ME720 - Modelos Lineares Generalizados

Parte 9 - Inferência em MLGs

Profa. **Larissa Avila Matos**

Estimação dos parâmetros

Estimação dos parâmetros

Uma vez definido cada componente do modelo, obteremos expressões gerais para a função de verossimilhança e para as distribuições assintóticas dos estimadores de máxima verossimilhança dos parâmetros para os MLGs.

Para n observações independentes, temos que a função log-verossimilhança do modelo é dada por $\ell(\boldsymbol{\beta}, \phi) = \sum_{i=1}^{n} \ell_i(\boldsymbol{\beta}, \phi)$, onde

$$\ell_i(\boldsymbol{\beta}, \phi) = \frac{[\mathrm{y}_i \theta_i - b(\theta_i)]}{a(\phi)} + c(\mathrm{y}_i, \phi).$$

Então,

$$\ell(\boldsymbol{\beta}, \phi) = \sum_{i=1}^{n} \frac{[\mathrm{y}_i \theta_i - b(\theta_i)]}{a(\phi)} + \sum_{i=1}^{n} c(\mathrm{y}_i, \phi).$$

Para um GLM $\eta_i = \sum_{j=1}^{p} \beta_j x_{ij} = g(\mu_i)$, com função de ligação g. Portanto, o sistema de equações de verossimilhança para $\boldsymbol{\beta}$ é dado por

$$\frac{\partial \ell(\boldsymbol{\beta}, \phi)}{\partial \beta_j =} \sum_{i=1}^{n} \frac{\partial \ell_i}{\partial \beta_j} = 0, \quad \forall j.$$

Para diferenciar a log-verossimilhança, usamos a regra da cadeia,

$$\frac{\partial \ell_i}{\partial \beta_j} = \frac{\partial \ell_i}{\partial \theta_i} \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j}.$$

Uma vez que,

$$\frac{\partial \ell_i}{\partial \theta_i} = \frac{[\mathrm{y}_i - b'(\theta_i)]}{a(\phi)} \quad \text{e} \quad \mu_i = b'(\theta_i), \quad \mathrm{Var}(Y_i) = b''(\theta_i) a(\phi),$$

temos que

$$\frac{\partial \ell_i}{\partial \theta_i} = \frac{(\mathrm{y}_i - \mu_i)}{a(\phi)} \quad \text{e} \quad \frac{\partial \mu_i}{\partial \theta_i} = b''(\theta_i) = \frac{\mathrm{Var}(Y_i)}{a(\phi)}.$$

Também uma vez que $\eta_i = \sum_{j=1}^{p} \beta_j x_{ij}$, então $\frac{\partial \eta_i}{\partial \beta_j} = x_{ij}$.

Finalmente, $\eta_i = g(\mu_i) \Rightarrow \frac{\partial \mu_i}{\partial \eta_i}$, depende da função de ligação para o modelo escolhido.

Resumindo, temos que

$$\begin{aligned} \frac{\partial \ell_i}{\partial \beta_j} &= \frac{\partial \ell_i}{\partial \theta_i} \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j} \\ &= \frac{(\mathrm{y}_i - \mu_i)}{a(\phi)} \frac{a(\phi)}{\mathrm{Var}(Y_i)} \frac{\partial \mu_i}{\partial \eta_i} x_{ij} = \frac{(\mathrm{y}_i - \mu_i)\, x_{ij}}{\mathrm{Var}(Y_i)} \frac{\partial \mu_i}{\partial \eta_i} \end{aligned}$$

A soma das n observações produz o sistema de equações de verossimilhança para um MLG.

Equações de verossimilhança para um MLG

Equações de verossimilhança para um MLG:

$$
\begin{aligned}
\frac{\partial \ell(\boldsymbol{\beta}, \phi)}{\partial \beta_j} &= \sum_{i=1}^{n} \frac{(\mathrm{y}_i - \mu_i)\, x_{ij}}{\mathrm{Var}(\mathrm{Y}_i)} \frac{\partial \mu_i}{\partial \eta_i} = 0, \quad j = 1, 2, \ldots, p, \quad \text{e} \\
\frac{\partial \ell(\boldsymbol{\beta}, \phi)}{\partial \phi} &= 0,
\end{aligned}
$$

onde $\eta_i = \sum_{j=1}^{p} \beta_j x_{ij} = g(\mu_i)$ para uma função de ligação g.

Seja $\boldsymbol{V}$ a matriz diagonal de variâncias das n observações, e seja $\boldsymbol{D}$ uma matriz diagonal com os elementos de $\frac{\partial \mu_i}{\partial \eta_i}$.

