MLG

Curso de Modelos Lineares Generalizado - DEST/UFMG
Marcos Oliveira Prates

16 de outubro de 2017

Família Exponencial

- Seja Y a variável de resposta.
- A distribuição de Y é membro da família exponencial se sua densidade for da forma

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

  onde θ é chamado parâmetro natural e ϕ é o parâmetro de escala.
- As funções a, b e c determinam uma distribuição em particular na família, tais como binomial, normal, Poisson, etc.

- Para a família exponencial é possível mostrar que

$$\mu = E[Y] = b'(\theta) = \frac{db(\theta)}{d\theta}$$

e

$$\text{Var}(Y) = b''(\theta)a(\phi) = \frac{d^2 b(\theta)}{d\theta^2} a(\phi).$$

- **Distribuição Bernoulli**:

$$f_Y(y|p) = p^y (1-p)^{1-y}, \;\; y = 0, 1,$$

onde $y = 0$ ($y = 1$) é "falha"("sucesso").

- Então, temos

$$f_Y(y|p) = \exp\left\{ y \log\left(\frac{p}{1-p}\right) - [-\log(1-p)] \right\},$$

- Dessa forma vemos

$$\theta = \log\left(\frac{p}{1-p}\right),\ b(\theta) = \log(1+\exp(\theta)),\ a(\phi) = 1,\ \text{e}\ c(y,\phi) = 0.$$

- Nesse é caso verificar que

$$b'(\theta) = \frac{\exp(\theta)}{1+\exp(\theta)} = p \ \text{ e } \ b''(\theta) = \frac{\exp(\theta)}{[1+\exp(\theta)]^2} = p(1-p).$$

- **Distribuição Binomial**:

$$f_Y(y|p) = \binom{n}{y} p^y (1-p)^{n-y},\ y = 0, 1, \ldots, n.$$

temos que

$$f_Y(y|p) = \exp\left\{ y \log\left(\frac{p}{1-p}\right) - [-n\log(1-p)] + \log\binom{n}{y} \right\},$$

- Portanto, temos

$$\theta = \log\left(\frac{p}{1-p}\right),\ b(\theta) = n\log(1+\exp(\theta)),\ a(\phi) = 1,\ \text{e}\ c(y,\phi) = \log\binom{n}{y}.$$

- É simples verificar que

$$b'(\theta) = n\frac{\exp(\theta)}{1+\exp(\theta)} = np \quad \text{e} \quad b''(\theta) = n\frac{\exp(\theta)}{[1+\exp(\theta)]^2} = np(1-p).$$

- **Distribuição Poisson**

$$f_Y(y|\lambda) = \frac{\lambda^y}{y!}\exp(-\lambda) = \exp\left\{y\log\lambda - \lambda + (-\log y!)\right\}$$

  for $y = 0, 1, 2, \ldots$.

- Logo,

$$\theta = \log\lambda,\ b(\theta) = \lambda = \exp(\theta),\ a(\phi) = 1,\ \text{e}\ c(y,\phi) = -\log y!.$$

- Novamente podemos verificar que $b'(\theta) = \exp(\theta) = \lambda$ e $b''(\theta) = \exp(\theta) = \lambda$.

- **Distribuição Normal**

$$f_Y(y|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{(y-\mu)^2}{2\sigma^2} \right\}$$
$$= \exp\left\{ \frac{y\mu - \frac{\mu^2}{2}}{\sigma^2} + \left[-\frac{y^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2) \right] \right\}.$$

- Assim, $\theta = \mu$, $b(\theta) = \frac{\mu^2}{2} = \frac{\theta^2}{2}$, $\phi = \sigma^2$, $a(\phi) = \sigma^2$, e $c(y,\phi) = -\frac{y^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)$.
- Além disso, temos $E(Y) = b'(\theta) = \theta = \mu$ e $\text{Var}(Y) = b''(\theta)a(\phi) = a(\phi) = \sigma^2$.

- Seja $X_1, X_2, \ldots, X_k$ um conjunto de covariáveis.
- A média da variável de resposta,

$$\mu = E[Y]$$

é relacionada com $X_1, X_2, \ldots, X_k$ através de uma função de ligação g.

- A relação entre μ e η é definida por

$$g(\mu) = g(E[Y]) = \eta = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k.$$

- As funções de ligações tradicionalmente usadas, são as funções canônicas, ou seja, $\theta = \eta$.

Modelos Lineares Generalizados para funções de ligação canônica

- **Modelo Linear Generalizado Normal**
  Sob a função de ligação canônica, temos que

$$\theta = g(\mu) = \eta = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k.$$

- Lembre-se que o modelo de regressão múltipla (MRM) pode ser escrito como

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon,\ \varepsilon \sim N(0, \sigma^2).$$

- Portanto, o MRM com erro normal é um modelo linear generalizado normal com função de ligação canônica.
- Nesse caso vemos que a função de ligação é

$$g(E[Y]) = g(\mu) = \mu = \eta,$$

chamado de ligação identidade.

- **Modelo Linear Generalizado Binário**
  Suponha que uma resposta binária Y dado $\mathbf{X} = (1, X_1, X_2, \ldots, X_k)'$ segue uma distribuição de Bernoulli(p).
  Sob a função de ligação canônica, temos

$$\theta = \log\left(\frac{p}{1-p}\right) = \eta = \mathbf{X}'\beta,$$

  onde $\beta = (\beta_0, \beta_1, \ldots, \beta_k)'$.
- Isso quer dizer,

$$p(\mathbf{X}) = \frac{\exp(\mathbf{X}'\beta)}{1 + \exp(\mathbf{X}'\beta)},$$

  chamado de modelo de *regressão logística*.
- Sob esse modelo, a ligação é da forma

$$g(E[Y]) = g(p) = \log\left(\frac{p}{1-p}\right),$$

  chamada de ligação *logit*.

