Método da máxima verossimilhança na presença de má especificação e Estatísticas Robustas.

Francisco Felipe de Queiroz

Centro de Ciências Exatas e da Terra
Universidade Federal do Rio Grande do Norte - UFRN
Departamento de Estatística
Seminário de Tópicos Especiais-STE

6 de abril de 2016

Contents

Sumário

Introdução

A teoria da máxima verossimilhança baseia-se na utilização da amostra para obter informações sobre características da população. Para isso, é definido um modelo estatístico para os dados e todas as inferências são baseadas na suposição de que o modelo escolhido é o correto.

Quando estamos no contexto de má especificação, utilizar os procedimentos usuais para se fazer inferência para um parâmetro pode causar graves prejuízos.
O objetivo dessa apresentação é apresentar um pouco da teoria de verossimilhança na presença de má especificação e abordar algumas das principais referências nesse tema.

Sumário

Conceitos iniciais

Considere uma amostra aleatória $X_1, X_2, \cdots, X_n$ de uma variável aleatória $X \sim g(x; \boldsymbol{\theta})$ em que $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_p)^T$, $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. Associado a densidade g, definimos as seguintes quantidades:

$$\ell_n(\boldsymbol{\theta}) = \ell_n(\boldsymbol{\theta}, \mathbf{x}) = \sum_{l=1}^{n} \log g(x_l; \boldsymbol{\theta})$$

$$\mathbf{U}_n(\boldsymbol{\theta}) = \frac{\partial \ell_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$$

$$\mathbf{K}_n(\boldsymbol{\theta}) = n\mathbf{K}(\boldsymbol{\theta}) = E\{\mathbf{U}_n(\boldsymbol{\theta})\mathbf{U}_n(\boldsymbol{\theta})^T\} = -E\left\{\frac{\partial \mathbf{U}_n(\boldsymbol{\theta})^T}{\partial \boldsymbol{\theta}}\right\},$$

em que $\mathbf{K}(\boldsymbol{\theta})$ é a matriz da informação de Fisher para uma única observação.

O estimador de máxima verossimilhança é então obtido resolvendo a equação 1 para $\hat{\boldsymbol{\theta}}$

$$\mathbf{U}_n(\hat{\boldsymbol{\theta}}) = \mathbf{0} \tag{1}$$

desde que a matriz hessiana satisfaça algumas condições.

Testes de Hipóteses

Suponha que o interesse é testar a hipótese nula $\mathcal{H}_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ contra a hipótese alternativa $\mathcal{H}_a : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$. Neste caso, pode-se usar as estatísticas da razão de verossimilhança (LR), de Wald (W), a Escore de Rao (R) e a Gradiente (T) que são dadas, respectivamente, por:

$$LR = 2\{\ell_n(\hat{\boldsymbol{\theta}}) - \ell_n(\boldsymbol{\theta}_0)\},$$

$$W = n(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^T \mathbf{K}(\hat{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0),$$

$$R = n^{-1}\mathbf{U}_n(\boldsymbol{\theta}_0)^T \mathbf{K}(\boldsymbol{\theta}_0)^{-1}\mathbf{U}_n(\boldsymbol{\theta}_0),$$

$$T = \mathbf{U}_n(\boldsymbol{\theta}_0)^T(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0).$$

Estes testes, no entanto, funcionam bem quando o modelo escolhido para os dados é "bem especificado". A presença de má especificação causa sérios prejuízos à eficiência destas estatísticas.

Sumário

Definições iniciais

Assumimos que temos um conjunto de n observações independentes e identicamente distribuidas $\mathbf{x} = (x_1, \cdots, x_n)^T$ da variável aleatória X com função densidade de probabilidade desconhecida $g(x)$, mas iremos analisar estes dados usando o modelo paramétrico $\{f(x;\boldsymbol{\theta}); \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$.
Definimos a função quase log-verossimilhança de $\boldsymbol{\theta}$ para uma amostra observada $\mathbf{x} = (x_1, \cdots, x_n)^T$ por

$$\ell_n(\boldsymbol{\theta}) = \ell_n(\boldsymbol{\theta}, \mathbf{x}) = \sum_{l=1}^{n} \log f(x_l; \boldsymbol{\theta})$$

e o estimador de quase máxima verossimilhança (QMLE) como o vetor de parâmetros $\hat{\boldsymbol{\theta}}$ que resolve o problema

$$\max_{\theta \in \Theta} \ell_n(\boldsymbol{\theta})$$

A fim de não carregar algumas definições, grande parte das suposições necessárias para os resultados aqui apresentados serão omitidas. Para mais detalhes, pode-se consultar White (1982).

