

# ML e GMM

(rascunho de notas de aula)

**Victor Gomes**

*Universidade de Brasilia*

03/05/2019

## Estimadores Extremos

No livro de Hayashi é definida a classe de estimadores extremos: ML, NLLS, CDE e GMM.

Um estimador  $\hat{\theta}$  é chamado de *estimador extremo* se existe uma função objetivo escalar  $Q_n(\theta)$  tal que

$$\max_{\hat{\theta}} \{Q_n(\theta)\} \text{ s.a. } \theta \in \Theta \subset \mathbb{R}^p \quad (1)$$

tal que  $\Theta$  (o espaço dos parâmetros) é o conjunto dos valores possíveis para os parâmetros. Restrição no conjunto de parâmetros: subconjunto do espaço Euclidiano com dimensão finita  $\mathbb{R}^p$ .  $Q_n(\theta)$  não depende apenas de  $\theta$  mas também depende da amostra  $(w_1, w_2, \dots, w_n)$ , tal que  $w_t$  é a observação  $t$  e  $n$  é o tamanho da amostra. O subscrito  $n$  sinaliza a dependência do tamanho da amostra.

## Estimadores Extremos: Medindo $\hat{\theta}$

O problema de maximização (1) pode não ter uma solução. Relembrar o seguinte fato de cálculo:

Faça  $h : \mathbb{R}^p \rightarrow \mathbb{R}$  ser contínua e  $A \subset \mathbb{R}^p$  ser um conjunto compacto (fechado e limitado). Então  $h$  possui um máximo no conjunto  $A$ . Isto é, existe um  $x^* \in A$  tal que  $h(x^*) \geq h(x)$  para todo  $x$  em  $A$ .

Portanto, se  $Q_n(\theta)$  é contínuo em  $\theta$  para qualquer amostra  $(w_1, w_2, \dots, w_n)$  e  $\Theta$  é compacto, então existe um  $\theta$  que soluciona o problema de maximização em (1) para qualquer amostra.

Quando se tem soluções múltiplas, poderíamos escolher uma delas. Então  $\hat{\theta}$  é uma função dos dados. Ser uma função do vetor dos dados

$(w_1, w_2, \dots, w_n)$  não é suficiente para fazer  $\hat{\theta}$  uma variável aleatória bem definida.  $\hat{\theta}$  precisa ser uma função mensurável de  $(w_1, w_2, \dots, w_n)$ .

**Lema 7.1 (existência de estimadores extremos):**

Suponha que (i) o espaço dos parâmetros  $\Theta$  é um subconjunto compacto de  $\mathbb{R}^p$ , (ii)  $Q_n(\theta)$  é contínua em  $\theta$  para qualquer  $(w_1, w_2, \dots, w_n)$  e (iii)  $Q_n(\theta)$  é uma função mensurável dos dados para todo  $\theta$  em  $\Theta$ . Então existe uma função mensurável  $\hat{\theta}$  dos dados que soluciona (1).

Comentário: na maioria das aplicações não se sabe o limite superior ou inferior do parâmetro verdadeiro. Mesmo se soubessemos estes limites não estão incluídos no espaço dos parâmetros, isso significa que o espaço dos parâmetros não é fechado. Então a hipótese de conjunto compacto para  $\Theta$  é algo que se deseja evitar. Hayashi troca a hipótese de conjunto compacto por alguma condição que são satisfeitas em muitas aplicações.

## Duas Classes de Estimadores

**Estimadores M:** Um estimador extremo é um *estimador M* se a função objetivo é uma média amostral:

$$Q_n(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^n m(\mathbf{w}_t; \boldsymbol{\theta}) \quad (2)$$

tal que  $m$  é uma função de valor real de  $(\mathbf{w}_t; \boldsymbol{\theta})$ . Dois exemplos: máxima verossimilhança (ML) e mínimos quadrados não-lineares (NLLS).

**GMM:** Um estimador extremo é um *estimador GMM* se a função objetivo pode ser escrita como:

$$Q_n(\boldsymbol{\theta}) = -\frac{1}{2} g_n(\boldsymbol{\theta})' \hat{\mathbf{W}} g_n(\boldsymbol{\theta}) \quad (3)$$

$$g_n(\boldsymbol{\theta}) = \frac{1}{n} \sum_{t=1}^n g(\mathbf{w}_t; \boldsymbol{\theta})$$

$K \times 1$

tal que  $\hat{\mathbf{W}}$  é uma matriz  $K \times K$  simétrica e positiva definida que determina a distância  $g_n(\boldsymbol{\theta})$  de zero. Como visto no capítulo 3, maximizar esta função é equivalente a minimizar a distância da função objetivo (a deflação da distância por 2 simplifica a expressão para a derivada da função objetivo).

Classical minimum distance estimator (CMD) não cabe nestas classes. A função objetivo da CMD por ser escrita como  $-\frac{1}{2}g_n(\boldsymbol{\theta})'\hat{\mathbf{W}}g_n(\boldsymbol{\theta})$ , mas  $g_n(\boldsymbol{\theta})$  não é necessariamente uma média amostral. Entretanto, o teorema de consistência da próxima seção é geral o suficiente para cobrir esta classe de modelo.

## ML: Máxima Verossimilhança

Caso da ML para um conjunto de sequências  $\{w_t\}$  iid. Neste caso a densidade de  $\{w_t\}$  é um membro de um família de densidades indexada por um vetor de dimensão finita  $\theta$  :  $f(w_t; \theta), \theta \in \Theta$ .<sup>\*</sup> A fórmula funcional  $f(\cdot)$  é conhecida. O modelo é paramétrico porque o vetor de parâmetros  $\theta$  é de dimensão finita. No vetor verdadeiro de parâmetros,  $\theta_0$ , a densidade do PGD verdadeiro é  $f(w_t; \theta_0)$ . Se diz que o modelo é *corretamente especificado* se  $\theta_0 \in \Theta$ .

Dado que  $\{w_t\}$  é iid, a densidade conjunta dos dados  $(w_1, w_2, \dots, w_n)$  é

$$f(w_1, w_2, \dots, w_n; \theta_0) = \prod_{t=1}^n f(w_t; \theta_0) \quad (4)$$

<sup>\*</sup>Como  $\{w_t\}$  é iid, a forma funcional  $f(\cdot)$  não depende de  $t$ .

Com a distribuição completa dos dados, o método de estimação natural é ML. Quando se troca o vetor  $\theta_0$  pelo vetor hipotético  $\theta$ , a densidade é denominada *função de verossimilhança*. O estimador ML de  $\theta_0$  é o  $\theta$  que maximiza a função de verossimilhança. Uma vez que a transformação em log é monotônica, maximizar a função de verossimilhança é equivalente a maximizar a *função de log verossimilhança*:

$$\begin{aligned} \log f(\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_n; \theta) &= \log \left[ \prod_{t=1}^n f(\mathbf{w}_t; \theta) \right] = \\ &= \sum_{t=1}^n \log f(\mathbf{w}_t; \theta) \quad (5) \end{aligned}$$

Portanto, o estimador de  $\theta_0$  é um estimador-M com

$$m(\mathbf{w}_t; \theta) = \log f(\mathbf{w}_t; \theta) \quad (6)$$

$$\text{i.e. } Q_n(\theta) = \frac{1}{n} \sum_{t=1}^n \log f(\mathbf{w}_t; \theta)$$



- A log verossimilhança média não deve assumir uma forma simples como esta se  $\{w_t\}$  possui correlação seriada.
- ML não é a única forma de se estimar o vetor de parâmetros. Por exemplo, faça  $\mu(\theta)$  ser a expectativa de  $w_t$  implicada pela função de densidade  $f(w_t; \theta)$ . Esta é uma função conhecida de  $\theta$ . Por construção vale a seguinte condição de média-zero:

$$E[w_t - \mu(\theta_0)] = 0 \quad (7)$$

O vetor de parâmetros  $\theta_0$  pode ser estimado por GMM com  $g = (w_t; \theta)w_t - \mu(\theta_0)$  na função objetivo.

- Eficiência: como será mostrado, o estimador ML é consistente e assintoticamente normal sob condições suficientes. Se acredita que o ML é eficiente (i.e. atinge a

variância assintótica mínima) em uma classe de estimadores assintoticamente normal e consistente. Esta crença geral foi mostrada ser errada por contraexemplos. Apesar disso, ML é eficiente para uma classe geral de estimadores que são assintoticamente normal – uma dessas classes é GMM.

## Exemplo 7.1

Estimando a média de uma distribuição normal: faça o dado  $(w_1, \dots, w_n)$  ser uma sequência de escalar iid com distribuição de  $w_t$  dado por  $N(\mu, \sigma^2)$ . Então  $\theta = (\mu, \sigma^2)'$  e

$$f(w_t; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[ -\frac{(w_t - \mu)^2}{2\sigma^2} \right]$$

A log verossimilhança média dos dados  $(w_1, \dots, w_n)$  é

$$\begin{aligned} \frac{1}{n} \sum_{t=1}^n \log f(w_t; \mu, \sigma^2) &= \\ &= -\frac{1}{2} \log(2\pi) - \frac{1}{2} \log(\sigma^2) - \frac{1}{n} \sum_{t=1}^n \left[ \frac{(w_t - \mu)^2}{2\sigma^2} \right] \end{aligned} \quad (8)$$

O estimador ML de  $(\mu_0, \sigma_0^2)$  é um estimador extremo com  $Q_n(\theta)$  com a log verossimilhança acima. Por hipótese, assumo a variância como positiva, mas que não existe nenhuma restrição

a priori sobre o valor de  $\mu_0$ , então o espaço de parâmetro  $\Theta$  é  $\mathcal{R} \times \mathcal{R}_{++}$ , tal que  $\mathcal{R}_{++}$  é o conjunto positivo de números reais. O estimador ML de  $\mu_0$  é a média amostral de  $w_t$ . O estimador GMM de  $\mu_0$  baseado na condição de média zero  $E(w_t - \mu_0) = 0$ , também é a média amostral de  $w_t$ .

## ML Condicional

Na maioria das aplicações o vetor  $w_t$  é:  $y_t$  e  $x_t$ . O interesse do econométrico é saber como  $x_t$  influencia a distribuição condicional de  $y_t$  dado  $x_t$ .  $y_t$  é a variável dependente e  $x_t$  é o regressor.\*

Faça  $f(y_t | x_t; \theta_0)$  ser a densidade condicional de  $y_t$  dado  $x_t$ , e faça  $f(x_t; \psi_0)$  ser a densidade marginal de  $x_t$ . Então, a densidade conjunta de  $w_t = (y_t, x_t')$  é:

$$f(y_t, x_t; \theta_0, \psi_0) = f(y_t | x_t; \theta_0) f(x_t; \psi_0) \quad (9)$$

Suponha por enquanto que  $\theta_0$  e  $\psi_0$  não são funcionalmente relacionados. A log verossimi-

\*Os resultados valem para  $y_t$  sendo escalar ou vetor.

lança média dos dados  $(w_1, \dots, w_n)$  é

$$\frac{1}{n} \sum_{t=1}^n \log f(w_t; \theta, \psi) = \underbrace{\frac{1}{n} \sum_{t=1}^n \log f(y_t | x_t; \theta)}_{\text{log verossimilhança condicional}} + \frac{1}{n} \sum_{t=1}^n \log f(x_t; \psi) \quad (10)$$

O estimador ML condicional de  $\theta_0$  maximiza o primeiro termo de (10), ignorando o segundo termo. Este é um estimador-M com:

$$m(w_t; \theta) = \log f(y_t | x_t; \theta_0), \text{ isto é} \quad (11)$$

$$Q_n(\theta) = \frac{1}{n} \sum_{t=1}^n \log f(y_t | x_t)$$

O segundo termo de (11) é a log verossimilhança marginal média. Este termo não depende de  $\theta$ , então a estimativa ML condicional de  $\theta_0$  é numericamente a mesma se o segundo termo fosse incluído.

Relação funcional entre  $\theta_0$  e  $\psi_0$ : por exemplo,  $\theta_0$  e  $\psi_0$  podem ser divididos como:

- $\theta_0 = (\alpha'_0, \beta'_0)'$

- $\psi_0 = (\beta'_0, \gamma'_0)'$

Neste caso o estimador completo e o condicional não são mais numericamente iguais. Neste caso, o estimador ML condicional de  $\theta_0$  é menos eficiente do que o ML completo obtido da maximização conjunta. Em várias aplicações a perda de eficiência é inevitável porque não se pode especificar a forma paramétrica para  $f(\mathbf{x}_t; \psi)$ .

## Exemplo 7.2

Regressão linear com erros normalmente distribuídos. Considere um modelo com erros homocedásticos e normalmente distribuídos, então

$\{y_t, \mathbf{x}_t\}$  é iid.

$$y_t = \mathbf{x}_t' \boldsymbol{\beta}_0 + \varepsilon_t \quad (12)$$

$$\varepsilon_t \mid \mathbf{x}_t \sim N(0, \sigma_0^2)$$

A versossimilhança de  $y_t \mid \mathbf{x}_t$ ,  $f(y_t \mid \mathbf{x}_t; \boldsymbol{\theta})$ , é a função de densidade de  $N(\mathbf{x}_t' \boldsymbol{\beta}, \sigma^2)$ . Dado  $\boldsymbol{\theta} = (\boldsymbol{\beta}', \sigma^2)'$  e  $\mathbf{w}_t = (y_t, \mathbf{x}_t')'$ , então a função  $m$  é

$$m(\mathbf{w}_t; \boldsymbol{\theta}) = \log f(y_t \mid \mathbf{x}_t; \boldsymbol{\beta}, \sigma^2) \quad (13)$$

$$\begin{aligned} &= \log \left\{ \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[ -\frac{(y_t - \mathbf{x}_t' \boldsymbol{\beta})^2}{2\sigma^2} \right] \right\} \\ &= -\frac{1}{2} \log(2\pi) - \log(\sigma) - \frac{1}{2} \left( \frac{y_t - \mathbf{x}_t' \boldsymbol{\beta}}{\sigma} \right)^2 \end{aligned} \quad (14)$$



O espaço dos parâmetros  $\Theta$  é  $\mathfrak{R}^K \times \mathfrak{R}_{++}$ , tal que  $K$  é a dimensão de  $\beta$  e  $\mathfrak{R}_{++}$  é o conjunto de números positivos reais refletindo a restrição  $\sigma^2 > 0$ .

