, f_r)^\top$ e ν, em que $f_j = f(t_j^0)$ com $t_j^0 < t_2^0 < \cdots < t_r^0$ sendo os valores distintos de $t_1, \ldots, t_n$, para $r \leq n$. Obter também as variâncias e covariância assintóticas $\mathrm{Var}(\widehat{\mathbf{f}})$, $\mathrm{Var}(\widehat{\nu})$ e $\mathrm{Cov}(\widehat{\mathbf{f}}, \widehat{\nu})$. Como estimar os graus de liberdade efetivos e o parâmetro de suavização λ?

5. No arquivo **aep** do `gamlss` é descrito um estudo com 1383 pacientes internados no Hospital del Mar de Barcelona nos anos de 1988 e 1990. São descritas as seguintes variáveis: (i) `los` (total de dias internados no hospital), (ii) `noinap` (número de dias inapropriados de internação), (iii) `loglogs` : log(los/10), (iv) `sex` (gênero, 1:masculino ou 0:feminino), (v) `ward` (tipo de internação, 1:tratamento médico, 2:cirurgia ou 3:outro tipo), (vi) `year` (ano da internação, 1988 ou 1990), (vii) `age` (idade do paciente subtraída de 55 anos) e (viii) `y` (vetor de respostas em que na primeira coluna está noinap e na segunda coluna los - noinap).

   Considere como resposta o número de dias inapropriados de internação. Fazer inicialmente uma análise descritiva, por exemplo com tabelas de contingência entre o número de dias de internação contra as variáveis categóricas sex, ward e year. Contruir também o diagrama de dispersão (com tendência) entre prop=noinap/los contra age. Comente.

   Compare os ajustes entre modelos logísticos com respostas binomial e beta-binomial para o número de dias inapropriados de internação. Considere apenas as variáveis explicativas sex, ward, year e age. Para fazer os ajustes use os comandos

   ```
   fit1.aep = gamlss(y ~ sex + ward + year + age, family=BI,
   ```

`data=aep)`

`fit2.aep = gamlss(y ∼ sex + ward + year + age, family=BB,`

`data=aep).`

Para o melhor ajuste tente melhorar o modelo, por exemplo incluindo interações de 1ª ordem. Para o modelo final interpretar os coeficientes estimados através de razões de chances, apresentar os gráficos de resíduos e interpretar o `term.plot`.

6. Considere novamente 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 resposta log(BigMac) e cada uma das variáveis explicativas. Comente.

Ajustar inicialmente um modelo com resposta gama e ligação logarítmica no `gamlss` considerando splines cúbicos para ajustar as variáveis com tendência não linear com a log(resposta). Por exemplo, se forem escolhidas as variáveis `Bread`, `Rice`, `Bus` e `TeachNI` o comnado fica dado por

`fit1.bigmac = gamlss(BigMac` $\sim$ `FoodIndex + Apt + TaxRate + TeachHours + cs(Bread) + cs(Rice) + cs(Bus) + cs(TeachNI), 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)`.

Comente as tendências das variáveis selecionadas através do comando `term.plot(fit2.bigmac)`.

7. No arquivo **aids** do `gamlss` são descritas as seguintes variáveis: (i) `y`, número de casos trimestrais de aids na Inglaterra e País de Gales, (ii) `x`, tempo (em meses) desde janeiro de 1983 e (iii) `qrt`, trimestre para controlar a sazonalidade. O arquivo pode ser disponibilizado diretamente no gamlss através dos comandos

   `require(gamlss)`

   `attach(aids).`

   Fazer inicialmente uma análise descritiva dos dados, tais como densidade e boxplot da variável resposta, diagrama de dispersão (com tendência) entre o tempo e a resposta e boxplots da variável resposta segundo o trimestre.

Inicialmente, comparar o ajuste entre modelos com respostas de Poisson e resposta binomial negativa supondo trimestre como variável categórica (fator) e o tempo de forma quadrática. Posteriormente, comparar modelos com as mesmas respostas, porém substituindo o termo quadrático por um spline cúbico. Escolher o melhor modelo através de análise de resíduos.