- **Modelo Linear Generalizado Poisson**
  Sob a distribuição de Poisson, temos que

$$\theta = \log \lambda$$

  e sob a ligação canônica, devemos ter

$$\log \lambda = \eta = \mathbf{X}'\beta.$$

- Assim, fazemos

$$g([Y]) = g(\lambda) = \log \lambda,$$

  e temos

$$g(\lambda) = \log \lambda = \eta.$$

- Esse g é chamado de ligação *log*.
- Portanto, sob a ligação canônica, o modelo de generalizado de Poisson é da forma:

$$f(y|\mathbf{X}, \beta) = \frac{\exp(y\mathbf{X}'\beta)}{y!} \exp\{-\exp(\mathbf{X}'\beta)\}.$$

- Resumo
  - A função de ligação *identitidade* é a função de ligação canônica para o modelo Normal.
  - A função de ligação *logit* é a função de ligação canônica para o modelo binário.
  - A função de ligação *log* é a função de ligação canônica para o modelo Poisson.
- **Modelo Linear Generalizado Binomial**
  Para regressão generalizada binomial, sob a função de ligação canônica, temos

$$\theta = \log\left(\frac{p}{1-p}\right) = \eta = \mathbf{X}'\beta.$$

- Nesse caso, escolhemos

$$g(E[Y]) = g(np) = \log\left(\frac{np}{n-np}\right) = \log\left(\frac{p}{1-p}\right) = \eta.$$

  Portanto, o modelo generalizado binomial é umas simples extensão do modelo generalizado binário.

Modelo Generalizado Binário sob diferentes ligações

- Assuma que

$$y_i \sim \text{binomial}(n_i, p_i)$$

para $i = 1, 2, ..., n$, onde n_i é conhecido e p_i depende do vetor de covariável $\mathbf{X}_i = (1, X_{i1}, X_{i2}, \ldots, X_{ik})'$.

- O modelo generalizado binomial geral assume que

$$F^{-1}(p_i) = \mathbf{X}_i'\beta,$$

onde $\beta = (\beta_0, \beta_1, \beta_2, \ldots, \beta_k)'$ é o vetor de coeficientes, F^{-1} é o inverso de uma função acumulada (cdf) F.

- Isso implica que

$$p_i = F(\mathbf{X}_i'\beta).$$

- Na regressão binomial generalizada, F^{-1} é comumente chamada de *função de ligação*.
- Note que quando $n_i \equiv 1$, a regressão binomial reduz a regressão binária.

Funções de Ligação

- Vamos apresentar quatro funções de ligação comuns, chamadas, probit, logit, log-log complementar (C log-log) e ligação-t.
- Como é bem conhecido, as funções ligação probit e logit são as funções de ligação simétricas mais utilizadas em diversas áreas de aplicação para modelar dados de resposta binários ou binomiais.
- A ligação C log-log é uma das funções de ligações assimétricas mais populares.

Funções de Ligação

- A família de ligação-t é uma classe de ligações simétricas, incluindo a ligação probit como caso especial, e a ligação logit também pode ser aproximada por um membro da dessa família ($\nu \approx 7$). Além disso, a ligação-t pode gerar funções de ligação de calda pesada com os graus de liberdade são pequenos.
- As cdf usadas para as ligações probit, logit, C log-log e ligação-t são a distribuição normal padrão, a distribuição logistic, a distribuição de valor extremo padrão, e a distribuição t, respectivamente.

Table 1: Resumo das Quatro Ligações

Tipo de Ligação	Forma da Ligação $F^{-1}(p_i)$	Distribuição Função $F(\eta_i)$	Média	Variância
logit	$\log\left(\frac{p_i}{1-p_i}\right)$	$\frac{e^{\eta_i}}{1+e^{\eta_i}}$	0	$\frac{\pi^2}{3}$
probit	$\Phi^{-1}(p_i)$	$\Phi(\eta_i)$	0	1
C log-log	$\log(-\log(1-p_i))$	$1-e^{-e^{\eta_i}}$	-0.577	$\frac{1}{6}\pi^2$
t_ν	$F_{t_\nu}^{-1}(p_i)$	$F_{t_\nu}(\eta_i)$	0	$\frac{\nu}{\nu-2}$ $(\nu > 2)$

Na Tabela 1, Φ^{-1} é o inverso da cdf da normal padrão Φ, $F_{t_\nu}(\eta_i)$ e $F_{t_\nu}^{-1}(\pi_i)$ são respectivamente a cdf e cdf inversa da distribuição t com ν graus de liberdade, e $\eta_i = \mathbf{X}_i'\beta$.

- **Notas**
  Lembre-se que $\theta_i = \log\left(\frac{p_i}{1-p_i}\right)$ e $g(E[Y_i]) = g(n_i p_i)$.
- Quando $n_i = 1$, temos $g(E[Y_i]) = g(p_i)$. Logo, sob a ligação probit, podemos tomar

$$g(p_i) = \Phi^{-1}(p_i) = \eta_i.$$

  Nesse caso,

$$\theta_i = \log\left(\frac{\Phi(\eta_i)}{1-\Phi(\eta_i)}\right) \neq \eta_i.$$

  Logo, a ligação probit **NÃO** é a função de ligação canônica.
- Da mesma forma, a ligação C log-log e ligação-t não são funções de ligação canônicas.
- SOMENTE a ligação logit é canônica.

- A probabilidade $p_i = F(\mathbf{X}_i'\beta)$ como função de $\mathbf{X}_i'\beta$ é mostrada para cada uma das funções de ligação na Figura GLM 1. A probabilidade $p_i = F(\mathbf{X}_i'\beta)$ também é plotada para duas funções de ligação-t na Figura GLM 2.
- Através das Figuras GLM 1 e GLM 2, podemos ver a diferenças entre as funções de ligação assimétrica e simétricas olhando para a taxa com que a probabilidade p_i se aproxima de 1 ou 0.
  - A taxa com que a probabilidade p_i aproxima de 1 é a mesma com que a taxa de aproxima 0 para as ligações simétricas logit, probit e t_3.
  - A probabilidade aproxima de 1 em uma taxa muito maior do que se aproxima de 0 para a ligação assimétrica C log-log.
  - As curvas na Figura GLM 2 são simétricas em torno de 0.5. A taxa com que a probabilidade se aproxima de 1 ou 0 para t_3 é mais rápida do que para t_1.