### Exemplo 7.3 (Probit)

No modelo probit, a variável escalar dependente  $y_t$  é binária ( $y_t \in \{0, 1\}$ ). Por exemplo,  $y_t = 1$  se a mulher decide entrar no mercado de trabalho e  $y_t = 0$  caso contrário.  $\mathbf{x}_t$  pode incluir a renda do marido. A probabilidade condicional de  $y_t$  dado um vetor de regressores  $\mathbf{x}_t$  é dado por

$$f(y_t = 1 \mid \mathbf{x}_t; \boldsymbol{\theta}_0) = \Phi(\mathbf{x}'_t \boldsymbol{\theta}_0) \quad (15)$$

$$f(y_t = 0 \mid \mathbf{x}_t; \boldsymbol{\theta}_0) = 1 - \Phi(\mathbf{x}'_t \boldsymbol{\theta}_0) \quad (16)$$

onde  $\Phi(\cdot)$  é a função de densidade cumulativa da distribuição normal padronizada. Isto pode ser escrito compactamente como

$$f(y_t \mid \mathbf{x}_t; \boldsymbol{\theta}_0) = \Phi(\mathbf{x}'_t \boldsymbol{\theta}_0)^{y_t} [1 - \Phi(\mathbf{x}'_t \boldsymbol{\theta}_0)]^{(1-y_t)} \quad (17)$$

Se não existe nenhuma restrição a priori sobre  $\boldsymbol{\theta}_0$ , o espaço de parâmetros  $\Theta$  é  $\mathbb{R}^p$ . O estimador ML de  $\boldsymbol{\theta}_0$  é um estimador-M com  $m$  função

dada por

$$\begin{aligned} m(\mathbf{w}_t; \boldsymbol{\theta}) &= \log f(y_t | \mathbf{x}_t; \boldsymbol{\theta}) = \\ &= y_t \log \Phi(\mathbf{x}'_t \boldsymbol{\theta}) + (1 - y_t) \log[1 - \Phi(\mathbf{x}'_t \boldsymbol{\theta})] \quad (18) \end{aligned}$$

## Invariância do ML

Os estimadores ML condicional e completo possuem um número de propriedades desejáveis. Uma delas é a *invariância* – que vale para amostras finitas.

Descrição: considere reparametrizar o modelo por um mapeamento ou função  $\lambda = \tau(\theta)$  definido em  $\Theta$ . Faça  $\Lambda$  ser o intervalo do mapeamento:

$$\Lambda \equiv \tau(\Theta) \equiv \{\lambda \mid \lambda = \tau(\theta) \text{ para algum } \theta \in \Theta\} \quad (19)$$

Por definição, para todo  $\lambda$  em  $\Lambda$ , existe pelo menos um  $\theta$  tal que  $\lambda = \tau(\theta)$ . O mapeamento

$$\tau : \Theta \rightarrow \Lambda$$

é chamado de reparametrização se ele é *um para um* – para todo  $\lambda$  em  $\Lambda$ , existe apenas um  $\theta$  tal que  $\lambda = \tau(\theta)$ . Este único  $\theta$  é uma função de  $\lambda$ .

Este mapeamento de  $\Theta$  em  $\Lambda$  é chamada a inversa de  $\tau$  e representado por  $\tau^{-1}$ . Se diz que um estimador extremo  $\hat{\theta}$  é *invariante* a reparametrização  $\tau$  se o estimador para o modelo reparametrizado é  $\tau(\hat{\theta})$ .

Faça  $\tilde{Q}_n(\lambda)$  ser a função objetivo associada com o modelo reparametrizado. Um estimador extremo é invariante se e somente se

$$\tilde{Q}_n(\lambda) = Q_n(\tau^{-1}(\lambda)) \text{ para todo } \lambda \in \Lambda \quad (20)$$

Para ver isto faça  $\hat{\lambda} = \tau(\hat{\theta})$ . Para qualquer  $\lambda \in \Lambda$ , temos  $\tilde{Q}_n(\lambda) = Q_n(\tau^{-1}(\lambda)) \leq Q_n(\hat{\theta})$ , porque  $\hat{\theta}$  maximiza  $Q_n(\hat{\theta})$  sobre  $\Theta$  e  $\tau^{-1}$  está em  $\Theta$ . Mas  $Q_n(\hat{\theta}) = Q_n(\tau^{-1}(\hat{\lambda})) = \tilde{Q}_n(\hat{\lambda})$ . Então  $\tilde{Q}_n(\lambda) \leq \tilde{Q}_n(\hat{\lambda})$  para todo  $\lambda \in \Lambda$ .

ML é invariante a reparametrização porque a verossimilhança após a transformação é  $f(\cdot; \lambda) = f(\cdot; \tau^{-1}(\lambda))$ , que produz a função objetivo que satisfaz (20). (GMM não é invariante.)

## NLS

Assuma o vetor particionado:  $\mathbf{w}_t = (y_t, \mathbf{x}_t')'$ . O modelo no NLS é um conjunto de processos estocásticos  $\{y_t, \mathbf{x}_t\}$ , tal que a média condicional  $E(y_t | \mathbf{x}_t)$ , que é função de  $\mathbf{x}$ , é um membro da família paramétrica de funções  $\varphi(\mathbf{x}_t; \boldsymbol{\theta})$ ,  $\boldsymbol{\theta} \in \Theta$ . A forma funcional de  $\varphi(\cdot; \cdot)$  é conhecida. Se  $\boldsymbol{\theta}_0$  é o parâmetro verdadeiro, então  $E(y_t | \mathbf{x}_t) = \varphi(\mathbf{x}_t; \boldsymbol{\theta}_0)$  para o PGD verdadeiro  $\mathbf{w}_t = (y_t, \mathbf{x}_t')'$ . Se definirmos  $\varepsilon_t = y_t - E(y_t | \mathbf{x}_t)$ , então o modelo corretamente especificado pode ser escrito como

$$y_t = \varphi(\mathbf{x}_t; \boldsymbol{\theta}) + \varepsilon_t \quad (21)$$

tal que  $E(\varepsilon_t | \mathbf{x}_t) = 0$  e  $\boldsymbol{\theta}_0 \in \Theta$ . O método largamente utilizado para estimar este modelo é o *mínimos quadrados* (minimizar a soma do quadrado dos resíduos). O estimador NLS é o mínimo dos quadrados aplicado a (21). NLS é um estimador-M com:

$$m(\mathbf{w}_t; \mathbf{w}_t) = -[y_t - E(y_t | \mathbf{x}_t)]^2 \quad (22)$$

i.e.

$$Q_n(\boldsymbol{\theta}) = -\frac{1}{n} \sum_{t=1}^n [y_t - \mathbb{E}(y_t | \boldsymbol{x}_t)]^2$$

A maximização de  $Q_n(\boldsymbol{\theta})$  é a mesma da minimização do soma dos quadrados dos resíduos.

## Exemplo 7.4 (CES)

Função de produção CES com erros aditivos. Considere a função de produção CES:

$$Q_t = A_0 \left[ \delta_0 K_t^{-\rho_0} + (1 - \delta_0) L_t^{-\rho_0} \right]^{-1/\rho_0} + \varepsilon_t \quad (23)$$

tal que  $Q_t$  é o produto no período  $t$ ,  $K_t$  é o estoque de capital,  $L_t$  é o trabalho e  $\varepsilon_t$  é o choque aditivo com  $E(\varepsilon_t | K_t, L_t) = 0$

Esse é um modelo de média condicional (21) com  $y_t = Q_t$ ,  $\mathbf{x}_t = (K_t, L_t)'$ ,  $\boldsymbol{\theta}_0 = (A_0, \delta_0, \rho_0)'$  e  $\varphi(\mathbf{x}_t; \boldsymbol{\theta}) = A_0 \left[ \delta_0 K_t^{-\rho_0} + (1 - \delta_0) L_t^{-\rho_0} \right]^{-1/\rho_0}$ . As propriedades usuais de função de produção (monotonicidade, concavidade) são satisfeitas se  $A > 0$ ,  $0 < \delta < 1$  e  $-1 < \rho$ , o que determina o espaço dos parâmetros  $\Theta$ .



## GMM linear e não-linear

O modelo GMM linear é:

$$y_t = z_t' \theta_0 + \varepsilon_t \quad (24)$$

com  $x_t$  sendo o vetor de instrumentos, as condições de ortogonalidade são:

$$E[x_t \cdot (y_t - z_t' \theta_0)] = 0 \quad (25)$$

O modelo corretamente especificado aqui é um conjunto de processos ergódicos estacionários  $w_t = (y_t, z_t', x_t)'$ , tal que estas condições de média-zero valem para  $\theta_0 \in \Theta$ . O estimador GMM linear de  $\theta_0$  é um estimador GMM com a função  $g$  da função objetivo GMM (3) dada por:

$$g(w_t; \theta) = x_t \cdot (y_t - z_t' \theta) = x_t \cdot y_t - x_t \cdot z_t' \theta \quad (26)$$

O estimador GMM pode ser aplicado para equações não-lineares. Suponha que a equação é não-linear:

$$a(y_t, z_t; \theta_0) = \varepsilon_t \quad (27)$$

Apenas faça

$$g(\mathbf{w}_t; \boldsymbol{\theta}) = \mathbf{x}_t \cdot a(y_t, \mathbf{z}_t; \boldsymbol{\theta}) \quad (28)$$

Este estimador é denominado *estimador de variáveis instrumentais generalizado*. Este ainda é um caso especial de GMM porque a função  $g(\mathbf{w}_t; \boldsymbol{\theta})$  pode ser escrita como um produto do vetor de instrumentos e do termo de erro.

## Exemplo 7.5: Hansen-Singleton 1982

Equação (não-linear) de Euler do consumo.

A equação de Euler do problema de otimização do consumidor é:

$$\mathbb{E} \left[ R_{t+1} \frac{\beta_0 u'(c_{t+1})}{u'(c_t)} \mid I_t \right] = 1 \quad (29)$$

$R_{t+1}$  é a taxa de retorno bruta ( $1 +$  taxa de retorno),  $c_t$  é o consumo de não-duráveis,  $\beta_0$  é o fator de desconto,  $u'(c)$  é a utilidade marginal e  $I_t$  é a informação disponível. Assuma a seguinte função para a utilidade:  $u(c) = c^{1-\alpha_0}/(1-\alpha_0)$ . Então  $u'(c) = c^{-\alpha_0}$ . A equação de Euler então é:

$$\mathbb{E} \left[ a \left( R_{t+1}, \frac{c_{t+1}}{c_t}; \alpha_0, \beta_0 \right) \mid I_t \right] = 0$$

$$a \left( R_{t+1}, \frac{c_{t+1}}{c_t}; \alpha_0, \beta_0 \right) = R_{t+1} \beta_0 \left( \frac{c_{t+1}}{c_t} \right)^{-\alpha_0} - 1 \quad (30)$$

Se  $\mathbf{x}_t$  é um vetor de variáveis cujos valores são conhecidos na data  $t$ , então  $\mathbf{x}_t \in I_t$ . Utilizando a equação anterior, temos a condição de ortogonalidade:

$$E \left[ \mathbf{x}_t \cdot a \left( R_{t+1}, \frac{c_{t+1}}{c_t}; \alpha_0, \beta_0 \right) \mid I_t \right] = \mathbf{0} \quad (31)$$

Tomando expectativas incondicional em ambos os lados e usando a lei total das expectativas, obtemos as condições de ortogonalidade  $E[g(\mathbf{w}_t; \boldsymbol{\theta}_0)] = \mathbf{0}$ , tal que

$$g(\mathbf{w}_t; \boldsymbol{\theta}_0) = \mathbf{x}_t \cdot a \left( R_{t+1}, \frac{c_{t+1}}{c_t}; \alpha_0, \beta_0 \right) \quad (32)$$

com

$$\mathbf{w}_t = \begin{bmatrix} \mathbf{x}_t \\ \frac{c_{t+1}}{c_t} \\ R_{t+1} \end{bmatrix}, \boldsymbol{\theta}_0 = \begin{bmatrix} \beta_0 \\ \alpha_0 \end{bmatrix}$$

## Consistência

A função objetivo  $Q_n(\cdot)$  é uma função aleatória porque para cada  $\theta$  o valor  $Q_n(\theta)$  é uma variável aleatória, pois  $Q_n(\theta)$  depende de  $(w_1, \dots, w_n)$ .

Idéia para consistência:  $Q_n(\theta)$  converge em probabilidade para  $Q_0(\theta)$ , e o parâmetro verdadeiro  $\theta$  soluciona o “problema limite” de maximização da função limite  $Q_0(\theta)$ , então o limite do máximo  $\hat{\theta}$  deve ser  $\theta_0$ .

Necessário: (i) convergência em probabilidade uniforme; (ii)  $\Theta$  compacto. Problema: o espaço dos parâmetros não é compacto na maioria das aplicações.

## Consistência: convergência é uniforme

- (extensão natural da sequência de funções aleatórias)
- *Convergência em probabilidade pontual* de  $Q_n(\cdot)$  para alguma função não aleatória  $Q_0(\cdot)$ , simplesmente significa  $\text{plim}_{n \rightarrow \infty} Q_n(\boldsymbol{\theta}) = Q_0(\boldsymbol{\theta})$  para todo  $\boldsymbol{\theta}$  – a sequência de variáveis aleatórias  $|Q_n(\boldsymbol{\theta}) - Q_0(\boldsymbol{\theta})| (n = 1, 2, \dots)$  converge em probabilidade para 0 para cada  $\boldsymbol{\theta}$ .
- *Convergência uniforme em probabilidade* é mais forte. A convergência tem que ocorrer uniformemente sobre o espaço dos parâmetros  $\Theta$  no seguinte sentido:

$$\sup_{\boldsymbol{\theta} \in \Theta} |Q_n(\boldsymbol{\theta}) - Q_0(\boldsymbol{\theta})| \rightarrow_p 0 \text{ enquanto } n \rightarrow \infty \quad (33)$$

- A expansão do resultado acima para vetor se dá pelo requerimento de que convergência uniforme para cada elemento. Uma sequência de vetor de funções aleatórias  $\{\mathbf{h}_n(\cdot)\}$  converge uniformemente em probabilidade para uma função não aleatória  $\mathbf{h}_0(\cdot)$  se cada elemento converge uniformemente. Esta convergência elemento por elemento é equivalente a convergência na norma:

$$\sup_{\boldsymbol{\theta} \in \Theta} \|\mathbf{h}_n(\boldsymbol{\theta}) - \mathbf{h}_0(\boldsymbol{\theta})\| \xrightarrow{p} 0 \text{ enquanto } n \rightarrow \infty \quad (34)$$

tal que  $\|\cdot\|$  é a norma Euclidiana.

## Consistência com espaço dos parâmetros compacto

**Proposição 7.1 (Consistência):** Suponha que

(i)  $\Theta$  é um subconjunto compacto de  $\mathfrak{R}^p$ , (ii)  $Q_n(\theta)$  é contínua em  $\theta$  para qualquer dado  $(w_1, \dots, w_n)$ , e (iii)  $Q_n(\theta)$  é função mensurável do dado para todo  $\theta$  em  $\Theta$ . Se existe uma função  $Q_0(\theta)$  tal que

1. (identificação)  $Q_0(\theta)$  é unicamente maximizado sobre  $\Theta$  em  $\theta_0 \in \Theta$ .
2. (convergência uniforme)  $Q_n(\cdot)$  converge uniformemente em probabilidade para  $Q_0(\cdot)$ , então  $\hat{\theta} \rightarrow_p \theta_0$ .



