Econometria I

Estimador de Máxima Verossimilhança

Victor Oliveira
Núcleo de Economia Internacional e Desenvolvimento Econômico

PPGDE – 2023

Sumário I

Modelo Linear Normal

- Um exemplo importante da estimação de máxima verossimilhança de um modelo de resposta contínua é o modelo linear normal.
- O modelo de regressão linear é usualmente escrito como

$$y_i = x_i'\beta + u_i \quad i = 1, \ldots, n \tag{1}$$

em que $x_i = (1, x_{i1}, \ldots, x_{ik})'$ é um vetor de variáveis explicativas com dimensão $(k+1) \times 1$ e β é um vetor de parâmetros.

- Sob os pressupostos de $\mathbb{E}(x_i u_i) = 0$ e $\mathbb{E}(u_i^2|x_i) = \sigma^2$, os parâmetros da regressão do modelo podem ser estimados por MQO e o estimador resultante é BLUE.

- Vamos assumir que u_i é normalmente distribuído com média zero e variância σ^2.
- O resultado é um **modelo linear normal** cujo formato segue os pressupostos acima

$$y_i|x_i \sim \mathcal{N}(x_i'\beta, \sigma^2) \tag{2}$$

- Vejamos como os parâmetros desse modelo, β e σ^2, podem ser estimados pelo método de máxima verossimilhança.

- A função densidade para cada observação pode ser escrita explicitamente como

$$f(y_i|x_i;\beta,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}}\exp\left[-\frac{1}{2}\left(\frac{(y_i - x_i'\beta)^2}{\sigma^2}\right)\right] \tag{3}$$

- Assumindo uma amostra aleatória de n pares de observações (y_i, x_i), o log da função de verossimilhança é

$$\ell(\beta,\sigma^2;y_i|x_i) = \sum_{i=1}^{n} \log f(y_i|x_i;\beta,\sigma^2) \tag{4}$$

- Podemos escrever como

$$
\begin{aligned}
\ell(\beta, \sigma^2; y_i|x_i) &= \sum_{i=1}^{n} \left[-\frac{1}{2}\log(2\pi) - \frac{1}{2}\log(\sigma^2) - \frac{1}{2}\left(\frac{(y_i - x_i'\beta)^2}{\sigma^2} \right) \right] \\
&= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - x_i'\beta)^2
\end{aligned}
\tag{5}
$$

- A CPO para maximizar esta log verossimilhança é dada por:

$$
\frac{\partial \ell}{\partial \beta} = \frac{1}{\sigma^2}\sum_{i=1}^{n} x_i(y_i - x_i'\beta) = 0 \tag{6}
$$

$$
\frac{\partial \ell}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(y_i - x_i'\beta)^2 = 0 \tag{7}
$$

- A derivada parcial $\dfrac{\partial \ell}{\partial \beta}$ é um vetor de dimensão $(k+1) \times 1$:
    1. O primeiro elemento é: $\dfrac{\partial \ell}{\partial \beta_0} = \sigma^{-2} \sum_{i=1}^{n} (y_i - x_i' \beta)$
    2. o segundo elemento é: $\dfrac{\partial \ell}{\partial \beta_1} = \sigma^{-2} \sum_{i=1}^{n} x_{i1}(y_i - x_i' \beta)$ e assim por diante

- O estimador de máxima verossimilhança de $\widehat{\beta}$ pode ser obtido da equação (6):

$$\sum_{i=1}^{n} x_i y_i = \sum_{i=1}^{n} x_i x_i' \widehat{\beta} \tag{8}$$

tal que

$$\widehat{\beta}_{MV} = \left(\sum_{i=1}^{n} x_i x_i'\right)^{-1} \left(\sum_{i=1}^{n} x_i y_i\right) = \widehat{\beta}_{MQ} \tag{9}$$

- O estimador de MV para $\widehat{\beta}_{MV}$ na eq. (9) é o mesmo obtido obtido por meio de MQO. Para o vetor de coeficientes, não há diferença entre estimação por MV e por MQO.
- Para encontrar o estimador para σ^2, basta substituir β na eq. (8) pelo seu estimador $\widehat{\beta}$ de MV e definir o resíduo, $\widehat{u}_i = y_i - x_i'\widehat{\beta}$. Assim,

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\widehat{u}_i^2 \tag{10}$$

- Note que a expressão (10) difere do familiar estimador de variância uma vez que no denominador temos n e não $n-k-1$.
- Por isso, **o estimador da variância é viesado para pequenas amostras. Contudo, para amostras grandes, o viés se torna irrelevante**.