Para as expressões de um MLG $\boldsymbol{\eta} = \boldsymbol{X\beta}$, com matriz de planejamento $\boldsymbol{X}$, as equações de verossimilhança têm a forma

$$\boldsymbol{XDV}^{-1}(\mathbf{y} - \boldsymbol{\mu}) = \mathbf{0}.$$

Apesar de $\boldsymbol{\beta}$ não aparecer nessas equações, ele aparece implicitamente através de $\mu_i = g^{-1}(\sum_{j=1}^{p} \beta_j x_{ij})$.

Diferentes funções de ligação gera diferentes conjuntos de equações.

Essas equações são funções não lineares de $\boldsymbol{\beta}$, e esse problema deve ser resolvido iterativamente. Há várias opções de algoritmos numéricos para resolução de sistemas não lineares: Newton-Raphson, Escore de Fisher (EF), Nelder-Mead, BFGS entre outros.

Exercício

- Encontrar as equações de máxima verossimilhança para o Modelo Poisson Log Linear.

Relação entre média e variância

Os sistema de equações de máxima verossimilhança depende da distribuição de Y_i somente pela média ($\mathbb{E}(Y_i)$) e pela variância ($\text{Var}(Y_i)$).

Além disso, a variância depende da média pela forma

$$\text{Var}(Y_i) = V(\mu_i),$$

para alguma função $V(\cdot)$.

Ou seja, a relação entre a média e a variância caracteriza a distribuição de Y_i.

Exemplo: Se Y_i tem distribuição pertencente a família exponencial e $\text{Var}(Y_i) = \mu_i$, então necessariamente Y_i tem distribuição de Poisson.

Distribuição Assintótica de $\widehat{\boldsymbol{\beta}}$

Através das propriedades de máxima verossimilhança, e sob condições de regularidade, para n grande o estimador de máxima verossimilhança $\widehat{\boldsymbol{\beta}}$ de $\boldsymbol{\beta}$ para um MLG é eficiente e tem distribuição Normal.

Temos que a matrix de informação $\boldsymbol{I}$, tem elementos dados por

$$
\begin{aligned}
\mathbb{E}\left(-\frac{\partial^2 \ell_i}{\partial\beta_j \partial\beta_k}\right) &= \mathbb{E}\left(\frac{\partial \ell_i}{\partial\beta_j}\frac{\partial \ell_i}{\partial\beta_k}\right) \\
&= \mathbb{E}\left(\frac{(Y_i-\mu_i)\,x_{ij}}{\mathrm{Var}(Y_i)}\frac{\partial\mu_i}{\partial\eta_i}\frac{(Y_i-\mu_i)\,x_{ik}}{\mathrm{Var}(Y_i)}\frac{\partial\mu_i}{\partial\eta_i}\right) \\
&= \frac{x_{ij}x_{ik}}{\mathrm{Var}(Y_i)^2}\left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2 \underbrace{\mathbb{E}\left((Y_i-\mu_i)^2\right)}_{=\mathrm{Var}(Y_i)} = \frac{x_{ij}x_{ik}}{\mathrm{Var}(Y_i)}\left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2.
\end{aligned}
$$

A matriz $\boldsymbol{I}$ é chamada de matriz de informação esperada.

Então,

$$\mathbb{E}\left(-\frac{\partial^2 \ell}{\partial\beta_j\partial\beta_k}\right) = \sum_{i=1}^{n} \frac{x_{ij}x_{ik}}{\mathrm{Var}(Y_i)}\left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2$$

Seja, $\boldsymbol{W}$ uma matriz diagonal com elementos dados por

$$w_i = \frac{(\partial\mu_i/\partial\eta_i)^2}{\mathrm{Var}(Y_i)}$$

$$\Rightarrow \boldsymbol{I} = \boldsymbol{X}'\boldsymbol{W}\boldsymbol{X}$$

A forma de $\boldsymbol{W}$ depende da função de ligação $g(\cdot)$, uma vez que $\partial\mu_i/\partial\eta_i = g'(\mu_i)$.

Portanto, a matriz de covariância de $\widehat{\boldsymbol{\beta}}$ é dada pela inversa da matriz de informação $\boldsymbol{I}$.

Distribuição Assintótica de $\widehat{\beta}$ para um MLG $\boldsymbol{\eta} = \boldsymbol{X}\beta$:

$\widehat{\boldsymbol{\beta}}$ tem distribuição aproximadamente normal $N(\boldsymbol{\beta}, (\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1})$,

onde $\boldsymbol{W}$ é a matriz diagonal com elementos $w_i = \frac{(\partial\mu_i/\partial\eta_i)^2}{\mathrm{Var}(Y_i)}$.

