Máxima Verossimilhança

Bibliografia: Greene

Claudio Lucinda

FEA/USP

Overview

Máxima Verossimilhança

- O ponto de partida do Método da Máxima Verossimilhança é a imposição de uma pdf que assumimos que represente o processo gerador dos dados que estão na nossa amostra.
- Essa PDF é condicional a um vetor de parâmetros, e a densidade conjunta de n iid observações que seguem esta pdf é o produto das densidades individuais;

$$f\left(y_1, \ldots, y_n \mid \boldsymbol{\theta}\right) = \prod_{i=1}^{n} f\left(y_i \mid \boldsymbol{\theta}\right) = L(\boldsymbol{\theta} \mid \mathbf{y})$$

Essa pdf conjunta é chamada de Função Verossimilhança, e $\mathbf{y}$ é a representação da amostra. Geralmente é mais simples trabalhar com o logaritmo da função verossimilhança, ou a log-verossimilhança:

$$\ln L(\boldsymbol{\theta} \mid \mathbf{y}) = \sum_{i=1}^{n} \ln f\left(y_i \mid \boldsymbol{\theta}\right)$$

Variáveis Independentes

- Como exemplo, vamos usar o modelo de Regressão Linear Clássico.
- Suponha que o termo erro seja normalmente distribuído
- Ou seja, condicional aos $\mathbf{x}_i$, y_i é normalmente distribuída com média $\mu_i = \mathbf{x}_i'\boldsymbol{\beta}$ e variância σ^2.
- A verossimilhança fica sendo então:

$$\ln L(\boldsymbol{\theta} \mid \mathbf{y}, \mathbf{X}) = \sum_{i=1}^{n} \ln f\left(y_i \mid \mathbf{x}_i, \boldsymbol{\theta}\right) = -\frac{1}{2} \sum_{i=1}^{n} \left[\ln \sigma^2 + \ln(2\pi) + \left(y_i - \mathbf{x}_i'\boldsymbol{\beta}\right)^2 / \sigma^2\right]$$

Identificação em ML

Definition

O Vetor de Parâmetros $\boldsymbol{\theta}$ é identificado (=estimável) se, para qualquer outro vetor $\boldsymbol{\theta}^* \neq \boldsymbol{\theta}$, para algum banco de dados $\mathbf{y}$, $L\left(\boldsymbol{\theta}^* \mid \mathbf{y}\right) \neq L(\boldsymbol{\theta} \mid \mathbf{y})$.

Propriedades do ML - Condições de Regularidade

- R1. As primeiras três derivadas de $\ln f(y_i \mid \boldsymbol{\theta})$ com respeito a $\boldsymbol{\theta}$ são finitas e contínuas para quase todos os y_i e para todos os θ. Isso assegura a existência de uma aproximãção de série de Taylor e garante uma variância finita das derivadas de $\ln L$.
- R2. As condições necessárias para se obter as esperanças das primeiras e segundas derivadas de $\ln f(y_i \mid \boldsymbol{\theta})$ são atendidas.
- R3. Para todos os valores de $\boldsymbol{\theta}$, $\left|\partial^3 \ln f(y_i \mid \boldsymbol{\theta}) / \partial\theta_j \partial\theta_k \partial\theta_l\right|$ é menor que uma função com esperança finita. Isso vai permitir que a série de Taylor seja truncada no segundo termo.

Propriedades do ML

- M1. Consistência: $\text{plim}\, \hat{\boldsymbol{\theta}} = \boldsymbol{\theta}_0$.
- M2. Normalidade Assintótica: $\hat{\boldsymbol{\theta}} \overset{a}{\sim} N\left[\boldsymbol{\theta}_0, \{\mathbf{I}(\boldsymbol{\theta}_0)\}^{-1}\right]$, em que

$$\mathbf{I}(\boldsymbol{\theta}_0) = -E_0\left[\partial^2 \ln L / \partial\boldsymbol{\theta}_0 \partial\boldsymbol{\theta}_0'\right].$$

- M3. Eficiência Assintótica: $\hat{\boldsymbol{\theta}}$ é assintoticamente eficiente e alcança o limite inferior de Cramér-Rao para estimadores consistentes.
- M4. Invariância: O estimador ML de $\gamma_0 = \mathbf{c}(\boldsymbol{\theta}_0)$ is $\mathbf{c}(\hat{\boldsymbol{\theta}})$ if $\mathbf{c}(\boldsymbol{\theta}_0)$ é uma função contínua e continuamente diferenciável.

