Econometria I

Máxima Verossimilhança

Victor Oliveira
Núcleo de Economia Internacional e Desenvolvimento Econômico

PPGDE – 2023

Sumário I

- Um **modelo paramétrico** para X é uma função de probabilidade completa que depende de um vetor de parâmetro desconhecido θ.
- Para o caso contínuo, podemos escrever ela como uma função densidade $f(x|\theta)$. O parâmetro θ pertence ao um conjunto Θ que é chamado de **espaço de parâmetros**.
- Um modelo paramétrico especifica uma distribuição da população que pertence a uma coleção específica de distribuições. Costumamos chamar de **família paramétrica**.
- Modelo paramétrico: x é distribuído exponencialmente com densidade $f(x|\lambda) = \lambda^{-1}\exp\left(-\frac{x}{\lambda}\right)$ com parâmetro $\lambda > 0$.
- Um modelo paramétrico especifica a distribuição de todas as observações.

Definição

Um **modelo** para uma amostra aleatória é o pressuposto que x_i, $i = 1, \ldots, n$ são *i.i.d.* com função densidade conhecida $f(x|\theta)$ ou função massa $\pi(x|\theta)$ com parâmetro $\theta \in \Theta$.

Definição

Definição 2: Um modelo é corretamente especificado quando há um único valor do parâmetro $\theta \in \Theta$ tal que $f(x|\theta_0) = f(x)$, a verdadeira distribuição dos dados. O valor do parâmetro θ_0 é chamado de **verdadeiro valor do parâmetro**. O parâmetro θ é **único** se não existe outro θ tal que $f(x|\theta_0) = f(x|\theta)$. Um modelo é **mal-especificado** se não existe nenhum valor de parâmetro $\theta \in \Theta$ tal que $f(x|\theta) = f(x)$.

- A **teoria da verossimilhança** é desenvolvida sob o pressuposto que o modelo é corretamente especificado.

Definição

Um modelo é **corretamente especificado** quando há um valor único para o parâmetro $\theta_0 \in \Theta$ tal que $f(x|\theta_0) = f(x)$, a verdadeira distribuição. Este parâmetro θ_0 é chamado de **verdadeiro parâmetro**.

- O parâmetro θ_0 é **único** se não há nenhum outro θ tal que $f(x|\theta_0) = f(x)$.

- A **verossimilhança** é a densidade conjunta das observações calculadas usando o modelo. Independência das observações significa que a densidade conjunta é o produto das densidades individuais. Distribuições idênticas significa que todas as densidades são idênticas. Isso significa que a densidade conjunta é igual a seguinte expressão:

$$f(x_1, \ldots, x_n|\theta) = f(x_1|\theta)f(x_2|\theta)\ldots f(x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta) \quad (1)$$

- A **função verossimilhança** é a densidade conjunta avaliada nos dados observados e vista como função de θ.

Definição

A **função de verossimilhança** para uma variável contínua é:

$$L_n(\theta) \equiv f(x_1, \ldots, x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta) \quad (2)$$

Definição

A **função de verossimilhança** para uma variável discreta é:

$$L_n(\theta) \equiv \prod_{i=1}^{n} \pi(x_i|\theta) \tag{3}$$

- A teoria da probabilidade usa a densidade para descreve a probabilidade de x assumir valores específicos. Na análise de verossimilhança, mudamos o uso.
- À medida que os dados nos são fornecidos, usamos a função de verossimilhança para descrever quais valores de θ são mais compatíveis com os dados. **O objetivo da estimação é encontrar o valor de θ que melhor descrever os dados.**

- Como a função densidade $f(x|\theta)$ nos mostra que valores de x são mais prováveis de ocorrer, dado um valor específico de θ a função de verossimilhança $\ell_n(\theta)$ nos mostra os valores de θ que são mais prováveis de gerar as observações.
- O valor de θ mais compatível com as observações é o valor que maximiza a verossimilhança. Este é um estimador razoável de θ.

