Modelos Lineares Generalizados - Análise de Resíduos

Erica Castilho Rodrigues

21 de Junho de 2017

Tipos de Resíduos

- Assim como em Regressão Linear, também precisamos fazer Análise de Resíduos para os MLG's.
- São semelhantes ao modelo clássico de regressão.
- Porém algumas adaptações são necessárias.
- Para verificar linearidade no modelo linear:
  - usamos os vetores

$$\mathbf{Y} \text{ e } \hat{\mathbf{Y}} .$$

- Para os MLG's devemos usar:
  - o valor estimado do preditor linear $\hat{\eta}$;
  - a variável dependente ajustada

$$\mathbf{z}_i = \sum_{k=1}^{p} x_{ik} b_k^{(m-1)} + (y_i - \mu_i) \left(\frac{\partial \eta_i}{\partial \mu_i} \right) .$$

- A matriz **H** para os modelos de regressão é dada por

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \, .$$

- Para os MLG's ela deve ser reponderada pela matriz **W**.
- Como **W** é definida?
- É uma matriz diagonal cujas entradas são dadas por

$$[\mathbf{W}]_{ii} = \frac{1}{Var(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2 \, .$$

- Para os MLG's devemos substituir

$$\mathbf{X} \text{ por } \mathbf{W}^{1/2}\mathbf{X}\,.$$

- A matriz $\mathbf{H}$ fica

$$\mathbf{H} = \mathbf{W}^{1/2}\mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}^{1/2}\,.$$

Tipos de Resíduos

- Os resíduos são importantes para identificar observações aberrantes.
- O resíduo R_i deve expressar:
  - a distância entre a observação y_i e seu valor ajustado $\hat{\mu}_i$.
- O resíduo é definido como

$$R_i = h_i(y_i, \hat{\mu}_i)$$

onde h_i é uma função escolhida tal que

- facilite a interpretação;
- estabilize a variância;
- induza a simetria na distibuição de R_i.

- Escolhas comuns da função $h_i()$

$$R_i = \frac{(y_i - \hat{\mu}_i)}{[Var(y_i)]^{1/2}} \qquad R_i = \frac{(y_i - \hat{\mu}_i)}{[Var(y_i - \hat{\mu}_i)]^{1/2}}$$

  onde as variâncias são estimadas segundo o modelo sob pesquisa.
- A escolha da função depende do tipo de anomalia que se deseja detectar.

- Anomalias que podem ser dectadas:
  - uma falsa distribuição populacional para a variável resposta;
  - uma ou mais observações não pertencendo hà distribuição proposta para os dados;
  - algumas observações que se mostram dependentes ou exibindo alguma forma de correlação serial;
  - um parâmetro importante que esteja sendo omitido no modelo.
- Temos diferentes tipos de resíduos que podem ser usados.

Resíduos ordinários

- São os tipos mais usados nos MLG's.
- Igual ao resíduo usado em Regressão Linear
- São definidos por

$$r_i = y_i - \hat{y}_i \, .$$

Resíduo de Person

- É dado pelas conponentes da Estatísitca de Pearson Generalizada.
- É definido por

$$r_i^P = \frac{y_i - \hat{y}_i}{V(\hat{Y}_i)} .$$

- Uma desvantagem desse tipo de resíduo é que ele é geralmente assimétrico para distribuições não normais.

Exemplos:

- Consider um modelo tal que

$$Y_i \sim Poisson(\lambda_i) \quad \lambda_i = exp^{\mathbf{x}_i^T \boldsymbol{\beta}} .$$

- O Resíduo de Pearson nesse caso é dado por

$$r_i^P = \frac{y_i - \hat{y}_i}{V(\hat{Y}_i)} = \frac{y_i - \hat{y}_i}{\hat{y}_i} .$$

Componentes do Desvio

- Vimos que a função Deviance é dada por

$$D = 2[l(\mathbf{b}_{max}; \mathbf{y}) - l(\mathbf{b}, \mathbf{y})] = 2 \sum_i [l(\mathbf{b}_{max}; y_i) - l(\mathbf{b}, y_i)] \, .$$

- As componentes do Desvio são dadas por

$$r_{D_i} = sinal(y_i - \hat{y}_i)\sqrt{di}$$

onde

$$d_i = [l(\mathbf{b}_{max}; y_i) - l(\mathbf{b}, y_i)]$$

são cada uma das componentes da Deviance e

$$sinal(y_i - \hat{y}_i)$$

retorna o sinal da diferença $(y_i - \hat{y}_i)$.