## Consistência sem espaço dos parâmetros compacto

**Proposição 7.2 (Consistência):** Suponha que

(i) o vetor verdadeiro dos parâmetros  $\theta_0$  é um elemento do interior de espaço de parâmetros convexo  $\Theta (\subset \mathbb{R}^p)$ , (ii)  $Q_n(\theta)$  é côncava sobre o espaço dos parâmetros para qualquer dado  $(w_1, \dots, w_n)$ , e (iii)  $Q_n(\theta)$  é função mensurável do dado para todo  $\theta$  em  $\Theta$ . Faça  $\hat{\theta}$  ser o estimador extremo definido anteriormente (7.1). Se existe uma função  $Q_0(\theta)$  tal que

1. (identificação)  $Q_0(\theta)$  é unicamente maximizado sobre  $\Theta$  em  $\theta_0 \in \Theta$ .
2. (convergência pontual)  $Q_n(\cdot)$  converge em probabilidade para  $Q_0(\cdot)$  para todo  $\theta_0 \in \Theta$ , então a medida que  $n \rightarrow \infty$ ,  $\hat{\theta}$

existe com probabilidade aproximando 1  
e  $\hat{\theta} \rightarrow_p \theta_0$ .

Na maioria das aplicações  $\Theta$  é um conjunto convexo aberto, então a condição (i) é satisfeita. Esta proposição requer que  $Q_n(\theta)$  seja côncava para qualquer dado, mas em várias aplicações esta condição é satisfeita.

## Consistência de Estimadores-M

A função objetivo de um estimador-M é

$$Q_n(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^n m(\mathbf{w}_t; \boldsymbol{\theta})$$

Se  $\{\mathbf{w}_t\}$  é estacionária ergódica, o Teorema Ergódico implica em convergência pontual para  $Q_n(\boldsymbol{\theta})$ :

$Q_n(\boldsymbol{\theta})$  converge em probabilidade pontual para cada  $\boldsymbol{\theta} \in \Theta$  para  $Q_0(\boldsymbol{\theta})$  dado por

$$Q_0(\boldsymbol{\theta}) = E[m(\mathbf{w}_t; \boldsymbol{\theta})] \quad (35)$$

Para aplicar a Proposição 7.1 de consistência a estimadores-M, precisamos mostrar que a convergência de  $\frac{1}{n} \sum_{i=1}^n m(\mathbf{w}_t; \boldsymbol{\theta})$  para  $E[m(\mathbf{w}_t; \boldsymbol{\theta})]$  é uniforme.

**Lema 7.2 (Lei uniforme dos grandes números):**

Faça  $\{\mathbf{w}_t\}$  ser um processo estacionário e

ergódico. Suponha que (i) o conjunto  $\Theta$  é compacto, (ii)  $m(\mathbf{w}_t; \theta)$  é contínua em  $\theta$  e (iii)  $m(\mathbf{w}_t; \theta)$  é mensurável em  $\mathbf{w}_t$  para todo  $\theta$  em  $\Theta$ . Além disso suponha:

**condição de dominância:** existe uma função  $d(\mathbf{w}_t)$  (algumas vezes chamada de "função dominante") tal que  $|m(\mathbf{w}_t; \theta)| \leq d(\mathbf{w}_t)$  para todo  $\theta \in \Theta$  e  $E[d(\mathbf{w}_t)] < \infty$

Então  $\frac{1}{n} \sum_{i=1}^n m(\mathbf{w}_t; \theta)$  converge uniformemente em probabilidade para  $E[m(\mathbf{w}_t); \theta]$  sobre  $\Theta$ . Além disso,  $E[m(\mathbf{w}_t); \theta]$  é uma função contínua de  $\theta$ .

A Lei Uniforme dos Grandes Números pode ser ampliada para vetor de funções aleatórias:

**Condição de dominância:** Faça  $\{\mathbf{w}_t\}$  ser um processo estacionário ergódico. Suponha

que (i) o espaço de parâmetros  $\Theta$  é compacto, (ii)  $h(w_t; \theta)$  é contínuo em  $\theta$  para todo  $w_t$  e (iii)  $h(w_t; \theta)$  é mensurável em  $w_t$  para todo  $\theta$  em  $\Theta$ . Além disso, suponha

**condição de dominância:**  $E[\sup_{\theta \in \Theta} \|h(w_t; \theta)\|] < \infty$

Então  $E[h(w_t; \theta)]$  é uma função contínua de  $\theta$  e  $\frac{1}{n} \sum_{t=1}^n h(w_t; \cdot)$  converge uniformemente em probabilidade para  $E[h(w_t; \cdot)]$  sobre  $\Theta$ .

## Consistência para estimadores-M com espaço dos parâmetros compacto

**Proposição 7.3 (Consistência):** Faça  $\{w_t\}$  ser um processo estacionário ergódico. Suponha que (i)  $\Theta$  é um subconjunto compacto de  $\mathbb{R}^p$ , (ii)  $m(w_t; \theta)$  é contínua em  $\theta$  para qualquer  $w_t$ , e (iii)  $m(w_t; \theta)$  é mensurável  $w_t$  para todo  $\theta$  em  $\Theta$ . Faça  $\hat{\theta}$  ser o estimador-M definido por (7.1.1) e (7.1.2). Suponha além disso:

1. (identificação)  $E[m(w_t; \theta)]$  é unicamente maximizado sobre  $\Theta$  em  $\theta_0 \in \Theta$ .
2. (dominância)  $E[\sup_{\theta \in \Theta} |h(w_t; \theta)|] < \infty$

Então  $\hat{\theta} \rightarrow_p \theta_0$

## Consistência para estimadores-M sem espaço dos parâmetros compacto

A função objetivo  $Q_n(\boldsymbol{\theta})$  é côncava se  $m(\boldsymbol{w}_t; \boldsymbol{\theta})$  é côncava em  $\boldsymbol{\theta}$ .

**Proposição 7.4 (Consistência):** Faça  $\{\boldsymbol{w}_t\}$  ser um processo estacionário ergódico. Suponha que (i) o vetor verdadeiro dos parâmetros  $\boldsymbol{\theta}_0$  é um elemento do interior de espaço de parâmetros convexo  $\Theta (\subset \mathbb{R}^p)$ , (ii)  $m(\boldsymbol{w}_t; \boldsymbol{\theta})$  é côncava sobre o espaço dos parâmetros para todo  $\boldsymbol{w}_t$ , e (iii)  $m(\boldsymbol{w}_t; \boldsymbol{\theta})$  é mensurável em  $\boldsymbol{w}_t$  para todo  $\boldsymbol{\theta}$  em  $\Theta$ . Faça  $\hat{\boldsymbol{\theta}}$  ser o estimador-M definido por (7.1.1) e (7.1.2). Suponha além disso:

1. (identificação)  $E[m(\boldsymbol{w}_t; \boldsymbol{\theta})]$  é unicamente maximizado sobre  $\Theta$  em  $\boldsymbol{\theta}_0 \in \Theta$ .
2. (convergência pontual)  $E[|m(\boldsymbol{w}_t; \boldsymbol{\theta})|] < \infty$  para todo  $\boldsymbol{\theta}$  em  $\Theta$ .

Então, a medida que  $n \rightarrow \infty$ ,  $\hat{\theta}$  existe com probabilidade aproximando 1 e  $\hat{\theta} \rightarrow_p \theta_0$ .



## Concavidade após Reparametrização

A Proposição 7.4 pode ser ampliada para estimadores de função objetivo que são côncavas após reparametrização. Suponha que todas as condições da proposição sejam satisfeitas exceto concavidade e que existe um mapeamento contínuo 1-1  $\tau(\theta) : \Theta \rightarrow \Lambda \equiv \tau(\Theta)$  tal que

$$\tilde{m}(\mathbf{w}_t; \boldsymbol{\lambda}) \equiv m(\mathbf{w}_t; (\tau)^{-1}\boldsymbol{\lambda})$$

é côncava em  $\boldsymbol{\lambda}$  e  $\Lambda = \boldsymbol{\lambda}(\Theta)$  é um conjunto convexo. Faça

$$\tilde{Q}_n(\boldsymbol{\lambda}) \equiv \frac{1}{n} \sum_{i=1}^n \tilde{m}(\mathbf{w}_t; \boldsymbol{\lambda})$$

seja a função objetivo após a reparametrização.

O estimador é consistente porque

$$\begin{aligned} \text{plim}_{n \rightarrow \infty} \hat{\boldsymbol{\theta}} &= \text{plim}_{n \rightarrow \infty} \tau^{-1}(\hat{\boldsymbol{\lambda}}) = \\ &= \tau^{-1}(\text{plim}_{n \rightarrow \infty} \hat{\boldsymbol{\lambda}}) = \tau^{-1}(\boldsymbol{\lambda}_0) = \boldsymbol{\lambda}_0 \end{aligned} \quad (36)$$

## Identificação NLS

Mean square error:

$$\begin{aligned} E[\{y_t - h(\mathbf{x}_t)\}^2] &= \\ &= E[\{y_t - E(y_t | \mathbf{x}_t)\}^2] + E[\{E(y_t | \mathbf{x}_t) - h(\mathbf{x}_t)\}^2] \geq \\ &\geq E[\{y_t - E(y_t | \mathbf{x}_t)\}^2] \quad (37) \end{aligned}$$

Fazendo  $h(\mathbf{x}_t) = \varphi(\mathbf{x}_t; \boldsymbol{\theta})$  e observando que  $\varphi(\mathbf{x}_t; \boldsymbol{\theta}) = E(y_t | \mathbf{x}_t)$ , se conclui que

$$\begin{aligned} E[\{y_t - \varphi(\mathbf{x}_t; \boldsymbol{\theta})\}^2] &> E[\{y_t - \varphi(\mathbf{x}_t; \boldsymbol{\theta}_0)\}^2] \\ &\text{se } \varphi(\mathbf{x}_t; \boldsymbol{\theta}) \neq \varphi(\mathbf{x}_t; \boldsymbol{\theta}_0) \quad (38) \end{aligned}$$

$$\begin{aligned} E[\{y_t - \varphi(\mathbf{x}_t; \boldsymbol{\theta})\}^2] &= E[\{y_t - \varphi(\mathbf{x}_t; \boldsymbol{\theta}_0)\}^2] \\ &\text{se } \varphi(\mathbf{x}_t; \boldsymbol{\theta}) = \varphi(\mathbf{x}_t; \boldsymbol{\theta}_0) \quad (39) \end{aligned}$$

**Identificação média condicional:** a condição de identificação funciona se

$$\varphi(\mathbf{x}_t; \boldsymbol{\theta}) \neq \varphi(\mathbf{x}_t; \boldsymbol{\theta}_0) \text{ para todo } \boldsymbol{\theta} \neq \boldsymbol{\theta}_0 \quad (40)$$

## Identificação para ML Condicional

O papel da expressão em (37) para NLS é feita pela *desigualdade de informação Kullback-Leibler*.

A densidade condicional hipotética  $f(y_t | \mathbf{x}_t; \boldsymbol{\theta})$ , sendo uma função de  $(y_t, \mathbf{x}_t)$ , é uma variável aleatória de qualquer  $\boldsymbol{\theta}$ . O valor esperado da variável aleatória  $\log f(y_t | \mathbf{x}_t; \boldsymbol{\theta})$  pode ser escrito como:

$$\begin{aligned} E[\log f(y_t | \mathbf{x}_t; \boldsymbol{\theta})] &= \\ &= \int \log f(y_t | \mathbf{x}_t; \boldsymbol{\theta}) f(y_t, \mathbf{x}_t | \mathbf{x}_t; \boldsymbol{\theta}_0, \phi_0) dy_t d\mathbf{x}_t \end{aligned} \quad (41)$$

A desigualdade informacional Kullback-Leibler sobre funções de densidade afirma que:

$$\begin{aligned} E[\log f(y_t | \mathbf{x}_t; \boldsymbol{\theta})] &> E[\log f(y_t | \mathbf{x}_t; \boldsymbol{\theta}_0)] \\ &\text{se } \log f(y_t | \mathbf{x}_t; \boldsymbol{\theta}) \neq \log f(y_t | \mathbf{x}_t; \boldsymbol{\theta}_0) \end{aligned} \quad (42)$$

$$\begin{aligned} E[\log f(y_t | \mathbf{x}_t; \boldsymbol{\theta})] &= E[\log f(y_t | \mathbf{x}_t; \boldsymbol{\theta}_0)] \\ \text{se } \log f(y_t | \mathbf{x}_t; \boldsymbol{\theta}) &= \log f(y_t | \mathbf{x}_t; \boldsymbol{\theta}_0) \end{aligned} \quad (43)$$

Segue imediatamente que a condição de identificação (*identificação densidade condicional*) é satisfeita se

$$f(y_t | \mathbf{x}_t; \boldsymbol{\theta}) \neq f(y_t | \mathbf{x}_t; \boldsymbol{\theta}_0) \quad (44)$$

para todo  $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$ .

A seguir são apresentadas as versões das proposições 7.3 e 7.4 para ML condicional.

## Consistência para ML condicional com espaço dos parâmetros compacto

**Proposição 7.5 (Consistência):** Faça  $\{y_t, w_t\}$  serem processos estacionários e ergódicos com densidade condicional  $f(y_t | x_t; \theta)$  e faça  $\hat{\theta}$  ser o estimador ML condicional, que maximiza a média do log da versossimilhança condicional:

$$\hat{\theta} = \operatorname{argmax}_{\theta} \frac{1}{n} \sum_{i=1}^n \log f(y_t | x_t; \theta)$$

Suponha modelo corretamente especificado, então  $\theta_0 \in \Theta$ . Suponha também que (i)  $\Theta$  é um subconjunto compacto de  $\mathbb{R}^p$ , (ii)  $f(y_t | x_t; \theta)$  é contínua em  $\theta$  para qualquer  $\{y_t, w_t\}$ , e (iii)  $f(y_t | x_t; \theta)$  é mensurável em  $\{y_t, w_t\}$  para todo  $\theta$  em  $\Theta$ . Suponha além disso:

1. (identificação)  $\operatorname{Prob}[f(y_t | x_t; \theta) \neq f(y_t | x_t; \theta_0)] > 0$  para todo  $\theta \neq \theta_0 \in \Theta$ .

2. (dominância)  $E[\sup_{\theta \in \Theta} |\log f(y_t | x_t; \theta)|] < \infty$

Então  $\hat{\theta} \rightarrow_p \theta_0$

## Consistência para ML condicional sem espaço dos parâmetros compacto

**Proposição 7.6 (Consistência):** Faça  $\{y_t, \mathbf{w}_t\}$  serem processos estacionários e ergódicos com densidade condicional  $f(y_t | \mathbf{x}_t; \boldsymbol{\theta})$  e faça  $\hat{\boldsymbol{\theta}}$  ser o estimador ML condicional, que maximiza a média do log da versossimilhança condicional:

$$\hat{\boldsymbol{\theta}} = \operatorname{argmax}_{\boldsymbol{\theta}} \frac{1}{n} \sum_{i=1}^n \log f(y_t | \mathbf{x}_t; \boldsymbol{\theta})$$

Suponha modelo corretamente especificado, então  $\boldsymbol{\theta}_0 \in \Theta$ . Suponha também que (i)  $\boldsymbol{\theta}_0$  é um elemento do interior de um conjunto espaço de parâmetros convexo  $\Theta \subset \mathbb{R}^p$ , (ii)  $f(y_t | \mathbf{x}_t; \boldsymbol{\theta})$  é côncava em  $\boldsymbol{\theta}$  para qualquer  $\{y_t, \mathbf{w}_t\}$ , e (iii)  $f(y_t | \mathbf{x}_t; \boldsymbol{\theta})$  é mensurável em  $\{y_t, \mathbf{w}_t\}$  para todo para todo  $\boldsymbol{\theta}$  em  $\Theta$ . Suponha além disso:

1. (identificação)  $\text{Prob}[f(y_t | x_t; \theta) \neq f(y_t | x_t; \theta_0)] > 0$  para todo  $\theta \neq \theta_0 \in \Theta$ .
2. (dominância)  $E[|\log f(y_t | x_t; \theta)|] < \infty$  para todo  $\theta_0 \in \Theta$

Então, a medida que  $n \rightarrow \infty$ ,  $\hat{\theta}$  existe com probabilidade aproximando 1 e  $\hat{\theta} \rightarrow_p \theta_0$



## Identificação ML: Exemplo 7.8

Identificação no modelo de regressão linear. Considere o modelo de regressão linear com erro normal, o log da verossimilhança condicional para a observação  $t$  é

$$\log f(y_t | \mathbf{x}_t; \boldsymbol{\theta}) = -\frac{1}{2} \log(2\pi) - \frac{1}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} (y_t - \mathbf{x}_t' \boldsymbol{\beta})^2 \quad (45)$$

Uma vez que a função log é estritamente monotônica, a identificação da densidade condicional é equivalente a condição que  $\log f(y_t | \mathbf{x}_t; \boldsymbol{\theta}) \neq \log f(y_t | \mathbf{x}_t; \boldsymbol{\theta}_0)$  com probabilidade positiva se  $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$ .