- Para obter a matriz de informação precisamos das condições de segunda ordem:

$$\frac{\partial^2 \ell}{\partial \beta \partial \beta'} = -\frac{1}{\sigma^2} \sum_{i=1}^{n} x_i x_i' \tag{11}$$

$$\frac{\partial^2 \ell}{\partial (\sigma^2)^2} = \frac{n}{2\sigma^4} - \frac{1}{\sigma^6} \sum_{i=1}^{n} (y_i - x_i'\beta)^2 \tag{12}$$

$$\frac{\partial^2 \ell}{\partial \beta \partial \sigma^2} = -\frac{1}{\sigma^4} \sum_{i=1}^{n} x_i (y_i - x_i'\beta) \tag{13}$$

- Por causa da simetria, $\dfrac{\partial^2 \ell}{\partial \beta \partial \sigma^2} = \dfrac{\partial^2 \ell}{\partial \sigma^2 \partial \beta}$.

- A partir da condição de segunda ordem, podemos construir a matriz Hessiana a seguir:

$$H(\beta,\sigma^2;y,x)=\begin{bmatrix} -\dfrac{\sum\limits_{i=1}^{n} x_i x_i'}{\sigma^2} & -\dfrac{\sum\limits_{i=1}^{n} x_i(y_i - x_i'\beta)}{\left(\sigma^2\right)^2} \\ \\ -\dfrac{\sum\limits_{i=1}^{n} x_i(y_i - x_i'\beta)}{\left(\sigma^2\right)^2} & \dfrac{n}{2\left(\sigma^2\right)^2} - \dfrac{\sum\limits_{i=1}^{n} (y_i - x_i'\beta)^2}{\left(\sigma^2\right)^3} \end{bmatrix} \quad (14)$$

- A matriz hessinana é negativa definida e os x_i's são bem comportados e não colinear. O log da função de verossimilhança do modelo linear normal é globalmente côncavo, e $\widehat{\beta}$ e $\widehat{\sigma}^2$ são de fato os valores que maximizam.
- A matriz de informação contém o negativo dos valores esperados da matriz hessiana.

- Assim,

$$H(\beta, \sigma^2) = \mathfrak{I}_\theta = \begin{bmatrix} \dfrac{1}{\sigma^2} \sum_{i=1}^{n} x_i x_i' & 0 \\ 0 & \dfrac{n}{2\sigma^4} \end{bmatrix} \tag{15}$$

- A sua inversa fornece a matriz de covariância assintótica do estimador de máxima verossimilhança no modelo de regressão linear normal.

$$\left[H(\beta, \sigma^2)\right]^{-1} = \mathfrak{I}_\theta^{-1} = \begin{bmatrix} \sigma^2 \left(\sum_{i=1}^{n} x_i x_i' \right)^{-1} & 0 \\ 0 & \dfrac{2\sigma^4}{n} \end{bmatrix} = \begin{bmatrix} \mathrm{var}[\widehat{\beta}] & 0 \\ 0 & \mathrm{var}[\widehat{\sigma}^2] \end{bmatrix} \tag{16}$$

- Finalmente, temos que

$$\begin{pmatrix} \widehat{\beta} \\ \widehat{\sigma}^2 \end{pmatrix} \overset{approx}{\sim} \mathcal{N}\left[\begin{pmatrix} \beta \\ \sigma^2 \end{pmatrix}, \begin{pmatrix} \sigma^2 \left(\sum_{i=1}^{n} x_i x_i'\right)^{-1} & 0 \\ 0 & \dfrac{2\sigma^4}{n} \end{pmatrix}\right] \tag{17}$$

- **Observação**: sob os pressupostos do modelo linear normal, MQO e MV fornecem o mesmo estimador para β, enquanto o estimador para σ^2 difere. O estimador $\widehat{\sigma}^2_{MV}$ é viesado para pequenas amostras, mas consistente e assintoticamente eficiente.

Estimador de Máxima Verossimilhança

- O estimador de máxima verossimilhança (EMV) é:
  1. Consistente
  2. Assintoticamente normal
  3. Eficiente.
- **Observação:** Pouco da validade geral pode ser dito sobre as propriedades do EMV para pequenas amostras.