Momentos das Derivadas do ML

- D1. $\ln f(y_i \mid \boldsymbol{\theta})$, $\mathbf{g}_i = \partial \ln f(y_i \mid \boldsymbol{\theta}) / \partial \boldsymbol{\theta}$, and $\mathbf{H}_i = \partial^2 \ln f(y_i \mid \boldsymbol{\theta}) / \partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'$, $i = 1, \ldots, n$, todas são amostras aleatórias de variáveis aleatórias (porque assumimos amostragem aleatória). A notação $\mathbf{g}_i(\boldsymbol{\theta}_0)$ and $\mathbf{H}_i(\boldsymbol{\theta}_0)$ indica a derivada avaliada em $\boldsymbol{\theta}_0$.
- D2. $E_0[\mathbf{g}_i(\boldsymbol{\theta}_0)] = \mathbf{0}$.
- D3. $\text{Var}[\mathbf{g}_i(\boldsymbol{\theta}_0)] = -E[\mathbf{H}_i(\boldsymbol{\theta}_0)]$.

Equação da Verossimilhança

A função verossimilhança é

$$\ln L(\boldsymbol{\theta} \mid \mathbf{y}) = \sum_{i=1}^{n} \ln f\left(y_i \mid \boldsymbol{\theta}\right)$$

O vetor de primeiras derivadas, ou vetor de score, é dado por

$$\mathbf{g} = \frac{\partial \ln L(\boldsymbol{\theta} \mid \mathbf{y})}{\partial \boldsymbol{\theta}} = \sum_{i=1}^{n} \frac{\partial \ln f\left(y_i \mid \boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta}} = \sum_{i=1}^{n} \mathbf{g}_i$$

Como estamos apenas somando termos, decorre das propriedades D1 e D2 do slide anterior que em $\boldsymbol{\theta}_0$,

$$E_0\left[\frac{\partial \ln L\left(\boldsymbol{\theta}_0 \mid \mathbf{y}\right)}{\partial \boldsymbol{\theta}_0}\right] = E_0\left[\mathbf{g}_0\right] = \mathbf{0}.$$

A Igualdade da Matriz de Informação - I

O Hessiano da log verossimilhança é

$$\mathbf{H} = \frac{\partial^2 \ln L(\boldsymbol{\theta} \mid \mathbf{y})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} = \sum_{i=1}^{n} \frac{\partial^2 \ln f\,(y_i \mid \boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} = \sum_{i=1}^{N} \mathbf{H}_i.$$

Avaliando em $\boldsymbol{\theta}_0$ e definindo

$$E_0\left[\mathbf{g}_0 \mathbf{g}_0'\right] = E_0\left[\sum_{i=1}^{n} \sum_{j=1}^{n} \mathbf{g}_{0i} \mathbf{g}_{0j}'\right]$$

A Igualdade da Matriz de Informação – II

Devido a D1, podemos tirar os termos com subscritos diferentes e chegar em

$$E_0\left[\mathbf{g}_0\mathbf{g}_0'\right]=E_0\left[\sum_{i=1}^{n}\mathbf{g}_{0i}\mathbf{g}_{0i}'\right]=E_0\left[\sum_{i=1}^{n}\left(-\mathbf{H}_{0i}\right)\right]=-E_0\left[\mathbf{H}_0\right]$$