Teorema (Existência)

Sob algumas condições, existe o QMLE, $\hat{\boldsymbol{\theta}}$.

O teorema acima garante que o QMLE existe, mas nada é dito a respeito de sua unicidade.

Sabemos que, quando f contém a verdadeira estrutura g, isto é

$$g(x) = f(x; \boldsymbol{\theta}_0),$$

o estimador de máxima verossimilhança é consitente para $\boldsymbol{\theta}_0$, sob algumas condições de regularidade (ver LeCam, 1953).

Se isso não acontece, Akaike (1973) notou que, desde que $\ell_n(\boldsymbol{\theta})$ seja um estimador natural para $E(\log f(X, \boldsymbol{\theta}))$, então $\hat{\boldsymbol{\theta}}$ é um estimador natural para $\boldsymbol{\theta}(g)$, o vetor de parâmetros que minimiza o critério de informação de Kullback Leibler.

Critério de Informação de Kullback - Leibler

Considere duas densidades $f(x;\theta)$ e $g(x)$. O critério de informação de Kullback-Leibler (KLIC) de $g : f$ com respeito a θ é definido por

$$\begin{aligned} I(g:f,\theta) &= E\left(\log\left[\frac{g(X)}{f(X;\theta)}\right]\right) \\ &= \int \log g(x) g(x) dx - \int \log f(x;\theta) g(x) dx \end{aligned}$$

Intuitivamente, $I(g : f, \theta)$ mensura a nossa ignorância sobre a verdadeira estrutura dos dados.
Obs.:

- $I(g : f, \theta) \geq 0, \forall \ \ \theta \in \Theta$;
- $I(g : f, \theta) = 0 \Rightarrow g(u) = f(u, \theta_0)$.

Critério de Informação de Fraser

A informação de Fraser de $f(x;\theta)$ sob $g(x)$ é definida como sendo

$$F(\theta) = E(\log f(X;\theta)) = \int \log f(X;\theta) g(x) dx.$$

Note que minimizar o KLIC $\Rightarrow$ maximizar a informação de Fraser.

Consistência

Teorema (Consistência)

Dadas algumas condições, $\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}(g)$ *quando* $n \to \infty$*, isto é,* $\hat{\boldsymbol{\theta}}$ *converge em probabilidade para* $\boldsymbol{\theta}(g)$ *.*

Com isso, notamos que o QMLE converge para $\boldsymbol{\theta}(g)$, o vetor de parâmetros que minimiza o KLIC. Ou seja, estamos "minimizando nossa ignorância" sobre a verdadeira estrutura dos dados.

Normalidade Assintótica

Sob algumas condições, defini-se as matrizes

$$\mathbf{J}(\boldsymbol{\theta}) = \int_{-\infty}^{\infty} \mathbf{u}(\boldsymbol{\theta})\mathbf{u}(\boldsymbol{\theta})^T g(x)dx, \quad \mathbf{H}(\boldsymbol{\theta}) = -\int_{-\infty}^{\infty} \frac{\partial \mathbf{u}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} g(x)dx,$$

e a matriz

$$C(\boldsymbol{\theta}) = \mathbf{H}(\boldsymbol{\theta})^{-1}\mathbf{J}(\boldsymbol{\theta})\mathbf{H}(\boldsymbol{\theta})^{-1}$$

com

$$\mathbf{u}(\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} \log f(x;\boldsymbol{\theta}) = \left(\frac{\partial}{\partial \theta_j} \log f(x;\boldsymbol{\theta})\right)_{j=1,\cdots,p}.$$

Note que a matriz $\mathbf{J}(\boldsymbol{\theta})$ é a covariância da função escore, e a matriz $\mathbf{H}(\boldsymbol{\theta})$ é o valor esperado da derivada da função escore. Ainda, perceba que $\mathbf{J}(\boldsymbol{\theta}) \neq \mathbf{H}(\boldsymbol{\theta})$, ou seja, a primeira identidade de Bartlett não se verifica.

