CE225 - Modelos Lineares Generalizados

Cesar Augusto Taconeli

11 de setembro, 2018

Aula 6 - Estimação em modelos lineares generalizados

Estimação em modelos lineares generalizados

- Seja y_i uma única observação de uma distribuição na forma:

$$f(y_i;\theta_i,\phi) = exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i;\phi)\right\}. \quad (1)$$

- A log-verossimilhança correspondente a essa observação fica dada por:

$$l_i = l(\theta_i;\phi,y_i) = log\left[f(y_i;\theta_i,\phi)\right] = \frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i;\phi). \quad (2)$$

- Considerando n observações independentes $y_1, y_2, ..., y_n$, a log-verossimilhança fica dada por:

$$l(\boldsymbol{\theta};\phi,\boldsymbol{y}) = \sum_{i=1}^{n} l_i = \sum_{i=1}^{n} l(\theta_i;\phi,y_i) = \sum_{i=1}^{n}\left[\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i;\phi)\right]. \quad (3)$$

Estimação em modelos lineares generalizados

- Considere um modelo linear generalizado com função de ligação $g(\cdot)$ e preditor linear $\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p$:

$$g(\mu) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p. \tag{4}$$

- A estimação de $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_p)'$ por máxima verossimilhança baseia-se na determinação de $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_p)'$ tal que:

$$\begin{aligned}
&\frac{\partial l(\boldsymbol{\beta})}{\partial \beta_0}\Big|_{\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_p} = 0 \\
&\frac{\partial l(\boldsymbol{\beta})}{\partial \beta_1}\Big|_{\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_p} = 0 \\
&\vdots \\
&\frac{\partial l(\boldsymbol{\beta})}{\partial \beta_p}\Big|_{\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_p} = 0
\end{aligned} \tag{5}$$

Estimação em modelos lineares generalizados

- Como $y_1, y_2, ..., y_n$ são independentes:

$$\frac{\partial l(\boldsymbol{\beta})}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial l_i(\boldsymbol{\beta})}{\partial \beta_j} = 0, \tag{6}$$

para todo j.

- Observe que estamos denotando a log-verossimilhança por $l(\boldsymbol{\beta})$ uma vez que $\mu_i = b'(\theta_i)$; $g(\mu_i) = \eta_i$ e $\eta_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}$.
- Assim, pela regra da cadeia:

$$\frac{\partial l_i}{\partial \beta_j} = \frac{\partial l_i}{\partial \theta_i} \times \frac{\partial \theta_i}{\partial \mu_i} \times \frac{\partial \mu_i}{\partial \eta_i} \times \frac{\partial \eta_i}{\partial \beta_j}, \tag{7}$$

para $j = 0, 1, 2, ..., p$.

Estimação em modelos lineares generalizados

- Usando a definição e as propriedades de modelos lineares generalizados:

$$\frac{\partial l_i}{\partial \beta_j} = \frac{(y_i - \mu_i)}{var(y_i)} \times \frac{\partial \mu_i}{\partial \eta_i} \times x_{ij}. \tag{8}$$

- Somando para as n observações, as equações de log-verossimilhança ficam dadas por:

$$\frac{\partial l(\beta)}{\partial \beta_j} = \sum_{i=1}^{n} \frac{(y_i - \mu_i)}{var(y_i)} \times \frac{\partial \mu_i}{\partial \eta_i} \times x_{ij}, \quad j = 0, 1, 2, ..., p, \tag{9}$$

onde $\eta_i = \sum_{j=0}^{p} \beta_j x_{ij} = g(\mu_i)$.

Estimação em modelos lineares generalizados

- Uma forma equivalente de escrever as equações de log-verossimilhança é a seguinte:

$$\boldsymbol{X}'\boldsymbol{D}\boldsymbol{V}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})=\boldsymbol{0}, \tag{10}$$

em que $\boldsymbol{V}$ é a matriz diagonal das variâncias das observações; $\boldsymbol{X}$ é a matriz do modelo; $\boldsymbol{D}$ é a matriz diagonal com entradas $\partial\mu_i/\partial\eta_i$ e $\boldsymbol{y}$ e $\boldsymbol{\mu}$ são os vetores de observações e de médias, respectivamente.

- As equações de log-verossimilhança são funções não lineares dos $\beta's$.
- Assim, a determinação das estimativas de máxima verossimilhança requer o uso de métodos iterativos. Vamos discutir mais adiante dois desses métodos: o **método de Newton-Raphson** e o **método Score de Fisher**.

