Modelos Lineares Generalizados

Fernando Lucambio

Departamento de Estatística
Universidade Federal do Paraná

Abril, 2021

Estrutura:

- http://leg.ufpr.br/∼lucambio/CE225/20211S/CE225.html
- terça-feira 20:45h, quinta-feira 19h
- 4 trabalhos assíncronos

Bibliografia:

- http://leg.ufpr.br/∼lucambio/GLM/GLM.html
- outras referências mencionadas na página da disciplina

Devido originalmente a Nelder and Wedderburn (1972), os modelos lineares generalizados são uma síntese e extensão notáveis de modelos de regressão familiares, como os modelos lineares. Modelos lineares generalizados tornaram-se tão centrais para a análise de dados estatísticos eficazes, entretanto, que vale a pena o esforço adicional necessário para adquirir um conhecimento básico do assunto.

Como suporte computacional utilizamos a linguagem de programação e ambiente de desenvolvimento integrado para cálculos estatísticos e gráficos R, versão 4.0.2 (2020-06-22) "Taking Off Again", especialmente a função **glm** e o pacote **gamlss**.

Estrutura:

I. A Estrutura dos modelos lineares generalizados
I.1. Estimando e testando modelos lineares generalizados

II. Modelos lineares generalizados para contagens
II.1. Modelos para dados de contagem com superdispersão
II.2. Modelo loglinear para tabelas de contingência

III. Teoria Estatística para modelos lineares generalizados
III.1. Família exponencial
III.2. Estimação por máxima verossimilhança
III.3. Testes de Hipóteses
III.4. Mostrando efeitos

IV. Diagnóstico para modelos lineares generalizados
IV.1. Diagnóstico de outliers, alavancagem e influência
IV.2. Diagnóstico de não linearidade

V. Exemplos: modelos contínuos e discretos

Um modelo linear generalizado (ou GLM) consiste em três componentes:

Um componente aleatório:

- Especificar a distribuição condicional da variável de resposta Y_i, para o i-ésimo de n observações amostradas independentemente, dados os valores das variáveis explicativas no modelo. Na formulação original, a distribuição de Y_i é membro de uma família exponencial, como a Gaussiana (normal), binomial, Poisson, gama ou famílias de distribuições gaussianas inversas. O trabalho subsequente, no entanto, estendeu os GLMs para famílias exponenciais multivariadas, como a distribuição multinomial, a certas famílias não exponenciais, como a distribuição binomial negativa de dois parâmetros e para algumas situações em que a distribuição de Y_i não é especificada completamente. A maioria dessas ideias é desenvolvida posteriormente neste texto.

Um modelo linear generalizado (ou GLM) consiste em três componentes:

Um preditor linear:

- Uma função linear de regressores

$$\eta_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik}.$$

Como no modelo linear, os regressores X_{ij} são funções pré-especificadas das variáveis explicativas e, portanto, podem incluir variáveis explicativas quantitativas, transformações de variáveis explicativas quantitativas, regressores polinomiais, regressores dummy, interações e outras. Na verdade, uma das vantagens dos GLMs é que a estrutura do preditor linear é a estrutura familiar de um modelo linear.

Um modelo linear generalizado (ou GLM) consiste em três componentes:

Função de ligação:

- Uma função de ligação linearizante suave e invertível $g(\cdot)$, que transforma a esperança da variável resposta, $\mu_i = \mathrm{E}(Y_i)$, no preditor linear:

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

Como a função de ligação é invertível, podemos escrever

$$\mu_i = g^{-1}(\eta_i) = g^{-1}(\alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik})$$