Curiosidade

As identidades de Bartlett são equanções que estabelecem certas condições de regularidade que facilitam a obtenção dos momentos conjuntos das derivadas de $\ell(\theta)$ e de seus cumulantes. As principais identidades são

$$\mu_r = k_r = 0 \quad k_{rs} + k_{r,s} = 0$$

em que
$\mu_r = E(U_r)$, $k_{rs} = E(\partial^2 \ell / \partial\theta_r \partial\theta_s)$ e $k_{r,s} = E(U_r U_s)$, $U_r = \partial\ell/\partial\theta_r$, $r, s = 1, ..., p$.
Para mais detalhes, ver Cordeiro (1999).

Teorema (Normalidade Assintótica)

Dadas algumas suposições,

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_*) \dot{\sim} N(0, \mathbf{C}(\boldsymbol{\theta}_*))$$

em que $\mathbf{C}(\boldsymbol{\theta}_*) = \mathbf{H}(\boldsymbol{\theta}_*)^{-1}\mathbf{J}(\boldsymbol{\theta}_*)\mathbf{H}(\boldsymbol{\theta}_*)^{-1}$ *e* $\boldsymbol{\theta}_* = \boldsymbol{\theta}(g)$

Como, em geral, $\mathbf{H}(\boldsymbol{\theta})$ e $\mathbf{J}(\boldsymbol{\theta})$ são desconhecidos, pode-se usar os estimadores consistentes definidos em Lemonte (2013) por

$$\hat{\mathbf{J}}_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{l=1}^{n}\mathbf{u}^{(l)}(\boldsymbol{\theta})\mathbf{u}^{(l)}(\boldsymbol{\theta})^T, \quad \hat{\mathbf{H}}_n(\boldsymbol{\theta}) = -\frac{1}{n}\sum_{l=1}^{n}\frac{\partial \mathbf{u}^{(l)}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T},$$

e,

$$\hat{\mathbf{C}}_n(\boldsymbol{\theta}) = \hat{\mathbf{H}}_n(\boldsymbol{\theta})^{-1}\hat{\mathbf{J}}_n(\boldsymbol{\theta})\hat{\mathbf{H}}_n(\boldsymbol{\theta})^{-1}$$

em que,

$$\mathbf{u}^{(l)}(\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}}\log f(x_l;\boldsymbol{\theta}) = \left(\frac{\partial}{\partial \theta_j}\log f(x_l;\boldsymbol{\theta})\right)_{j=1,\cdots,p}, l = 1,\cdots,n,$$

e,

$$\mathbf{u}_n(\boldsymbol{\theta}) = \sum_{l=1}^{n}\mathbf{u}^{(l)}(\boldsymbol{\theta}).$$

Nesse contexto, White (1982) mostrou que

$$\hat{\mathbf{C}}_n(\hat{\boldsymbol{\theta}}) \xrightarrow{q.c.} \mathbf{C}(\boldsymbol{\theta}_*).$$

Note que se o modelo assumido é correto, isto é, $g(x) = f(x; \boldsymbol{\theta}_0)$ para algum $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}$, então o teorema em LeCam (1953), que fala da normalidade assintótica do estimador de máxima verossimilhança, é um caso particular do teorema anterior.

Em White (1982) é ainda abordado um teste para má especificação utilizando a matriz de informação, o qual não será abordado aqui.

Sumário

Testes de Hipóteses

Suponha que o nosso interesse é testar a hipótese nula $\mathcal{H}_0 : \boldsymbol{\theta}(g) = \boldsymbol{\theta}_0$ contra a hipótese alternativa $\mathcal{H}_a : \boldsymbol{\theta}(g) \neq \boldsymbol{\theta}_0$.

- Testes Robustos
- Em Kent (1982) é sugerido alternativas para testar hipóteses sob o modelo mal especificado.
- O tipo de modelo mal especificado que estamos considerando é quando a densidade $g(x)$ não pertence a família paramétrica $\{f(x;\boldsymbol{\theta});\boldsymbol{\theta} \in \boldsymbol{\Theta}\}$, mas satisfaz $\boldsymbol{\theta}(g) = \boldsymbol{\theta}_0$.