Escore Eficiente

- Uma propriedade crucial do método de MV é que $\mathbb{E}[S(\theta); y]$, o escore eficiente (ou escore esperado), se avaliado no verdadeiro parâmetro θ_0, é igual a zero.
- Se $\mathbb{E}[S(\theta); y]$ é um vetor, isto significa que cada elemento do vetor é igual a zero.
- Como iremos ver, essa propriedade do escore eficiente zero implica consistência do estimador de máxima verossimilhança.

Consistência

- É intuitivamente claro porque o EMV é consistente.
- Sob uma amostra aleatória, a função escore é a soma de componentes independentes. Pela LGN, o escore da amostra converge em probabilidade para o seu valor esperado quando a amostra aumenta.
- A regra da MV recomenda escolher o EMV tal que o escore da amostra seja igual a zero.
- Uma vez que a condição de escore eficiente é satisfeita no verdadeiro valor, deve ser o caso em que, no limite $\widehat{\theta} = \theta_0$.

Igualdade da Matriz de Informação

- Para irmos além da consistência e analisar a variância e o limite em distribuição do EMV, precisamos do resultado da **matriz de informação de Fisher** e a sua relação com a segunda derivada da função log-verissimilhança, uma matriz se θ for um vetor.

$$H(\theta; y) = \frac{\partial^2 \log L(\theta; y)}{\partial\theta\partial\theta'} = \sum_{i=1}^{n} \frac{\partial^2 \log f(y_i; \theta)}{\partial\theta\partial\theta'} \tag{18}$$

- *A matriz de informação de uma amostra é simplesmente definida como o negativo da esperança da matriz hessiana.*

- Em termos algébricos

$$I(\theta) = -\mathbb{E}[H(\theta; y)] \tag{19}$$

- A matriz de informação, $I(\theta)$, é importante de várias formas para o desenvolvimento da metodologia de MV.
    1. $I(\theta)$ pode ser usada para avaliar se a função de verossimilhança é "bem comportada";
    2. $I(\theta)$ é o inverso da variância do estimador de máxima verossimilhança;
    3. A $I(\theta)$ conecta resultados da estimação de máxima verossimilhança a um importante resultado sobre precisão dos estimadores, o Limite Inferior de Cramér-Rao.

Eficiência Assintótica

- Como uma consequência, uma vez que o EMV alcança o limite inferior de Cramér-Rao, ele é **assintoticamente eficiente**.
- Um resultado final referente a $I(\theta)$ é a chamada igualdade da matriz de informação. Esta igualdade estabelece que a matriz de informação pode ser derivada de duas formas, mas ambos avaliado no verdadeiro parâmetro θ_0:
  1. Negativo do hessiano esperado: $I(\theta) = -\mathbb{E}[H(\theta; y)]$
  2. Variância da função escore: $\text{var}[S(\theta_0; y)] = -\mathbb{E}[H(\theta; y)]$

Distribuição Assintótica

- As três propriedades do estimador de máxima verossimilhança, consistência, eficiência e normalidade assintótica, podem ser sintetizadas em um único resultado sobre a convergência em distribuição.
- Seja $\widehat{\theta}$ o estimador de máxima verossimilhança, θ o verdadeiro parâmetro e $I(\theta)$ a matriz de informação da amostra. Então

$$\sqrt{n}(\widehat{\theta} - \theta) \xrightarrow{D} \mathcal{N}\left(0, nI(\theta)^{-1}\right) \tag{20}$$

- Para amostras grandes porém finitas, podemos aproximar a distribuição de $\widehat{\theta}$ como

$$\widehat{\theta} \overset{app}{\sim} \mathcal{N}\left(\theta, -\mathbb{E}[H(\theta; y)]^{-1}\right) \tag{21}$$

- Com isso, o estimador de máxima verossimilhança é **normalmente distribuído**.
- Como o valor esperado do limite em distribuição é o verdadeiro parâmetro θ_0, o estimador de máxima verossimilhança é **consistente**;
- Uma vez que sua variância assintótica é o inverso da matriz de informação, o estimador de máxima verossimilhança é **eficiente**.