Distribuição assintótica dos estimadores dos parâmetros de um MLG

- Os estimadores de máxima verossimilhança dos parâmetros de um MLG atendem às propriedades gerais de estimadores de máxima verossimilhança.
- Assim, assintoticamente:

$$\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \boldsymbol{\jmath}^{-1}), \tag{11}$$

em que $\boldsymbol{\jmath}$ é a matriz informação de Fisher (ou matriz informação esperada), com entradas $-E(\partial^2 l(\boldsymbol{\beta})/\partial\beta_r\partial\beta_s)$.

Distribuição assintótica dos estimadores dos parâmetros de um MLG

- Usando o fato que, sob condições de regularidade atendidas pela família exponencial de distribuições:

$$E\left(-\frac{\partial^2 l(\beta)}{\partial\beta_r\partial\beta_s}\right) = E\left[\left(\frac{\partial l(\beta)}{\partial\beta_r}\right)\left(\frac{\partial l(\beta)}{\partial\beta_s}\right)\right] \tag{12}$$

chega-se a:

$$E\left(-\frac{\partial^2 l(\beta)}{\partial\beta_r\partial\beta_s}\right) = \sum_{i=1}^{n}\frac{x_{ir}x_{is}}{var(y_i)}\left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2. \tag{13}$$

Distribuição assintótica dos estimadores dos parâmetros de um MLG

- Seja $\boldsymbol{W}$ a matriz diagonal com elementos:

$$\omega_i = \frac{(\partial \mu_i / \partial \eta_i)^2}{var(y_i)}. \tag{14}$$

- Então, generalizando para toda a matriz de informação, temos:

$$\jmath = \boldsymbol{X}'\boldsymbol{W}\boldsymbol{X}, \tag{15}$$

em que $\boldsymbol{X}$ é a matriz do modelo. A matriz $\jmath$ depende da função de ligação, uma vez que $\partial \eta_i / \partial \mu_i = g'(\mu_i)$.

Distribuição assintótica dos estimadores dos parâmetros de um MLG

- Assim, a distribuição assintótica de $\hat{\beta}$ é dada por:

$$\hat{\beta} \sim N(\beta, (\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}). \tag{16}$$

- A matriz de covariância assintótica é estimada por

$$\widehat{var}(\hat{\beta}) = (\boldsymbol{X}'\hat{\boldsymbol{W}}\boldsymbol{X})^{-1}, \tag{17}$$

sendo que $\hat{\boldsymbol{W}}$ é $\boldsymbol{W}$ avaliado em $\hat{\beta}$.

Método de Newton-Raphson

- O método de Newton-Raphson é aplicado na solução de equações não lineares (no caso, na determinação do ponto em que a função assume seu máximo);
- O método inicia com um valor incial como primeira aproximação para a solução;
- Na sequência, uma segunda aproximação é obtida aproximando a função, na vizinhança do valor inicial, por um polinômio de segundo grau, e encontrando o ponto de máximo do polinômio.
- Após a repetição de uma sequência de aproximações, o processo converge para a locação do máximo se a função é bem comportada e a aproximação inicial é boa.

Método de Newton-Raphson

- Formalizando: desejamos determinar $\hat{\beta}$ que maximiza $L(\beta)$. Seja:

$$\boldsymbol{S} = \left(\frac{\partial l(\beta)}{\partial \beta_0}, \frac{\partial l(\beta)}{\partial \beta_1}, ..., \frac{\partial l(\beta)}{\partial \beta_p} \right)' . \tag{18}$$

- Seja $\boldsymbol{H}$ a matriz hessiana, definida pelas deriavdas parciais de segunda ordem:

$$\boldsymbol{H} = \begin{bmatrix} \frac{\partial^2 l(\beta)}{\partial \beta_0 \partial \beta_0} & \frac{\partial^2 l(\beta)}{\partial \beta_0 \partial \beta_1} & \cdots & \frac{\partial^2 l(\beta)}{\partial \beta_0 \partial \beta_p} \\ \frac{\partial^2 l(\beta)}{\partial \beta_0 \partial \beta_1} & \frac{\partial^2 l(\beta)}{\partial \beta_1 \partial \beta_1} & \cdots & \frac{\partial^2 l(\beta)}{\partial \beta_1 \partial \beta_p} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 l(\beta)}{\partial \beta_p \partial \beta_0} & \frac{\partial^2 l(\beta)}{\partial \beta_p \partial \beta_1} & \cdots & \frac{\partial^2 l(\beta)}{\partial \beta_p \partial \beta_p} \end{bmatrix}$$

Método de Newton-Raphson

- Sejam $\boldsymbol{S}^{(t)}$ e $\boldsymbol{H}^{(t)}$, respectivamente, $\boldsymbol{S}$ e $\boldsymbol{H}$ avaliados em $\beta^{(t)}$, a aproximação no passo t para $\hat{\beta}$.