Tal que

$$\begin{aligned}\mathrm{Var}_0\left[\frac{\partial\ln L\left(\boldsymbol{\theta}_0\mid\mathbf{y}\right)}{\partial\boldsymbol{\theta}_0}\right]&=E_0\left[\left(\frac{\partial\ln L\left(\boldsymbol{\theta}_0\mid\mathbf{y}\right)}{\partial\boldsymbol{\theta}_0}\right)\left(\frac{\partial\ln L\left(\boldsymbol{\theta}_0\mid\mathbf{y}\right)}{\partial\boldsymbol{\theta}_0'}\right)\right]\\&=-E_0\left[\frac{\partial^2\ln L\left(\boldsymbol{\theta}_0\mid\mathbf{y}\right)}{\partial\boldsymbol{\theta}_0\partial\boldsymbol{\theta}_0'}\right].\end{aligned}$$

Normalidade Assintótica

No estimador ML, o gradiente da função verossimilhança é igual a zero por construção, então

$$\mathbf{g}(\hat{\boldsymbol{\theta}}) = \mathbf{0}$$

Fazendo uma expansão de Taylor de segunda ordem em volta dos valores veradeiros $\boldsymbol{\theta}_0$. Vamos usar o teorema do valor médio para truncar a expansão no segundo termo.

$$\mathbf{g}(\hat{\boldsymbol{\theta}}) = \mathbf{g}\left(\boldsymbol{\theta}_0\right) + \mathbf{H}(\overline{\boldsymbol{\theta}})\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) = \mathbf{0}.$$

Normalidade Assintótica II

O Hessiano é avaliado em um ponto $\overline{\boldsymbol{\theta}}$ entre $\hat{\boldsymbol{\theta}}$ e $\boldsymbol{\theta}_0$ $\left(\overline{\boldsymbol{\theta}} = w\hat{\boldsymbol{\theta}} + (1-w)\boldsymbol{\theta}_0\right.$ para algum $0 < w < 1$). Reorganizando a função e multiplicando o resultado por $\sqrt{n}$:

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) = [-\mathbf{H}(\overline{\boldsymbol{\theta}})]^{-1}\left[\sqrt{n}\mathbf{g}\left(\boldsymbol{\theta}_0\right)\right]$$

Uma vez que plim $\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) = \mathbf{0}$, plim$(\hat{\boldsymbol{\theta}} - \overline{\boldsymbol{\theta}}) = 0$ também. As derivadas segundas são funções contínuas.

Normalidade Assintótica III

Se existir uma distribuição limite, ela seria

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) \xrightarrow{d} \left[-\mathbf{H}\left(\boldsymbol{\theta}_0\right)\right]^{-1}\left[\sqrt{n}\mathbf{g}\left(\boldsymbol{\theta}_0\right)\right]$$

Dividindo $\mathbf{H}\left(\boldsymbol{\theta}_0\right)$ e $\mathbf{g}\left(\boldsymbol{\theta}_0\right)$ por n, temos

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) \xrightarrow{d} \left[-\frac{1}{n}\mathbf{H}\left(\boldsymbol{\theta}_0\right)\right]^{-1}\left[\sqrt{n}\bar{\mathbf{g}}\left(\boldsymbol{\theta}_0\right)\right]$$

Normalidade Assintótica IV

Aplicando o Teorema Central do Limite de Lindberg-Lévy a $\left[\sqrt{n}\overline{\mathbf{g}}\left(\boldsymbol{\theta}_0\right)\right]$, uma vez que é $\sqrt{n}$ vezes a média de uma amostra aleatória.
A variância limite de $\left[\sqrt{n}\overline{\mathbf{g}}\left(\boldsymbol{\theta}_0\right)\right]$ is $-E_0\left[(1/n)\mathbf{H}\left(\boldsymbol{\theta}_0\right)\right]$, so

$$\sqrt{n}\overline{\mathbf{g}}\left(\boldsymbol{\theta}_0\right) \xrightarrow{d} N\left\{\mathbf{0}, -E_0\left[\frac{1}{n}\mathbf{H}\left(\boldsymbol{\theta}_0\right)\right]\right\}$$