Por exemplo, para ajustar um modelo com resposta binomial negativa e spline cúbico para o tempo, aplicar os comandos

```
fit.aids = gamlss(y ~ qrt + cs(x), family=NBI)

summary(fit.aids)

plot(fit.aids)

rqres.plot(fit.aids, howmany=8, type="wp").
```

Para o modelo selecionado interpretar os resultados respondendo se há efeito de trimestre e quantos graus de liberdade foram gastos com o ajuste aditivo. Apresentar também o gráfico da curva ajustada ao longo do tempo.

8. No arquivo **Milk** do `gamlss` são apresentados dados referentes a um experimento longitudinal desenvolvido na Austrália com 79 vacas que foram aleatorizadas segundo três dietas e foi observado semanalmente a quantidade de proteína no leite de cada animal. O objetivo principal do estudo é verificar se há diferenças significativas entre as quantidades médias semanais de proteína sob as três dietas. Os dados estão descritos na seguinte ordem: (i) `protein` (quantidade de proteína), (ii) `Time` (semana), (iii) `Cow` (identificação do animal) e (iv) `Diet` (cevada (barley), cevada+tremoços e tremoços (lupins)).

   Fazer inicialmente uma análise descritiva dos dados, por exemplo apresentando os boxplots, perfis individuais, perfis médios e perfis médios com erros padrão ao longo do tempo para as três dietas (sugestão de códigos em anexo). Comente.

   Para ler o arquivo use os comandos:

   ```
   require(gamlss)

   summary(Milk)
   ```

```
attach(Milk).
```

Escrever o modelo de equações de estimação generalizadas EEG-gama com estruturas de correlação do tipo AR(M) (para M=1,2,3) com dieta como fator e o tempo como variável explicativa contínua até ordem quadrática. Para realizar esses ajustes através da biblioteca `glmtoolbox` usar os comandos.

```
require(glmtoolbox)
```

```
fitM.milk = glmgee(protein ~ Diet + Time + I(Time^2), id = Cow,
family = Gamma("log"), corstr="AR-M-dependent(M)", data=Milk)
```

```
summary(fitM.milk).
```

Comparar os três ajustes fit1.milk, fit2.milk e fit3.milk através da medida de informação QIC, gráfico do resíduo de Pearson contra o valor ajustado e gráfico de sensibilidade (por exemplo influência local).

Sugestão de comandos que podem ser modificados:

```
QIC(fit1.milk, fit2.milk, fit3.milk)
```

```
par(mfrow=c(1,3))
```

```
residuals(fit1.milk,type="pearson",plot.it=TRUE, ylab="Resíduo
de Pearson", xlab="Valor Ajustado")
```

```
residuals(fit2.milk,type="pearson",plot.it=TRUE, ylab="Resíduo
de Pearson", xlab="Valor Ajustado")
```

```
residuals(fit3.milk,type="pearson",plot.it=TRUE, ylab="Resíduo
de Pearson", xlab="Valor Ajustado")
```

```
l1=localInfluence(fit1.milk, type="local", perturbation="cw-clusters",
coefs="Diet", plot.it=FALSE)
```

```
l2=localInfluence(fit2.milk, type="local", perturbation="cw-clusters",
coefs="Diet", plot.it=FALSE)
```

```
l3=localInfluence(fit3.milk, type="local", perturbation="cw-clusters",
coefs="Diet", plot.it=FALSE)
```

```
plot(l1, pch=16, ylim=c(0,0.4), ylab=dmax|", xlab="Índice")
```

```
plot(l2, pch=16, ylim=c(0,0.4), ylab=dmax|", xlab="Índice")
```

```
plot(l3, pch=16, ylim=c(0,0.4), ylab=dmax|", xlab="Índice").
```

Escolher um dos ajustes (use o princípio da parcimônia). Para o ajuste escolhido interpretar os parâmetros estimados. Há diferenças significativas entre as dietas? Apresentar as curvas ajustadas para o valor esperado da quantidade de proteína no leite para cada dieta.