Definição

O **estimador de máxima verossimilhança** $\widehat{\theta}$ de θ é o valor que maximiza $L_n(\theta)$:

$$\widehat{\theta} = \underset{\theta \in \Theta}{\arg\max} L_n(\theta) \tag{4}$$

Exemplo

- Considere a densidade dada por $f(x|\lambda) = \frac{1}{\lambda}\exp\left(-\frac{x}{\lambda}\right)$. A função de verossimilhança é:

$$L_n(\lambda) = \prod_{i=1}^{n}\left(\frac{1}{\lambda}\exp\left(-\frac{X_i}{\lambda}\right)\right) = \frac{1}{\lambda^n}\exp\left(-\frac{n\overline{X}_n}{\lambda}\right) \quad (5)$$

- A CPO para maximização é dada por:

$$\frac{\partial L_n(\lambda)}{\partial\lambda} = 0 \Leftrightarrow \frac{-n}{\lambda^{n+1}}\exp\left(-\frac{n\overline{X}_n}{\lambda}\right) + \frac{n\overline{X}_n}{\lambda^n\lambda^2}\exp\left(-\frac{n\overline{X}_n}{\lambda}\right) = 0 \quad (6)$$

- Cancelando os termos comuns e resolvendo, encontramos uma solução úncia que é um EMV para λ:

$$\widehat{\lambda} = \overline{X}_n \quad (7)$$

- Em alguns casos é mais conveniente calcular e maximizar o logaritmo da função.
- Duas razões:
  1. É mais conveniente porque o log da verossimilhança é a soma dos log da densidade individual
  2. Em muitos modelos paramétricos o log da densidade é computacionalmente mais robusto (menos intensivo)

Definição

O **log da função de verossimilhança** é dado por:

$$\ell_n(\theta) \equiv \log L_n(\theta) = \sum_{i=1}^{n} \log f(x_i|\theta) \tag{8}$$

Teorema

$\widehat{\theta} = \underset{\theta \in \Theta}{\arg\max} \quad \ell_n(\theta)$

Exemplo

- Considere a densidade dada por $f(x|\lambda) = \frac{1}{\lambda}\exp\left(-\frac{x}{\lambda}\right)$.
- O log da densidade é $\log f(x|\lambda) = -\log\lambda - \frac{x}{\lambda}$.
- O log da função de verossimilhança é:

$$\ell_n(\theta) \equiv \log L_n(\theta) = \sum_{i=1}^{n}\left(-\log\lambda - \frac{X_i}{\lambda} = -n\log\lambda - \frac{n\overline{X}_n}{\lambda}\right) \tag{9}$$

- A CPO é dada por:

$$\frac{\partial \ell_n(\lambda)}{\partial \lambda} = 0 \quad \Longleftrightarrow \quad -\frac{n}{\lambda} + \frac{n\overline{X}_n}{\lambda^2} = 0 \tag{10}$$

- A solução é única e dada por $\widehat{\lambda} = \overline{X}_n$.
- A CSO é:

$$\frac{\partial^2 \ell_n(\lambda)}{\partial \lambda^2} = \frac{n}{\widehat{\lambda}^2} - 2\frac{n\overline{X}_n}{\widehat{\lambda}^3} = -\frac{n}{\overline{X}_n^2} < 0 \tag{11}$$

Exemplo

- Considere a função massa dada por $\pi(x|p) = p^x(1-p)^{1-x}$. O log da função massa é $\log \pi(x) = x \log p + (1-x)\log(1-p)$. O log da função de verossimilhança é

$$
\begin{aligned}
\ell_n(p) &= \sum_{i=1}^{n} X_i \log p + (1 - X_i)\log(1-p) \\
&= n\overline{X}_n \log p + n(1-\overline{X}_n)\log(1-p)
\end{aligned}
\tag{12}
$$

- A CPO da eq. (12) é dada por:

$$
\frac{\partial \ell_n(p)}{\partial p} = 0 \quad \Longleftrightarrow \quad \frac{n\overline{X}_n}{p} - \frac{n(1-\overline{X}_n)}{1-p} = 0 \tag{13}
$$

- A solução é única e dada por $\widehat{p} = \overline{X}_n$.

- A condição de segunda ordem é:

$$\frac{\partial^2 \ell_n(\widehat{p})}{\partial p^2} = -\frac{n\overline{X}_n}{\widehat{p}^2} - \frac{n(1-\overline{X}_n)}{(1-\widehat{p})^2} = -\frac{n}{\overline{X}_n} < 0 \qquad (14)$$

EMV

- Agora vamos mostrar que o EMV é um análogo amostral do verdadeiro parâmetro. Vamos definir a esperança do log da função densidade:

$$\ell_n(\lambda) = \mathbb{E}\left[\log f(x|\theta)\right] \tag{15}$$

que é uma função do parâmetro θ.