Esta condição é satisfeita se  $E(\mathbf{x}_t \mathbf{x}_t')$  é não singular e portanto positiva definida. Para ver porque, faça  $\boldsymbol{\theta} = (\boldsymbol{\beta}', \sigma^2)'$  e  $\boldsymbol{\theta}_0 = (\boldsymbol{\beta}_0', \sigma_0^2)'$ . Claramente,  $\log f(y_t | \mathbf{x}_t; \boldsymbol{\theta}) \neq \log f(y_t | \mathbf{x}_t; \boldsymbol{\theta}_0)$  com probabilidade positiva se  $\sigma^2 \neq \sigma_0^2$ . Então,

considerando o caso  $\sigma^2 = \sigma_0^2$  mas com  $\beta \neq \beta_0$ .  
Então

$$\begin{aligned} E[(\mathbf{x}'_t\beta - \mathbf{x}'_t\beta_0)^2] &= E[\{\mathbf{x}'_t(\beta - \beta_0)\}^2] = \\ &= (\beta - \beta_0)'E[\mathbf{x}_t\mathbf{x}'_t](\beta - \beta_0) > 0 \quad (46) \end{aligned}$$

Conseqüentemente,  $\mathbf{x}'_t\beta \neq \mathbf{x}'_t\beta_0$  com probabilidade positiva; se  $\mathbf{x}'_t\beta = \mathbf{x}'_t\beta_0$  com probabilidade 1, então  $E[(\mathbf{x}'_t\beta - \mathbf{x}'_t\beta_0)^2]$  será zero. Portanto,  $y_t - \mathbf{x}'_t\beta \neq y_t - \mathbf{x}'_t\beta_0$  com probabilidade positiva, implicando que a condição de identificação seja satisfeita. A mesma condição sobre  $E(\mathbf{x}_t\mathbf{x}'_t)$  também implica condição (b) da Proposição 7.6 porque

$$\begin{aligned} E[(\mathbf{x}'_t\beta - \mathbf{x}'_t\beta_0)^2] &= E[\{\varepsilon_t + \mathbf{x}'_t(\beta_0 - \beta)\}^2] = \\ &= E(\varepsilon_t)^2 + (\beta_0 - \beta)'E(\mathbf{x}_t\mathbf{x}'_t)(\beta_0 - \beta) \end{aligned}$$

O último termo é finito se  $E(\mathbf{x}_t\mathbf{x}'_t)$  é finito. Estes resultados, junto com o fato observado anteriormente de que a log verossimilhança é

côncava após a reparametrização, implica que o estimador ML condicional do modelo de regressão linear com erro normal é consistente se  $\{y_t \mathbf{x}_t\}$  é estacionário ergódico e  $E(\mathbf{x}_t \mathbf{x}_t')$  é não singular.

## Identificação ML: Exemplo 7.9

Identificação no modelo probit. Valem as mesmas conclusões do modelo de regressão linear para o probit. A verossimilhança condicional para a observação  $t$  é

$$f(y_t | \mathbf{x}_t; \boldsymbol{\theta}) = \Phi(\mathbf{x}'_t \boldsymbol{\theta})^{y_t} \Phi(-\mathbf{x}'_t \boldsymbol{\theta})^{1-y_t} \quad (47)$$

O mesmo argumento do exemplo anterior vale aqui. Assumindo não singularidade de  $E(\mathbf{x}_t \mathbf{x}'_t)$ , os argumentos implicam que  $\mathbf{x}'_t \boldsymbol{\theta} \neq \mathbf{x}'_t \boldsymbol{\theta}_0$  com probabilidade positiva se  $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$ . Como  $\Phi(v)$  é estritamente monotônica,  $\mathbf{x}'_t \boldsymbol{\theta} \neq \mathbf{x}'_t \boldsymbol{\theta}_0$  com probabilidade positiva implica  $\Phi(\mathbf{x}'_t \boldsymbol{\theta}) \neq \Phi(\mathbf{x}'_t \boldsymbol{\theta}_0)$  e  $\Phi(-\mathbf{x}'_t \boldsymbol{\theta}) \neq \Phi(-\mathbf{x}'_t \boldsymbol{\theta}_0)$  com probabilidade positiva. Portanto, a condição de identificação (a) da Proposição 7.6 é satisfeita se  $E(\mathbf{x}_t \mathbf{x}'_t)$  é não-singular.

A não singularidade de  $E(\mathbf{x}_t \mathbf{x}'_t)$  também implica que a condição (b) da Proposição 7.6 é

satisfeita para o modelo probit. É fácil verificar que

$$|\log \Phi(v)| \leq |\log \Phi(0)| + |v| + |v|^2 \text{ para todo } v \quad (48)$$

Combinando este limite para  $|\log \Phi(v)|$  e o fato de que  $y_t$  e  $1 - y_t$  são menos do que ou igual 1 em valor absoluto, isto é fácil mostrar que  $E[|\log f(y_t | \mathbf{x}_t; \boldsymbol{\theta})|] < \infty$  se  $E(\mathbf{x}_t \mathbf{x}_t')$  existe e é finito (condição implicada pela não-singularidade). Se conclui que o estimador ML probit é consistente se  $\{y_t, \mathbf{x}_t\}$  é estacionária ergódica e  $E(\mathbf{x}_t \mathbf{x}_t')$  é não-singular.

## Consistência do GMM

Aplicar a Proposição 7.1 para GMM. A função objetivo GMM é (7.1.3). A continuidade de  $Q_n(\boldsymbol{\theta})$  em  $\boldsymbol{\theta}$  é satisfeita se  $\mathbf{g}(\mathbf{w}_t; \boldsymbol{\theta})$  é contínua em  $\boldsymbol{\theta}$  para todo  $\mathbf{w}_t$ . Se  $\{\mathbf{w}_t\}$  é estacionária ergódica, então  $g_n(\boldsymbol{\theta}) (\equiv \frac{1}{n} \sum_{t=1}^n g(\mathbf{w}_t; \boldsymbol{\theta})) \xrightarrow{p} E[g(\mathbf{w}_t; \boldsymbol{\theta})]$ . Então a função limite de  $Q_0(\boldsymbol{\theta})$  é

$$Q_0(\boldsymbol{\theta}) = -\frac{1}{n} E[g(\mathbf{w}_t; \boldsymbol{\theta})]' \mathbf{W} E[g(\mathbf{w}_t; \boldsymbol{\theta})] \quad (49)$$

Esta função é não positiva se  $\mathbf{W}$  é positiva definida. Ela possui um máximo de zero em  $\boldsymbol{\theta}_0$  porque  $E[g(\mathbf{w}_t; \boldsymbol{\theta}_0)] = \mathbf{0}$  pelas condições de ortogonalidade. Portanto, a condição de identificação (que  $Q_0(\boldsymbol{\theta})$  seja unicamente maximizado em  $\boldsymbol{\theta}_0$ ) na Proposição 7.1 é satisfeita se  $E[g(\mathbf{w}_t; \boldsymbol{\theta})] \neq \mathbf{0}$  para  $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$ . Levando em conta a convergência uniforme de  $Q_n(\boldsymbol{\theta})$  para  $Q_0(\boldsymbol{\theta})$ , não é difícil de mostrar que esta condição é satisfeita com se  $\mathbf{g}_n(\cdot)$  converge uniformemente para  $E[g(\mathbf{w}_t; \cdot)]$ .

A versão multivariada da Lei Uniforme dos Grandes Números (Lema 7.2) fornece uma condição suficiente para convergência uniforme.

## Consistência do GMM com espaço de parâmetros compacto

**Proposição 7.7:** Faça  $\{w_t\}$  ser estacionária ergódica e faça  $\hat{\theta}$  ser o estimador GMM definido por

$$\hat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmin}} \left[ \frac{1}{n} \sum_{t=1}^n g(w_t; \theta) \right]' \hat{W} \left[ \frac{1}{n} \sum_{t=1}^n g(w_t; \theta) \right] \quad (50)$$

tal que a matriz simétrica  $\hat{W}$  é assumida convergir em probabilidade para alguma matriz simétrica positiva definida  $W$ . Suponha que o modelo é corretamente especificado tal que  $E[g(w_t; \theta_0)] = \mathbf{0}$  (i.e. valem as condições de ortogonalidade) para  $\theta_0 \in \Theta$ . Suponha que (i) o espaço dos parâmetros  $\Theta$  é um subconjunto compacto de  $\mathbb{R}^p$ , (ii)  $g(w_t; \theta)$  é contínuo em  $\theta$  para todo  $w_t$  e (iii)  $g(w_t; \theta)$  é mensurável em  $w_t$  para todo  $\theta$  em  $\Theta$  (então  $\hat{\theta}$  é uma variável aleatória



bem definida pelo Lema 7.1). Além disso, suponha que

1. (identificação)  $E[g(\mathbf{w}_t; \boldsymbol{\theta})] \neq \mathbf{0}$  para todo  $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$  em  $\Theta$
2. (dominância)  $E[\sup_{\boldsymbol{\theta} \in \Theta} \|g(\mathbf{w}_t; \boldsymbol{\theta})\|] < \infty$

Então  $\hat{\boldsymbol{\theta}} \rightarrow_p \boldsymbol{\theta}_0$ .

Quando  $g(\mathbf{w}_t; \boldsymbol{\theta})$  é não-linear em  $\boldsymbol{\theta}$ , especificar as condições primitivas para identificação e dominância na Proposição 7.7 é geralmente muito difícil. Então na maioria das aplicações se assumem as condições da Proposição 7.7.

## Normalidade assintótica para Estimador-M\*

A função objetivo do estimador-M é

$$Q_n(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^n m(\mathbf{w}_t; \boldsymbol{\theta}) \quad (51)$$

É conveniente fornecer símbolos ao gradiente (vetor das primeiras derivadas) e ao Hessiano (matriz das segundas derivadas) da função  $m$  como

$$\underset{p \times 1}{s(\mathbf{w}_t; \boldsymbol{\theta})} = \frac{\partial m(\mathbf{w}_t; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \quad (52)$$

$$\underset{p \times p}{\mathbf{H}(\mathbf{w}_t; \boldsymbol{\theta})} = \frac{\partial s(\mathbf{w}_t; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}'} = \frac{\partial^2 m(\mathbf{w}_t; \boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \quad (53)$$

$s(\mathbf{w}_t; \boldsymbol{\theta})$  é denominado *vetor score para observação  $t$* .  $\mathbf{H}(\mathbf{w}_t; \boldsymbol{\theta})$  será definido como o *Hessiano para a observação  $t$* . (obs: não confundir

\*Prova de normalidade assintótica em separado para estimador-M e GMM.

com outra definição de *score* e Hessiano para  $Q_n(\boldsymbol{\theta})$ ).

Hipóteses necessárias para normalidade assintótica. Assuma que  $m(\boldsymbol{w}_t; \boldsymbol{\theta})$  é diferenciável em  $\boldsymbol{\theta}$  e que  $\hat{\boldsymbol{\theta}}$  está no interior de  $\Theta$ . Então, se  $\hat{\boldsymbol{\theta}}$  satisfaz as condições de primeira ordem:

$$\mathbf{0}_{p \times 1} = \frac{\partial Q_n(\hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}} = \frac{1}{n} \sum_{t=1}^n s(\boldsymbol{w}_t; \boldsymbol{\theta}) \quad (54)$$

Usar o seguinte resultado de cálculo:

**Teorema do Valor Médio:** Faça  $h : \mathbb{R}^p \rightarrow \mathbb{R}^q$  ser continuamente diferenciável. Então  $h(\boldsymbol{x})$  admite a expansão do valor médio

$$\mathbf{h}(\boldsymbol{x}) = \mathbf{h}(\boldsymbol{x}_0) + \frac{\partial \mathbf{h}(\bar{\boldsymbol{x}})}{\partial \boldsymbol{x}'} (\boldsymbol{x} - \boldsymbol{x}_0) \quad (55)$$

tal que  $\bar{\boldsymbol{x}}$  é o valor médio permanecendo entre  $\boldsymbol{x}$  e  $\boldsymbol{x}_0$ .

Fazendo  $q = p$ ,  $x = \hat{\theta}$ ,  $x_0 = \theta_0$  e  $h(\cdot) = \partial Q_n(\cdot)/\partial \theta$  no Teorema do Valor Médio, obtemos as seguintes expansões do valor médio:

$$\begin{aligned} \frac{\partial Q_n(\hat{\theta})}{\frac{\partial \theta}{p \times 1}} &= \frac{\partial Q_n(\theta_0)}{\frac{\partial \theta}{p \times 1}} + \frac{\partial^2 Q_n(\bar{\theta})}{\frac{\partial \theta \partial \theta'}{p \times p}} (\hat{\theta} - \theta_0) = (56) \\ &= \frac{1}{n} \sum_{t=1}^n s(\mathbf{w}_t; \theta_0) + \left[ \frac{1}{n} \sum_{t=1}^n \mathbf{H}(\mathbf{w}_t; \bar{\theta}) \right] (\hat{\theta} - \theta_0) \end{aligned}$$

tal que  $\bar{\theta}$  é o valor médio que está entre  $\hat{\theta}$  e  $\theta_0$ .<sup>†</sup> Combinando esta equação com a CPO:

$$\mathbf{0}_{p \times 1} = \frac{1}{n} \sum_{t=1}^n s(\mathbf{w}_t; \theta_0) + \left[ \frac{1}{n} \sum_{t=1}^n \mathbf{H}(\mathbf{w}_t; \bar{\theta}) \right] (\hat{\theta} - \theta_0) \quad (57)$$

assumindo que  $\frac{1}{n} \sum_{t=1}^n \mathbf{H}(\mathbf{w}_t; \bar{\theta})$  é não-singular, esta equação pode ser solucionada para  $\hat{\theta} - \theta_0$

<sup>†</sup>O requerimento de diferenciabilidade contínua é satisfeito se  $m(\mathbf{w}_t; \theta)$  é continuamente diferenciável duplamente com respeito a  $\theta$ .

para resultar

$$\sqrt{n} (\hat{\theta} - \theta_0) = - \left[ \frac{1}{n} \sum_{t=1}^n \mathbf{H}(w_t; \bar{\theta}) \right]^{-1} \sqrt{n} \frac{1}{n} \sum_{t=1}^n s(w_t; \theta_0) \quad (58)$$

A expressão para o erro amostral (multiplicado por  $\sqrt{n}$ ) é denominado de *expansão do valor médio para o erro amostral*. O score vector é solucionado no valor de parâmetro verdadeiro  $\theta_0$ .