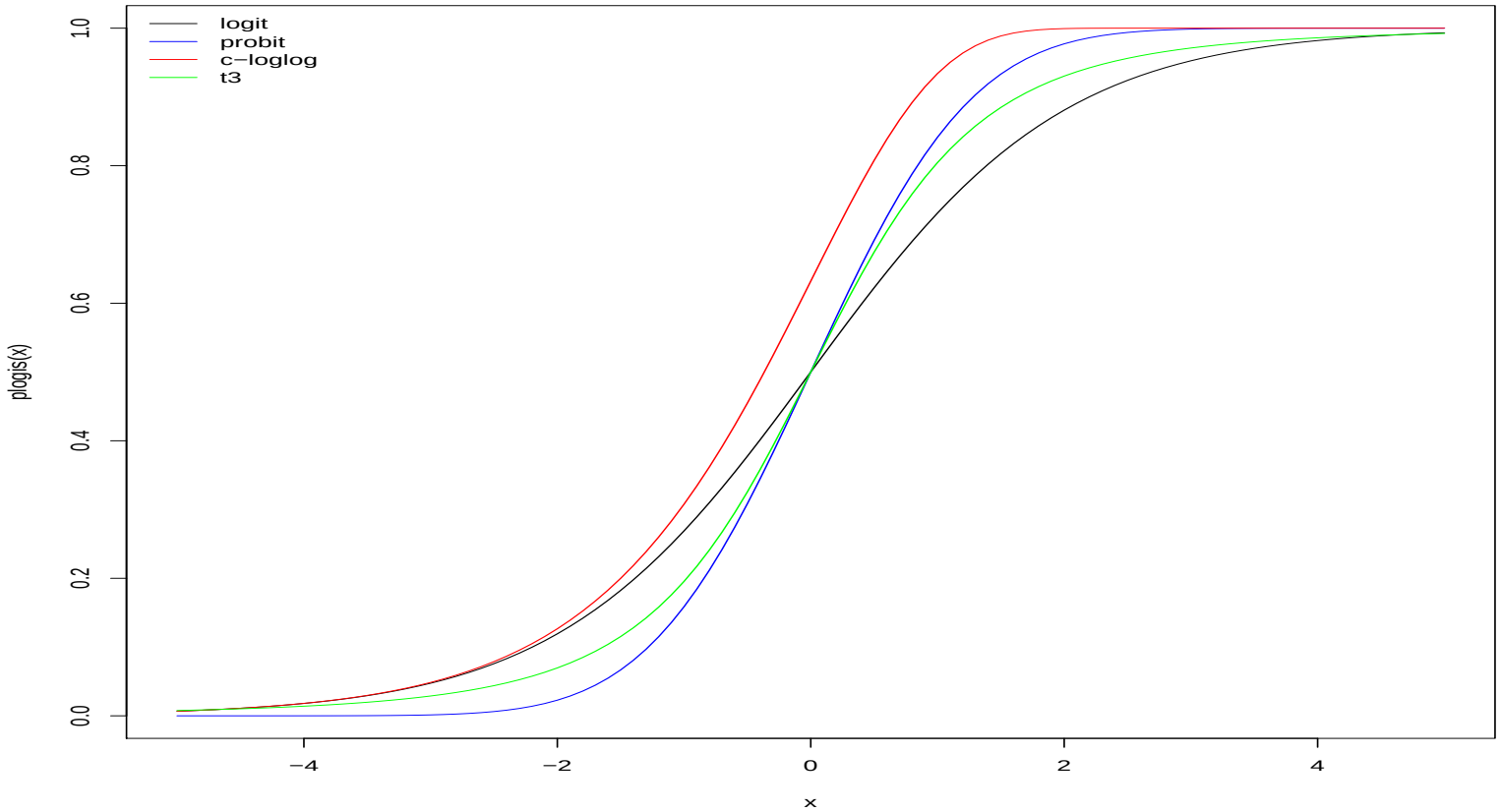

Figura: GLM 1: Gráfico de probabilidade onde a linha sólida, pontilhada, trastejada (traço curto) e trastejada (traço longo) correspondem as ligações probit, logit, C log-log e t_3, respectivamente.

