GET00211 - Modelos Lineares 2

Rafael Erbisti

Instituto de Matemática e Estatística
Universidade Federal Fluminense

Aula 03

Propriedades

Se a fdp é escrita na forma

$$f_Y(y; \theta, \phi) = \exp\{\phi[y\theta - b(\theta)] + c(y, \phi)\}\mathcal{I}_A(y), \tag{1}$$

então

$$E(Y) = \mu = b'(\theta), \quad V(Y) = \phi^{-1}V(\mu),$$

onde a função de variância é

$$V(\mu) = \frac{d\mu}{d\theta} = b''(\theta) = \frac{\partial^2 b(\theta)}{\partial \theta^2}.$$

- Vamos provar as duas propriedades acima, para o caso contínuo. A prova para o caso discreto é análoga, substituindo-se as integrais por somatórios.

Propriedades

Sob certas condições de regularidade, temos que:

$$E\left[\frac{\partial \log f_Y(Y)}{\partial \theta}\right] = 0 \quad \text{e} \quad E\left[\frac{\partial^2 \log f_Y(Y)}{\partial \theta^2}\right] = -E\left[\left(\frac{\partial \log f_Y(Y)}{\partial \theta}\right)^2\right].$$

Na primeira igualdade, temos:

$$\begin{aligned} E\left[\frac{\partial \log f_Y(Y)}{\partial \theta}\right] &= \int_A \frac{\partial \log f_Y(y)}{\partial \theta} f_Y(y)\,dy \\ &= \int_A \frac{\partial f_Y(y)}{\partial \theta} \cdot \frac{1}{f_Y(y)} \cdot f_Y(y)\,dy \\ &= \int_A \frac{\partial f_Y(y)}{\partial \theta}\,dy \overset{*}{=} \frac{\partial}{\partial \theta} \int_A f_Y(y)\,dy \\ &= \frac{\partial}{\partial \theta}(1) = 0. \end{aligned} \tag{2}$$

* sob as condições de regularidade.

Por outro lado, note que:

$$
\begin{aligned}
\frac{\partial^2 \log f_Y(y)}{\partial \theta^2} &= \frac{\partial}{\partial \theta}\left(\frac{\partial \log f_Y(y)}{\partial \theta}\right) \\
&= \frac{\partial}{\partial \theta}\left(\frac{1}{f_Y(y)} \cdot \frac{\partial f_Y(y)}{\partial \theta}\right) \quad \left(\text{pois } \frac{\partial \log f_Y}{\partial \theta} = \frac{f_Y'}{f_Y}\right) \\
&= \frac{\dfrac{\partial^2 f_Y(y)}{\partial \theta^2} \cdot f_Y(y) - \left(\dfrac{\partial f_Y(y)}{\partial \theta}\right)^2}{f_Y(y)^2} \\
&= \frac{\partial^2 f_Y(y)}{\partial \theta^2} \cdot \frac{1}{f_Y(y)} - \left(\frac{\partial f_Y(y)}{\partial \theta}\right)^2 \cdot \frac{1}{f_Y(y)^2} \\
&= \frac{\partial^2 f_Y(y)}{\partial \theta^2} \cdot \frac{1}{f_Y(y)} - \left(\frac{\partial \log f_Y(y)}{\partial \theta}\right)^2 \qquad (3)
\end{aligned}
$$

Propriedades

Na segunda igualdade, usando o resultado em (3), temos:

$$\begin{aligned}
E\left[\frac{\partial^2 \log f_Y(Y)}{\partial\theta^2}\right] &= \int_A \frac{\partial^2 \log f_Y(y)}{\partial\theta^2}\, f_Y(y)\,dy \\
&= \int_A \left\{\frac{\partial^2 f_Y(y)}{\partial\theta^2}\cdot\frac{1}{f_Y(y)} - \left(\frac{\partial \log f_Y(y)}{\partial\theta}\right)^2\right\} f_Y(y)\,dy \\
&= \int_A \frac{\partial^2 f_Y(y)}{\partial\theta^2}\,dy - \int_A \left(\frac{\partial \log f_Y(y)}{\partial\theta}\right)^2 f_Y(y)\,dy \\
&\overset{*}{=} \frac{\partial^2}{\partial\theta^2}\int_A f_Y(y)\,dy - E\left[\left(\frac{\partial \log f_Y(Y)}{\partial\theta}\right)^2\right] \\
&= \frac{\partial^2}{\partial\theta^2}(1) - E\left[\left(\frac{\partial \log f_Y(Y)}{\partial\theta}\right)^2\right] \\
&= -E\left[\left(\frac{\partial \log f_Y(Y)}{\partial\theta}\right)^2\right]
\end{aligned} \tag{4}$$

Propriedades

Por (2), temos

$$V\left(\frac{\partial \log f_Y(Y)}{\partial\theta}\right) = E\left[\left(\frac{\partial \log f_Y(Y)}{\partial\theta}\right)^2\right] - \left(E\left[\frac{\partial \log f_Y(Y)}{\partial\theta}\right]\right)^2$$
$$= E\left[\left(\frac{\partial \log f_Y(Y)}{\partial\theta}\right)^2\right] = -E\left[\frac{\partial^2 \log f_Y(Y)}{\partial\theta^2}\right] \qquad (5)$$