Como  $\bar{\theta}$  permanece entre  $\hat{\theta}$  e  $\theta_0$ ,  $\bar{\theta}$  é consistente para  $\theta_0$  se  $\hat{\theta}$  é. Se  $\{w_t\}$  é estacionária ergódica é natural conjecturar que

$$\frac{1}{n} \sum_{t=1}^n \mathbf{H}(w_t; \bar{\theta}) \rightarrow_p \mathbb{E}[\mathbf{H}(w_t; \theta_0)] \quad (59)$$

A estacionariedade ergódica de  $w_t$  e a consistência de  $\bar{\theta}$  não são suficientes para garantir o resultado acima. Também é necessário assumir convergência uniforme em (59) em uma

vizinhança de  $\theta_0$ ). Mas pelo Teorema da Convergência Uniforme, a convergência uniforme de  $\frac{1}{n} \sum_{t=1}^n \mathbf{H}(w_t; \bar{\theta})$  é satisfeita se a seguinte condição de dominância é satisfeita para o Hessiano: para alguma vizinhança  $\mathfrak{N}$  de  $\theta_0$ ,  $E[\sup_{\theta \in \mathfrak{N}} \|\mathbf{H}(w_t; \theta)\|] < \infty$ . Esta é uma condição suficiente para (59).

Finalmente, se (59) se sustenta e se

$$\frac{1}{\sqrt{n}} \frac{1}{n} \sum_{t=1}^n s(w_t; \theta_0) \rightarrow_d N(\mathbf{0}, \Sigma) \quad (60)$$

então pelo Teorema de Slutsky (Lema 2.4c) temos:

$$\sqrt{n} (\hat{\theta} - \theta_0) \rightarrow_d N \left( \mathbf{0}, (E[\mathbf{H}(w_t; \theta_0)])^{-1} \Sigma (E[\mathbf{H}(w_t; \theta_0)]) \right) \quad (61)$$

## Normalidade assintótica dos estimadores-M

**Proposição 7.8:** Suponha que as condições tanto da Proposição 7.3 ou da Proposição 7.4 são satisfeitas, tal que  $\{\mathbf{w}_t\}$  é estacionário ergódico e o estimador M  $\hat{\boldsymbol{\theta}}$  definido por (7.1.1) e (7.1.2) é consistente. Além disso, suponha que

1.  $\boldsymbol{\theta}_0$  é interior de  $\Theta$ ;
2.  $m(\mathbf{w}_t; \boldsymbol{\theta})$  é continuamente diferenciável duplamente em  $\boldsymbol{\theta}$  para qualquer  $\mathbf{w}_t$ ,
3.  $\frac{1}{n} \sum_{t=1}^n \mathbf{s}(\mathbf{w}_t; \boldsymbol{\theta}_0) \rightarrow_d N(\mathbf{0}, \boldsymbol{\Sigma})$ ,  $\boldsymbol{\Sigma}$  é positiva definida, tal que  $\mathbf{s}(\mathbf{w}_t; \boldsymbol{\theta})$  é definida em (7.3.2),
4. (condição de dominância local sobre o Hessiano) para alguma vizinhança de  $\aleph$

de  $\theta_0$ ),

$$E[\sup_{\theta \in \mathcal{N}} \|\mathbf{H}(\mathbf{w}_t; \theta)\|] < \infty$$

tal que para qualquer estimador consistente de  $\bar{\theta}$ ,  $\frac{1}{n} \sum_{t=1}^n \mathbf{H}(\mathbf{w}_t; \bar{\theta}) \rightarrow_p E[\mathbf{H}(\mathbf{w}_t; \theta_0)]$ , tal que  $\mathbf{H}(\mathbf{w}_t; \theta_0)$  é definido em (53).

5.  $E[\mathbf{H}(\mathbf{w}_t; \theta_0)]$  é não singular. Então  $\hat{\theta}$  é assintoticamente normal com

$$\text{Avar}(\hat{\theta}) = (E[\mathbf{H}(\mathbf{w}_t; \theta_0)])^{-1} \Sigma (E[\mathbf{H}(\mathbf{w}_t; \theta_0)])^{-1}$$



## Estimação da Variância Assintótica Consistente

Para usar este resultado da variância assintótica para teste de hipótese é necessária a estimativa consistente de

$$\text{Avar}(\hat{\theta}) = (\text{E}[H(w_t; \theta_0)])^{-1} \Sigma (\text{E}[H(w_t; \theta_0)])^{-1} \quad (62)$$

Uma vez que  $\hat{\theta} \rightarrow_p \theta_0$ , condição 4 da Proposição 7.8 implica que

$$\frac{1}{n} \sum_{t=1}^n H(w_t; \hat{\theta}) \rightarrow_p \text{E}[H(w_t; \theta_0)]$$

além disso, dado que existe um estimador consistente de  $\hat{\Sigma}$  de  $\Sigma$  disponível:

$$\widehat{\text{Avar}}(\hat{\theta}) = \left( \frac{1}{n} \sum_{t=1}^n H(w_t; \hat{\theta}) \right)^{-1} \hat{\Sigma} \left( \frac{1}{n} \sum_{t=1}^n H(w_t; \hat{\theta}) \right)^{-1} \quad (63)$$

é uma estimativa consistente da matriz de variância assintótica.

## Normalidade assintótica para ML condicional

Especializar a normalidade assintótica para ML. Considere as duas igualdades do item 3 da Proposição 7.9 a seguir.

A segunda igualdade é chamada de *matriz de igualdade de informação* dado que  $E[s(\mathbf{w}_t; \boldsymbol{\theta}_0)s(\mathbf{w}_t; \boldsymbol{\theta}_0)']$  é a matriz de informação para a observação  $t$ . Implicação destas duas igualdades para a Proposição 7.8: se  $\mathbf{w}_t$  é i.i.d., o Teorema do Limite Central Lindeberg-Levy e a primeira igualdade ( $E[s(\mathbf{w}_t; \boldsymbol{\theta}_0)] = \mathbf{0}$ ) implicam em

$$\frac{1}{n} \sum_{t=1}^n s(\mathbf{w}_t; \boldsymbol{\theta}_0) \rightarrow_d N(\mathbf{0}, \boldsymbol{\Sigma}) \quad (64)$$

tal que

$$\boldsymbol{\Sigma} = E[s(\mathbf{w}_t; \boldsymbol{\theta}_0)s(\mathbf{w}_t; \boldsymbol{\theta}_0)']$$

Isto e a matriz de igualdade de informação implicam que  $\text{Avar}(\hat{\theta})$  na Proposição 7.8 é simplificada em duas formas como:

$$\begin{aligned}\text{Avar}(\hat{\theta}) &= -\{E[\mathbf{H}(w_t; \theta_0)]\}^{-1} = & (65) \\ &= \{E[s(w_t; \theta_0)s(w_t; \theta_0)']\}^{-1}\end{aligned}$$

Esta é a prova da Proposição 7.9.

## Normalidade assitótica para ML condicional

**Proposição 7.9:** Faça  $w_t (= (y_t, \mathbf{x}'_t)')$  ser iid. Suponha que as condições tanto da Proposição 7.5 ou 7.6 são satisfeitas, tal que  $\hat{\boldsymbol{\theta}} \rightarrow_p \boldsymbol{\theta}_0$ . Além disso, suponha que

1.  $\boldsymbol{\theta}_0$  é um ponto interior de  $\Theta$ ;
2.  $f(y_t | \mathbf{x}_t; \boldsymbol{\theta})$  é continuamente diferenciável duplamente em  $\boldsymbol{\theta}$  para todo  $(y_t, \mathbf{x}_t)$ ;
3.  $E[s(\mathbf{w}_t; \boldsymbol{\theta}_0)] = 0$  e  $-E[\mathbf{H}(\mathbf{w}_t; \boldsymbol{\theta}_0)] = E[s(\mathbf{w}_t; \boldsymbol{\theta}_0)s'$  tal que as funções  $s$  e  $\mathbf{H}$  são definidas em (52) e (53);
4. (condição de dominância local para o Hessiano) para alguma vizinhança de  $\mathfrak{N}$  de  $\boldsymbol{\theta}_0$ )

$$E[\sup_{\boldsymbol{\theta} \in \mathfrak{N}} \|\mathbf{H}(\mathbf{w}_t; \boldsymbol{\theta})\|] < \infty$$

tal que para qualquer estimador consistente de  $\boldsymbol{\theta}$ ,  $\frac{1}{n} \sum_{t=1}^n \mathbf{H}(\mathbf{w}_t; \bar{\boldsymbol{\theta}}) \rightarrow_p \mathbf{E}[\mathbf{H}(\mathbf{w}_t; \boldsymbol{\theta}_0)]$ ;

5.  $\mathbf{E}[\mathbf{H}(\mathbf{w}_t; \boldsymbol{\theta}_0)]$  é não singular.

Então  $\hat{\boldsymbol{\theta}}$  é assintoticamente normal com  $\text{Avar}(\hat{\boldsymbol{\theta}})$  dada por

$$\begin{aligned} \text{Avar}(\hat{\boldsymbol{\theta}}) &= -\{\mathbf{E}[\mathbf{H}(\mathbf{w}_t; \boldsymbol{\theta}_0)]\}^{-1} = \\ &= \{\mathbf{E}[s(\mathbf{w}_t; \boldsymbol{\theta}_0)s(\mathbf{w}_t; \boldsymbol{\theta}_0)']\}^{-1} \end{aligned}$$

## Estimativa de $\text{Avar}(\hat{\theta})$

A estimação da variância assintótica pode seguir duas formas contidas em (65). Uma utilizando Hessiano e outro o vetor de primeira derivada.

$$1o \text{ estimador de } \text{Avar}(\hat{\theta}) = \left( \frac{1}{n} \sum_{t=1}^n \mathbf{H}(w_t; \hat{\theta}) \right)^{-1} \quad (66)$$

Como  $\hat{\theta} \rightarrow_p \theta_0$ , este estimador é consistente com a Proposição 7.9.

Outro estimador baseado na relação  $\text{Avar}(\hat{\theta}) = \{E[s(w_t; \theta_0)s(w_t; \theta_0)']\}^{-1}$  é

$$2o \text{ estimador de } \text{Avar}(\hat{\theta}) = \left\{ \frac{1}{n} \sum_{t=1}^n s(w_t; \hat{\theta})s(w_t; \hat{\theta})' \right\}^{-1} \quad (67)$$

Para garantir a consistência para este estimador é preciso uma hipótese adicional na Proposição 7.9 que não precisa ser demonstrada

pois é observada diretamente. A vantagem do segundo estimador é que se usa apenas a primeira derivada e para métodos de aproximação numérica isto é muito importante.\*

\*A vantagem da fórmula do Hessiano é quando se aplica o problema à amostras finitas.

## Exemplo 7.10

**Normalidade assintótica no modelo de regressão linear.** Para reproduzir o log da densidade condicional para observação  $t$  para a regressão linear com erro normal:

$$\log f(y_t | \mathbf{x}_t; \boldsymbol{\theta}) = -\frac{1}{2} \log(2\pi) - \frac{1}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} (y_t - \mathbf{x}'_t \boldsymbol{\beta})^2 \quad (68)$$

com  $\Theta = \Re^K \times \Re_{++}$  ( $K$  é a dimensão de  $\boldsymbol{\beta}$ ), condição 1 da Proposição 7.9 é satisfeita. Condição 2 é obviamente satisfeita. Para verificar a Condição 3:

$$\mathbf{s}(\mathbf{w}_t; \hat{\boldsymbol{\theta}}) = \begin{pmatrix} \frac{1}{\sigma^2} \mathbf{x}_t \cdot \hat{\varepsilon}_t \\ -\frac{1}{2\sigma^2} + \frac{1}{2\sigma^4} \hat{\varepsilon}_t^2 \end{pmatrix}$$

$$\mathbf{H}(\mathbf{w}_t; \hat{\boldsymbol{\theta}}) = \begin{pmatrix} -\frac{1}{\sigma^2} \mathbf{x}_t \cdot \mathbf{x}'_t & -\frac{1}{\sigma^4} \mathbf{x}_t \cdot \hat{\varepsilon}_t \\ -\frac{1}{\sigma^4} \mathbf{x}_t \cdot \hat{\varepsilon}_t & -\frac{1}{2\sigma^4} + \frac{1}{2\sigma^6} \hat{\varepsilon}_t^2 \end{pmatrix} \quad (69)$$

$$\mathbf{s}(\mathbf{w}_t; \hat{\boldsymbol{\theta}}) \mathbf{s}(\mathbf{w}_t; \hat{\boldsymbol{\theta}})' = \begin{pmatrix} -\frac{1}{\sigma^4} \mathbf{x}_t \mathbf{x}'_t \hat{\varepsilon}_t^2 & -\frac{1}{2\sigma^4} \mathbf{x}_t \cdot \hat{\varepsilon}_t + \frac{1}{2\sigma^6} \hat{\varepsilon}_t^3 \\ -\frac{1}{2\sigma^4} \mathbf{x}_t \cdot \hat{\varepsilon}_t + \frac{1}{2\sigma^6} \mathbf{x}_t \cdot \hat{\varepsilon}_t^3 & \frac{1}{4\sigma^4} - \frac{1}{2\sigma^6} \hat{\varepsilon}_t^2 \end{pmatrix}$$



tal que  $\mathbf{w}_t = (y_t, \mathbf{x}'_t)'$ ,  $\boldsymbol{\theta} = (\boldsymbol{\beta}', \sigma^2)'$  e  $\hat{\varepsilon}_t = y_t - \mathbf{x}'_t \boldsymbol{\beta}$  (este último é diferente de  $\varepsilon_t = y_t - \mathbf{x}'_t \boldsymbol{\beta}_0$ ). Para  $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ ,  $\hat{\varepsilon}_t$  pode ser substituído por  $\varepsilon_t$ .

No modelo de regressão linear,  $E(\varepsilon_t | \mathbf{x}_t) = 0$ . Também, dado que  $\varepsilon_t$  é  $N(0, \sigma_0^2)$ , temos  $E(\varepsilon_t^3) = 0$  e  $E(\varepsilon_t^4) = 3\sigma_0^4$ . Usando estas relações, é fácil verificar item 3 da Proposição 7.9. Em particular:

$$\begin{aligned} -E[\mathbf{H}(\mathbf{w}_t; \hat{\boldsymbol{\theta}}_0)] &= E[s(\mathbf{w}_t; \hat{\boldsymbol{\theta}})s(\mathbf{w}_t; \hat{\boldsymbol{\theta}})'] = \quad (70) \\ &= \begin{bmatrix} \frac{1}{\sigma_0^2} E[\mathbf{x}_t \mathbf{x}'_t] & \mathbf{0} \\ \mathbf{0}' & \frac{1}{2\sigma_0^4} \end{bmatrix} \end{aligned}$$

Se  $E[\mathbf{x}_t \mathbf{x}'_t]$  é não-singular, então  $E[\mathbf{H}(\mathbf{w}_t; \hat{\boldsymbol{\theta}}_0)]$  é não-singular e o item 5 é satisfeito.