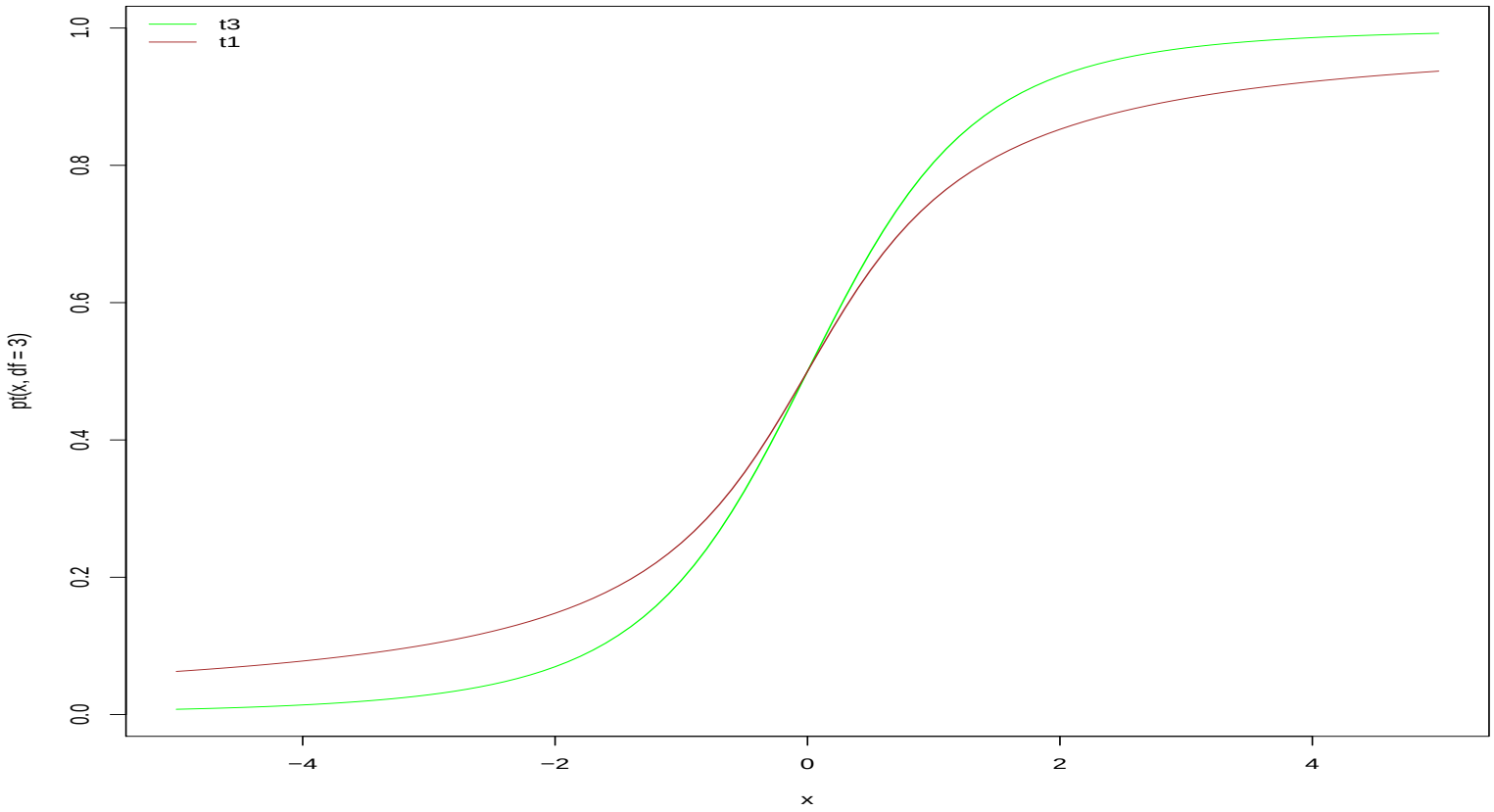

Figura: GLM 2: Gráfico de probabilidade para a ligação-t, onde a linha sólida representa t_3 e a linha pontilhada corresponde à t_1.

Utilização na Prática

- **Modelos de Regressão Logística**
  O modelo de regressão logística é talvez o modelo mais popular para respostas categóricas em estudos biomédicos. Também tem sido amplamente utilizado em ciências sociais e marketing.
- Recentemente, regressão logística se transformou numa ferramente popular em aplicação de negócios. Algumas aplicações de credit-scoring utilizam regressão logística para modelar a probabilidade de um cliente ser merecedor de crédito.

Modelo Logística

- Nesse caso assumimos que

$$y_i \sim \text{bin}(n_i, p_i)$$

para $i = 1, 2, ..., n$, onde n_i é conhecido e π_i depende do vetor de covariáveis $\mathbf{X}_i = (1, X_{i1}, X_{i2}, \ldots, X_{ik})'$.

- O modelo de regressão logística assume que

$$\text{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = \mathbf{X}_i'\beta,$$

onde $\beta = (\beta_0, \beta_1, \ldots, \beta_k)'$ é o vetor de coeficientes

- Logo

$$p_i = \frac{\exp(\mathbf{X}_i'\beta)}{1+\exp(\mathbf{X}_i'\beta)} = \frac{1}{1+\exp(-\mathbf{X}_i'\beta)}.$$

Interpretação dos Parâmetros do Modelo

1. **Inclinações**

   O logit de p_i é log odds, ou seja,

$$\text{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = \log\left(\frac{\exp(\mathbf{X}_i'\beta)/[1+\exp(\mathbf{X}_i'\beta)]}{1-\exp(\mathbf{X}_i'\beta)/[1+\exp(\mathbf{X}_i'\beta)]}\right) = \mathbf{X}_i'\beta,$$

   e os coeficientes são as razões de log-odds, quer dizer, β_j é o aumento aditivo no log-odds resultante do acréscimo de uma unidade em x_{ij}.

2. Especificamente, quando outro x_{ij^*}'s, $j^* \neq j$, são mantidos fixos,

$$\begin{aligned}
&\text{logit}(p_i)\Big|_{x_{ij}+1} - \text{logit}(p_i)\Big|_{x_{ij}} \\
=&(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_{j-1}x_{i,j-1} + \beta_j(x_{ij}+1) + \beta_{j+1}x_{i,j+1} + \cdots + \beta_{ik}x_{ik}) \\
&-(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_{j-1}x_{i,j-1} + \beta_j x_{ij} + \beta_{j+1}x_{i,j+1} + \cdots + \beta_{ik}x_{ik}) \\
=&\beta_j
\end{aligned}$$

   para $j = 1, 2, \ldots, k$.

- **Intercepto**
  Comumente, o intercepto β_0 não é de muito interesse.
- Porém, se centramos as covariáveis X_{ij} em torno de 0, i.e., substituímos X_j por $(X_j - \bar{X}_j)$, onde $\bar{X}_j = \frac{1}{n}\sum_{i=1}^{n} X_{ij}$, então,

$$p(X_1, \ldots, X_k) = \frac{\exp(\beta_0 + \beta_1(X_1 - \bar{X}_1) + \cdots + \beta_1(X_k - \bar{X}_k))}{1 + \exp(\beta_0 + \beta_1(X_1 - \bar{X}_1) + \cdots + \beta_1(X_k - \bar{X}_k))},$$

  e, portanto, β_0 é o logit na média, i.e.,

$$p(\bar{X}_1, \ldots, \bar{X}_k) = \frac{e^{\beta_0}}{1 + e^{\beta_0}}.$$

- Note que assim como na regressão linear tradicional, centrar também ajuda a reduzir a correlação entre o intercepto e as inclinações acelerando a convergência do estimador de Gibbs em inferência Bayesiana.