- O termo

$$d_i = [l(\mathbf{b}_{max}; y_i) - l(\mathbf{b}, y_i)]$$

é chamado **Deviance Residual**.

- Se d_i é grande, isso significa que:
  - a observação i contribui em excesso para a Função Deviance;
  - isso indica que alguma das hipóteses sobre o modelo está sendo violada.
- Se o modelo é bem ajustado, pode-se mostrar que

$$d_i \approx 1 \ .$$

- Valores tais que

$$d_i >> 1$$

indicam que a i-ésima observação está contribuindo para o mau ajuste do modelo.

- Como $D \sim \chi^2_{n-p}$,

$$d_i \sim \chi^2_{(n-p)/n} \, .$$

- Temos então que distribuição aproximada de d_i é

$$\sqrt{d_i} \sim N(0,1) \, .$$

Exemplo:

- Considere uma amostra aleatória

$$Y_1, Y_2, \ldots, Y_m$$

tal que $Y_i \sim Bin(n_i, \pi_i)$.

- Vimos que a função Deviance é dada por

$$D = 2\left[\sum_i y_i \log\left(\frac{y_i}{\hat{y}_i}\right) - y_i \log\left(\frac{n_i - \hat{y}_i}{n_i - y_i}\right) + n_i \log\left(\frac{n_i - \hat{y}_i}{n_i - y_i}\right)\right].$$

- O Desvio Residual fica

$$d_i = y_i \log\left(\frac{y_i}{\hat{y}_i}\right) - y_i \log\left(\frac{n_i - \hat{y}_i}{n_i - y_i}\right) + n_i \log\left(\frac{n_i - \hat{y}_i}{n_i - y_i}\right).$$

Exemplo: (continuação)

- A Componente do Desvio fica

$$r_{D_i} = sinal(y_i - \hat{y}_i)\sqrt{y_i \log\left(\frac{y_i}{\hat{y}_i}\right) - y_i \log\left(\frac{n_i - \hat{y}_i}{n_i - y_i}\right) + n_i \log\left(\frac{n_i - \hat{y}_i}{n_i - y_i}\right)}$$

- Podemos ainda usar uma forma padronizada das Componentes do Desvio.
- Esse tipo de resíduo é chamado Desvio Studentizado

$$r_{D_i'} = \frac{r_{D_i}}{\sqrt{1 - h_{ii}}}$$

onde h_{ii} é o termo da diagonal da matriz **H**.

- O R calcula os r_{D_i} automaticamente.
- Basta então encontrar a matriz **H** e padronizá-los.

- O script a seguir calcula as Componentes do Desvio Padronizadas

```
# Pega a matriz X
X <- model.matrix(fit.model)
# Defini a matriz W
w <- fit.model$weights
W <- diag(w)
```

```
# Calcula a matriz H
H < - solve(t(X)%*%W%*%X)
H < - sqrt(W)%*%X%*%H%*%t(X)%*%sqrt(W)
# Pega somente os termos da diagonal
h < - diag(H)
# Retorna o desvio residual
rd < - resid(fit.model, type= ''deviance'')
# Padroniza as componentes do desvio
td < - rd*sqrt(1/(1-h))
```

- Para verificar se o modelo está bem ajustado podemos plotar os valores $r_{D_i'}$.
- Esses valores, porém, não seguem uma distribuição específica.
- Uma solução: encontrar a distribuição empírica desses resíduos.
- Ajustamos o modelo.
- Encontramos as estimativas dos parâmetros.