Considerando a condição 4, faça  $\tilde{\varepsilon}_t = y_t - \mathbf{x}'_t \tilde{\boldsymbol{\beta}}$  para algum estimador consistente  $\tilde{\boldsymbol{\beta}}$  e  $\tilde{\sigma}^2$ .

Condição (59) neste exemplo é

$$\left( \begin{array}{cc} -\frac{1}{\sigma^2} \frac{1}{n} \sum_{t=1}^n \mathbf{x}_t \cdot \mathbf{x}'_t & -\frac{1}{\sigma^4} \frac{1}{n} \sum_{t=1}^n \mathbf{x}_t \cdot \hat{\varepsilon}_t \\ -\frac{1}{\sigma^4} \frac{1}{n} \sum_{t=1}^n \mathbf{x}_t \cdot \hat{\varepsilon}_t & -\frac{1}{2\sigma^4} + \frac{1}{2\sigma^6} \frac{1}{n} \sum_{t=1}^n \hat{\varepsilon}_t^2 \end{array} \right) \quad (71)$$

$$\rightarrow_p \begin{bmatrix} \frac{1}{\sigma_0^2} \mathbf{E}[\mathbf{x}_t \mathbf{x}'_t] & \mathbf{0} \\ \mathbf{0}' & \frac{1}{2\sigma_0^4} \end{bmatrix}$$

é direto mostrar que  $\tilde{\varepsilon}_t = \varepsilon_t - \mathbf{x}'_t(\tilde{\beta} - \beta_0)$ . Concluimos que todas as condições da Proposição 7.9 são satisfeitas se  $\mathbf{E}[\mathbf{x}_t \mathbf{x}'_t]$  é não singular.

## Exemplo 7.11

**Normalidade assintótica do ML probit.** Reproduzindo a log verossimilhança condicional do modelo probit

$$\log f(y_t | \mathbf{x}_t; \boldsymbol{\theta}) = y_t \log \Phi(\mathbf{x}'_t \boldsymbol{\theta}) + (1 - y_t) \log [1 - \Phi(\mathbf{x}'_t \boldsymbol{\theta})] \quad (72)$$

Com  $\Theta = \mathbb{R}^p$ , condição 1 da Proposição 7.9 é satisfeita. Condição 2 é obviamente satisfeita. Para verificar 3, se tem:

$$\mathbf{s}(\mathbf{w}_t; \hat{\boldsymbol{\theta}}) = \frac{(y_t - \Phi(\mathbf{x}'_t \hat{\boldsymbol{\theta}})) \Phi(\mathbf{x}'_t \hat{\boldsymbol{\theta}})}{(1 - \Phi(\mathbf{x}'_t \hat{\boldsymbol{\theta}})) \Phi(\mathbf{x}'_t \hat{\boldsymbol{\theta}})} \mathbf{x}_t \quad (73)$$

$$\begin{aligned} \mathbf{H}(\mathbf{w}_t; \hat{\boldsymbol{\theta}}) = & \left\{ - \left[ \frac{y_t - \Phi(\mathbf{x}'_t \hat{\boldsymbol{\theta}})}{(1 - \Phi(\mathbf{x}'_t \hat{\boldsymbol{\theta}})) \Phi(\mathbf{x}'_t \hat{\boldsymbol{\theta}})} \right]^2 [\Phi(\mathbf{x}'_t \hat{\boldsymbol{\theta}})]^2 \right. \\ & \left. + \left[ \frac{y_t - \Phi(\mathbf{x}'_t \hat{\boldsymbol{\theta}})}{(1 - \Phi(\mathbf{x}'_t \hat{\boldsymbol{\theta}})) \Phi(\mathbf{x}'_t \hat{\boldsymbol{\theta}})} \right] \Phi'(\mathbf{x}'_t \hat{\boldsymbol{\theta}}) \right\} \mathbf{x}_t \mathbf{x}'_t \quad (74) \end{aligned}$$

Aqui  $\Phi'(\cdot)$  é função de densidade cumulativa de  $N(0, 1)$  e  $\phi(\cdot) = \Phi'(\cdot)$  é a função de densidade.

É fácil provar a condição 3:

$$E[s(\mathbf{w}_t; \boldsymbol{\theta}_0) \mid \mathbf{x}_t] = \mathbf{0} \text{ e}$$

$$-E[\mathbf{H}(\mathbf{w}_t; \boldsymbol{\theta}_0) \mid \mathbf{x}_t] = E[s(\mathbf{w}_t; \boldsymbol{\theta}_0)s(\mathbf{w}_t; \boldsymbol{\theta}_0)' \mid \mathbf{x}_t] \quad (75)$$

Então 3 é satisfeita pela Lei das Expectativas Totais. Em particular para o probit se pode mostrar que

$$\begin{aligned} -E[\mathbf{H}(\mathbf{w}_t; \boldsymbol{\theta}_0)] &= E[s(\mathbf{w}_t; \boldsymbol{\theta}_0)s(\mathbf{w}_t; \boldsymbol{\theta}_0)'] = \\ &= E[\lambda(\mathbf{x}'_t\boldsymbol{\theta}_0)\lambda(-\mathbf{x}_t\boldsymbol{\theta}_0)'\mathbf{x}_t\mathbf{x}'_t] \end{aligned} \quad (76)$$

tal que

$$\lambda(v) = \frac{\phi(v)}{1 - \Phi(v)} \quad (77)$$

é chamada de *inversa da razão de Mill* ou *hazard* para  $N(0, 1)$ . Pode ser mostrado que em (75) o termo entre chaves está entre 0 e 2. Então

$$\|\mathbf{H}(\mathbf{w}_t; \boldsymbol{\theta}_0)\| \leq 2 \|\mathbf{x}_t\mathbf{x}'_t\| \quad (78)$$

A norma Euclidiana  $\| \mathbf{x}_t \mathbf{x}'_t \|$  é a raiz quadrada da soma dos quadrados dos elementos de  $\mathbf{x}_t \mathbf{x}'_t$ . Além disso, a expectativa de  $\| \mathbf{x}_t \mathbf{x}'_t \|^2$  e assim de  $\| \mathbf{x}_t \mathbf{x}'_t \|$  são finitos se  $E[\mathbf{x}_t \mathbf{x}'_t]$  existe e é finito. Então a condição de *dominância local* do Hessiano (condição 4) é satisfeita se  $E[\mathbf{x}_t \mathbf{x}'_t]$  é não-singular. Se conclui que todas as condições são satisfeitas se  $E[\mathbf{x}_t \mathbf{x}'_t]$  é não-singular.

## Normalidade assintótica do GMM

A função objetivo do GMM é:

$$Q_n(\boldsymbol{\theta}) = -\frac{1}{2}g_n(\boldsymbol{\theta})'\widehat{\mathbf{W}}g_n(\boldsymbol{\theta}) \quad (79)$$

$$g_n(\boldsymbol{\theta}) = \frac{1}{n} \sum_{t=1}^n g(\mathbf{w}_t; \boldsymbol{\theta})$$

$K \times 1$                        $n$                        $K \times 1$

O Teorema do Valor Médio é aplicado em  $g_n(\boldsymbol{\theta})$  e não na primeira derivada. Em GMM, a função objetivo precisa ser diferenciável uma vez – não duas vezes como no ML. Isto ocorre porque a média amostral entra na função objetivo de forma diferente no caso GMM.

Assumindo que  $\hat{\boldsymbol{\theta}}$  é um ponto interior, a CPO para maximização da função objetivo é:

$$\mathbf{0}_{(p \times 1)} = \frac{\partial Q_n(\hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}_{(p \times 1)}} = -\mathbf{G}_n(\hat{\boldsymbol{\theta}})'_{(p \times K)} \widehat{\mathbf{W}}_{(K \times K)} \mathbf{g}_n(\hat{\boldsymbol{\theta}})_{(K \times 1)} \quad (80)$$

$G_n(\theta)$  é o Jacobiano de  $g_n(\theta)$

$$G_n(\theta) \equiv \frac{\partial g_n(\theta)}{\partial \theta} \quad (81)$$

Aplicando o Teorema do Valor Médio para  $g_n(\theta)$  se obtém:

$$\underset{(p \times 1)}{\mathbf{0}} = -\underset{(p \times K)}{G_n(\hat{\theta})'} \underset{(K \times K)}{\hat{W}} \underset{(K \times 1)}{g_n(\theta_0)} - \underset{(p \times K)}{G_n(\hat{\theta})'} \underset{(K \times K)}{\hat{W}} \underset{(K \times p)}{G_n(\bar{\theta})} (\hat{\theta} - \theta_0) \quad (82)$$

Solucionando para  $(\hat{\theta} - \theta_0)$  e substituindo  $g_n(\theta)$  por  $\frac{1}{n} \sum_{t=1}^n g(w_t; \theta)$ :

$$\underset{(p \times 1)}{(\hat{\theta} - \theta_0)} = - \left[ \underset{(p \times K)}{G_n(\hat{\theta})'} \underset{(K \times K)}{\hat{W}} \underset{(K \times p)}{G_n(\bar{\theta})} \right]^{-1} - \underset{(p \times K)}{G_n(\hat{\theta})'} \underset{(K \times K)}{\hat{W}} \frac{1}{\sqrt{n}} \quad (83)$$

Nesta expressão o Jacobiano  $G_n(\theta)$  é solucionado em dois pontos diferentes,  $\hat{\theta}$  e  $\bar{\theta}$ . Como  $g_n(\theta) = \frac{1}{n} \sum_{t=1}^n g(w_t; \theta)$ , o Jacobiano em qualquer estimador  $\tilde{\theta}$  dado, pode ser escrito como

$$G_n(\tilde{\theta}) \equiv \frac{\partial g_n(\theta)}{\partial \tilde{\theta}} \quad (84)$$

Se  $w_t$  é estacionário ergódico e  $\tilde{\theta}$  é consistente, é natural conjecturar que esta expressão converge para  $E[\frac{\partial g_n(\theta)}{\partial \theta_0}]$ . Como foi verdade para o estimador GMM, esta conjectura é verdade se  $\frac{\partial g_n(\theta)}{\partial \theta_0}$  satisfaz a condição de dominância do item 4 da Proposição 7.10 a seguir. A Proposição 7.8 para GMM é apresentada a seguir.



## Normalidade assintótica do GMM

**Proposição 7.10** Suponha que as condições tanto da Proposição 7.7 são satisfeitas, tal que  $\{w_t\}$  é estacionário ergódico,  $\hat{W}(K \times K)$  converge em probabilidade para uma matriz  $W$  simétrica positiva definida e o estimador GMM de  $\hat{\theta}$  com dimensão  $p$  é consistente. Além disso, suponha que

1.  $\theta_0$  é interior de  $\Theta$ ;
2.  $g(w_t; \theta)$  é continuamente diferenciável  $(K \times 1)$  duplamente em  $\theta$  para qualquer  $w_t$ ,
3.  $\frac{1}{n} \sum_{t=1}^n g(w_t; \theta_0) \rightarrow_d N(\mathbf{0}, S)$ ,  $S (K \times K)$  é positiva definida;
4. (condição de dominância local sobre  $\frac{\partial g_n(\theta)}{\partial \theta}$ ) para alguma vizinhança de  $\mathfrak{N}$  de  $\theta_0$ ,

$$E\left[\sup_{\theta \in \mathfrak{N}} \left\| \frac{\partial g_n(\theta)}{\partial \theta} \right\| \right] < \infty$$

tal que para qualquer estimador consistente de  $\theta$ ,  $\frac{1}{n} \sum_{t=1}^n \frac{\partial g_n(\theta)}{\partial \theta} \rightarrow_p E\left[\frac{\partial g_n(\theta)}{\partial \theta}\right]$ .

5.  $E\left[\frac{\partial g_n(\theta)}{\partial \theta}\right]$  é coluna posto completo (full column rank).

(a) (normalidade assintótica)  $\hat{\theta}$  é assintoticamente normal com

$$\text{Avar}(\hat{\theta}) = \left(\hat{G}' \hat{W} \hat{G}\right)^{-1} \left(\hat{G}' \hat{W} \hat{S} \hat{W} \hat{G}\right) \left(\hat{G}' \hat{W} \hat{G}\right)^{-1}$$

tal que  $\hat{G}_{K \times p} \equiv G_n(\hat{\theta}) = \frac{\partial g_n(\hat{\theta})}{\partial \theta}$

A hipótese de que  $S$  é positiva definida não é necessária neste ponto mas será usada a frente.

## GMM linear e não-linear

Comparação de GMM linear e não-linear: Proposição 3.1 vs Proposição 7.10.

No linear:

$$\mathbf{g}_n(\boldsymbol{\theta}) = \left( \frac{1}{n} \sum_{t=1}^n \mathbf{x}_t y_t \right) - \left( \frac{1}{n} \sum_{t=1}^n \mathbf{x}_t \mathbf{z}'_t \right) \boldsymbol{\theta} \quad (85)$$

Então  $\hat{\mathbf{G}} = \mathbf{G}_n(\hat{\boldsymbol{\theta}})$  na Proposição 7.10 se reduz a  $-\left(\frac{1}{n} \sum_{t=1}^n \mathbf{x}_t \mathbf{z}'_t\right)$  e  $\mathbf{G}$  se reduz a  $-\mathbb{E}[\mathbf{x}_t \mathbf{z}'_t]$ . A verdade da Proposição 7.10 é que a distribuição assintótica do estimador GMM não-linear pode ser obtido assumindo a aproximação linear do  $\mathbf{g}_n(\boldsymbol{\theta})$  em torno de parâmetro verdadeiro.

## GMM Não-linear: Hansen (2019, cap. 13)

GMM pode ser aplicado em qualquer situação em que um modelo implica em  $K \times 1$  momentos:

$$E(\mathbf{g}_i(\delta)) = 0 \quad (86)$$

aqui  $\mathbf{g}$  é uma possível função não-linear dos parâmetros  $\delta$ . Às vezes isso é tudo que é conhecido. Identificação requer que  $K \geq L$ . O estimador GMM minimiza:

$$J(\tilde{\delta}) = n \cdot \bar{\mathbf{g}}_n'(\delta) \widehat{W} \bar{\mathbf{g}}_n(\delta) \quad (87)$$

para uma matriz de pesos  $\widehat{W}$  e

$$\bar{\mathbf{g}}_n(\delta) = \frac{1}{n} \sum_{i=1}^n \mathbf{g}_n(\delta)$$

O estimador GMM eficiente pode ser construído determinando

$$\widehat{W} = \left( \frac{1}{n} \sum_{i=1}^n \hat{\mathbf{g}}_i \hat{\mathbf{g}}_i' - \bar{\mathbf{g}}_n \bar{\mathbf{g}}_n' \right)^{-1} \quad (88)$$

com  $\hat{g}_i = g(w_i, \tilde{\delta})$  construído usando um estimador preliminar de  $\tilde{\delta}$ , pode ser utilizando  $\widehat{W} = I_K$ .