Estatísticas Robustas

No caso de modelo má especificado, alternativas para a realização de testes de hipóteses são as versões robustas das estatísticas de Wald e de Rao.
A versão robusta da estatística de Wald é dada por

$$S_W = n(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^T \mathbf{H}(\boldsymbol{\theta}_0)\mathbf{J}(\boldsymbol{\theta}_0)^{-1}\mathbf{H}(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0),$$

e a versão robusta da estatística escore de Rao é dada por

$$S_R = n^{-1}\mathbf{u}_n(\boldsymbol{\theta}_0)^T \mathbf{J}(\hat{\boldsymbol{\theta}}_0)^{-1}\mathbf{u}_n(\boldsymbol{\theta}_0),$$

em que

$$\mathbf{u}_n(\boldsymbol{\theta}) = \frac{\partial \ell_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \sum_{l=1}^{n} \frac{\partial}{\partial \boldsymbol{\theta}} \log f(x_l; \boldsymbol{\theta}).$$

Não existe uma versão robusta para a estatística da razão de verossimilhança e, segundo Kent (1982), as estatísticas robustas S_W e S_R tem distribuição assintótica χ^2_p sob a hipótese nula.

Lemonte (2013) encontrou a versão robusta da estatística gradiente para testar a hipótese nula $\mathcal{H}_0 : \boldsymbol{\theta}(g) = \boldsymbol{\theta}_0$ sob modelo má especificado. A *estatística gradiente robusta* é dada por

$$S_T = \mathbf{u}_n(\boldsymbol{\theta}_0)^T \mathbf{J}(\hat{\boldsymbol{\theta}}_0)^{-1} \mathbf{H}(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0),$$

e, sob a hipótese nula, ele mostrou que ela também tem distribuição assintótica χ^2_p.

Estudo de Simulação

Em Lemonte (2013) são feitos estudos de simulação para diferentes modelos e verificado que, tanto sob má especificação como sob o modelo correto, as estatísticas robustas têm um bom desempenho.

Em uma reprodução do que foi feito em Lemonte (2013), geramos amostras aleatórias de tamanho 115 obtidas de uma distribuição Weibull com parâmetro de forma $a = 2$ e parâmetro de escala $b = 1$, mas assumimos que os dados são de uma distribuição Gama com parâmetro de forma $\alpha = \Gamma(1.5)$ e parametro de escala θ. Queremos testar a hipótese nula $\mathcal{H}_0 : \theta = 1$ contra a hipótese alternativa $\mathcal{H}_a : \theta \neq 1$.

A tabela abaixo mostra a taxa de rejeição da hipótese nula (dado que ela é verdadeira) para os níveis de significância de 10%,5% e 1%.

Tabela: Taxa de rejeição da hipótese nula para testar $\mathcal{H}_0 : \theta = 1$

	Modelo correto			Modelo má especificado		
	10%	5%	1%	10%	5%	1%
LR	10.17	4.91	1.00	0.08	0.00	0.00
W	10.46	5.18	1.52	0.16	0.02	0.00
R	10.12	4.81	1.02	0.09	0.00	0.00
T	10.12	4.81	1.02	0.09	0.00	0.00
Sw*	8.49	5.23	2.00	9.50	4.86	1.19
Sr*	10.82	5.56	1.65	10.21	4.98	0.88
St*	9.74	4.58	0.94	9.88	4.80	0.85

Percebemos que, sob o modelo mal especificado, as estatísticas usuais apresentam uma taxa de rejeição muito inferior a esperada. Já as versões robustas dessas estatísticas apresentam-se com um bom desempenho.

Sumário

Família de Posição e Escala

Considere $X_1, \cdots, X_n$ uma amostra aleatória de X, com X pertencente a família de posição e escala. Então, X têm densidade

$$f(x;\theta) = \frac{1}{\sigma} f_0\left(\frac{x-\theta}{\sigma}\right),$$

em que o parâmetro de escala σ é conhecido, $f_0(\cdot)$ é uma densidade e θ é o parâmetro de posição.
Com isso, para uma amostra observada $x_1, \cdots, x_n$, defini-se a função de verossimilhança para θ

$$L(\theta, \mathbf{x}) = \frac{1}{\sigma^n} \prod_{l=1}^{n} f_0(z_l), \quad z_l = \frac{x_l - \theta}{\sigma}.$$

Neste caso, pode-se mostrar que

$$\hat{\mathbf{H}}_n(\theta) = -\frac{1}{n\sigma^2}\sum_{l=1}^{n}\frac{\partial^2 \log f_0(z_l)}{\partial^2 z_l}$$

e

$$\hat{\mathbf{J}}_n(\theta) = -\frac{1}{n\sigma^2}\sum_{l=1}^{n}\left[\frac{\partial \log f_0(z_l)}{\partial z_l}\right]^2.$$