Função de Verossimilhança

- Dado: $\{(y_i, \mathbf{X}_i),\ \ i = 1, 2, \ldots, n\}$, onde $\mathbf{X}_i = (1, X_{i1}, X_{i2}, \ldots, X_{ik})'$.
- Parâmetros: $\beta = (\beta_0, \beta_1, \ldots, \beta_k)'$.
- **Modelo de Regressão Logística Binária**
  Seja y_i uma resposta binária tomando valores 0 ou 1, a função de verossimilhança é dada por

$$L(\beta) = \prod_{i=1}^{n} p_i^{y_i} (1 - p_i)^{1 - y_i} = \prod_{i=1}^{n} \frac{\exp(y_i \mathbf{X}_i' \beta)}{1 + \exp(\mathbf{X}_i' \beta)}$$

  e a função do log-verossimilhança é

$$l(\beta) = \log L(\beta) = \sum_{i=1}^{n} \Big\{ y_i \mathbf{X}_i' \beta - \log[1 + \exp(\mathbf{X}_i' \beta)] \Big\}.$$

- **Modelo de Regressão Logística Binomial**
  Seja y_i uma resposta binária tomando valores $0, 1, \ldots, n_i$, a função de verossimilhança é dada por

$$L(\beta) = \prod_{i=1}^{n} \binom{n_i}{y_i} p_i^{y_i} (1-p_i)^{n_i - y_i} = \prod_{i=1}^{n} \binom{n_i}{y_i} \frac{\exp(y_i \mathbf{X}_i' \beta)}{[1+\exp(\mathbf{X}_i' \beta)]^{n_i}}$$

  e a função do log-verossimilhança é

$$l(\beta) = \log L(\beta) = \sum_{i=1}^{n} \left\{ y_i \mathbf{X}_i' \beta - n_i \log[1+\exp(\mathbf{X}_i' \beta)] + \log \binom{n_i}{y_i} \right\}.$$

- **Modelo de Regressão Poisson**
  Seja y_i uma resposta binária tomando valores $0, 1, \ldots$, a função de verossimilhança é dada por

$$L(\beta) = \prod_{i=1}^{n} \frac{\exp(y_i \mathbf{X}_i \beta)}{y_i!} \exp(-\exp(\mathbf{X}_i' \beta)).$$

  e a função do log-verossimilhança é

$$l(\beta) = \log L(\beta) = \sum_{i=1}^{n} \left\{ y_i \mathbf{X}_i' \beta - \exp(\mathbf{X}_i' \beta) - \log(y_i!) \right\}.$$

Estimador de Máximo Verossimilhança

- Para MGL, não temos uma forma analítica para o EMV $\hat{\beta}$ disponível. O algoritmo de Newton-Raphson (NR) é comumente utilizado para calcular o EMV $\hat{\beta}$ de β.
- **Algoritmo de NR Geral**
  Seja $l(\beta)$ o log da verossimilhança. Em cada passo do algoritmo de Newton-Raphson, a estimativa $\beta^{(t)}$ é atualizada por

$$\beta^{(t+1)} = \beta^{(t)} + \left[-l''(\beta^{(t)})\right]^{-1} l'(\beta^{(t)}),$$

  onde $l'(\beta)$ é o vetor de derivadas de primeira ordem

$$l'(\beta) = (\partial l(\beta)/\partial\beta_0, \partial l(\beta)/\partial\beta_1, \ldots, \partial l(\beta)/\partial\beta_k)',$$

  também conhecido como vetor de score,

- e $l''(\beta^{(t)})$ é a matriz de segundas derivadas (também chamada de matriz Hessian).
- Ou seja, $l''(\beta^{(t)})$ é uma matriz de $(k+1) \times (k+1)$ elementos com o elemento $(j, j^*)^{th}$ igual a

$$\frac{\partial^2 l(\beta)}{\partial \beta_j \beta_{j^*}}.$$

- O processo é repetido até que convergência seja alcançada $\beta^{(t+1)} \approx \beta^{(t)}$.

Observações

- O método de Newton-Raphson tenta resolver a equação escore

$$l'(\beta) = 0$$

utilizando uma aproximação da função escore na vizinhança de $\beta^{(t)}$. Isso é feito, através de uma expansão de Taylor de primeira ordem,

$$l'(\beta) \approx l'(\beta^{(t)}) + l''(\beta^{(t)})(\beta - \beta^{(t)}).$$

O lado direito é zero em

$$\beta = \beta^{(t)} + \left[-l''(\beta^{(t)})\right]^{-1} l'(\beta^{(t)}).$$

- **Estimando a Matriz de Variância**
Após convergência, o inverso da matriz Hessian é uma estimativa para a matriz de covariância,

$$\widehat{\text{Var}}(\hat{\beta}) = \left[-l''(\hat{\beta})\right]^{-1},$$

o inverso da matriz de "informação observada".

Aplicando NR para regressão Binomial Logística

- Para o modelo logit, a verossimilhança é

$$l(\beta) = \log L(\beta) = \sum_{i=1}^{n} y_i \mathbf{X}_i' \beta - \sum_{i=1}^{n} n_i \log[1 + \exp(\mathbf{X}_i' \beta)].$$

- A primeira derivada de $\mathbf{X}_i' \beta$ com respeito a β_j é x_{ij} ($X_{i0} = 1$ corresponde ao intercepto), portanto, temos que

$$\frac{\partial l(\beta)}{\partial \beta_j} = \sum_{i=1}^{n} y_i x_{ij} - \sum_{i=1}^{n} n_i \left(\frac{1}{1 + \exp(\mathbf{X}_i' \beta)} \right) \exp(\mathbf{X}_i' \beta) x_{ij} = \sum_{i=1}^{n} (y_i - \mu_i) x_{ij},$$

onde $\mu_i = E[y_i] = n_i p_i = n_i [\exp(\mathbf{X}_i' \beta) / (1 + \exp(\mathbf{X}_i' \beta))]$.

- As segundas derivadas são

$$= -\sum_{i=1}^{n} n_i x_{ij} \frac{\partial}{\partial \beta_{j^*}} \left(\frac{\exp(\mathbf{X}_i'\beta)}{1+\exp(\mathbf{X}_i'\beta)} \right) = -\sum_{i=1}^{n} n_i p_i (1-p_i) x_{ij} x_{ij^*}.$$

Agora conseguimos expressas um passo do NR utilizando notação matricial semelhante a modelos de regressão linear. Seja

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} \mathbf{X}_1' \\ \mathbf{X}_2' \\ \vdots \\ \mathbf{X}_n' \end{pmatrix}, \quad \text{e} \quad \mu = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{pmatrix}.$$

- Veja que $\mathbf{Y}$ e μ são $n \times 1$, $\mathbf{X}$ é $n \times (k+1)$, e que os elementos de μ são funções não lineares de um valor dado para β. Também, defina

$$W = \text{Diag}(n_i p_i (1 - p_i)),$$

onde W é $n \times n$. Então, podemos mostrar quer

$$\begin{aligned} l'(\beta) &= \mathbf{X}'(\mathbf{Y} - \mu), \\ l''(\beta) &= -\mathbf{X}' W \mathbf{X}. \end{aligned}$$

- Para cada passo do Newton-Raphson, usamos $\beta^{(t)}$, a estimativa corrente para β, para calcular $\mu^{(t)}$ e $W^{(t)}$. A nova estimativa de β é então

$$\beta^{(t+1)} = \beta^{(t)} + \left(\mathbf{X}' W^{(t)} \mathbf{X}\right)^{-1} \mathbf{X}'(\mathbf{Y} - \mu^{(t)}).$$

- O processo é repetido até atingir convergência, i.e., até $\beta^{(t+1)}$ é suficientemente perto de $\beta^{(t)}$.
- Após convergência, $(\mathbf{X}' W \mathbf{X})^{-1}$ é a matriz de covariância estimada por $\hat{\beta} = \hat{\beta}$.