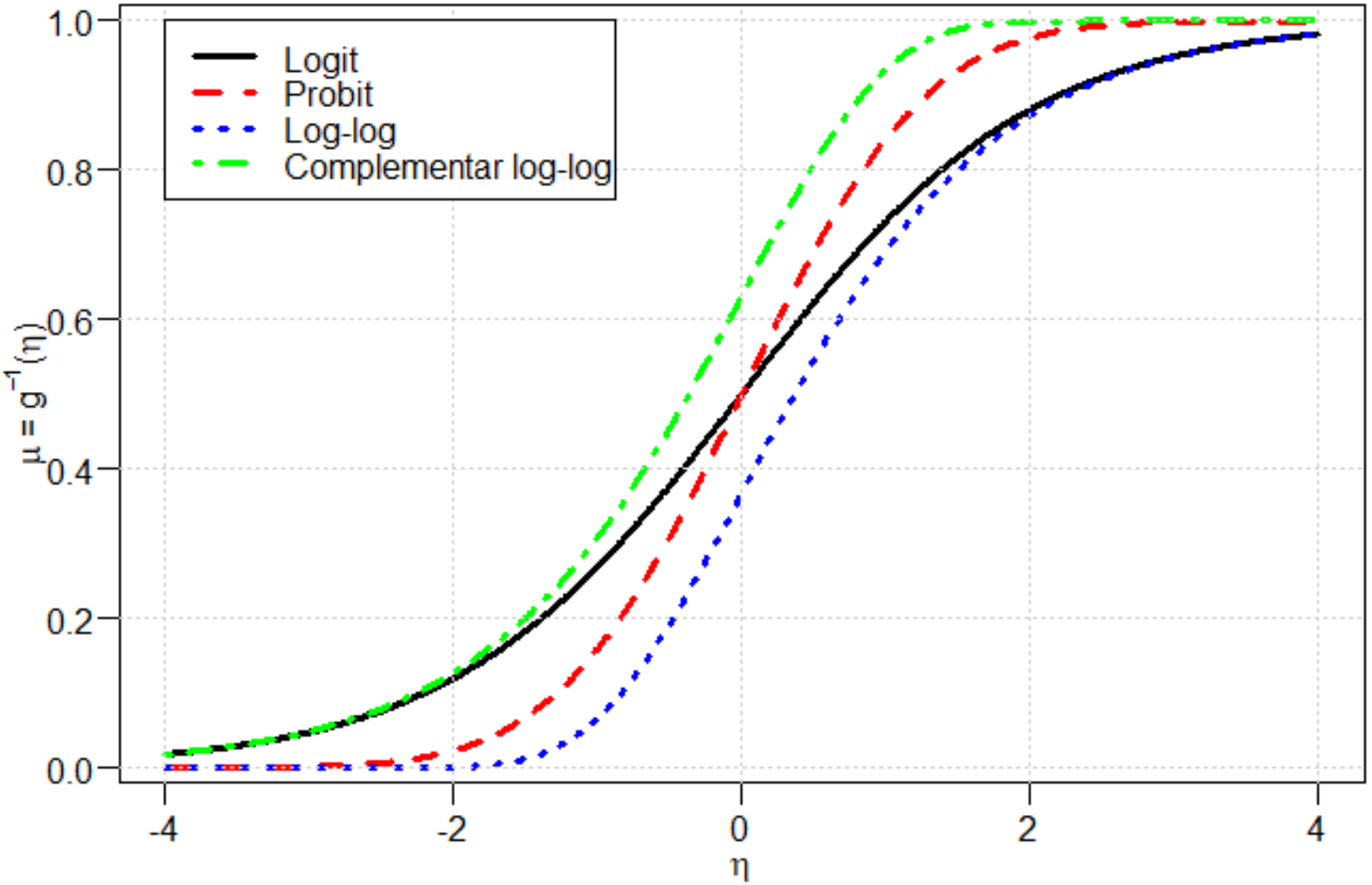
e, assim, o GLM pode ser pensado como um modelo linear para uma transformação da resposta esperada ou como um modelo de regressão não linear para a resposta. A ligação inversa $g^{-1}(\cdot)$ também é chamada de função média.

Tabela 1. Algumas funções de ligação comuns e seus inversos.