9. Considere o banco de dados **polypharm** da biblioteca `aplore3` do R, em que uma amostra $n = 500$ pacientes é analisada com relação ao uso de medicamentos para o tratamento de doenças mentais. O objetivo principal do estudo é relacionar 11 variáveis explicativas observadas em 7 anos (2002 a 2008) com o uso de medicamentos representada pela variável resposta `polypharmacy` (=0 uso de no máximo 3 medicamentos diferentes, =1 uso de mais de 3 medicamentos diferentes). Considere para análise apenas 6 variáveis explicativas: (i) `mhv4`, número de consultas ambulatoriais relacionadas à saúde mental (0: nenhuma, 1: uma a cinco, 2: seis a quatorze e 3: maior do que quatorze), (ii) `inptmhv3`, número de internações hospitalares relacionadas à saúde mental (0: nenhuma, 1: uma e 2: mais do que uma), (iii) `gender`, gênero (0: Feminino, 1: Masculino), (iv) `urban`, local de residência (0: Urbana, 1: Rural)) e (v) `comorbid`, existência de comorbidade (0: Não, 1: Sim) e (vi) `age`, idade em anos. Inicialmente disponibilizar o arquivo de dados

   ```
   require(aplore3)

   summary(polypharm)

   attach(polypharm).
   ```

   Fazer inicialmente uma análise descritiva dos dados, tais como tabelas de contigência entre a variável resposta e cada variável explicativa categórica e boxplots da idade, para cada ano. Para construir uma tabela de contigência entre `polypharmacy` e `mhv4` use o comando

   ```
   table(polypharmacy,mhv4).
   ```

   A partir dessa tabela calcular as proporções de sucesso para cada nível de `mhv4`, ou seja, as proporções para cada coluna da tabela. Comente se há indícios de diferenças entre essas proporções. Assim, sucessivamente para as demais tabelas.

Agora transforme a variável resposta em variável numérica binária para fins de ajuste, através do comando

```
resp=as.numeric(polypharmacy) - 1.
```

Ajustar equações de estimação generalizadas EEG-Bernoulli com estrutura de correlação do tipo AR(M) (para M=1,2,3) e considere `tage=log(age/100)` como variável explicativa contínua. Ajustar apenas os efeitos principais e utilizar a biblioteca `glmtoolbox`. Sugestão de comandos:

```
require(glmtoolbox)
```

```
fitM.poly = glmgee(resp ~ mhv4 + inptmhv3 + gender + urban
+ comorbid + tage, id = id, family = binomial,
corstr="AR-M-dependent(M)")
```

```
summary(fitM.poly).
```

Comparar os três ajustes fit1.poly, fit2.poly e fit3.poly através da medida de informação QIC, gráfico da distância de Mahalanobis e gráfico de sensibilidade (por exemplo influência local).

Sugestão de comandos que podem ser modificados:

```
QIC(fit1.poly, fit2.poly, fit3.poly)
```

```
r1 = residuals(fit1.poly,type="mahalanobis",plot.it=FALSE)
```

```
r2 = residuals(fit2.poly,type="mahalanobis",plot.it=FALSE)
```

```
r3 = residuals(fit3.poly,type="mahalanobis",plot.it=FALSE)
```

```
par(mfrow=c(1,3))
```

```
plot(r1, pch=16, xlab="Índice", ylab="Distância de Mahalanobis",
ylim=c(0,10))
```

```
plot(r2, pch=16, xlab="Índice", ylab="Distância de Mahalanobis",
ylim=c(0,10))
```

```
plot(r3, pch=16, xlab="Índice", ylab="Distância de Mahalanobis",
ylim=c(0,10))
```

```
l1 = localInfluence(fit1.poly, type="local", perturbation="cw-clusters",
coefs="all", plot.it=FALSE)
```

```
l2 = localInfluence(fit2.poly, type="local", perturbation="cw-clusters",
coefs="all", plot.it=FALSE)
```

```
l3 = localInfluence(fit3.poly, type="local", perturbation="cw-clusters",
coefs="all", plot.it=FALSE)
```

```
plot(l1, pch=16, ylim=c(0,0.3), ylab=dmax|", xlab="Índice")
```

```
plot(l2, pch=16, ylim=c(0,0.3), ylab=dmax|", xlab="Índice")
```

```
plot(l3, pch=16, ylim=c(0,0.3), ylab=dmax|", xlab="Índice").
```

Escolher um dos ajustes (use o princípio da parcimônia). Para o modelo escolhido selecione as variáveis explicativas através de um critério `backward` com valor-p de 10% de entrada e saída de varíaveis em cada passo. Use comandos do tipo:

```
stepCriterion(fit2.poly,direction="backward",criterion="p-value",
test="wald", levels=c(0.10,0.10)).
```

Para o modelo final, obter as estimativas intervalares de 95% para as razões de chances das variáveis selecionadas. Comente.