Maximização da Verossimilhança

- O sistema de equações gerado a partir das condições de primeira ordem é quase sempre não-linear. Isso obriga que a maximização seja realizada por algum processo numérico.
- Os procedimentos de otimização numéricos funcionam de forma recursiva, sendo o valor dos parâmetros na iteração $t+1$ uma função dos valores deste na iteração t.
- O algoritmo numérico consiste em tentar um valor para o parâmetro, e depois corrigi-lo continuamente até que algum critério de convergência seja atendido, quando então tem-se um máximo para a função de verossimilhança.
- Há casos em que não ocorre convergência, o processo de iteração tem de ser interrompido.

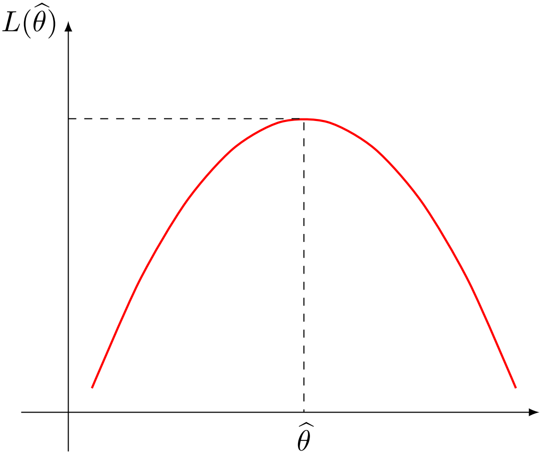
- As recursões são em geral da forma:

$$\widehat{\theta}_{t+1} = \widehat{\theta}_t + \lambda_t d_t \tag{22}$$

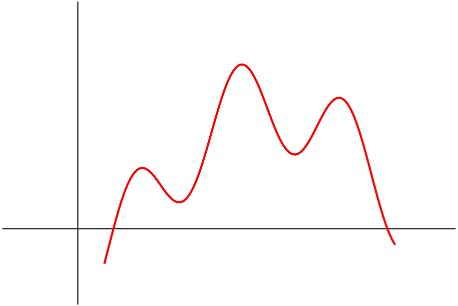
em que

$$\lambda_t d_t = \lambda_t \left[I^*(\widehat{\theta}) \right]^{-1} \left(\frac{\partial L}{\partial \theta} \Big|_{\theta = \widehat{\theta}} \right) \tag{23}$$

- Os diferentes métodos de otimização numérica variam quanto a forma de $I^*(\widehat{\theta})$, que pode ser a matriz de informação, como é o caso no método de score, ou alguma aproximação dela.
- Muitos problemas podem surgir a depender da forma da função de verossimilhança.



É o formato ideal para a função de verossimilhança.



Situação em que existe máximo local. Neste caso o algoritmo pode ficar preso no máximo local.

Algoritmo de Newton-Raphson

- Um método muito utilizado de aproximação quadrática é o método de Newton-Raphson.
- Dado uma estimativa inicial de um parâmetro, tipo $\widehat{\theta}^0$, podemos obter uma aproximação de segunda ordem do $\log(\theta)$ ao redor de $\widehat{\theta}^0$:

$$\begin{aligned} \log L^*(\theta) &= \log L\left(\widehat{\theta}^0\right) + S(\widehat{\theta}^0)'\left(\theta - \widehat{\theta}^0\right) \\ &+ \frac{1}{2}\left(\theta - \widehat{\theta}^0\right)' H\left(\widehat{\theta}^0\right)\left(\theta - \widehat{\theta}^0\right) \\ &\approx \log L(\theta) \end{aligned} \tag{24}$$

em que $S(\cdot)$ denota o escore e $H(\cdot)$ é o hessiano da função de verossimilhança. Podemos maximizar $\log L^*$ com respeito a θ, obtendo um novo parâmetro que chamaremos de $\widehat{\theta}^1$.

- A CPO deste problema é:

$$S\left(\widehat{\theta}^{0}\right)+H\left(\widehat{\theta}^{0}\right)\left(\widehat{\theta}^{1}-\widehat{\theta}^{0}\right)=0 \tag{25}$$

ou

$$\widehat{\theta}^{1}=\widehat{\theta}^{0}-\left[H\left(\widehat{\theta}^{0}\right)\right]^{-1}S\left(\widehat{\theta}^{0}\right) \tag{26}$$

- Assim, para um valor arbitrário $\widehat{\theta}^{0}$, a regra de atualização de Newton-Raphson é dada por:

$$\widehat{\theta}^{t+1}=\widehat{\theta}^{t}-\left[H\left(\widehat{\theta}^{t}\right)\right]^{-1}S\left(\widehat{\theta}^{t}\right) \qquad t=0,1,\ldots \tag{27}$$