A matriz de covariância assintótica é estimada por $\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}'\widehat{\boldsymbol{W}}\boldsymbol{X})^{-1}$, onde $\widehat{\boldsymbol{W}}$ é $\boldsymbol{W}$ avaliado em $\hat{\boldsymbol{\beta}}$.

Obs:

1. Para a FE com parâmetro de escala, θ e ϕ são parâmetros ortogonais.
2. $\widehat{\boldsymbol{\beta}}$ e $\widehat{\phi}$ são assintoticamente independentes.

Matriz de covariância para os valores ajustados

O preditor linear estimado é dado por

$$\widehat{\boldsymbol{\eta}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}.$$

Para n grande, temos

$$\text{Var}(\widehat{\boldsymbol{\eta}}) = \boldsymbol{X}'\text{Var}(\widehat{\boldsymbol{\beta}})\boldsymbol{X} \approx \boldsymbol{X}'(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}.$$

Podemos obter a variância assintótica de $\widehat{\boldsymbol{\mu}}$ ($\text{Var}(\widehat{\boldsymbol{\mu}})$) por $\text{Var}(\widehat{\boldsymbol{\eta}})$, através do método delta. Então,

$$\text{Var}(\widehat{\boldsymbol{\mu}}) \approx \boldsymbol{D}'\text{Var}(\widehat{\boldsymbol{\eta}})\boldsymbol{D} \approx \boldsymbol{D}\boldsymbol{X}'(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}\boldsymbol{D},$$

onde $\boldsymbol{D}$ é uma matriz diagonal com elementos $\partial\mu_i/\partial\eta_i$.

Exercício: Pesquisar sobre o método delta.

Exercício

- Voltando ao exemplo da Poisson, encontre os elementos da matriz $\boldsymbol{W}$.

Estimação de $\boldsymbol{\beta}$

Como encontramos os estimadores de máxima verossimilhança dos parâmetros de um MLG?

Problema: O sistema de equações são geralmente não-lineares em $\boldsymbol{\beta}$.

Solução: Métodos iterativos para resolver sistema de equações não lineares. Focaremos em dois métodos:

- *Newton-Raphson*
- *Escore de Fisher*

Método de Newton-Raphson

O algoritmo Newton-Raphson, método de Newton-Raphson, foi desenvolvido por Isaac Newton e Joseph Raphson e tem o objetivo estimar as raízes de uma função.

Suponha que queremos encontrar a solução da equação $g(x_0) = 0$, onde g é uma função diferenciável. Dado um número x próximo de x_0, segue da expansão em série de Taylor em torno de x que

$$0 = g(x_0) \approx g(x) + g'(x)(x_0 - x).$$

Resolvendo para x_0, temos

$$x_0 \approx x - \frac{g(x)}{g'(x)}.$$

Assim, dado um valor estimado x_t, então podemos ter um novo valor estimado x_{t+1} por

$$x_{t+1} \approx x_t - \frac{g(x_t)}{g'(x_t)}.$$

Este procedimento é repetido para $t = 1, 2, 3, \ldots$ até $|g(x_t)/g'(x_t)|$ ser suficientemente pequeno.

Método de Newton-Raphson

Voltando ao nosso problema, queremos encontrar a solução da equação $\frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \mathbf{0}$.

No passo t, o processo iterativo $(t = 0, 1, 2, \ldots)$ aproxima $\ell(\boldsymbol{\beta})$ próximo de $\boldsymbol{\beta}^{(t)}$ pela expansão em série de Taylor de segunda ordem,

$$\ell(\boldsymbol{\beta}) \approx \ell(\boldsymbol{\beta}^{(t)}) + \boldsymbol{u}^{(t)}(\boldsymbol{\beta} - \boldsymbol{\beta}^{(t)}) + \left(\frac{1}{2}\right)(\boldsymbol{\beta} - \boldsymbol{\beta}^{(t)})'\boldsymbol{H}^{(t)}(\boldsymbol{\beta} - \boldsymbol{\beta}^{(t)}),$$

onde $\boldsymbol{u}^{(t)}$ e $\boldsymbol{H}^{(t)}$ são $\boldsymbol{u}$ e $\boldsymbol{H}$ avaliados em $\boldsymbol{\beta}^{(t)}$ respectivamente, com

- $\boldsymbol{u} = \left(\frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_1}, \frac{\partial \ell(\beta_2)}{\partial \beta_p}, \ldots, \frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right)'$, e
- $\boldsymbol{H}$ a matriz *Hessiana*, onde $H_{ij} = \frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \beta_i \partial \beta_j}$.

Obs: $\boldsymbol{\beta}^{(0)}$ valor inicial (chute inicial).