Agora, para a família exponencial na forma canônica

$$f_Y(y;\theta,\phi) = \exp\{\phi[y\theta - b(\theta)] + c(y,\phi)\},$$

temos

$$\frac{\partial \log f_Y(y)}{\partial\theta} = \phi\,[\,y - b'(\theta)\,]$$
$$\frac{\partial^2 \log f_Y(y)}{\partial\theta^2} = -\phi\, b''(\theta)$$

Propriedades

Portanto, tomando esperança

$$
\begin{aligned}
E\left[\frac{\partial \log f_Y(Y)}{\partial \theta}\right] &= E\left[\phi\,(Y - b'(\theta))\right] = 0 \\
&\Rightarrow\ \phi\left[E(Y) - b'(\theta)\right] = 0 \\
&\Rightarrow\ E(Y) = b'(\theta). \qquad (5)
\end{aligned}
$$

Usando a identidade encontrada,

$$
V\left(\frac{\partial \log f_Y(Y)}{\partial \theta}\right) = -\,E\left[\frac{\partial^2 \log f_Y(Y)}{\partial \theta^2}\right],
$$

e substituindo as expressões da família exponencial, temos

$$
\begin{aligned}
V\left(\phi\,[Y - b'(\theta)]\right) &= -E[-\phi\, b''(\theta)] \\
\phi^2\, V(Y) &= \phi\, b''(\theta) \\
\Rightarrow \quad V(Y) &= b''(\theta)\,\phi^{-1}. \qquad (6)
\end{aligned}
$$

Estrutura do MLG

Na construção de um MLG, devemos responder três perguntas fundamentais:

1. Qual é a distribuição dos dados?
   O vetor $\mathbf{y} = (y_1, \ldots, y_n)'$, geralmente, consiste em uma amostra aleatória proveniente de uma distribuição de probabilidade da **família exponencial**.

2. Quais são os preditores?
   É o vetor de variáveis explicativas $\mathbf{x}_i$, $i = 1, \ldots, n$, associado ao vetor de coeficientes $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$, que possivelmente influenciam à variável resposta y_i.

3. Qual é a função da média que será modelada como função linear dos preditores?
   Aqui definimos a **função de ligação** $g(\mu_i)$, tal que

$$g(E[Y_i]) = \eta_i = \mathbf{x}_i'\boldsymbol{\beta},$$

   onde η_i é o preditor linear para a observação i, $i = 1, \ldots, n$ e $g(\cdot)$ é uma função monótona e diferenciável.

Notação

- **y**: vetor coluna $n \times 1$ de observações

$$\mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

- $\mathbf{y}' = (y_1, \ldots, y_n)$: vetor linha $1 \times n$ de observações.
- x_{ij}: elemento da matriz $\mathbf{X}$, onde i indica a observação e j a variável explicativa
- $\mathbf{x}_i = \begin{pmatrix} x_{i1} \\ \vdots \\ x_{ip} \end{pmatrix}$: vetor coluna $p \times 1$ de variáveis explicativas para a i-ésima observação.
- $\mathbf{x}_i' = (x_{i1}, \ldots, x_{ip})$: vetor linha $1 \times p$ de variáveis explicativas.
- $\boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}$: vetor $p \times 1$ de parâmetros.

Descrição do MLG

Um MLG possui três componentes:

1. **Componente aleatório:**
   Variáveis resposta $Y_1, \ldots, Y_n$, que possuem a mesma forma de distribuição pertencente à família exponencial.

2. **Componente sistemático:**
   Conjunto de parâmetros $\boldsymbol{\beta}$ e variáveis explicativas (preditor linear)

$$\eta_i = \mathbf{x}_i' \boldsymbol{\beta}, \qquad \mathbf{X} = \begin{pmatrix} \mathbf{x}_1' \\ \vdots \\ \mathbf{x}_n' \end{pmatrix} = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix}.$$

3. **Função de ligação:**
   Uma função monótona $g(\cdot)$ tal que

$$g(\mu_i) = \eta_i = \mathbf{x}_i' \boldsymbol{\beta}, \quad \text{onde } \mu_i = E(Y_i).$$

Exemplo 1: Modelo Linear Normal

Um caso particular de um MLG é o modelo linear normal, usualmente escrito como

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

onde $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)', \epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2), i = 1, \ldots, n$.

Este modelo é um MLG, pois possui as três componentes:

1. Componente aleatório: temos n variáveis aleatórias independentes com distribuição normal, que pertence à família exponencial.

2. Componente sistemático: o preditor linear é dado por

$$\eta_i = \mathbf{x}_i'\boldsymbol{\beta}.$$

3. Função de ligação: a função de ligação é a identidade,

$$g(\mu_i) = \mu_i = \eta_i = \mathbf{x}_i'\boldsymbol{\beta},$$

pois $E(Y_i) = \mu_i = \eta_i$.