- O método de Newton-Raphson aproxima $l(\beta)$ em torno de $\beta^{(t)}$ por meio de uma expansão em série de Taylor de segunda ordem:

$$l(\beta) \approx l(\beta^{(t)}) + \boldsymbol{S}^{(t)'}(\beta - \beta^{(t)}) + \left(\frac{1}{2}\right)(\beta - \beta^{(t)})'\boldsymbol{H}^{(t)}(\beta - \beta^{(t)}) \quad (19)$$

Método de Newton-Raphson

- Resolvendo $\partial L(\beta)/\partial\beta \approx \boldsymbol{S}^{(t)} + \boldsymbol{H}^{(t)}(\beta - \boldsymbol{\beta^{(t)}}) = \mathbf{0}$ para β, temos como aproximação para $\hat{\beta}$ no passo $t+1$:

$$\beta^{(t+1)} = \beta^{(t)} - (\boldsymbol{H}^{(t)})^{-1}\boldsymbol{S}^{(t)}, \tag{20}$$

assumindo que $\boldsymbol{H}^{(t)}$ é não singular.

- As iterações continuam até que a mudança em $\beta^{(t)}$ em passos sucessivos seja suficientemente pequena (convergência).

Método Score de Fisher

- A diferença do método Score de Fisher para o método de Newton Raphson é que o primeiro usa o valor esperado da matriz Hessiana, que é a matriz de *informação esperada*, enquanto o segundo usa a própria hessiana, que é a matriz de *informação observada*.

- No caso de MLGs com função de ligação canônica, os procedimentos baseados nas matrizes de informação observada e esperada são equivalentes.

- Seja $\mathcal{J}^{(t)}$ a aproximação no passo t para a matriz de informação esperada, com entradas $-E(\partial^2 l(\beta)/\partial\beta_r\partial\beta_s)$ avaliado em $\beta^{(t)}$. O algoritmo score de Fisher fica dado por:

$$\beta^{(t+1)} = \beta^{(t)} - (\mathcal{J}^{(t)})^{-1}\boldsymbol{S}^{(t)}. \tag{21}$$

Máxima verossimilhança e método de mínimos quadrados ponderados iterativamente

- Existe uma relação entre o algoritmo Score de Fisher, aplicado na estimação de máxima verossimilhança dos parâmetros de um MLG, e o método de mínimos quadrados ponderados.
- Após algumas passagens, as equações do algoritmo Score de Fisher podem ser expressas na seguinte forma:

$$(\boldsymbol{X}'\boldsymbol{W}^{(t)}\boldsymbol{X})\beta^{(t+1)} = \boldsymbol{X}'\boldsymbol{W}^{(t)}\boldsymbol{z}^{(t)}, \tag{22}$$

em que $\boldsymbol{z}^{(t)}$ é o vetor da variável dependente ajustada, dado por:

$$z_i^{(t)} = \sum_j x_{ij}\beta_j^{(t)} + (y_i - \mu_i^{(t)})\frac{\partial \eta_i^{(t)}}{\partial \mu_i^{(t)}} = \eta_i^{(t)} + (y_i - \mu_i^{(t)})\frac{\partial \eta_i^{(t)}}{\partial \mu_i^{(t)}}. \tag{23}$$

Máxima verossimilhança e método de mínimos quadrados ponderados iterativamente

- Isolando $\beta^{(t+1)}$:

$$\beta^{(t+1)} = (\boldsymbol{X}'\boldsymbol{W}^{(t)}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}^{(t)}\boldsymbol{z}^{(t)} \tag{24}$$

- O vetor $\boldsymbol{z}^{(t)}$ é uma forma linearizada da função de ligação g avaliada em $\boldsymbol{y}$:

$$g(y_i) \approx g(\mu_i^{(t)}) + (y_i - \mu_i^{(t)})g'(\mu_i^{(t)}) = \eta_i^{(t)} + (y_i - \mu_i^{(t)})\frac{\partial \eta_i^{(t)}}{\partial \mu_i^{(t)}}. \tag{25}$$

Máxima verossimilhança e método de mínimos quadrados ponderados iterativamente

- Desta forma, a estimação por máxima verossimilhança, em modelos lineares generalizados, requer apenas recursos computacionais que produzam estimativas de mínimos quadrados ponderados.