- Geramos observações do modelo ajustado.
- Temos então certeza que elas vieram do modelo ajustado.
- Ajustamos um modelo para essas variávei simuladas.
- Encontramos os Desvios Padronizados.
- Fazemos isso várias vezes.
- Teremos várias amostras de resíduos de modelos perfeitos (observações vêm de fato dele).
- Se os resíduos que observamos for muito discrepantes desses gerados, o nosso modelo não está bem ajustado.
- Vejamos o script que roda esse envelope. (Prof. Gilberto Paula)

Caso Binomial

```
# Define a matriz X
X <- model.matrix(fit.model)
n <- nrow(X)
p <- ncol(X)
# Calcula a matriz W
w <- fit.model$weights
W <- diag(w)
# Calcula a matriz H
H <- solve(t(X)%*%W%*%X)
H <- sqrt(W)%*%X%*%H%*%t(X)%*%sqrt(W)
h <- diag(H)
# Encontra as deviance padronizadas
td <- resid(fit.model,type="deviance")/sqrt(1-h)
# Essa matriz guarda os resíduos obtidos
# dos modelos gerados
```

```
e <- matrix(0,n,100)
#
for(i in 1:100){
# Gera os valores da Binomial
nresp <- rbinom(1, fitted(fit.model))
# Ajusta o modelo para os valores gerados
fit <- glm(nresp ~ X, family=binomial)
w <- fit$weights
W <- diag(w)
H <- solve(t(X)%*%W%*%X)
H <- sqrt(W)%*%X%*%H%*%t(X)%*%sqrt(W)
h <- diag(H)
e[,i] <- sort(resid(fit,type="deviance")/sqrt(1-h))
#
e1 <- numeric(n)
e2 <- numeric(n)
#
```

```
for(i in 1:n){
  eo <- sort(e[i,])
e1[i] <- (eo[2]+eo[3])/2
e2[i] <- (eo[97]+eo[98])/2}
#
med <- apply(e,1,mean)
faixa <- range(td,e1,e2)
par(pty="s")
qqnorm(td,xlab="Percentil da N(0,1)",
ylab="Componente do Desvio", ylim=faixa, pch=16)
#
par(new=T)
#
qqnorm(e1,axes=F,xlab="",ylab="",type="l",
ylim=faixa,lty=1)
par(new=T)
qqnorm(e2,axes=F,xlab="",ylab="", type="l",
```

```
ylim=faixa,lty=1)
par(new=T)
qqnorm(med,axes=F,xlab="", ylab="", type="l",
ylim=faixa,lty=2)
```

- Podemos fazer ainda gráficos semelhantes aos que eram feitos em Regressão Linear.
- São feitas algumas modificações.

Resíduos versus alguma função dos valores ajustados

- Esse gráfico consiste em plotar:
  - o desvio studentizado $r_{D_i'}$;
  - e o valor ajustado do preditor linear $\hat{\eta}_i$.
- Os resíduos devem ficar distribuídos aleatoriamente em torno de zero e com amplitude constante.

Resíduos versus varáveis explicativas não incluídas

- Igual ao gráfico de regressão linear.
- Serve para verificar se existe relação entre os resíduos e uma variável não incluída no modelo.
- Os resíduos devem ficar distribuidos aleatoriamente em torno de zero.
- Se esse gráfico apresentar algum padrão:
  - significa que a variável deve ser incluída no modelo

Resíduos versus varáveis explicativas não incluídas

- Igual ao gráfico de regressão linear.
- Serve para verificar se existe relação entre os resíduos e uma variável não incluída no modelo.
- Os resíduos devem ficar distribuidos aleatoriamente em torno de zero.
- Se esse gráfico apresentar algum padrão:
  - significa que a variável deve ser incluída no modelo

Resíduos versus varáveis explicativas incluídas

- Igual ao gráfico de regressão linear.
- Serve para verificar se existe relação entre os resíduos e uma variável incluída no modelo.
- Os resíduos devem ficar distribuidos aleatoriamente em torno de zero.