Teorema

Quando o modelo é corretamente especificado o parâmetro verdadeiro θ_0 maximiza a esperança do log da densidade $\ell_n(\theta)$.

$$\theta_0 = \underset{\theta\in\Theta}{argmax} \;\; \ell(\theta) \tag{16}$$

Exemplo

- Considere a função massa dada por $\pi(x|p) = p^x(1-p)^{1-x}$. O log da função massa é $\log \pi(x|p) = x \log p + (1-x)\log(1-p)$ que tem valor esperado

$$
\begin{aligned}
\ell_n(p) &= \mathbb{E}\left[\log \pi(X|p)\right] \\
&= \mathbb{E}\left[X \log p + (1-X)\log(1-p)\right] \\
&= p_0 \log p + (1-p_0)\log(1-p) \qquad (17)
\end{aligned}
$$

- A CPO da eq. (17) é dada por:

$$
\frac{p_0}{p} - \frac{1-p_0}{1-p} = 0 \qquad (18)
$$

- A solução é única e dada por $p = p_0$.
- A condição de segunda ordem é negativa. Portanto, o verdadeiro parâmetro p_0 é o máximo de $\mathbb{E}\left[\log \pi(X|p)\right]$.

Distribuição normal com σ^2 conhecido

- A densidade de x é

$$f(x|\mu) = \frac{1}{(2\pi\sigma_0^2)^{1/2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma_0^2}\right) \tag{19}$$

- O log da densidade é

$$\log f(x|\mu) = -\frac{1}{2}\log(2\pi\sigma_0^2) - \left(\frac{(x-\mu)^2}{2\sigma_0^2}\right) \tag{20}$$

- Assim,

$$\begin{aligned} \ell(\mu) &= -\frac{1}{2}\log(2\pi\sigma_0^2) - \left(\frac{\mathbb{E}[(x-\mu)^2]}{2\sigma_0^2}\right) \\ &= -\frac{1}{2}\log(2\pi\sigma_0^2) - \frac{(\mu_0-\mu)^2}{2\sigma_0^2} \end{aligned} \tag{21}$$

- Para obter o EMV consideramos as seguintes etapas:
  - Construir $f(x|\theta)$ como uma função de x e θ;
  - Tomar o log da função densidade: $\log f(x|\theta)$;
  - Avaliar em $x = X_i$ e somar em i: $\ell_n(\theta) = \sum_{i=1}^{n} \log f(x|\theta)$;
  - Resolver a CPO para encontrar o máximo;
  - Checar a CSO para verificar que é um máximo.

Distribuição normal com σ^2 conhecido

- A densidade de x é

$$f(x|\mu) = \frac{1}{(2\pi\sigma_0^2)^{1/2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma_0^2}\right) \tag{22}$$

- O log da densidade é

$$\log f(x|\mu) = -\frac{1}{2}\log(2\pi\sigma_0^2) - \left(\frac{(x-\mu)^2}{2\sigma_0^2}\right) \tag{23}$$

- O log da verossimilhança é dado por:

$$\ell_n(\mu) = -\frac{n}{2}\log(2\pi\sigma_0^2) - \frac{1}{2\sigma_0^2}\sum_{i=1}^{n}(x_i-\mu)^2 \tag{24}$$

- A CPO para $\widehat{\mu}$ é:

$$\frac{\partial}{\partial \mu}\ell_n(\widehat{\mu}) = \frac{1}{\sigma_0^2}\sum_{i=1}^{n}(x_i - \mu) \tag{25}$$

- A solução é dada por

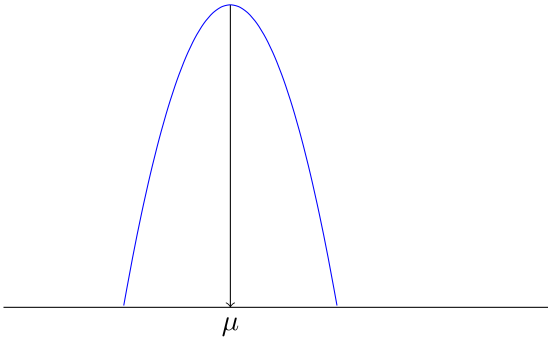
$$\widehat{\mu} = \overline{x}_n \tag{26}$$

- A CSO é dada por

$$\frac{\partial^2}{\partial \mu^2}\ell_n(\widehat{\mu}) = -\frac{n}{\sigma_0^2} < 0 \tag{27}$$

- Abaixo o log da verossimilhança para $\overline{x}_n = 1$. O EMV é indicado pela flecha.