### Distribuição do Estimador GMM Não-linear:

Sob condições de regularidade gerais,

$$\sqrt{1/n} (\hat{\delta}_{gmm} - \delta) \rightarrow_d N(0, V_\delta) \quad (89)$$

tal que

$$V_\delta = (Q'WQ)^{-1} (Q'WSWQ) (Q'WQ)^{-1} \quad (90)$$

$$S = E(g_i g_i') \quad (91)$$

e

$$Q = E\left(\frac{\partial}{\partial \delta'} g_i(\delta)\right) \quad (92)$$

$Q = G$  i.e. compatibilizando notações de Hansen e Hayashi.

Se a matriz eficiente de pesos é utilizada, então:

$$V_\delta = (Q'S^{-1}Q)^{-1} \quad (93)$$

## ML vs GMM

Tomando as condições de ortogonalidade como dadas, qual a escolha ótima da matriz de pesos  $W$ ?

Caso iid com densidade de  $w_t$  dada por  $f(w_t; \theta)$ . Faça  $\hat{\theta}$  ser o estimador GMM associado com as condições de ortogonalidade  $E[g(w_t; \theta_0)] = 0$ . Sua variância assintótica é dada na Proposição 7.10. Pode se mostrar que:

$$\text{Avar}(\hat{\theta}) \geq E[s(w_t; \theta_0)s(w_t; \theta_0)']^{-1} \quad (94)$$

tal que

$$s(w_t; \theta_0) = \frac{\partial g(w_t; \theta_0)}{\partial \theta}$$

Isto é, a inversa da matriz de informação  $E[ss']$  é o limite inferior para a variância assintótica dos estimadores GMM. Esta matriz é ativa sob as condições da Proposição 7.10 mais algumas

condições técnicas sobre  $f(\mathbf{w}_t; \boldsymbol{\theta})$  que permite a conexão entre diferenciação e integração.

Eficiência assintótica do ML sobre GMM depende deste resultado porque o limite inferior  $E[\mathbf{s}\mathbf{s}']^{-1}$  é a variância assintótica do estimador ML. A superioridade do ML não surpreende dado que ele explora o conhecimento da forma paramétrica da função de densidade  $f(\mathbf{w}_t; \boldsymbol{\theta})$ , enquanto o GMM não faz isso.

O GMM atinge seu valor inferior quando a função  $\mathbf{g}$  da condição de ortogonalidade é o *score* para a observação  $t$ :

$$\underset{(K \times 1)}{\mathbf{g}(\mathbf{w}_t; \boldsymbol{\theta})} = \underset{(p \times 1)}{\mathbf{s}(\mathbf{w}_t; \boldsymbol{\theta})} \equiv \frac{\partial \mathbf{g}(\mathbf{w}_t; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \quad (95)$$

Portanto, o estimador GMM com condições de ortogonalidade ótimas é assintoticamente equivalente ao ML. Na verdade eles são numericamente equivalentes.\*

\*Uma vez que as condições de ortogonalidade  $K =$

## Erro amostral em formato comum

Prova diferente da normalidade assintótica mais adequada para teste de hipótese. Considere primeiro estimadores-M.

Notando que  $\frac{\partial Q_n(\theta_0)}{\partial \theta} = \frac{1}{n} \sum_{t=1}^n s(\mathbf{w}_t; \theta)$ , a expansão do valor médio em (58) pode ser escrita como

$$\sqrt{n} (\hat{\theta} - \theta_0) = - \left[ \frac{1}{n} \sum_{t=1}^n \mathbf{H}(\mathbf{w}_t; \bar{\theta}) \right]^{-1} \sqrt{n} \frac{\partial Q_n(\theta_0)}{\partial \theta} \quad (96)$$

Pela condição 4 da Proposição 7.8  $\frac{1}{n} \sum_{t=1}^n \mathbf{H}(\mathbf{w}_t; \bar{\theta})$  converge em probabilidade para alguma matriz  $p \times p$  simétrica  $\Psi$  dada por

$$\Psi = E[\mathbf{H}(\mathbf{w}_t; \theta_0)] \quad (97)$$

$p$  número de parâmetros, então  $\mathbf{g}$  é escolhido otimamente como em (95) podendo ser escrito como  $\frac{1}{n} \sum s(\mathbf{w}_t; \hat{\theta}) = \mathbf{0}$ . Isto é simplesmente a equação(s) de máxima verossimilhança (FOC para ML) (e levando em conta se a solução é única ou não).



Então a equação para o erro amostral pode ser escrita como:

$$\begin{aligned} \sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) &= -[\boldsymbol{\Psi}]^{-1} \sqrt{n} \frac{\partial Q_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} \\ &- \left\{ \left[ \frac{1}{n} \sum_{t=1}^n \mathbf{H}(\mathbf{w}_t; \bar{\boldsymbol{\theta}}) \right]^{-1} - [\boldsymbol{\Psi}]^{-1} \right\} \sqrt{n} \frac{\partial Q_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} \end{aligned} \quad (98)$$

Por construção o termo entre chaves converge em probabilidade para zero. Pela condição 3 da Proposição 7.8  $\sqrt{n} \frac{\partial Q_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} (= 1/n \sum s(\mathbf{w}_t; \hat{\boldsymbol{\theta}}))$  converge para uma variável aleatória. Então o último termo converge para zero (pelo Lema 2.4(b)). Este fato pode ser escrito como uma expansão de Taylor:

$$\sqrt{n} \underset{(p \times 1)}{(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)} = - \underset{(p \times p)}{[\boldsymbol{\Psi}]^{-1}} \underset{(p \times 1)}{\sqrt{n} \frac{\partial Q_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}}} + \underset{(p \times 1)}{o_p} \quad (99)$$

tal que o termo  $o_p$  significa “alguma variável aleatória que converge para zero em probabilidade.” A exata expressão para  $o_p$  depende

do contexto. Aqui é igual ao último termo de (99). O que importa aqui é que o termo  $o_p$  desaparece. Uma vez que a diferença  $\sqrt{n}(\hat{\theta} - \theta_0)$  e  $-[\Psi]^{-1} \sqrt{n} \frac{\partial Q_n(\theta_0)}{\partial \theta}$  desaparecem, a distribuição assintótica de  $\sqrt{n}(\hat{\theta} - \theta_0)$  é a mesma que  $-\Psi^{-1} \sqrt{n} \frac{\partial Q_n(\theta_0)}{\partial \theta}$  pelo Lema 2.4(a). Então  $\sqrt{n}(\hat{\theta} - \theta_0)$  converge para uma distribuição normal com média zero e variância assintótica dado por:

$$\text{Avar}(\hat{\theta}) = [\Psi]^{-1} \Sigma [\Psi]^{-1} \quad (100)$$

tal que

$$\Sigma_{(p \times p)} \equiv \text{Avar} \left( \frac{\partial Q_n(\theta_0)}{\partial \theta} \right)$$

Para estimadores-M,  $\sqrt{n} \frac{\partial Q_n(\theta_0)}{\partial \theta} = \frac{1}{n} \sum s(\mathbf{w}_t; \hat{\theta})$  e  $\Sigma$  é a variância de longo prazo de  $s(\mathbf{w}_t; \theta_0)$ . Fazendo  $\Psi = E[\mathbf{H}(\mathbf{w}_t; \theta_0)]$  fornece uma proposição para a variância na Proposição 7.8.

Agora calculamos para a variância para o GMM.

No caso GMM:

$$\sqrt{n} \frac{\partial Q_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} = -[\mathbf{G}_n(\boldsymbol{\theta}_0)]' \widehat{\mathbf{W}} \frac{1}{\sqrt{n}} \sum_{t=1}^n \mathbf{g}(w_t; \hat{\boldsymbol{\theta}}_0) \quad (101)$$

Uma vez que  $\mathbf{G}_n(\boldsymbol{\theta}_0) = \frac{\partial \mathbf{g}(w_t; \hat{\boldsymbol{\theta}}_0)}{\partial \boldsymbol{\theta}'} \rightarrow_p \mathbf{G}$ ,  $\widehat{\mathbf{W}} \rightarrow_p \mathbf{W}$  e  $\frac{1}{n} \sum_{t=1}^n \mathbf{g}(w_t; \hat{\boldsymbol{\theta}}_0) \rightarrow_p N(0, \mathbf{S})$  sob as condições da Proposição 7.10, temos  $\sqrt{n} \frac{\partial Q_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}}$  convergindo para uma distribuição normal com média zero e

$$\begin{aligned} \Sigma_{(p \times p)} &\equiv \text{Avar} \left( \frac{\partial Q_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} \right) = \\ &= \begin{matrix} \mathbf{G}' & \mathbf{W} & \mathbf{S} & \mathbf{W} & \mathbf{G} \\ (p \times K) & (K \times K) & (K \times K) & (K \times K) & (K \times p) \end{matrix} \quad (102) \end{aligned}$$

Agora reescreva a expansão do valor médio para o erro amostral para GMM como:

$$\sqrt{n} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = -\mathbf{B}^{-1} \mathbf{c} \quad (103)$$

$$\mathbf{B} \equiv -\mathbf{G}_n(\hat{\boldsymbol{\theta}})' \widehat{\mathbf{W}} \mathbf{G}_n(\bar{\boldsymbol{\theta}})$$

$$\mathbf{c} \equiv -\mathbf{G}_n(\hat{\boldsymbol{\theta}})' \widehat{\mathbf{W}} \frac{1}{\sqrt{n}} \sum_{t=1}^n \mathbf{g}(w_t; \hat{\boldsymbol{\theta}}_0)$$

Esta equação pode ser escrita na forma de expansão de Taylor (como no caso ML), com a matriz  $\Psi$  dada por

$$\underset{(p \times p)}{\Psi} = - \underset{(p \times K)}{\mathbf{G}'} \underset{(K \times K)}{\mathbf{W}} \underset{(K \times K)}{\mathbf{S}} \quad (104)$$

Substituindo as equações (104) e (102) em (100) se obtém a variância assintótica GMM da Proposição 7.10.

Observação: na próxima seção  $\Sigma = -\Psi$  para ML (com obs iid) e GMM eficiente (com  $\mathbf{W} = \mathbf{S}^{-1}$ ). Para GMM eficiente fazendo  $\mathbf{W} = \mathbf{S}^{-1}$  temos  $\Sigma = \mathbf{G}'\mathbf{S}^{-1}\mathbf{G}$  (que é a negativa de  $\Psi$ ).

## Teste de hipóteses

Trio de estatísticas para ML: Wald, LM e LR.

Como considerado anteriormente desejamos testar um conjunto de  $r$  restrições possivelmente não-lineares. Faça  $\theta_0$  ser o parâmetro do modelo com dimensão  $p$ . A hipótese nula pode ser expressa como

$$H_0 : \underset{(r \times 1)}{a(\theta_0)} = \mathbf{0} \quad (105)$$

Assumimos que  $a(\cdot)$  é continuamente diferenciável. Faça

$$\underset{(r \times p)}{A(\theta)} = \frac{\partial a(\theta)}{\partial \theta'} \quad (106)$$

ser o Jacobiano de  $a(\theta)$ . Assuma que

$$A_0 = A(\theta_0) \text{ é posto completo.} \quad (107)$$

Isto garante que as  $r$  restrições não são redundantes. Esta condição de posto implica em  $r \leq p$ .

Faça  $\hat{\theta}$  ser o estimador extremo em questão: ML ou GMM. Ele é solução para o problema de maximização (irrestrito) em (1). A estatística de Wald utiliza o estimador não-restrito. Por outro lado a estatística LM utiliza o estimador restrito ( $\tilde{\theta}$ ) que soluciona

$$\max_{\theta \in \Theta} Q_n(\theta) \text{ s.a. } a(\theta) = 0 \quad (108)$$

O parâmetro verdadeiro  $\theta_0$  soluciona o “problema limite irrestrito” onde  $Q_n$  é o problema de otimização irrestrito (1) que é substituído pela função limite  $Q_0$ . Isto também soluciona o “problema limite restrito” tal que  $Q_n$  no problema acima é substituído por  $Q_0$ , porque  $\theta_0$  satisfaz a restrição. A convergência uniforme em probabilidade de  $Q_n(\cdot)$  em  $Q_0(\cdot)$  garante que o limite do estimador restrito  $\tilde{\theta}$  é a solução limite do problema restrito, que é  $\theta_0$ .

## A trindade

**Proposição 7.11 (A trindade):** Faça  $\hat{\theta}$  o estimador extremo definido em (1). Considere a hipótese nula em (105) tal que  $a(\cdot)$  é continuamente diferenciável (então o Jacobiano é contínuo) e a condição de posto (107) é satisfeita. Faça  $\tilde{\theta}$  ser o estimador extremo restrito definido em (108). Assuma:

- as hipóteses da Proposição 7.9, se o estimador extremo é ML condicional,
- as hipóteses apropriadamente modificadas como indicadas abaixo da Proposição 7.9, no caso do ML condicional,
- as hipóteses da Proposição 7.10 com  $W = S^{-1}$  e dado que é disponível estimador consistente  $\hat{S}$  de  $S$  construído

de  $\hat{\theta}$  e  $\hat{S}$  construído de  $\tilde{\theta}$ , no caso do estimador GMM eficiente.

Defina as estatísticas Wald, LM e LR como:

$$\mathbf{Wald:} \quad n \mathbf{a}(\hat{\theta})' \begin{bmatrix} \mathbf{A}(\hat{\theta}) \hat{\Sigma}^{-1} \mathbf{A}(\hat{\theta})' \\ (1 \times r) \quad (r \times p) \quad (p \times p) \quad (p \times r) \end{bmatrix}^{-1} \mathbf{a}(\hat{\theta}) \quad (r \times 1)$$

$$\mathbf{LM:} \quad n \begin{pmatrix} \frac{\partial Q_n(\tilde{\theta})}{\partial \theta} \\ (1 \times p) \end{pmatrix}' \hat{\Sigma}^{-1} \begin{pmatrix} \frac{\partial Q_n(\tilde{\theta})}{\partial \theta} \\ (p \times 1) \end{pmatrix}$$

$$\mathbf{LR:} \quad 2n[Q_n(\hat{\theta}) - Q_n(\tilde{\theta})]$$

Os termos para substituição são:

$$Q_n(\theta): \bullet \text{ ML: } \frac{1}{n} \sum_{t=1}^n \log f(y_t | \mathbf{x}_t; \theta);$$

$$\bullet \text{ GMM eficiente: } -\frac{1}{2} \mathbf{g}_n(\theta)' \hat{S}^{-1} \mathbf{g}_n(\theta)$$

$$\hat{\Sigma}: \bullet \text{ ML: } -\frac{\partial^2 Q_n(\hat{\theta})}{\partial \theta \theta'} \text{ ou } \frac{1}{n} \sum \mathbf{s}(w_t; \hat{\theta}) \mathbf{s}(w_t; \hat{\theta})'$$

$$\bullet \text{ GMM: } \hat{G}' \hat{S}^{-1} \hat{G}, \quad \hat{G} \underset{(K \times p)}{=} \mathbf{G}_n(\hat{\theta})$$



$\tilde{\Sigma}$ : • ML: substituir  $\hat{\theta}$  por  $\tilde{\theta}$  em  $\hat{\Sigma}$

• GMM:  $\tilde{G}'\tilde{S}^{-1}\tilde{G}$ ,  $\underset{(K \times p)}{\tilde{G}} = G_n(\tilde{\theta})$

Portanto:

1. O estimador extremo restrito  $\tilde{\theta}$  é consistente e assintoticamente normal,
2. Todas as três estatísticas convergem em distribuição para  $\chi^2(r)$  sob a hipótese nula tal que  $r$  é o número de restrições na hipótese nula,
3. além disso, a diferença numérica entre as três estatísticas convergem em probabilidade para zero sob a hipótese nula.

