Modelos Lineares Generalizados

Teoria Estatística

Fernando Lucambio

Departamento de Estatística
Universidade Federal do Paraná

13 de maio de 2021

Como muito mais nas estatísticas modernas, a perspicácia de que muitas das distribuições mais importantes na estatística poderiam ser expressas na seguinte forma linear-exponencial comum foi devido a R.A. Fisher:

$$p(y;\theta,\phi) = \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right)$$

onde

- $p(y;\theta,\phi)$ é a função de probabilidade para a variável aleatória discreta Y ou a função de densidade de probabilidade para Y contínua.
- $a(\cdot)$, $b(\cdot)$ e $c(\cdot)$ são funções conhecidas que variam de uma família exponencial para outra; veja exemplos abaixo.

- $\theta = g_c(\mu)$, o parâmetro canônico para a família exponencial em questão, é uma função da esperança $\mu = \mathrm{E}(Y)$ de Y.
- a função de ligação canônica $g_c(\cdot)$ não depende de ϕ.
- $\phi > 0$ é um parâmetro de dispersão que, em algumas famílias, assume um valor fixo conhecido, enquanto em outras famílias de distribuições é um parâmetro desconhecido a ser estimado a partir dos dados junto com θ.

Considere, por exemplo, a distribuição normal ou gaussiana com média μ e variância σ^2.

Colocar a distribuição normal na forma da equação acima requer alguma manipulação algébrica, eventualmente produzindo

$$p(y;\theta,\phi) = \exp\left(\frac{y\theta - \theta^2/2}{\phi} - \frac{1}{2}\left(\frac{y^2}{\phi} + \log_e(2\pi\phi)\right)\right),$$

onde $\theta = g_c(\mu) = \mu$, $\phi = \sigma^2$, $a(\phi) = \phi$, $b(\theta) = \theta^2/2$ e

$$c(y,\phi) = -\frac{1}{2}\left(y^2/\phi + \log_e(2\pi\phi)\right).$$

Agora considere a distribuição binomial, onde Y é a proporção de sucessos em n tentativas binárias independentes e μ é a probabilidade de sucesso em uma tentativa individual.

Escrito depois da ginástica algébrica como uma família exponencial

$$p(y;\theta,\phi) = \exp\left(\frac{y\theta - \log_e\left(1+e^{\theta}\right)}{1/n} + \log_e\binom{n}{ny}\right),$$

onde $\theta = g_c(\mu) = \log_e\left(\mu/(1-\mu)\right)$, $\phi = 1$, $a(\phi) = 1/n$,

$$b(\theta) = \log_e\left(1+e^{\theta}\right)$$

e

$$c(y,\phi) = \log_e\binom{n}{ny}.$$

Tabela 9. Funções $a(\cdot)$, $b(\cdot)$ e $c(\cdot)$ para construir as famílias exponenciais.

Família	$a(\phi)$	$b(\theta)$	$c(y,\phi)$
Gaussiana	ϕ	$\theta^2/2$	$-\frac{1}{2}\left(y^2/\phi + \log_e\left(2\pi\phi\right)\right)$
Binomial	$1/n$	$\log_e\left(1+e^\theta\right)$	$\log_e\binom{n}{ny}$
Poisson	1	e^θ	$-\log_e(y!)$
Gama	ϕ	$-\log_e(-\theta)$	$\phi^{-2}\log_e(y/\phi) - \log_e(y) - \log_e\left(\Gamma(\phi^{-1})\right)$
Gaussiana invera	ϕ	$-\sqrt{-2\theta}$	$-\frac{1}{2}\left(\log_e\left(\pi\phi y^3\right) + (\phi y)^{-1}\right)$

NOTA: Nesta tabela, n é o número de observações binomiais e $\Gamma(\cdot)$ é a função gama.