Resolvendo, $\frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \approx \boldsymbol{u}^{(t)} + \boldsymbol{H}^{(t)}(\boldsymbol{\beta} - \boldsymbol{\beta}^{(t)}) = \mathbf{0}$, temos a seguinte aproximação

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

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

O procedimento descrito é repetido até que mudanças em $\ell(\boldsymbol{\beta}^{(t)})$ entre ciclos sucessivos são suficientemente pequenas.

Para muitos MLGs, a matriz Hessian é negativa definida, e a log verossimilhança é uma função estritamente côncava. Então, as estimativas de máxima verossimilhança dos parâmetros do modelo existem e são únicas sob condições bastante gerais. A convergência de $\boldsymbol{\beta}^{(t+1)}$ para $\widehat{\boldsymbol{\beta}}$ na vizinhança de $\widehat{\boldsymbol{\beta}}$ é rápida.

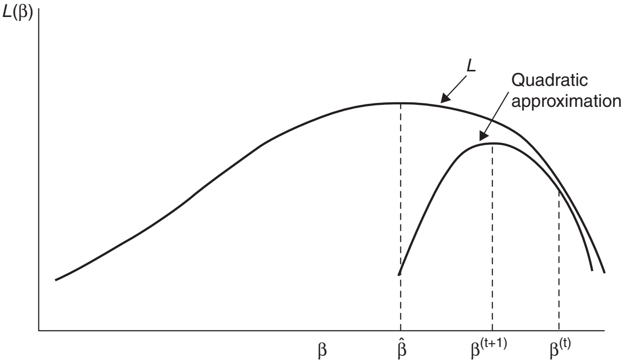


Figure 4.2 Illustration of a cycle of the Newton–Raphson method.

Figura ilustrativa do livro texto (Agresti, A. (2015)).

Método de Escore de Fisher

O método de Escore de Fisher é um método iterativo alternativo para resolver o sistema de equações de verossimilhança.

A diferença entre o método de Escore de Fisher e o método de Newton-Raphson está na maneira como escolhemos a matriz *Hessiana*.

O método de Escore de Fisher usa o valor esperado da matriz *Hessiana*, chamada de matriz de informação esperada, enquanto método de Newton-Raphson usa a própria matriz *Hessiana*, chamada de matriz de informação observada.

Portanto, temos a seguinte aproximação

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} + (\boldsymbol{I}^{(t)})^{-1}\boldsymbol{u}^{(t)},$$

onde $\boldsymbol{I}^{(t)}$ é $\boldsymbol{I}$ avaliado em $\boldsymbol{\beta}^{(t)}$, ou seja $\boldsymbol{I}^{(t)}$ tem elementos $-\mathbb{E}(\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial\beta_i\partial\beta_j})$ avaliado em $\boldsymbol{\beta}^{(t)}$.

Os algoritmos de Escore de Fisher e de Newton-Raphson são idênticos para os MLGs que usam a função de ligação canônica (Nelder e Wedderburn, 1972).

Newton-Raphson e Escore de Fisher

Newton-Raphson e Escore de Fisher para o parâmetro de uma binomial

Exemplo para o método de Newton-Raphson e Escore de Fisher de um problema mais simples para o qual sabemos a resposta.

Newton-Raphson e Escore de Fisher

Newton-Raphson e Escore de Fisher para o parâmetro de uma binomial

Exemplo para o método de Newton-Raphson e Escore de Fisher de um problema mais simples para o qual sabemos a resposta.

Seja Y uma a.a de uma distribuição $Bin(n, p)$, a função de log-verossimilhança é dada por

$$\ell(p) = \log(p^{n\mathrm{y}}(1-p)^{n-n\mathrm{y}}) = n\mathrm{y}\log(p) + (n - n\mathrm{y})\log(1-p).$$

Sabemos que

$$u = \frac{\partial \ell}{\partial p} = \frac{(n\mathrm{y} - np)}{p(1-p)} \quad \text{e} \quad H = \frac{\partial^2 \ell}{\partial p^2} = -\left[\frac{n\mathrm{y}}{p^2} + \frac{n - n\mathrm{y}}{(1-p)^2}\right].$$

Maximizando a log-verossimilhança temos que o estimador de MV para p é $\hat{p} = \mathrm{y}$.

Cada passo do algoritmo de Newton-Raphson é dado por

$$p^{(t+1)} = p^{(t)} + \left[\frac{n\mathrm{y}}{(p^{(t)})^2} + \frac{n - n\mathrm{y}}{(1 - p^{(t)})^2} \right]^{-1} \frac{(n\mathrm{y} - np^{(t)})}{p^{(t)}(1 - p^{(t)})}.$$

- Se, $p^{(0)} = \frac{1}{2} \Rightarrow p^{(1)} = \mathrm{y}$.
- Quando $p^{(t)} = \mathrm{y}$, temos que $p^{(t+1)} = \mathrm{y}$, ou seja, não temos nenhuma alteração da iteração t para $t+1$, o qual é o estimador de MV.