Figura 1: Log da Função de Verossimilhança



Distribuição normal com μ conhecida

- A densidade de x é

$$f(x|\sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left(-\frac{(x-\mu_0)^2}{2\sigma^2}\right) \tag{28}$$

- O log da densidade é

$$\log f(x|\sigma^2) = -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log(\sigma^2) - \frac{(x-\mu_0)^2}{2\sigma^2} \tag{29}$$

- O log da verossimilhança é dado por:

$$\ell_n(\sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{\sum_{i=1}^{n}(x_i-\mu_0)^2}{2\sigma^2} \tag{30}$$

- A CPO para $\widehat{\sigma^2}$ é:

$$-\frac{n}{2\widehat{\sigma^2}} + \frac{\sum_{i=1}^{n}(x_i - \mu_0)^2}{2\left(\widehat{\sigma^2}\right)^2} = 0 \tag{31}$$

- A solução é dada por

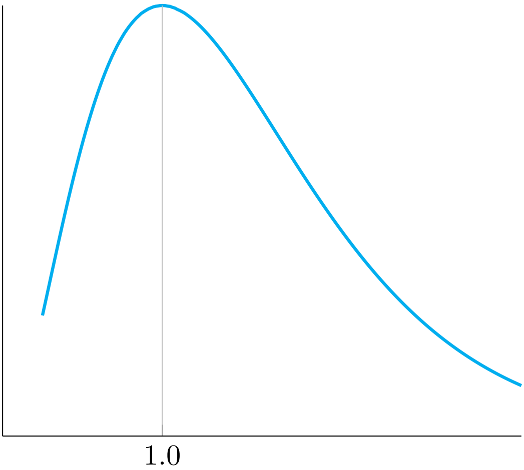
$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_0)^2 \tag{32}$$

- A CSO é dada por

$$\frac{\partial^2 \ell_n(\widehat{\sigma}^2)}{\partial(\sigma^2)^2} = \frac{n}{2\left(\widehat{\sigma^2}\right)^2} - \frac{\sum_{i=1}^{n}(x_i - \mu_0)^2}{\left(\widehat{\sigma^2}\right)^3} = -\frac{n}{2\widehat{\sigma}^3} < 0 \tag{33}$$

- Abaixo o log da verossimilhança para $\widehat{\sigma}^2 = 1$. O EMV é indicado pela flecha.

Figura 2: Log da Função de Verossimilhança



Função Score

- Considere a função log verossimilhança dada por:

$$\ell_n(\theta) = \sum_{i=1}^{n} \log f(x_i|\theta) \tag{34}$$

- Vamos assumir que $\log f(x_i|\theta)$ é diferenciável com respeito a θ. O **score da verossimilhança** é a derivada da função de verossimilhança. Quando θ é um vetor o score eficiente é um vetor de derivadas parciais

$$S_n(\theta) = \frac{\partial}{\partial\theta}\ell_n(\theta) = \sum_{i=1}^{n} \frac{\partial}{\partial\theta}\log f(x_i|\theta) \tag{35}$$

- O escore nos diz quão sensível é o log-verossimilhança ao vetor de parâmetros. Uma propriedade algébrica é que ele é zero no EMV: $S_n(\widehat{\theta}) = 0$ quando $\widehat{\theta}$ é uma solução interior.

- Vamos assumir que θ inclui um total de p parâmetros $\theta = (\theta_1, \ldots, \theta_p)'$. Assim definimos $S(\theta; y)$ como vetor coluna $(p \times 1)$ com a seguinte propriedade:

$$S(\theta; y) = \frac{\partial \ell(\theta; y)}{\partial \theta} = \begin{pmatrix} \dfrac{\partial \ell(\theta; y)}{\partial \theta_1} \\ \\ \dfrac{\partial \ell(\theta; y)}{\partial \theta_2} \\ \vdots \\ \dfrac{\partial \ell(\theta; y)}{\partial \theta_p} \end{pmatrix} \tag{36}$$

Matriz Hessiana

- O hessiano da verossimilhança é a segunda derivada negativa da função de verossimilhança. Quando θ é um vetor, o hessiano é uma matriz das segundas derivadas parciais:

$$H_n(\theta) = -\frac{\partial^2}{\partial\theta\partial\theta'}\ell_n(\theta) = -\sum_{i=1}^{n} \frac{\partial^2}{\partial\theta\partial\theta'} \log f(x_i|\theta) \tag{37}$$

- O hessiano indica o grau de curvatura no log-verossimilhança. Valores grandes indica que a verossimilhança é mais curva, enquanto valores menores indicam que a verossimilhança é mais achatada.