Tabela 1: Variável dependente ajustada (z) e pesos (ω) para alguns MLGs

Distribuição	Ligação	z	ω
Normal	Identidade	y	1
Binomial(m, μ)	Logito	$\eta + (y-\mu)/m\mu(1-\mu)$	$m\mu(1-\mu)$
Gamma	Inversa	$\eta - (y-\mu)/\mu^2$	μ^2
Gamma	Log	$\eta + (y-\mu)/\mu$	1
Poisson	Log	$\eta + (y-\mu)/\mu$	μ
Normal inversa	μ^{-2}	$\eta - 2(y-\mu)/\mu^3$	μ^3

Máxima verossimilhança e método de mínimos quadrados ponderados iterativamente

- O algoritmo de mínimos quadrados ponderados iterativamente para o ajuste por máxima verossimilhança de modelos lineares generalizados fica definido da seguinte forma:

1. Defina valores iniciais para μ_i e calcule $\eta_i = g_{(\mu_i)}$, $i = 1, 2, ..., n$, denotados por $\mu_i^{(0)}$ e $\eta_i^{(0)}$. Uma escolha simples é $\mu_i^{(0)} = y_i$ e $\eta_i^{(0)} = g(y_i)$;
2. Calcule os elementos do vetor $\boldsymbol{z}$ e a matriz $\boldsymbol{W}$ com base nos valores de μ_i atribuídos no passo anterior:

$$z_i^{(0)} = \eta_i^{(0)} + (y_i - \mu_i^{(0)}) \times g'(\mu_i^{(0)}); \tag{26}$$

$$\omega_{ii} = \frac{1}{V(\mu_i^{(0)}) \times [g'(\mu_i^{(0)})]^2}. \tag{27}$$

Máxima verossimilhança e método de mínimos quadrados ponderados iterativamente

3. Calcular a aproximação de β no próximo passo por:

$$\beta^{(1)} = (\boldsymbol{X}'\boldsymbol{W}^{(0)}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}^{(0)}\boldsymbol{z}^{(0)} \tag{28}$$

4. Repetir os passos 2 e 3, com a aproximação atual de β, até verificar convergência.

- Um critério de convergência que pode ser considerado é o seguinte:

$$\sum_{j=0}^{p}\left(\frac{\beta_j^{(t)} - \beta_j^{(t-1)}}{\beta_j^{(t-1)}}\right)^2. \tag{29}$$

Estimação em modelos lineares generalizados

- Uma vez obtidos os estimadores dos parâmetros de um modelo linear generalizado, o ajuste pode ser apresentado na escala do preditor:

$$g(\hat{\mu}) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + ... + \hat{\beta}_p x_p, \tag{30}$$

ou na escala da média (denominaremos de escala da resposta ao longo do curso):

$$\hat{\mu} = g^{-1}(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + ... + \hat{\beta}_p x_p). \tag{31}$$

Estimação em modelos lineares generalizados

Exercício 1

Retome o Exemplo 4 da primeira aula (referente ao acasalamento de elefantes). Os dados estão disponíveis no pacote ***Sleuth2*** *(sob o nome* ***case2201****). Vamos considerar um modelo de regressão Poisson com função de ligação logarítmica (canônica). Ajuste o MLG programando o algoritmo de estimação, maximizando a log- verossimilhança usando um otimizador do R e via função glm, conforme visto em aula.*

Estimação do parâmetro de dispersão

- Nas situações em que ϕ é desconhecido, precisamos estimá-lo para avaliação dos erros das estimativas, construção de intervalos de confiança...
- Um estimador consistente para ϕ baseia-se na estatística X^2 de Pearson:

$$X^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)}, \tag{32}$$

onde $V(\mu)$ é a função de variância, sendo definido por:

$$\hat{\phi} = \frac{X^2}{n-p}. \tag{33}$$

Robustez dos MLG's quanto à especificação incorreta do modelo

- Os estimadores dos parâmetros de modelos lineares generalizados são consistentes ainda que a distribuição especificada esteja incorreta, mas desde que a especificação do preditor linear e da função de ligação esteja correta;

- Entretanto, ao assumir uma distribuição incorreta, a função de variância também estará errada, de forma que $Var(\hat{\beta})$ (e os resultados subsequentes) estarão incorretos;

- Estudaremos adiante como contornar os problemas decorrentes da especificação incorreta da função de variância sem precisar, para isso, assumir um particular modelo probabilístico (abordagem de quasi-verossimilhança).