Calculando a esperança de H, temos que

$$\mathbb{1} = \frac{n}{p(1-p)},$$

e cada passo do algoritmo de Escore de Fisher é dado por

$$\begin{aligned} p^{(t+1)} &= p^{(t)} + \left[\frac{n}{p^{(t)}(1-p^{(t)})^2} \right]^{-1} \frac{(n\mathrm{y} - np^{(t)})}{p^{(t)}(1-p^{(t)})} \\ &= p^{(t)} + (\mathrm{y} - p^{(t)}) = \mathrm{y} \end{aligned}$$

- Ou seja, $p^{(t+1)}$ é o estimador de MV após apenas uma iteração e ficando nesse valor em todas as iterações.

Estimação do parâmetro de escala

1. Maximizar $\ell(\boldsymbol{\beta}, \phi)$ com respeito a ϕ. (Muito sensível a suposição da distribuição)

2. Sabemos que $\text{Var}(Y_i) = a(\phi)V(\mu_i)$, então

$$\frac{\mathbb{E}((Y_i - \mu_i)^2)}{V(\mu_i)} = a(\phi)$$

$$\Rightarrow \widehat{a}(\phi) = \frac{1}{n-p} \sum_{i=1}^{n} \frac{(Y_i - \widehat{\mu}_i)^2}{\text{Var}(\widehat{\mu}_i)}.$$

Teste de hipóteses

Teste de hipóteses

A inferência para MLGs tem três abordagens usando a função de verossimilhança:

1. Teste da razão de verossimilhanças;
2. Teste de Wald;
3. Teste de Escore.

Focaremos em testes

$$H_0 : \beta = \beta_0 \quad \text{vs.} \quad H_1 : \beta \neq \beta_0.$$

Teste da razão de verossimilhanças

O teste da razão de verossimilhanças usa a função de verossimilhança através da razão de (1) seu valor L_0 em β_0, e (2) seu valor máximo L_1 nos valores de β permitindo que H_0 ou H_1 sejam verdadeiras.

A razão $\Lambda = \frac{L_0}{L_1} \leq 1$, uma vez que L_0 resulta da maximização em um valor restrito de β. A estatística do teste da razão de verossimilhanças é dada por

$$-2\log(\Lambda) = -2\log\left(\frac{L_0}{L_1}\right) = -2(\ell_0 - \ell_1),$$

onde ℓ_0 e ℓ_1 denotam as funções de log-verossimilhança maximizadas. Sob condições de regularidade, $-2\log(\Lambda)$ possui distribuição qui-quadrado quando $n \to \infty$, com $g.l = 1$ Então,

$$-2\log(\Lambda) \sim \chi^2_1, \quad n \to \infty.$$

O $p - valor$ é a probabilidade da qui-quadrado ser maior igual ao valor da estatística do teste observado, ou seja, $p - valor = \mathbb{P}(\chi^2_1 \geq \chi_{obs})$.

Este teste se estende diretamente a vários parâmetros. Por exemplo, para $\boldsymbol{\beta} = (\boldsymbol{\beta}_0, \boldsymbol{\beta}_1)$, considere $H_0 : \boldsymbol{\beta}_0 = \mathbf{0}$.

Então, L_1 é a função de verossimilhança calculada no valor de $\boldsymbol{\beta}$ pelo qual os dados seriam mais prováveis e L_0 é a função de verossimilhança calculada no valor de $\boldsymbol{\beta}$ para o qual os dados seriam mais prováveis quando $\boldsymbol{\beta} = \mathbf{0}$.

O grau de liberdade da distribuição da estatística do teste é igual à diferença nas dimensões dos espaços paramétricos em $H_0 \cup H_1$ e em H_0.

O teste também se estende à hipótese linear geral

$$H_0 : \boldsymbol{C}\boldsymbol{\beta} = \mathbf{0},$$

uma vez que as restrições lineares implicam um novo modelo que é um caso especial do original.

Teste de Wald

Erros padrões obtidos a partir da inversa da matriz de informação dependem dos valores desconhecidos dos parâmetros.

Quando substituímos as estimativas irrestritas de MV (ou seja, não assumindo a hipótese nula), obtemos um erro padrão estimado (SE) de $\widehat{\beta}$.

Para $H_0 : \beta = \beta_0$ a estatística do teste usando esse erro padrão estimado não nulo,

$$z = \frac{\widehat{\beta} - \beta_0}{SE},$$

é chamada de estatística de Wald e segue aproximadamente uma distribuição normal padrão quando $\beta = \beta_0$.