- Definimos a matriz hessiana $H(\theta; y)$ como uma matriz de dimensão $p \times p$ com a seguinte propriedade:

$$H(\theta; y) = \frac{\partial^2 \ell(\theta; y)}{\partial\theta\partial\theta'} = \begin{pmatrix} \frac{\partial^2 \ell(\theta; y)}{\partial\theta_1\partial\theta_1} & \frac{\partial^2 \ell(\theta; y)}{\partial\theta_1\partial\theta_2} & \cdots & \frac{\partial^2 \ell(\theta; y)}{\partial\theta_1\partial\theta_p} \\ \frac{\partial^2 \ell(\theta; y)}{\partial\theta_2\partial\theta_1} & \frac{\partial^2 \ell(\theta; y)}{\partial\theta_2\partial\theta_2} & \cdots & \frac{\partial^2 \ell(\theta; y)}{\partial\theta_2\partial\theta_p} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 \ell(\theta; y)}{\partial\theta_p\partial\theta_1} & \frac{\partial^2 \ell(\theta; y)}{\partial\theta_p\partial\theta_2 1} & \cdots & \frac{\partial^2 \ell(\theta; y)}{\partial\theta_p\partial\theta_p} \end{pmatrix} \tag{38}$$

- Duas observações a respeito da matriz hessiana:
    1. A matriz hessiana é por definição simétrica já que as derivadas cruzadas são invariantes a ordem de diferenciação.
    2. Se a função log-verossimilhança é côncava em θ, $H(\theta; y)$ é dita para ser negativa definida. Para o caso em que $p = 1$, temos que a segunda derivada da função log-verossimilhança é negativa.

Score Eficiente

- É a derivada da função de log-verossimilhança para uma simples observação, avaliada no vetor aleatório x e dado o verdadeiro parâmetro

$$S = \frac{\partial}{\partial\theta} \log f(x|\theta_0) \tag{39}$$

- O escore eficiente tem um importante papel na distribuição assintótica.

Teorema

Se o modelo está corretamente especificado, o suporte de x não depende de θ, e θ_0 está no interior de Θ. Então o escore eficiente S satisfaz $\mathbb{E}[S] = 0$.

Informação de Fisher

- É a variância do escore eficiente

$$\mathfrak{I}_\theta = \mathbb{E}[SS'] \tag{40}$$

Hessino Esperado

- É dado por

$$H_\theta = -\frac{\partial^2}{\partial\theta\partial\theta'}\ell(\theta_0) \tag{41}$$

- Quando $f(x|\theta)$ é duas vezes diferenciável em θ e o suporte de x não depende de θ, o hessiano esperado iguala a esperança do Hessiano da verossimilhança para uma simples observação, isto é,

$$H_\theta = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta\partial\theta'}\log f(x|\theta)\right] \tag{42}$$

Teorema

Se o modelo está corretamente especificado e o suporte de x não depende de θ, então a informação de Fisher iguala ao hessiano esperado: $\mathfrak{I}_\theta = H_\theta$.

- O teorema nos informa que a curvatura na função de verossimilhança e a variância do escore são idênticas.
- É importante principalmente porque é usado para simplificar a fórmula para a variância assintótica do EMV e similarmente para a estimação da variância assintótica.

Limite Inferior de Cramér-Rao

- A matriz de informação fornece um limite inferior para a variância entre os estimadores não viesados.

Teorema (eorema para o Limite Inferior de Cramér-Rao (LIC-R))

Supondo que o modelo está corretamente especificado, que o suporte de x não depende de θ e que θ_0 está contido no interior de Θ, se $\tilde{\theta}$ é um estimador não viesado de θ, então $var[\tilde{\theta}] \geqslant (n\mathfrak{I}_\theta)^{-1}$, em que o Limite Inferior de Cramér-Rao é $(n\mathfrak{I}_\theta)^{-1}$.