A fim de simplificar a notação, façamos

$$\log f_0(z_l) \equiv s(z_l)$$

Com isso,

$$\frac{\partial \log f_0(z_l)}{\partial z_l} \equiv s'(z_l) \quad e \quad \frac{\partial^2 \log f_0(z_l)}{\partial^2 z_l} \equiv s''(z_l)$$

Assim,

$$\hat{\mathbf{H}}_n(\theta) = -\frac{1}{n\sigma^2}\sum_{l=1}^{n} s''(z_l) \quad e \quad \hat{\mathbf{J}}_n(\theta) = -\frac{1}{n\sigma^2}\sum_{l=1}^{n}\left[s'(z_l)\right]^2.$$

Suponha que o nosso interesse é testar a hipótese nula $\mathcal{H}_0 : \theta = \theta_0$ contra a hipótese alternativa $\mathcal{H}_a : \theta \neq \theta_0$.

Pode-se mostrar que, nestes casos, as estatísticas robustas de Wald, Escore de Rao e Gradiente são dadas pelas expressões à seguir.

$$S_W^* = n\left(\frac{\hat{\theta}-\theta_0}{\sigma}\right)^2 \frac{\left[\sum_{l=1}^{n} s''(z_l)\right]^2}{\sum_{l=1}^{n}\left[s'(z_l)\right]^2}, \quad z_l = \frac{x_l-\theta_0}{\sigma}.$$

$$S_R^* = \frac{-\left[\sum_{l=1}^{n} s'(z_l)\right]^2}{\sum_{l=1}^{n}\left[s'(z_l)\right]^2}, \quad z_l = \frac{x_l-\theta_0}{\sigma}.$$

$$S_T^* = \left(\frac{\theta_0 - \hat{\theta}}{\sigma}\right) \frac{\left[\sum_{l=1}^{n} s'(z_l)\right]\left[\sum_{l=1}^{n} s''(z_l)\right]}{\sum_{l=1}^{n} \left[s'(z_l)\right]^2}, \quad z_l = \frac{x_l - \theta_0}{\sigma}.$$

Exemplo

Considere um conjunto de n observações $\mathbf{x} = (x_1, \cdots, x_n)^T$ de uma distribuição normal com média $\theta \in \Re$ e variância conhecida σ^2. Então podemos escrever $f(x;\theta)$ como sendo

$$f(x;\theta) = \frac{1}{\sigma} f_0\left(\frac{x-\theta}{\sigma}\right)$$

em que

$$f_0(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}.$$

Neste caso, notamos que $f(\cdot)$ pertence a família de posição e escala, com σ conhecido.

Se o interesse é testar a hipótese nula $\mathcal{H}_0 : \theta(g) = \theta_0$ contra a hipótese alternativa $\mathcal{H}_a : \theta(g) \neq \theta_0$, pode-se mostrar que as estatísticas robustas são dadas por

$$S_W^* = S_R^* = S_T^* = n\frac{(\bar{d} - \theta_0)^2}{(\bar{d}_2 - 2\bar{d}\theta_0 + \theta_0^2)}$$

em que $\bar{d} = n^{-1}\sum_{l=1}^{n} x_l$ e $\bar{d}_2 = n^{-1}\sum_{l=1}^{n} x_l^2$

Sumário

Conclusões

Vimos que a presença de má especificação pode deturpar resultados obtidos a partir de inferências estatísticas. Nesse contexto, uma das principais alternativas é a utilização de estatísticas robustas pois elas funcionam bem ainda sob modelos mal especificados.

Sumário

Trabalhos Futuros

- Realizar mais algumas simulações para modelos da família de posição e escala com σ conhecido;
- Realizar simulações para modelos da família de posição e escala quando os dois parâmetros são desconhecidos;
- Analisar o desempenhos das estatísticas robustas em diferentes modelos da família de posição e escala.

Sumário

Referências

- LEMONTE, Artur J.. **On the gradient statistic under model misspecification**. Statistics And Probability Letters, v. 83, p.390-398, out. 2013.
- WHITE, Halbert. **Maximum Likelihood estimation of misspecification models**. Econometrica,v. 50, jan. 1982.
- Kent, T.J., 1982. **Robust properties of likelihood ratio tests**. Biometrika 69, 19?27.