Além disso, z^2 tem aproximadamente uma distribuição qui-quadrado com 1 grau de liberdade, $z^2 \sim \chi^2_1$.

Para vários parâmetros $\boldsymbol{\beta} = (\boldsymbol{\beta}_0, \boldsymbol{\beta}_1)$, para testar $H_0 : \boldsymbol{\beta}_0 = \mathbf{0}$, a estatística qui-quadrado de Wald é

$$\widehat{\boldsymbol{\beta}_0}' \left[\mathrm{Var}(\widehat{\boldsymbol{\beta}_0}) \right]^{-1} \widehat{\boldsymbol{\beta}_0},$$

onde $\widehat{\boldsymbol{\beta}_0}$ é a estimativa de máxima verossimilhança irrestrita de $\boldsymbol{\beta}_0$ e $\mathrm{Var}(\widehat{\boldsymbol{\beta}_0})$ é um bloco da matriz de covariância estimada irrestrita de $\widehat{\boldsymbol{\beta}}$.

Teste de Escore

O teste de Escore, usa a inclinação (ou seja, a função Escore) e a curvatura esperada da função de log-verossimilhança, avaliada no valor da hipótese nula β_0.

A estatística do teste é dada por

$$\frac{\left[\partial \ell(\beta)/\partial \beta_0\right]^2}{-\mathbb{E}\left[\partial \ell^2(\beta)/\partial \beta_0^2\right]} \sim \chi_1^2.$$

Para vários parâmetros $\boldsymbol{\beta} = (\boldsymbol{\beta}_0, \boldsymbol{\beta}_1)$, para testar $H_0 : \boldsymbol{\beta}_0 = \mathbf{0}$, a estatística qui-quadrado do teste de Escore é uma forma quadrática baseada no vetor de derivadas parciais da função de log-verossimilhança e na inversa da matriz de informação, ambas avaliadas nas estimativas sob H_0.

Função Desvio

Função Desvio ou Desvio

A função desvio é útil para termos uma idéia da adequabilidade do modelo (verificação da qualidade do ajuste).

Sem perda de generalidade, seja $\ell(\boldsymbol{\mu};\mathbf{y}) \equiv \ell(\boldsymbol{\beta},\phi;\mathbf{y})$ a log verossimilhança associada ao modelo em estudo (a log verossimilhança expressada em termos da média), então

$$\ell(\boldsymbol{\mu};\mathbf{y}) = \sum_{i=1}^{n} \ell(\mu_i;\mathrm{y}_i),$$

em que $\mu_i = g^{-1}(\eta_i)$ e $\eta_i = \boldsymbol{x}_i'\boldsymbol{\beta} = \sum_{j=1}^{p} x_{ij}\beta_j$.

Seja também $\ell(\mathbf{y};\mathbf{y})$ a log verossimilhança do modelo saturado $(n = p)$, em que cada média é representada por ela mesma.

A função $\ell(\boldsymbol{\mu};\mathbf{y})$ é estimada por

$$\ell(\mathbf{y};\mathbf{y}) = \sum_{i=1}^{n} \ell(\mathrm{y}_i;\mathrm{y}_i),$$

ou seja, a estimativa de máxima verossimilhança de μ_i fica nesse caso dada por $\hat{\mu}_i = \mathrm{y}_i$.

Quando $p < n$, denotamos a estimativa de $\ell(\boldsymbol{\mu};\mathbf{y})$ por $\ell(\hat{\boldsymbol{\mu}};\mathbf{y})$.

O modelo saturado, explica toda variação pelo preditor linear do modelo.

Um modelo perfeito parece bom, mas o modelo saturado não é útil. Ele não suaviza os dados ou tem as vantagens de um modelo mais simples devido à sua parcimônia.

No entanto, muitas vezes serve como linha de base para comparação com outros modelos, como para verificar a bondade de ajuste.

A qualidade do ajuste de um MLG é avaliada através da função desvio

$$D^{\star}(\mathbf{y};\hat{\boldsymbol{\mu}}) = 2\left[\ell(\mathbf{y};\mathbf{y}) - \ell(\hat{\boldsymbol{\mu}};\mathbf{y})\right],$$

que é uma distância entre o logaritmo da função de verossimilhança do modelo saturado (com n parâmetros) e do modelo sob investigação (com p parâmetros) avaliado na estimativa de máxima verossimilhança $\hat{\boldsymbol{\beta}}$.

Uma vez que $\ell(\hat{\boldsymbol{\mu}};\mathbf{y}) \leq \ell(\mathbf{y};\mathbf{y})$, então $D^{\star}(\mathbf{y};\hat{\boldsymbol{\mu}}) \geq 0$.