Ligação	$\eta_i = g(\mu_i)$	$\mu_i = g^{-1}(\eta_i)$
Identidade	μ_i	η_i
Log	$\log_e(\mu_i)$	e^{η_i}
Inversa	μ_i^{-1}	η_i^{-1}
Inversa quadrada	μ_i^{-2}	$\eta_i^{-1/2}$
Raiz quadrada	$\sqrt{\mu_i}$	η_i^2
Logit	$\log_e\left(\frac{\mu_i}{1-\mu_i}\right)$	$\frac{1}{1+\exp(-\eta_i)}$
Progit	$\Phi^{-1}(\mu_i)$	$\Phi(\eta_i)$
Log-log	$-\log_e(-\log_e(\mu_i))$	$\exp(-\exp(-\eta_i))$
Complementar log-log	$\log_e(-\log_e(1-\mu_i))$	$1-\exp(-\exp(\eta_i))$

NOTA: μ_i é o valor esperado da resposta; η_i é o preditor linear e $\Phi(\cdot)$ é a função de distribuição normal padrão.

Além do desejo geral de selecionar uma função de ligação que torne a regressão de *Y* nos *X*s linear, uma ligação promissora removerá as restrições no intervalo da resposta esperada. Esta é uma ideia familiar dos modelos logit e probit, onde o objetivo é modelar a probabilidade de sucesso, representada por i em nosso atual geral notação. Como probabilidade, μ_i está confinada ao intervalo unitário [0,1]. As ligações logit e probit mapeiam este intervalo para toda a linha real, de $-\infty, +\infty$. Da mesma forma, se a resposta *Y* for uma contagem, assumindo apenas valores inteiros não negativos, 0,1,2, e, consequentemente, μ_i é uma contagem esperada que, embora não necessariamente um número inteiro também não é negativa, a ligação logaritmo mapeia μ_i para toda a linha real. Isso não quer dizer que a escolha da função de ligação seja inteiramente determinada pelo intervalo da variável resposta.



Funções de ligação logit, probit, log-log e complementar log-log

Uma propriedade conveniente das distribuições na família exponencial é que a variância condicional de Y_i é uma função de sua média μ_i, digamos, $V(\mu_i)$ e, possivelmente, um parâmetro de dispersão ϕ. As funções de variância para as famílias exponenciais comumente usadas aparecem na Tabela 2. A variância condicional da resposta na família Gaussiana é uma constante ϕ, que é simplesmente uma notação alternativa para o que anteriormente denominamos variância do erro σ^2_ϵ.

Nas famílias binomial e Poisson, o parâmetro de dispersão é definido com o valor fixo $\phi = 1$.

A Tabela 2 também mostra a faixa de variação da variável resposta em cada família e a função de elo chamada canônica ou natural associada a cada componente na família. A ligação canônica simplifica o GLM, mas outras funções de ligação também podem ser usadas.

Na verdade, um dos pontos fortes do GLM - em contraste com as transformações da variável resposta na regressão linear - é que a escolha da transformação linearizante é parcialmente separada da distribuição da resposta, e a mesma transformação não precisa normalizar a distribuição de Y e fazer sua regressão linear nos X.

Tabela 2.Ligação canônica, intervalo de resposta e função de variância condicional para famílias exponenciais.

Família	**Ligação canônica**	**Intervalo da resposta**	$\mathrm{Var}(Y_i \mid \eta_i)$
Gaussiana	Identidade	$-\infty. +\infty)$	ϕ
Binomial	Logit	$0, 1/n_i, \cdots, n_i/n_i$	$\mu_i(1-\mu_i)/n_i$
Poisson	Log	$0, 1, 2, \cdots$	μ_i
Gama	Inversa	$(0, +\infty)$	$\phi\mu_i^2$
Normal inversa	Inversa quadrada	$(0, +\infty)$	$\phi\mu_i^3$

NOTA: ϕ é; o parâmetro de dispersão, η_i é o preditor linear e μ_i é a esperança de Y_i (a resposta). Na família binomial, n_i é o número de tentativas independentes.

Há também esta diferença mais sutil: quando transformamos Y e regredimos a resposta transformada nos X, nós estamos modelando a esperança da resposta transformada,

$$\mathrm{E}\big(g(Y)\big) = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik}.$$

Em um modelo linear generalizado, em contraste, modelamos a esperança transformada da resposta,

$$g\big(\mathrm{E}(Y)\big) = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik}.$$

Embora semelhante em espírito, isso não é exatamente a mesma coisa quando a função de ligação $g(\cdot)$ é não linear.