Aplicando NR para regressão de Poisson

- Para a regressão de Poisson, o log da verossimilhança é

$$l(\beta) = \log L(\beta) = \sum_{i=1}^{n} y_i \mathbf{X}_i' \beta - \sum_{i=1}^{n} [\exp(\mathbf{X}_i' \beta) - \log y_i!].$$

- Seja

$$W = \text{Diag}(\exp(\mathbf{X}_i' \beta)),$$

  novamente $n \times n$. É possível mostrar que

$$\begin{aligned} l'(\beta) &= \mathbf{X}'(\mathbf{Y} - \mu), \\ l''(\beta) &= -\mathbf{X}' W \mathbf{X}, \end{aligned}$$

  onde $\mu = (\mu_1, \mu_2, \ldots, \mu_n)'$ e $\mu_i = \exp(\mathbf{X}_i' \beta)$ for $i = 1, 2, \ldots, n$.
- O passo de Newton-Raphson para a regressão de Poisson é exatamente o mesmo a regressão binomial.
- Novamente, após convergir, $(\mathbf{X}' W \mathbf{X})^{-1}$ é a matriz de variância estimada para $\hat{\beta}$.
- O desvio padrão de $\hat{\beta}_j$ é a raiz quadrada de $(j+1)^{th}$ elemento diagonal de $(\mathbf{X}' W \mathbf{X})^{-1}$.

Um pouco sobre a Estatística de Teste para testar H_0: $\beta_j = 0$

- Diferentemente da regressão linear, a estatística de teste para H_0: $\beta_j = 0$ é

$$\chi_j^2 = \left(\frac{\hat{\beta}_j}{se(\hat{\beta}_j)}\right)^2 \overset{assimptoticamente}{\sim} \chi_1^2,$$

  i.e., a distribuição qui-quadrado com um grau de liberdade. A estatística, χ_j^2, é chamada de estatística qui-quadrado de Wald. O teste correspondente é conhecido como **Wald Test**.

- **Intervalo de confiança para um** β_j
  Utilizando a estatística de Wald, um intervalo de confiança $100(1-\alpha)\%$ para β_j é dado por

$$\hat{\beta}_j \pm z_{1-\alpha/2} se(\hat{\beta}_j),$$

  onde $z_{1-\alpha/2}$ é o $(1-\alpha/2)^{th}$ quartil da distribuição normal padrão $N(0,1)$.

Algoritmo Score de Fisher

- Fisher scoring (FS) é idêntico ao Newton-Raphson (NR), exceto que no FD a matriz Hessian $l''(\beta)$ é substituída pelo seu valor esperado, $E[l''(\beta)]$.
- Isso quer dizer que utilizamos a informação de Fisher ou informação “esperada”ao invés da informação “observada”.
- Ambos NR e FS convergem para o mesmo valor de β, o valor no qual $l'(\beta) = 0$. Após convergir, ambos oferecem assimptoticamente estimativas equivalentes de $\text{Var}(\hat{\beta})$.
- Em problemas comuns, utilizar FS ao invés de NR é simplesmente por conveniência com implicações estatísticas minímas.

Algoritmo Score de Fisher

- **FS Algorithm**
  Para uma iteração do Fisher scoring, dado o estado corrente estimado de $\beta^{(t)}$, a estimativa atualizada é então

$$\beta^{(t+1)} = \beta^{(t)} + \left[-E(l''(\beta^{(t)})) \right]^{-1} l'(\beta^{(t)}).$$

- O processo é repetido até convergência, i.e., até $\beta^{(t+1)}$ ser suficientemente perto de $\beta^{(t)}$.
- Após convergência, $\left[-E(l''(\hat{\beta})) \right]^{-1}$ é a estimativa da matriz de covariância de $\hat{\beta}$.

Valores Ajustados para Regressão Logística

- Suponha $y_i \sim \text{bin}(n_i, p_i)$. Então, um função importante não linear na regressão logística binomial é

$$p_i = p_i(\beta) = \frac{\exp(\mathbf{X}_i'\beta)}{1 + \exp(\mathbf{X}_i'\beta)},$$

onde $\mathbf{X}_i = (X_{i0}, X_{i1}, \ldots, X_{ik})'$ e $X_{i0} = 1$ corresponde ao intercepto.

- Podemos verificar que

$$\frac{\partial p_i}{\partial \beta_j} = p_i(1 - p_i)X_{ij}$$

e portanto

$$\frac{\partial p_i}{\partial \beta} = p_i(1 - p_i)\mathbf{X}_i,$$

onde $\mathbf{X}_i = (1, X_{i1}, \ldots, X_{ik})'$.

- O valor ajustado é

$$\hat{p}_i = \frac{\exp(\mathbf{X}_i'\hat{\beta})}{1 + \exp(\mathbf{X}_i'\hat{\beta})}.$$

- Uma estimativa da variância da probabilidade ajustada é

$$\widehat{\text{Var}}(\hat{p}_i) = [p_i(1-p_i)]^2 \mathbf{X}_i'(\mathbf{X}'W\mathbf{X})^{-1}\mathbf{X}_i,$$

  onde $W = \text{Diag}(n_i p_i(1-p_i))$.
- Um intervalo de confiança de 95% é dado por

$$\frac{1}{1+\exp\{-[\mathbf{X}_i'\hat{\beta} \pm 1.96 se(\mathbf{X}_i'\hat{\beta})]\}},$$

  onde $se(\mathbf{X}_i'\hat{\beta}) = \sqrt{\mathbf{X}_i'(\mathbf{X}'\widehat{W}\mathbf{X})^{-1}\mathbf{X}_i}$, $\widehat{W} = \text{Diag}(n_i\hat{p}_i(1-\hat{p}_i))$, e $\hat{\beta}$ é o EMV de β.