Pelo Teorema de Chebyshev (Teorema D.2 do Greene),
$\operatorname{plim}\left[-(1/n)\mathbf{H}\left(\boldsymbol{\theta}_0\right)\right] = -E_0\left[(1/n)\mathbf{H}\left(\boldsymbol{\theta}_0\right)\right]$. Como é uma matriz constante, podemos reorganizar

$$\left[-\frac{1}{n}\mathbf{H}\left(\boldsymbol{\theta}_0\right)\right]^{-1}\sqrt{n}\overline{\mathbf{g}}\left(\boldsymbol{\theta}_0\right) \xrightarrow{d}$$

$$N\left[\mathbf{0}, \left\{-E_0\left[\frac{1}{n}\mathbf{H}\left(\boldsymbol{\theta}_0\right)\right]\right\}^{-1}\left\{-E_0\left[\frac{1}{n}\mathbf{H}\left(\boldsymbol{\theta}_0\right)\right]\right\}\left\{-E_0\left[\frac{1}{n}\mathbf{H}\left(\boldsymbol{\theta}_0\right)\right]\right\}^{-1}\right]$$

Normalidade Assintótica V

Voltando

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) \xrightarrow{d} N\left[\mathbf{0}, \left\{-E_0\left[\frac{1}{n}\mathbf{H}\left(\boldsymbol{\theta}_0\right)\right]\right\}^{-1}\right]$$

Que nos dá a distribuição assintótica do MLE:

$$\hat{\boldsymbol{\theta}} \overset{a}{\sim} N\left[\boldsymbol{\theta}_0, \left\{\mathbf{I}\left(\boldsymbol{\theta}_0\right)\right\}^{-1}\right]$$

Eficiência Assintótica e o Cramér-Rao Lower Bound

Theorem

Supondo que a densidade de y_i atenda às condições de regularidade $R1 - R3$, a variância assintótica de um estimador normalmente distribuído consistente e assintoticamente normalmente distribuído $\boldsymbol{\theta}_0$ será sempre pelo menos tão grande quanto:

$$[\mathbf{I}(\boldsymbol{\theta}_0)]^{-1} = \left(-E_0\left[\frac{\partial^2 \ln L(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}_0\partial\boldsymbol{\theta}_0'}\right]\right)^{-1} = \left(E_0\left[\left(\frac{\partial \ln L(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}_0}\right)\left(\frac{\partial \ln L(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}_0}\right)'\right]\right)^{-1}.$$

Variância Assintótica do ML

Se a forma dos valores esperados das derivadas segundas da log verossimilhança é conhecida, então podemos avaliar esta fórmula em $\hat{\theta}$ e encontrar a matriz VC do ML:

$$[\mathbf{I}(\boldsymbol{\theta}_0)]^{-1} = \left\{ -E_0 \left[\frac{\partial^2 \ln L(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_0 \partial \boldsymbol{\theta}_0'} \right] \right\}^{-1}$$

Como isso quase nunca é fácil, existem alternativas. A primeira delas é

$$[\hat{\mathbf{I}}(\hat{\boldsymbol{\theta}})]^{-1} = \left(- \frac{\partial^2 \ln L(\hat{\boldsymbol{\theta}})}{\partial \hat{\boldsymbol{\theta}} \partial \hat{\boldsymbol{\theta}}'} \right)^{-1}.$$

Ou seja, avaliando as derivadas segundas na função verossimilhança em torno das estimativas de ML.

Variância Assintótica do Estimador ML

Outro estimador é baseado no resultado abaixo:

$$[\hat{\hat{\mathbf{I}}}(\hat{\theta})]^{-1} = \left[\sum_{i=1}^{n} \hat{\mathbf{g}}_i \hat{\mathbf{g}}_i'\right]^{-1} = \left[\hat{\mathbf{G}}'\hat{\mathbf{G}}\right]^{-1},$$

em que

$$\hat{\mathbf{g}}_i = \frac{\partial \ln f\left(\mathbf{x}_i, \hat{\theta}\right)}{\partial \hat{\theta}}$$

e

$$\hat{\mathbf{G}} = [\hat{\mathbf{g}}_1, \hat{\mathbf{g}}_2, \ldots, \hat{\mathbf{g}}_n]'$$