Um valor pequeno para a função desvio indica que, para um número menor de parâmetros, obtemos um ajuste tão bom quanto o ajuste com o modelo saturado.

Denotando por $\hat{\theta}_i = \theta_i(\hat{\mu}_i)$ e $\tilde{\theta}_i = \theta_i(\tilde{\mu}_i)$ as estimativas de máxima verossimilhança de θ_i para os modelos com p parâmetros ($p < n$) e saturado ($p = n$), respectivamente, temos que a função $D^\star(\mathbf{y}; \boldsymbol{\mu})$ fica, alternativamente, dada por

$$\begin{aligned} D^\star(\mathbf{y}; \hat{\boldsymbol{\mu}}) &= 2\sum_{i=1}^{n}\left[\frac{\mathrm{y}_i\tilde{\theta}_i - b(\tilde{\theta}_i)}{a(\phi)}\right] - 2\sum_{i=1}^{n}\left[\frac{\mathrm{y}_i\hat{\theta}_i - b(\hat{\theta}_i)}{a(\phi)}; \mathbf{y})\right] \\ &= 2\sum_{i=1}^{n}\left[\frac{\mathrm{y}_i(\tilde{\theta}_i - \hat{\theta}_i) - b(\tilde{\theta}_i) + b(\hat{\theta}_i)}{a(\phi)}\right]. \end{aligned}$$

Usualmente $a(\phi) = \frac{\phi}{w_i}$, então

$$\frac{D(\mathbf{y}; \hat{\boldsymbol{\mu}})}{\phi} = 2 \sum_{i=1}^{n} w_i \left[\mathrm{y}_i (\tilde{\theta}_i - \hat{\theta}_i) - b(\tilde{\theta}_i) + b(\hat{\theta}_i) \right],$$

$\frac{D(\mathbf{y}; \hat{\boldsymbol{\mu}})}{\phi}$ é chamado de desvio escalonado. A estatística $D(\mathbf{y}; \hat{\boldsymbol{\mu}})$ é chamada de desvio.

- **Obs.**: O desvio no R sai com o nome de deviance após o ajuste do modelo e o número de graus de liberdade correspondente é dado por $n - p$.

Exemplo Livro

O que afeta o preço de venda de uma casa?

Os dados a seguir mostram as observações sobre as vendas de casas recentes em Gainesville, Flórida. O arquivo de dados para as 100 vendas de casas estão no site do livro texto.

As variáveis listadas são:

- o preço de venda (em milhares de dólares),
- o tamanho da casa (em pés quadrados),
- a conta anual do imposto predial (em dólares),
- o número de quartos,
- o número de banheiros, e
- se a casa é nova.

Como essas 100 observações são de uma única cidade, não podemos usá-las para fazer inferências sobre as relações em geral.

Mas, para fins ilustrativos, os tratamos como uma amostra aleatória de uma população conceitual de vendas de casas nesse mercado e analisamos como o preço de venda parece se relacionar com essas características.

A seguir são apresentados os dados de 5 casas do conjunto de dados e algumas análises iniciais.

```r
Houses <- as.data.frame(read.table('Houses.txt', header = T))
Houses[1:5,]
```

```
  case taxes beds baths new price size
1    1  3104    4     2   0 279.9 2048
2    2  1173    2     1   0 146.5  912
3    3  3076    4     2   0 237.7 1654
4    4  1608    3     2   0 200.0 2068
5    5  1454    3     3   0 159.9 1477
```

```r
str(Houses)
```