Resíduos de Pearson e Estatística Qui-Quadrado de Pearson

- Os resíduos de Pearson são definidos como

$$r_i = \frac{y_i - \hat{E}[Y_i]}{\sqrt{\widehat{\text{Var}}(\hat{E}[Y_i])}}$$

$$= \begin{cases} \frac{y_i - n_i\hat{p}_i}{\sqrt{n_i\hat{p}_i(1-\hat{p}_i)}} & \text{para regressão binomial} \\ \frac{y_i - \exp(\mathbf{x}_i'\hat{\beta})}{\sqrt{\exp(\mathbf{x}_i'\hat{\beta})}} & \text{para regressão de Poisson,} \end{cases}$$

onde $\hat{\beta}$ é o EMV de β.

- A estatística Qui-Quadrado de Pearson é

$$X^2 = \sum_{i=1}^{n} r_i^2,$$

que possui distribuição assimptoticamente χ^2_{n-k-1}.

- Quando o índice binomial n_i é grande ou $\exp(\mathbf{X}_i'\hat{\beta})$ é grande, o resíduo de Pearson r_i possui uma distribuição aproximadamente normal
- Quando o modelo é verdadeiro, a estatística tem valor esperado aproximadamente zero mas uma variabilidade menor que uma variável normal padrão.
- Se o numero de parâmetros do modelo é pequeno comparado ao tamanho da amostra n, os resíduos de Pearson são tradados como desvios de uma normal padrão, com valores absolutos maiores que 2 indicando uma possível falta de ajuste “lack of fit”.

Deviance

- **Outra Expressão para o Log da Função de Verossimilhança**
  Para um modelo GLM, seja

$$\mu = (\mu_1, \mu_2, \ldots, \mu_n)'$$

  o vetor de resposta das médias tal que

$$\mu_i = E[Y_i], \;\; i = 1, 2, \ldots, n.$$

- Note que no GLM, nós temos $\mu_i = b'(\theta_i)$. Então podemos expressar a função do log da verossimilhança em termos do vetor de parâmetros da média μ ao invés de $\theta = (\theta_1, \ldots, \theta_n)'$.
- Denotamos então

$$l(\mu, \phi; \mathbf{Y}) = \sum_{i=1}^{n} l(\mu_i, \phi; y_i) \equiv \sum_{i=1}^{n} l_i,$$

  onde $\mathbf{Y} = (y_1, y_2, \ldots, y_n)'$ e $l_i = l(\mu_i, \phi; y_i)$ é o log da função de verossimilhança para a i-ésima observação.

- Para $y_i \sim \text{bin}(y_i, p_i)$, temos $\mu_i = n_i p_i$ e

$$\begin{aligned} l_i =& y_i \log\left(\frac{p_i}{1-p_i}\right) - [-n_i \log(1-p_i)] + \log\binom{n_i}{y_i} \\ =& y_i \log\left(\frac{\mu_i}{n_i - \mu_i}\right) - [-n_i \log(n_i - \mu_i)] - n_i \log n_i + \log\binom{n_i}{y_i}. \end{aligned}$$

- Para $y_i \sim \text{Poisson}(\lambda_i)$, temos $\mu_i = \lambda_i$ e

$$l_i = y_i \log \lambda_i - \lambda_i + (-\log y_i!) = y_i \log \mu_i - \mu_i + (-\log y_i!).$$

Deviance and Deviance Escalada

- **Deviance Escalada**
  A deviance é definida por

$$D^*(\mathbf{Y}, \hat{\mu}) = 2[l(\mathbf{Y}, \phi; \mathbf{Y}) - l(\hat{\mu}, \phi; \mathbf{Y})],$$

  onde $\hat{\mu} = (\hat{\mu}_1, \hat{\mu}_2, \ldots, \hat{\mu}_n)'$ são os valores ajustados para o vetor μ, que são comumente os EMV's de μ.

- **Deviance**
  Note que de forma geral o GLM é dado por

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

- A deviance escalada é definida por

$$D(\mathbf{Y}, \hat{\mu}) = \frac{D^*(\mathbf{Y}, \hat{\mu})}{\phi}.$$

- Assim, o deviance é uma medida de de capacidade de ajuste "goodness-of-fit", que compara os valores ajustados (preditos) $\hat{\mu}_i$'s com os valores observados y_i's.
- Além disso, como

$$l(\mu, \phi; \mathbf{Y}) = \sum_{i=1}^{n} l(\mu_i, \phi; y_i),$$

podemos reescrever $D(\mathbf{Y}, \hat{\mu})$ como

$$D(\mathbf{Y}, \hat{\mu}) = \frac{D^*(\mathbf{Y}, \hat{\mu})}{\phi} = \sum_{i=1}^{n} d_i,$$

onde

$$d_i = 2[l(y_i, \phi; y_i) - l(\hat{\mu}_i, \phi; y_i)]/\phi.$$

for $i = 1, 2, \ldots, n$.

- Para ambos os modelos, $\phi = 1$, o que implica a deviance escalada e a deviance são a mesma.
- **Binomial**

$$\begin{aligned} d_i =& 2\Big\{ y_i \log\left(\frac{y_i}{n_i - y_i}\right) - [-n_i \log(n_i - y_i)] - n_i \log n_i + \log \binom{n_i}{y_i} \\ & - [y_i \log\left(\frac{\hat{\mu}_i}{n_i - \hat{\mu}_i}\right) - [-n_i \log(n_i - \hat{\mu}_i)] - n_i \log n_i + \log \binom{n_i}{y_i}] \Big\} \\ =& 2\Big\{ y_i \log(y_i/\hat{\mu}_i) + (n_i - y_i) \log[(n_i - y_i)/(n_i - \hat{\mu}_i)] \Big\}. \end{aligned}$$