## Instrumentos ótimos: Hansen (2019, cap. 13)

A versão de GMM não-linear de Hansen é mais simples do que estudada até aqui. A Proposição 7.10 fornece um resultado completo do estimador.

Aqui impomos a restrição de momentos condicional do estimador GMM:

$$E(e_i(\beta) | z_i) = 0 \quad (109)$$

aqui  $e$  é uma possível função não-linear ( $s \times 1$ ) dos parâmetros  $\beta$  (em Hayashi era  $\theta$ ). A variável  $z_i$  são os instrumentos.

A restrição de momentos condicional é muito mais poderosa e restritiva do que a condição de momentos incondicional.

Exemplos. O modelo linear com  $y_i = x_i' \beta + e_i$  com instrumentos  $z_i$  se aplica a esta classe

de modelos com  $E(e_i | z_i) = 0$ . Nesse caso  $e_i(\beta) = y_i - x_i' \beta$ . No modelo não-linear  $e_i(\beta) = y_i - g(x_i'; \beta)$ . Em um modelo conjunto de média condicional  $E(y_i | x_i) = x_i \beta$  e  $\text{Var}(y_i | x_i) = f(x_i)' \gamma$ , então ( $s = 2$ )

$$e_i(\beta, \gamma) = \begin{cases} y_i - x_i' \beta \\ (y_i - x_i' \beta)^2 - f(x_i)' \gamma \end{cases}$$

Dada uma restrição de momento condicional, uma restrição de momento incondicional pode ser sempre construída. Isto é, para qualquer função  $\phi(z_i'; \beta)$  ( $l \times 1$ ) podemos determinar  $g_i(\beta) = \phi(z_i'; \beta) e_i(\beta)$  que satisfaz  $E(g_i(\beta)) = 0$  e um modelo de equação com momento incondicional.

O problema (óbvio?) é que essa classe de funções  $\phi(\cdot)$  é infinita. Qual função deve ser selecionada?

Uma solução é construir uma lista infinita de instrumentos potenciais e então usar os primeiros  $k$  instrumentos. Como  $k$  deveria ser determinado? Esta área de pesquisa ainda está em desenvolvimento e não pode não haver solução pois o vetor de instrumentos pode ser intratável. Sobre esta estratégia veja por exemplo Newey (1990) e Donald e Newey (2001).

Outra estratégia de solução é denominado de *instrumentos ótimos* e foi proposto por Chamberlain (1987). Assuma o caso com  $s = 1$ . Faça

$$\mathbf{R}_i = \mathbb{E} \left( \frac{\partial e_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \mid \mathbf{z}_i \right) \quad (110)$$

e

$$\sigma_i^2 = \mathbb{E}(e_i(\boldsymbol{\beta})^2 \mid \mathbf{z}_i) \quad (111)$$

Então o instrumento ótimo é:

$$\mathbf{A}_i = \left[ \frac{\partial e_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \mid \mathbf{z}_i \right] T(\mathbf{z}_i) = \mathbf{R}_i T(\mathbf{z}_i) \quad (112)$$

Na solução geral de Chamberlein (1987)  $T(z_i)T(z_i)' = E[g_i(\beta)g_i(\beta)']^{-1}$  com esta matriz normalizada  $A_i$  é o conjunto de instrumentos ótimos. Hansen (2019) assume  $T(z_i) = -\sigma_i^{-2}$ , implicando que

$$A_i = -\sigma_i^{-2}R_i \quad (113)$$

Nesse caso o momento ótimo é:

$$g_i(\beta) = A_i e_i(\beta) \quad (114)$$

fazendo  $g_i(\beta)$  ser esta escolha isto resulta no melhor estimador GMM possível. Na prática  $A_i$  é desconhecido, mas a sua forma pode nos ajudar sobre a construção de instrumentos ótimos. Por exemplo, no modelo linear  $e_i(\beta) = y_i - x_i'\beta$ , observe que

$$R_i = -E(x_i | z_i)$$

$$\sigma_i^2 = E(e_i^2 | z_i)$$

Portanto,

$$A_i = \sigma_i^{-2}E(x_i | z_i)$$

No caso da regressão linear,  $\mathbf{x}_i = \mathbf{z}_i$ , então  $\mathbf{A}_i = \sigma_i^{-2} \mathbf{z}_i$ . Portanto, o estimador GMM eficiente é equivalente ao GLS ótimo.

No caso das variáveis endógenas, observe que o instrumento eficiente  $\mathbf{A}_i$  contém a estimação da média condicional de  $\mathbf{x}_i$  dado  $\mathbf{z}_i$ . Em outras palavras, para conseguir os melhores instrumentos para  $\mathbf{x}_i$ , é necessário o melhor modelo da média condicional de  $\mathbf{x}_i$  dado  $\mathbf{z}_i$ , não apenas uma projeção linear arbitrária. O instrumento eficiente é também inversamente proporcional a variância condicional  $e_i$ . Isto funciona como o estimador GLS – o aumento de eficiência pode ser obtido se as observações são ponderadas inversamente a variância condicional dos erros.

## **Bootstrapping SE**

(Hansen, 2019, sec. 10.8) A distribuição bootstrap é obtida pela estimação sobre amostras independentes criadas por amostragem (com reposição) i.i.d. a partir do conjunto de dados original.

## Otimização numérica

Em muitas aplicações não existe forma fechada para a solução numérica (e.g. forma quadrática do GMM linear), portanto algoritmos numéricos precisam ser aplicados para localizar o máximo. Hayashi cobre dois algoritmos importantes.

### Newton-Raphson

Considere estimadores-M, cuja função objetivo  $Q_n(\boldsymbol{\theta})$  é (2). Uma vez que  $Q_n(\boldsymbol{\theta})$  é duas vezes continuamente diferenciável para os estimadores-M, existe uma expansão de Taylor de segunda ordem

$$Q_n(\boldsymbol{\theta}) \cong Q_n(\hat{\boldsymbol{\theta}}_j) + s_n(\hat{\boldsymbol{\theta}}_j)'(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_j) + \frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_j)' \mathbf{H}_n(\hat{\boldsymbol{\theta}}_j)(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_j) \quad (115)$$



tal que  $\hat{\theta}_j$  é a estimativa na iteração  $j$  do processo iterativo a ser descrito e  $s_n$  e  $H_n$  são o gradiente e o Hessiano da função objetivo:

$$s_n = \frac{\partial Q_n(\theta)}{\partial \theta'} \quad (116)$$

$$H_n = \frac{\partial^2 Q_n(\theta)}{\partial \theta \partial \theta'} \quad (117)$$

O estimador  $\hat{\theta}_{j+1}$  da iteração  $j + 1$  é o maximizador da função quadrática sobre o lado direito de (115). Ele é dado por:

$$\hat{\theta}_{j+1} = \hat{\theta}_j - [H_n(\hat{\theta}_j)]^{-1} s_n(\hat{\theta}_j) \quad (118)$$

Este processo iterativo é chamado de *algoritmo Newton-Raphson*. Se a função objetivo é côncava, o algoritmo geralmente converge rapidamente para o máximo global.

Para o ML, quando o máximo global é obtido, a estimativa da  $Avar(\hat{\theta})$  é obtida como  $-[H_n(\hat{\theta})]^{-1}$ .\*

$$*H_n(\hat{\theta})^{-1} = \frac{1}{n} \sum H(w_t; \theta).$$

## Gauss-Newton

No GMM a função objetivo é  $Q_n(\boldsymbol{\theta}) = -\frac{1}{2}\mathbf{g}_n(\boldsymbol{\theta})'\widehat{\mathbf{W}}\mathbf{g}_n(\boldsymbol{\theta})$ . Como na derivação da distribuição assintótica, é utilizada uma linearização da função  $\mathbf{g}_n(\boldsymbol{\theta})$ . A expansão de Taylor de primeira ordem de  $\mathbf{g}_n(\boldsymbol{\theta})$  em torno de  $\widehat{\boldsymbol{\theta}}_j$  é

$$\begin{aligned} \underset{(K \times 1)}{\mathbf{g}_n(\boldsymbol{\theta})} &\cong \underset{(K \times 1)}{\mathbf{g}_n(\widehat{\boldsymbol{\theta}}_j)} + \underset{(K \times p)}{\mathbf{G}_n(\widehat{\boldsymbol{\theta}}_j)} \underset{(p \times 1)}{(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_j)} \\ &= [\mathbf{g}_n(\widehat{\boldsymbol{\theta}}_j) - \mathbf{G}_n(\widehat{\boldsymbol{\theta}}_j)\widehat{\boldsymbol{\theta}}_j] - [-\mathbf{G}_n(\widehat{\boldsymbol{\theta}}_j)]\boldsymbol{\theta} \\ &= \mathbf{v}_j - \mathbf{G}_j\boldsymbol{\theta} \end{aligned} \quad (119)$$

tal que

$$\mathbf{v}_j \equiv \mathbf{g}_n(\widehat{\boldsymbol{\theta}}_j) - \mathbf{G}_n(\widehat{\boldsymbol{\theta}}_j)\widehat{\boldsymbol{\theta}}_j, \quad \mathbf{G}_j \equiv -\mathbf{G}_n(\widehat{\boldsymbol{\theta}}_j) \quad (120)$$

$$\mathbf{G}_n(\boldsymbol{\theta}) \equiv \frac{\partial \mathbf{g}_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}$$

Se a função  $\mathbf{g}_n(\boldsymbol{\theta})$  (produto erro-instrumento) na expressão para a função objetivo do GMM dada por  $\mathbf{v}_j - \mathbf{G}_j\boldsymbol{\theta}$ , então a função objetivo

deve ser quadrática em  $\theta$  e o maximizador (ou o minimizador da distância GMM) deve ser o estimador GMM linear:

$$\hat{\theta}_{j+1} = (G_j' \widehat{W} G_j)^{-1} G_j' \widehat{W} v_j \quad (121)$$

Esta é a estimação da rodada  $j + 1$  no *algoritmo Gauss-Newton*. De forma distinta do Newton-Raphson, não há necessidade de calcular as segunda derivadas.

## Newton-Raphson e Gauss-Newton em formato comum

Duas similaridades entre os algoritmos. Substituindo (120) em (121) e reorganizando se obtém:

$$\hat{\theta}_{j+1} = -[-G_n(\hat{\theta}_j)' \widehat{W} G_n(\hat{\theta}_j)]^{-1} [-G_n(\hat{\theta}_j)' \widehat{W} g_n(\hat{\theta}_j)] \quad (122)$$

O segundo termo entre colchetes não é o gradiente solucionado em  $\hat{\theta}_j$  para a função objetivo GMM  $Q_n(\theta) = -\frac{1}{2} g_n(\theta)' \widehat{W} g_n(\theta)$ . O papel

do Hessiano na função objetivo em (??) é realizado aqui pelo termo entre colchetes, que é a estimativa baseada em  $\hat{\theta}_j$  de  $G'WG$ . Se  $g_n$  é linear, então o equivalente ao Hessiano é o Hessiano da função objetivo GMM. A analogia aqui é exata: o algoritmo Gauss-Newton coincide com o algoritmo Newton-Raphson.

## Equação apenas não-linear nos parâmetros

No algoritmo Gauss-Newton (122), quando  $g_n(\theta) = 1/n \sum g(w_t; \theta)$  é não linear em  $\theta$  é necessário usar médias sobre as observações para resolver o gradiente e o Hessiano equivalente em cada iteração (isto também é verdade se a função em Newton-Raphson se a função objetivo não é quadrática em  $\theta$ ).

Existe um caso onde a média sobre observações em cada iteração não é necessária. Isto ocorre em uma classe de estimação por variáveis instrumentais não-linear generalizado. Este é um caso especial do GMM não-linear tal que  $g(y_t, z_t, \theta)$  pode ser escrito como  $a(y_t, z_t; \theta)x_t$ . Suponha que a equação  $a(y_t, z_t; \theta)$  assuma a forma:

$$a(y_t, z_t; \theta) = a_0(y_t, z_t) + \underset{(1 \times q)}{a_1(y_t, z_t)'} \underset{(q \times 1)}{\alpha(\theta)} \quad (123)$$

tal que  $a_0(\cdot)$  e  $a_1(\cdot)$  são funções conhecidas de  $(y_t, z_t)$  (mas não são funções de  $(\theta)$ ). Uma

função dessa forma é dita ser *apenas não linear nos parâmetros* ou *linear nas variáveis*. Isto pode ser visto:

$$\begin{aligned}
 \mathbf{g}_n(\boldsymbol{\theta}) &\equiv \frac{1}{n} \sum_{t=1}^n \mathbf{g}(y_t, \mathbf{z}_t; \boldsymbol{\theta}) = \frac{1}{n} \sum_{t=1}^n \mathbf{a}(y_t, \mathbf{z}_t; \boldsymbol{\theta}) \mathbf{x}_t = \\
 &= \left( \frac{1}{n} \sum_{t=1}^n a_0(y_t, \mathbf{z}_t) \mathbf{x}_t \right)_{(K \times 1)} + \left( \frac{1}{n} \sum_{t=1}^n \mathbf{x}_t \mathbf{a}_1(y_t, \mathbf{z}_t)' \right)_{(K \times q)} \boldsymbol{\alpha}(\boldsymbol{\theta})_{(q \times 1)} \\
 & \hspace{20em} (124)
 \end{aligned}$$

$$\begin{aligned}
 \mathbf{G}_n(\boldsymbol{\theta}) &\equiv \frac{\partial \mathbf{g}_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'} = \left( \frac{1}{n} \sum_{t=1}^n \mathbf{x}_t \mathbf{a}_1(y_t, \mathbf{z}_t)' \right)_{(K \times q)} \frac{\partial \boldsymbol{\alpha}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}_{(q \times p)} \\
 & \hspace{20em} (125)
 \end{aligned}$$

Nestas equações está claro que é necessário calcular as médias uma vez antes de iniciar as iterações.

## Referências

**Chamberlein, Gary** “Asymptotic efficiency in estimation with conditional moment restrictions,” *Journal of Econometrics*, 34 (3), 1987.

**Donald e Newey** 2001.

**Hansen, Bruce** *Econometrics*. 2019.

**Hansen, Lars Peter** “Large Sample Properties of Generalized Method of Moments Estimators,” *Econometrica*, 50 (4), 1982.

**Hayashi, Fumio** *Econometrics*. Princeton University Press, 2000.

**Newey** 1990.

**Stock, James H. e Mark W. Watson** *Introduction to Econometrics*. 3a ed. Pearson, 2015.