A vantagem de expressar diversas famílias de distribuições na forma exponencial comum é que as propriedades gerais das famílias exponenciais podem então ser aplicadas aos casos individuais. Por exemplo, é verdade em geral que

$$b'(\theta) = \frac{\mathrm{d}b(\theta)}{\mathrm{d}\theta} = \mu$$

e que

$$V(Y) = a(\phi)b''(\theta) = a(\phi)\frac{\mathrm{d}^2 b(\theta)}{\mathrm{d}\theta^2} = a(\phi)\nu(\mu)$$

levando aos resultados na Tabela 2.

Observe que $b(\cdot)$ é o inverso da função de ligação canônica. Por exemplo, para a distribuição normal,

$$\begin{aligned} b'(\theta) &= \frac{\mathrm{d}(\theta^2/2)}{\mathrm{d}\theta} = \theta = \mu, \\ a(\phi)b''(\theta) &= \phi \times 1 = \sigma^2, \\ \nu(\mu) &= 1 \end{aligned}$$

e para a distribuição binomial,

$$\begin{aligned} b'(\theta) &= \frac{\mathrm{d}\log_e\left(1+e^\theta\right)}{\mathrm{d}\theta} = \frac{e^\theta}{1+e^\theta} = \frac{1}{1+e^{-\theta}} = \mu, \\ a(\phi)b''(\theta) &= \frac{1}{n}\left(\frac{e^\theta}{1+e^\theta} - \left(\frac{e^\theta}{1+e^\theta}\right)^2\right) = \frac{\mu(1-\mu)}{n}, \\ \nu(\mu) &= \mu(1-\mu)\cdot \end{aligned}$$

O logaritmo da função de verossimilhança para uma observação individual Y_i assume a forma

$$\log_e\left(L(\theta_i, \phi; y_i)\right) = \frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi).$$

Para n observações independentes, temos

$$\log_e\left(L(\theta, \phi; y)\right) = \sum_{i=1}^{n}\left(\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi)\right),$$

onde $\theta = (\theta_1, \theta_2 \cdots, \theta_n) = \{\theta_i\}_{i=1}^n$ e $y = (y_1, y_2, \cdots, y_n) = \{y_i\}_{i=1}^n$.

Suponha que um modelo linear generalizado use a função de ligação $g(\cdot)$, é notacionalmente conveniente escrever β_0 para a constante de regressão α; de modo que,

$$g(\mu_i) = \eta_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik}.$$

O modelo, portanto, expressa os valores esperados das n observações em termos de um número muito menor de parâmetros de regressão.

Para obter equações de estimação para os parâmetros de regressão, temos que diferenciar o logaritmo da verossimilhança em relação a cada coeficiente por sua vez.

Consideremos que ℓ_i represente a i-ésima componente da log verossimilhança. Então, pela regra da cadeia,

$$\frac{\partial \ell_i}{\partial \beta_j} = \frac{\partial \ell_i}{\partial \theta_i} \times \frac{\mathrm{d}\theta_i}{\mathrm{d}\mu_i} \times \frac{\mathrm{d}\mu_i}{\mathrm{d}\eta_i} \times \frac{\partial \eta_i}{\partial \beta_j}, \qquad \text{para} \qquad j = 0, 1, \cdots, k.$$

Depois de algum trabalho, podemos reescrever a equação acima como

$$\frac{\partial \ell_i}{\partial \beta_j} = \frac{y_1 - \mu_i}{a_i(\phi)\nu(\mu_i)} \times \frac{\mathrm{d}\mu_i}{\mathrm{d}\eta_i} \times x_{ij}.$$

Somando as observações e definindo a soma sendo zero, produz as equações de estimação de máxima verossimilhança para os modelos lineares generalizados,

$$\sum_{i=1}^{n} \frac{y_i - \mu_i}{a_i \nu(\mu_i)} \times \frac{\mathrm{d}\mu_i}{\mathrm{d}\eta_i} \times x_{ij} = 0, \qquad \text{para} \qquad j = 0, 1, \cdots, k,$$

onde $a_i = a - i(\phi)/\phi$ não depende do parâmetro de dispersão, que é constante nas observações.