Assume-se $0 \log 0 = 0$. Então, quando $n_i = 1$, o que corresponde a regressão de modelo binária, e temos

$$\begin{aligned} d_i &= -2\Big[y_i \log \hat{\mu}_i + (1 - y_i) \log(1 - \hat{\mu}_i) \Big] \\ &= -2\Big[y_i \log \hat{p}_i + (1 - y_i) \log(1 - \hat{p}_i) \Big]. \end{aligned}$$

- **Poisson**

$$\begin{aligned} d_i =& 2\Big[y_i \log y_i - y_i + (-\log y_i!) - [y_i \log \hat{\mu}_i - \hat{\mu}_i + (-\log y_i!)]\Big\} \\ =& 2\Big[y_i \log(y_i/\hat{\mu}_i) - (y_i - \hat{\mu}_i)\Big]. \end{aligned}$$

- Para regressão binomial,

$$\hat{\mu}_i = n_i \hat{p}_i = \frac{\exp(\mathbf{X}_i'\hat{\beta})}{1 + \exp(\mathbf{X}_i'\hat{\beta})},$$

e para regressão de Poisson,

$$\hat{\mu}_i = \hat{\lambda}_i = \exp(\mathbf{X}_i'\hat{\beta}),$$

onde $\hat{\beta}$ é o EMV de β.

- **Assimptótico**
  Para regressão binomial e Poisson,

$$D(\mathbf{Y},\hat{\mu}) \overset{\text{assimptoticamente}}{\sim} \chi^2_{n-k-1}$$

  quando o modelo ajusta.
- Então,

$$E\left[D(\mathbf{Y},\hat{\mu})/(n-k-1)\right] \approx 1,$$

  o que implica que na média, o deviance por grau de liberdade é aproximadamente 1.

Resíduos Deviance

- Seja

$$\text{sign}(w) = \begin{cases} 1 & \text{se } w > 0 \\ 0 & \text{se } w = 0 \\ -1 & \text{se } w < 0. \end{cases}$$

- o resíduo de deviance é definido por

$$dev_i = \text{sign}(y_i - \hat{\mu}_i)\sqrt{d_i}, \quad i = 1, 2, \ldots, n.$$

- Assim como os resíduos de Pearson, os resíduos de deviance são distribuídos aproximadamente pela distribuição normal mas (quando o modelo vale) possui uma variabilidade um pouco menor que variáveis de normal padrão

Superdisperção

- Para um modelo especificado corretamente, a estatística qui-quadrado de Pearson e o deviance divididos pelos graus de liberade, deve ser aproximadamente iguais a um.
- Quando os seus valores observados são muito maiores do que um, a suposição de variabilidade binomial (Poisson) pode não ser válida e os dados são ditos possuir superdisperção.
- Subdisperção, quando o resultado da razão é menor do que um. Ocorre muito menos na prática.
- Quando ajustando o modelo, existem vários problemas que podem fazer a estatística de goodness-of-fit a exceder os seus graus de liberdade. Entre esses, problemas tipo outliers nos dados, utilizar a função de ligação incorreta, omitir termos importantes no modelo, e a necessidade de transformar preditores. Esses problemas podem ser eliminados utilizando uma variância apropriada no modelo.

Possíveis Razões para Falta de Ajuste

- Mesmo que os testes X^2 ou G^2 revelem uma falta de ajuste, possíveis razões incluem:
  (a) Covariáveis omitidas;
  (b) Superdisperção;
  (c) Erro na escolha da função de ligação;
  (d) Um ou mais outliers que destoam do modelo;
- Diagnósticos podem nos ajudar a identificar (c) ou (d). Mas se não possuímos mas covariáveis para examinar é difícil distinguir (a) de (b).

Ajustando Superdisperção

- Superdisperção quer dizer que os dados mostram evidência de que a variância das respostas y_i é maior do que deveria sob o modelo escolhido. (Subdisperção é também teoricamente possível, mas rara na prática.)
- Superdisperção pode ser acomodada em 2 maneiras distintas. Uma maneira é especificar um modelo paramétrico mais rico, onde a distribuição da variável resposta é assuma ser alguma coisa mais dispersa que a usual.
- Os exemplos mais comuns são:
  - mudar o modelo binomial para o modelo beta-binomial;
  - mudar o modelo Poisson para o modelo binomial negativo;

- Outra maneira possível é especificar a função da média e da variância.
- A função da média determina como $\mu_i = E[y_i]$ é relacionado com as covariáveis. A função da variância determina a relação entre a variância da variável resposta e sua média.
- Para o modelo binomial, a função de variância é $\mu_i(n_i - \mu_i)/n_i$. Mas para acomodar superdisperção, vamos incluir um novo fator ϕ chamado de parâmetro de escala tal que

$$\text{Var}(y_i) = \phi\mu_i(n_i - \mu_i)/n_i$$

para o modelo binomial e

$$V(\mu_i) = \phi\mu_i$$

para o modelo Poisson.
- No caso $\phi > 1$ representa superdisperção e $\phi < 1$ indica subdisperção.

Testando Modelos Aninhados

- Suponha que gostaríamos de comparar o ajuste de dois modelos aninhados, ou seja, um modelo é um caso especial do outro. Quer dizer, queremos testar:

$$H_0: \text{ modelo simples é verdadeiro}$$
$$H_1: \text{ modelo mais complicado é verdadeiro}$$

- O teste de razão de verossimilhança, denotado por G^2, é

$$G^2 = -2\log\left[\frac{L(H_0)}{L(H_1)}\right],$$

onde $L(H_j)$ denota a função de verossimilhança maximizada sob H_j para $j = 0, 1$.

- Alternativamente, podemos encontrar D_0, a deviance do modelo nulo, e D_1, a deviance do modelo alternativo. Então pode se mostrar que a estatística de teste G^2 derivada pelo teste da razão de verossimilhança é igual a

$$G^2 = D_0 - D_1.$$

  quando $\phi = 1$ sob GLM.
- Os graus de liberdade para esse teste são

$$df = df_0 - df_1,$$

  a diferença entra o numero de parâmetros entre os dois modelos.
- Note que $df_0 = n - k_0 - 1$ e $df_1 = n - k_1 - 1$, onde k_0 e k_1 denotam os números de covariáveis sob os modelos H_0 e H_1, respectivamente.
- Sob H_0, G^2 possui distribuição assimptótica χ^2_{df}.