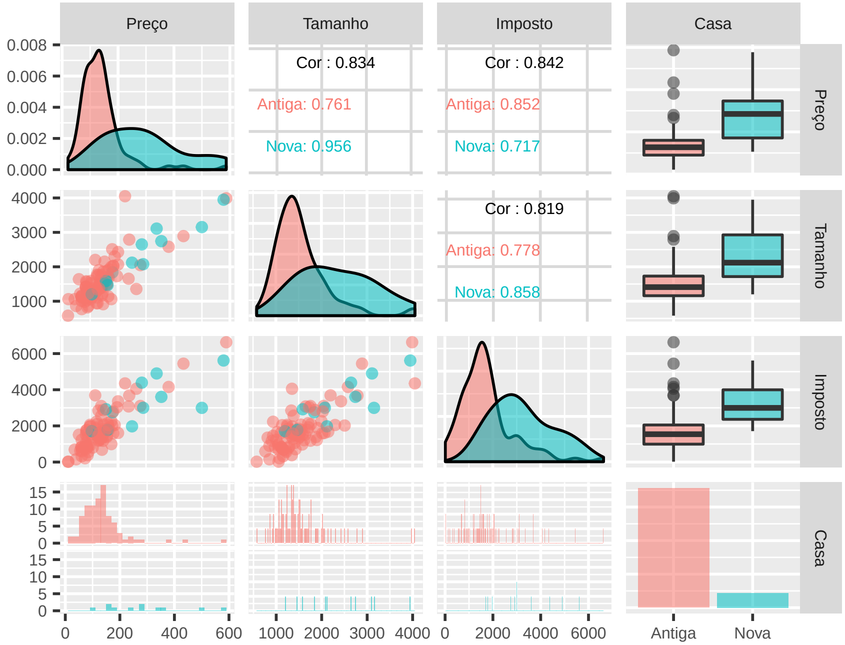
```
'data.frame':   100 obs. of  7 variables:
 $ case : int  1 2 3 4 5 6 7 8 9 10 ...
 $ taxes: int  3104 1173 3076 1608 1454 2997 4054 3002 6627 320 ...
 $ beds : int  4 2 4 3 3 3 3 3 5 3 ...
 $ baths: int  2 1 2 2 3 2 2 2 4 2 ...
 $ new  : int  0 0 0 0 0 1 0 1 0 0 ...
 $ price: num  280 146 238 200 160 ...
 $ size : int  2048 912 1654 2068 1477 3153 1355 2075 3990 1160 ...
```

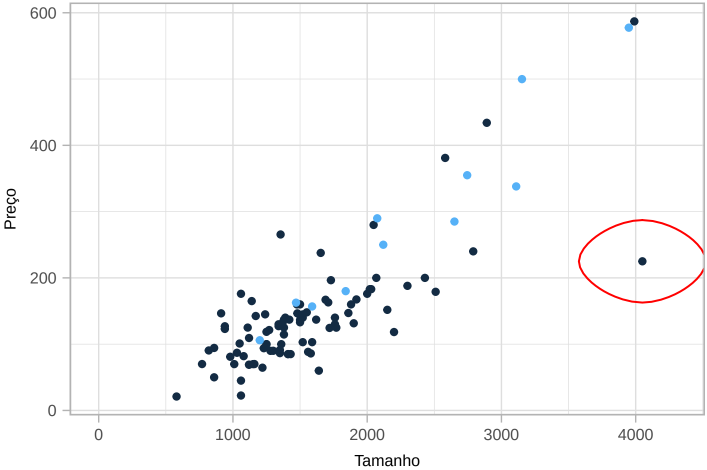
```r
cbind(mean(price), sd(price), mean(size), sd(size))
```

```
        [,1]     [,2]    [,3]     [,4]
[1,] 155.331 101.2622 1629.28 666.9417
```

```r
table(new)
```

```
new
 0  1 
89 11 
```


Preço
Tamanho
Imposto
Casa
Cor : 0.834
Antiga: 0.761
Nova: 0.956
Cor : 0.842
Antiga: 0.852
Nova: 0.717
Cor : 0.819
Antiga: 0.778
Nova: 0.858
Antiga
Nova



Temos aproximadamente uma tendência linear crescente para o preço de venda em função do tamanho. Uma exceção é um preço de venda relativamente baixo para uma residência muito grande que não era nova (observação 64 no campo dos dados). Apenas 11 casas da amostra eram novas, portanto o impacto dessa variável é pouco claro.

Em seguida, ajustamos o modelo $E(Y_i) = \beta_0 + \beta_1 \text{ size}_{i1} + \beta_2 \text{ size}_{i2}$, tendo efeitos aditivos dessas variáveis explicativas.

```r
fit <- lm(Houses$price ~ Houses$size + Houses$new)
summary(fit)
```

```
Call:
lm(formula = Houses$price ~ Houses$size + Houses$new)

Residuals:
     Min       1Q   Median       3Q      Max 
-205.102  -34.374   -5.778   18.929  163.866 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -40.230867  14.696140  -2.738  0.00737 ** 
Houses$size   0.116132   0.008795  13.204  < 2e-16 ***
Houses$new   57.736283  18.653041   3.095  0.00257 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 53.88 on 97 degrees of freedom
Multiple R-squared:  0.7226,	Adjusted R-squared:  0.7169 
F-statistic: 126.3 on 2 and 97 DF,  p-value: < 2.2e-16
```

Resultando num ajuste: $\hat{\boldsymbol{\mu}} = -40.231 + 0.116 \text{ size}_{i1} + 57.736 \text{ size}_{i2}$.

Referência

- Notas de aula do Prof. Gilberto de Paula.
- Agresti, A. (2015). *Foundations of Linear and Generalized Linear Models*. Wiley series in probability and statistics.