- O processo iterativo finaliza quando um **critério de convergência** pré-determinado é satisfeito.
- Possíveis critérios:
  1. As mudanças no valor da estimativa $\left(\widehat{\theta}^{t+1} - \widehat{\theta}^{t}\right)$;
  2. As mudanças no valor da função de log-verossimilhança $\left(\log L\left(\widehat{\theta}^{t+1}\right) - \log L\left(\widehat{\theta}^{t}\right)\right)$;
  3. Valor do gradiente na estimativa $S\left(\widehat{\theta}^{t}\right)$.
- Para ver como isso estaria funcionando, vamos fazer um exemplo para um polinômio de ordem 3.

Exemplo

- Suponha que desejamos encontrar um máximo da função: $f(x) = x^3 - 3x + 1$.
- Assim

$$f(x)' = 3x^2 - 3 \tag{28}$$
$$f^{''}(x) = 6x \tag{29}$$

- Resolvendo a CPO temos dois candidatos: -1 e $+1$.
- *Como $f^{''}(-1) < 0$, temos que o primeiro dos dois candidatos é um* ***máximo local*** *da função.*
- *Para $f^{''}(+1) > 0$, temos que o segundo dos dois candidatos é um* ***mínimo local*** *da função.*

- Vamos usar o algoritmo de Newton-Raphson e a regra de atualização dada anteriormente dada por:

$$x_{t+1} = x_t - \frac{f'(x_t)}{f''(x_t)} \tag{30}$$

$$= x_t - \frac{3x_t^2 - 3}{6x_t} \tag{31}$$

- Iniciando com $x_0 = -2$ obtemos uma sequência atualizada: -2, $-1,25$, $-1,025, \ldots$ que converge rapidamente para o máximo local.

$$x_{t+1} = -2 - \frac{3(-2)^2 - 3}{6(-2)} = -1,25 \tag{32}$$

$$x_{t+2} = -1,25 - \frac{3(-1,25)^2 - 3}{6(-1,25)} = -1,025 \tag{33}$$

$$x_{t+3} = -1,025 - \frac{3(-1,025)^2 - 3}{6(-1,025)} = -1,00 \tag{34}$$

- Se iniciarmos a iteração com $x_0 = 2$, o algoritmo não encontra o máximo local, mas finaliza no mínimo local.

Testes de Hipóteses

- Estamos interessado em testar um vetor de restrições nos parâmetros. A hipótese nula pode ser representada como:

$$H_0 \colon h(\theta) = 0 \tag{35}$$

- Esta notação cobre hipóteses mais simples como $\theta_1 = 0$, restrições lineares, como $\theta_1 + \theta_2 - 1 = 0$, e restrições não lineares, tais como $\theta_3 + \theta_1\theta_2 = 0$.
- Com base no princípio da máxima verossimilhança, há três testes distintos, porém assintoticamente equivalentes.

- Teste da Razão da Verossimilhança (LR).
- Teste de Wald.
- Teste do Multiplicador de Lagrange ou Escore Eficiente (LM).
- A escolha entre esses três testes vai depender, em cada caso, do conhecimento de suas propriedades para pequena amostra, quando disponível, e da conveniência computacional

Teste da Razão da Verossimilhança

- **Condição**: o teste **LR** requer a estimação do modelo restrito e sem restrição.
- Vamos definir $\widetilde{\theta}$ como o vetor de parâmetros restrito e $\widehat{\theta}$ como o vetor de parâmetros não restrito.
- A **hipótese a ser testada** é: $H_0 : h\left(\widetilde{\theta}\right) = 0$.
- Precisamos calcular o valor da função de verossimilhança no ponto de máximo **com e sem restrição**: $\boldsymbol{L(\widetilde{\theta})}$ e $\boldsymbol{L(\widehat{\theta})}$.
- **Intuição**: se a restrição for verdadeira, o valor da função de verossimilhança avaliada em $\widetilde{\theta}$ e $\widehat{\theta}$ devem estar "próximos", indicando que os dados estão dando suporte a restrição.

- O teste **LR** está baseado no ln da razão entre as duas verossimilhanças. Isto é a diferença entre o $\ln L(\widetilde{\theta})$ e $\ln L(\widehat{\theta})$.
- A estatística de teste é dada por:

$$LR = -2\left[\ln L(\widetilde{\theta}) - \ln L(\widehat{\theta})\right] \sim \chi^2_g \tag{36}$$

em que g é o número de restrições. O teste é distribuído assintoticamente como uma qui-quadrado com g graus de liberdade.