Exemplo 2: Taxa de mortalidade

Suponha que mortes por doenças não contagiosas sejam eventos independentes.

- O número de mortes Y por uma dada doença pode ser modelado por uma distribuição de Poisson:

$$f_Y(y;\lambda) = \frac{\lambda^y e^{-\lambda}}{y!}, \quad y = 0, 1, \ldots,$$

  onde $\lambda = E(Y)$ é o número esperado de mortes em um período de tempo específico (por exemplo, um ano).

- O parâmetro λ pode ser modelado como

$$E(Y) = \lambda = n \cdot f(\mathbf{x}'\boldsymbol{\beta}),$$

  onde
  - n é o tamanho da população,
  - $f(\mathbf{x}'\boldsymbol{\beta})$ é a taxa de mortalidade por 100 mil pessoas por ano, que pode depender de características da população (idade, sexo, histórico médico) através do componente linear $\mathbf{x}'\boldsymbol{\beta}$.

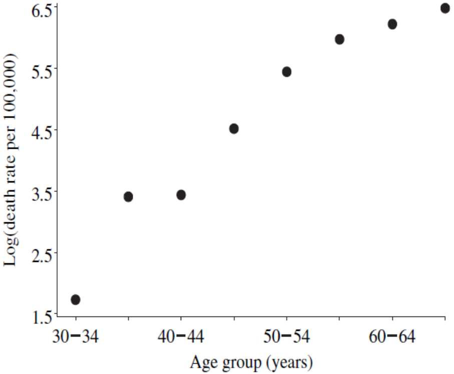
Exemplo 2: Taxa de mortalidade

Número de mortes por doenças do coração e tamanho da população para grupos etários de homens da região de Hunter, New South Wales, Austrália, em 1991.

Grupo etário	**nº mortes** (y_i)	**Pop** (z_i)	**Taxa** $(y_i/z_i) \times 100,000$
30–34	1	17.742	5,6
35–39	5	16.554	30,2
40–44	5	16.059	31,1
45–49	12	13.083	91,7
50–54	25	10.784	231,8
55–59	38	9.645	394,0
60–64	54	10.706	504,4
65–69	65	9.933	654,4

Exemplo 2: Taxa de mortalidade

- O objetivo é modelar as mudanças na mortalidade com a idade.
- Considere variáveis aleatórias independentes $Y_1, \ldots, Y_n$ representando o número de mortes de homens que ocorrem em diversos grupos etários.
- A figura a seguir mostra como a taxa de mortalidade $y_i/n_i \times 100.000$, na escala logarítmica, cresce com a idade.



Exemplo 2: Taxa de mortalidade

- Um possível modelo para esses dados é:

$$E(Y_i) = \lambda_i = n_i e^{i\theta}, \quad Y_i \sim \text{Poisson}(\lambda_i),$$

onde $i = 1$ para o grupo 30–34 anos, $i = 2$ para 35–39 anos, ..., $i = 8$ para 65–69 anos.

- Podemos escrevê-lo como um MLG usando a função de ligação logarítmica:

$$g(\lambda_i) = \log \lambda_i = \log n_i + i\theta$$

com preditor linear $\mathbf{x}_i'\boldsymbol{\beta}$, em que

$$\mathbf{x}_i' = [\log n_i \quad i], \quad \boldsymbol{\beta} = \begin{bmatrix} 1 \\ \theta \end{bmatrix}.$$

Função de Ligação

- A função de ligação relaciona os componentes sistemático e aleatório, ou seja, relaciona o preditor linear à média:

$$g(\mu_i) = \eta_i,$$

onde $g(\cdot)$ é uma função monótona e diferenciável.

- Se a função de ligação é escolhida de forma que $g(\mu_i) = \theta_i$, o preditor linear modela diretamente o parâmetro natural.

- Esta função é chamada de **função de ligação canônica**.

Distribuição	**Ligação Canônica**
Normal	Identidade: $\eta = \mu$
Poisson	Logarítmica: $\eta = \log \lambda$
Binomial	Logística: $\eta = \log \frac{\pi}{1-\pi}$
Gama	Recíproca: $\eta = \frac{1}{\mu}$

Função de Ligação

- Para o modelo linear normal, a função de ligação é chamada identidade, pois o preditor linear é igual à média. Esta função é adequada, pois η e μ podem assumir qualquer valor real.

- Para a distribuição de Poisson existe a restrição $\lambda > 0$, e uma função de ligação adequada deve respeitá-la.
  - A identidade, por exemplo, não seria apropriada, pois η pode assumir qualquer valor real dependendo dos valores estimados de $\hat{\boldsymbol{\beta}}$.

- Deve-se sempre observar a adequação da função de ligação, considerando as possíveis restrições sobre a média da distribuição da variável resposta.

- Outro exemplo é a distribuição Binomial, cujo parâmetro π deve pertencer ao intervalo $(0, 1)$. Uma função de ligação apropriada deve transformar este intervalo na reta real.

- A ligação canônica é conveniente, mas não deve ser utilizada no lugar de outra ligação que apresente melhor ajuste para os dados analisados.