Simplificação adicional pode ser alcançada quando $g(\cdot)$ é a ligação canônica. Neste caso, as equações de estimação de máxima verossimilhança tornam-se

$$\sum_{i=1}^{n} \frac{y_i x_{ij}}{a_i} = \sum_{i=1}^{n} \frac{\mu_i x_{ij}}{a_i},$$

definindo a soma observada à esquerda da equação para a soma esperada à direita.

Observamos esse padrão nas equações de estimação para modelos de regressão logística anteriormente. No entanto, mesmo aqui as equações de estimação são, exceto no caso da família Gaussiana, emparelhados com o ligação de identidade, funções não lineares dos parâmetros de regressão e geralmente requerem métodos iterativos para sua solução.

Seja

$$Z_i = \eta_i + (y_i - \mu_i)\frac{\mathrm{d}\eta_i}{\mathrm{d}\mu_i} = \eta_i + (y_i - \mu_i)g'(\mu_i).$$

Então

$$\mathrm{E}(Z_i) = \eta_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik},$$

e

$$\mathrm{Var}(Z_i) = \big(g'(\mu_i)\big)^2 a_i \nu(\mu_i).$$

Se, portanto, pudéssemos calcular o Z_i, seríamos capazes de ajustar o modelo por regressão de mínimos quadrados ponderados de Z nos Xs, usando os inversos da $\mathrm{Var}(Z_i)$ como pesos.

Claro, este não é o caso porque não sabemos os valores de μ_i e η_i, que, de fato, dependem dos coeficientes de regressão que desejamos estimar - ou seja, o argumento é essencialmente circular. Esta observação sugeriu a Nelder and Wedderburn (1972) a possibilidade de estimar os modelos lineares generalizados por mínimos quadrados ponderados iterativos (IWLS), habilmente transformando a circularidade em um procedimento iterativo:

- Comece com as estimativas iniciais de $\widehat{\mu}_i$ e $\widehat{\eta}_i = g(\widehat{\mu}_i)$, denotados $\widehat{\mu}_i^{(0)}$ e $\widehat{\eta}_i^{(0)}$. Uma escolha simples é definir $\widehat{\mu}_i^{(0)} = y_i$. Em certas configurações, começar com $\widehat{\mu}_i^{(0)} = y_i$ pode causar dificuldades computacionais. Para um modelo linear generalizado binomial, onde $y_i = 0$, podemos tomar $\widehat{\mu}_i^{(0)} = 0.5/n_i$, e onde $y_i = 1$, podemos tomar $\widehat{\mu}_i^{(0)} = (n_i - 0.5)/n_i$. Para dados binários, então, todos os $\widehat{\mu}_i^{(0)}$ são 0.5.

- Em cada iteração l, calcule a variável resposta de trabalho Z usando os valores de $\widehat{\mu}$ e $\widehat{\eta}$ da iteração anterior,

$$Z_i^{(l-1)} = \eta_i^{(l-1)} + (y_i - \mu_i^{(l-1)}) g'\left(\mu_i^{(l-1)}\right)$$

junto com os pesos

$$W_i^{(l-1)} = 1/\left(g'\left(\mu_i^{(l-1)}\right)\right)^2 a_i \nu\left(\mu_i^{(l-1)}\right).$$

- Ajuste uma regressão de mínimos quadrados ponderados de $Z^{(l-1)}$ nos Xs, usando $W^{(l-1)}$ como pesos. Ou seja, computar

$$\mathbf{b}^{(l)} = \left(\mathbf{X}^\top \mathbf{W}^{(l-1)} \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{W}^{(l-1)} \mathbf{z}^{(l-1)},$$

onde $\mathbf{b}^{(l)}_{(k+1)\times 1}$