- **Se o valor da estatística for maior que o valor crítico ao nível de significância desejado, rejeitamos H_0.**

Teste de Wald

- **Condição**: o teste de Wald requer apenas a estimação do modelo sem restrição.
- **Lógica do teste**: investigar se a estimativa sem restrição está perto de cumprir a restrição. Utiliza-se $\widehat{\theta}$, o vetor estimado sem restrição, para se testar se $h(\widehat{\theta})$ está próximo de zero.
- Caso tenhamos $h(\widehat{\theta}) = 0$, a restrição estará sendo satisfeita pelos dados.
- Para realizar o teste necessitamos calcular primeiramente a variância de $h(\widehat{\theta})$ que é dada por:

$$\text{var}[h(\widehat{\theta})] = \boldsymbol{J}'\text{var}(\widehat{\theta})\boldsymbol{J} \tag{37}$$

- O vetor $\boldsymbol{J}$ é obtido como

$$\boldsymbol{J}' = \left[\frac{\partial h_i}{\partial \widehat{\theta}_i}\right] \tag{38}$$

- A estatística do teste de Wald é dada por

$$\boldsymbol{W} = h(\widehat{\theta})' \left[\text{var}\left[h(\widehat{\theta})\right]\right]^{-1} h(\widehat{\theta}) \sim \chi^2_g \tag{39}$$

em que g é o número de restrições. O teste é distribuído assintoticamente como uma chi-quadrada com g graus de liberdade.

Teste do Multiplicador de Lagrange

- **Condição**: o teste LM requer apenas a estimação do modelo com restrição.
- **Lógica do teste**: precisamos resolver um problema de maximização condicionada.
- Se a restrição for verdadeira, isto é H_0 for verdadeira, podemos esperar que $S\left(\widetilde{\theta}\right)$, o escore eficiente avaliado em $\widetilde{\theta}$, seja "próximo" de zero. *Lembre-se que no ponto de máximo temos* $S(\theta) = 0$.
- O valor do escore eficiente avaliado em $\widetilde{\theta}$ é que determina a aceitação ou não da restrição.

- A estatística do teste é dada por:

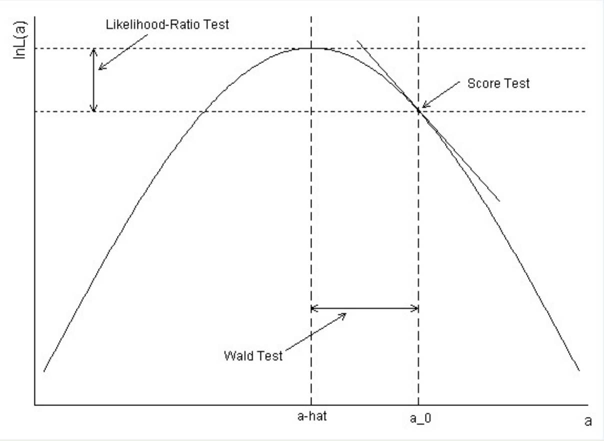
$$\boldsymbol{LM} = S\left(\widetilde{\theta}\right)'\left[I\left(\widetilde{\theta}\right)\right]^{-1} S\left(\widetilde{\theta}\right) \sim \chi_g^2 \qquad (40)$$

em que g é o número de restrições. O teste é distribuído assintoticamente como uma chi-quadrada com g graus de liberdade

- **Em síntese**:
  1. O Teste de Wald procura medir a distância entre $\widetilde{\theta}$ e $\widehat{\theta}$.
  2. O teste da Razão da Verossimilhança ocupa-se da distância entre $\ln(\widetilde{\theta})$ e $\ln(\widehat{\theta})$.
  3. O Teste do Multiplicador de Lagrange compara as tangentes nos pontos $\widetilde{\theta}$ e $\widehat{\theta}$.

- Os três testes são assintoticamente equivalentes. Assim, **quando um dos testes aceita a restrição, os demais também aceitam**.

Figura 1: Relação entre os Testes



Econometria I

Estimador de Máxima Verossimilhança

Victor Oliveira

Núcleo de Economia Internacional e Desenvolvimento Econômico

PPGDE – 2023