- Um estimador $\tilde{\theta}$ é dito Cramér-Rao eficiente se ele é não viesado para θ e $\text{var}[\tilde{\theta}] = (n\mathfrak{I}_\theta)^{-1}$.
- O Limite Inferior de Cramér-Rao diz que na classe dos estimadores não viesados, a menor variância possível é a inversa da informação de Fisher escalonada pelo tamanho da amostra. **Assim, a informação de Fisher fornece um limite sobre a precisão da estimação.**

$X \sim N(\mu, \sigma^2)$ com σ^2 conhecida

- A segunda derivada do log da densidade é:

$$\frac{\partial^2 \log f(x|\mu)}{\partial \mu^2} = \frac{\partial^2}{\partial \mu^2}\left(-\frac{\log(2\pi\sigma^2)}{2} - \frac{(x-\mu)^2}{2\sigma^2}\right) = -\frac{1}{\sigma^2} \quad (43)$$

- Portanto $\mathfrak{I}_\mu = \sigma^{-2}$.
- Assim, o Limite Inferior de Cramér-Rao é $\dfrac{\sigma^2}{n}$. O EMV é $\widehat{\mu} = \overline{x}_n$, que é não viesado e possui variância $\text{var}[\widehat{\mu}] = \dfrac{\sigma^2}{n}$ que se iguala ao LIC-R. Portanto o EMV é Cramér-Rao eficiente.

$X \sim N(\mu, \sigma^2)$ com μ e σ^2 desconhecidos

- Precisamos calcular a matriz de informação para o vetor de parâmetros $\theta = (\mu, \sigma^2)$. O log da densidade é:

$$\log f(x|\mu, \sigma^2) = -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log(\sigma^2) - \frac{(x-\mu)^2}{2\sigma^2} \tag{44}$$

- As primeiras derivadas são

$$\frac{\partial}{\partial\mu}\log f(x|\mu, \sigma^2) = \frac{x-\mu}{\sigma^2} \tag{45}$$

$$\frac{\partial}{\partial\sigma^2}\log f(x|\mu, \sigma^2) = -\frac{1}{2\sigma^2} + \frac{(x-\mu)^2}{2(\sigma^2)^2} \tag{46}$$

- As segundas derivadas são

$$\frac{\partial^2}{\partial\mu^2}\log f(x|\mu,\sigma^2) = -\frac{1}{\sigma^2} \tag{47}$$

$$\frac{\partial}{\partial(\sigma^2)^2}\log f(x|\mu,\sigma^2) = \frac{1}{2(\sigma^2)^2} - \frac{(x-\mu)^2}{(\sigma^2)^3} \tag{48}$$

$$\frac{\partial}{\partial\mu\partial\sigma^2}\log f(x|\mu,\sigma^2) = -\frac{x-\mu}{(\sigma^2)^2} \tag{49}$$

- A matriz de informação de Fisher esperada é

$$\mathfrak{I}_{\theta} = -\mathbb{E}\begin{bmatrix} -\dfrac{1}{\sigma^2} & -\dfrac{x-\mu}{\left(\sigma^2\right)^2} \\[2ex] -\dfrac{x-\mu}{\left(\sigma^2\right)^2} & \dfrac{1}{2(\sigma^2)^2} - \dfrac{(x-\mu)^2}{(\sigma^2)^3} \end{bmatrix} \\ = \begin{bmatrix} \dfrac{1}{\sigma^2} & 0 \\[2ex] 0 & \dfrac{1}{2\left(\sigma^2\right)^2} \end{bmatrix} \tag{50}$$

- O limite inferior é

$$LICR = (n\mathfrak{I}_\theta)^{-1} = \begin{bmatrix} \dfrac{\sigma^2}{n} & 0 \\ \\ 0 & \dfrac{2\left(\sigma^2\right)^2}{n} \end{bmatrix} \tag{51}$$

- Dois aspectos são importantes nesse resultado:
  - A matriz de informação é diagonal. Isto significa que a informação de μ e σ^2 são não relacionadas.
  - Os termos da diagonal principal são idênticos ao LIC-R dos casos mais simples em que σ^2 e μ são conhecidos.

Econometria I

Máxima Verossimilhança

Victor Oliveira
Núcleo de Economia Internacional e Desenvolvimento Econômico

PPGDE – 2023