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

$$\max_{\widehat{\boldsymbol{\theta}}} \{Q_n(\boldsymbol{\theta})\} \text{ s.a. } \boldsymbol{\theta} \in \Theta \subset \Re^p \tag{1}$$

tal que Θ (o espaço dos parâmetros) é o conjunto do valores possíveis para os parâmetros. Restrição no conjunto de parâmetros: subconjunto do espaço Euclidiano com dimensão finita $\Re^p$. $Q_n(\boldsymbol{\theta})$ não depende apenas de $\boldsymbol{\theta}$ mas também depende da amostra $(\boldsymbol{w}_1, \boldsymbol{w}_2, ..., \boldsymbol{w}_n)$, tal que $\boldsymbol{w}_t$ é a observação t e n é o tamanho da amostra. O subescrito n sinaliza a dependência do tamanho da amostra.

Estimadores Extremos: Medindo $\widehat{\boldsymbol{\theta}}$

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

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

Portanto, se $Q_n(\boldsymbol{\theta})$ é contínuo em $\boldsymbol{\theta}$ para qualquer amostra $(\boldsymbol{w}_1, \boldsymbol{w}_2, ..., \boldsymbol{w}_n)$ e $\boldsymbol{\Theta}$ é compacto, então existe um $\boldsymbol{\theta}$ que soluciona o problema de maximização em (1) para qualquer amostra.

Quando se tem soluções múltiplas, poderíamos escolher uma delas. Então $\widehat{\boldsymbol{\theta}}$ é uma função dos dados. Ser uma função do vetor dos dados $(\boldsymbol{w}_1, \boldsymbol{w}_2, ..., \boldsymbol{w}_n)$ não é suficiente para fazer $\widehat{\boldsymbol{\theta}}$ uma variável aleatória bem definida. $\widehat{\boldsymbol{\theta}}$ precisa ser uma função mensurável de $(\boldsymbol{w}_1, \boldsymbol{w}_2, ..., \boldsymbol{w}_n)$.

Lema 7.1 (existência de estimadores extremos): Suponha que (i) o espaço dos parâmetros Θ é um subconjunto compacto de $\Re^p$, (ii) $Q_n(\boldsymbol{\theta})$ é contínua em $\boldsymbol{\theta}$ para qualquer $(\boldsymbol{w}_1, \boldsymbol{w}_2, ..., \boldsymbol{w}_n)$ e (iii) $Q_n(\boldsymbol{\theta})$ é uma função mensurável dos dados para todo $\boldsymbol{\theta}$ em Θ. Então existe uma função mensurável $\widehat{\boldsymbol{\theta}}$ 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 incluidos no espaço dos parâmetros, isso significa que o espaço dos parâmetros não é fechado. Então a hipótese de conjunto compacto para Θ 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:

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

tal que m é uma função de valor real de $(\boldsymbol{w}_t; \boldsymbol{\theta})$. 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:

$$Q_n(\boldsymbol{\theta}) = -\frac{1}{2} g_n(\boldsymbol{\theta})' \widehat{\boldsymbol{W}} g_n(\boldsymbol{\theta}) \tag{3}$$

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

tal que $\hat{\boldsymbol{W}}$ 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 estimador (CMD) não cabe nestas classes. A função objetivo da CMD por ser escrita como $-\frac{1}{2}g_n(\boldsymbol{\theta})'\hat{\boldsymbol{W}}g_n(\boldsymbol{\theta})$, 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 $\{\boldsymbol{w}_t\}$ iid. Neste caso a densidade de $\{\boldsymbol{w}_t\}$ é um membro de um família de densidades indexada por um vetor de dimensão finita $\boldsymbol{\theta}$: $f(\boldsymbol{w}_t;\boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta$.* A fórmula funcional $f(\cdot)$ é conhecida. O modelo é paramétrico porque o vetor de parâmetros $\boldsymbol{\theta}$ é de dimensão finita. No vetor verdadeiro de parâmetros, $\boldsymbol{\theta}_0$, a densidade do PGD verdadeiro é $f(\boldsymbol{w}_t;\boldsymbol{\theta}_0)$. Se diz que o modelo é *corretamente especificado* se $\boldsymbol{\theta}_0 \in \Theta$.

Dado que $\{\boldsymbol{w}_t\}$ é iid, a densidade conjunta dos dados $(\boldsymbol{w}_1, \boldsymbol{w}_2, ..., \boldsymbol{w}_n)$ é

$$f(\boldsymbol{w}_1, \boldsymbol{w}_2, ..., \boldsymbol{w}_n; \boldsymbol{\theta}_0) = \prod_{t=1}^{n} f(\boldsymbol{w}_t; \theta_0) \tag{4}$$

*Como $\{\boldsymbol{w}_t\}$ 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 $\boldsymbol{\theta}_0$ pelo vetor hipotético $\boldsymbol{\theta}$, a densidade é denominada *função de verossimilhança*. O estimador ML de $\boldsymbol{\theta}_0$ é o $\boldsymbol{\theta}$ 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*:

$$\log f(\boldsymbol{w}_1, \boldsymbol{w}_2, ..., \boldsymbol{w}_n; \boldsymbol{\theta}) = \log \left[\prod_{t=1}^{n} f(\boldsymbol{w}_t; \boldsymbol{\theta}) \right] = \\ = \sum_{t=1}^{n} \log f(\boldsymbol{w}_t; \boldsymbol{\theta}) \quad (5)$$

Portanto, o estimador de $\boldsymbol{\theta}_0$ é um estimador-M com

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

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

- A log verossimilhança média não deve assumir uma forma simples como esta se $\{\boldsymbol{w}_t\}$ possui correlação seriada.

- ML não é a única forma de se estimar o vetor de parâmetros. Por exemplo, faça $\boldsymbol{\mu}(\boldsymbol{\theta})$ ser a expecativa de $\boldsymbol{w}_t$ implicada pela função de densidade $f(\boldsymbol{w}_t; \boldsymbol{\theta})$. Esta é uma função conhecida de $\boldsymbol{\theta}$. Por construção vale a seguinte condição de média-zero:

$$\mathrm{E}[\boldsymbol{w}_t - \boldsymbol{\mu}(\boldsymbol{\theta}_0)] = \mathbf{0} \tag{7}$$

O vetor de parâmetros $\boldsymbol{\theta}_0$) pode ser estimado por GMM com $\boldsymbol{g} = (\boldsymbol{w}_t; \boldsymbol{\theta})\boldsymbol{w}_t - \boldsymbol{\mu}(\boldsymbol{\theta}_0)$ 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 $(w_1, ..., w_n)$ ser uma sequência de escalar iid com distribuição de w_t dado por $N(\mu, \sigma^2)$. Então $\boldsymbol{\theta} = (\mu, \sigma^2)'$ 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 $(w_1, ..., w_n)$ é

$$\frac{1}{n}\sum_{t=1}^{n} log f(w_t; \mu, \sigma^2) =$$

$$= -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log(\sigma^2) - \frac{1}{n}\sum_{t=1}^{n}\left[\frac{(w_t - \mu)^2}{2\sigma^2}\right] \quad (8)$$

O estimador ML de (μ_0, σ_0^2) é um estimador extremo com $Q_n(\boldsymbol{\theta})$ com a log verossimilhança acima. Por hipótese, assuma a variância como positiva, mas que não existe nenhuma restrição a priori sobre o valor de μ_0, então o espaço de parâmetro Θ é $\Re \times \Re_{++}$, tal que $\Re_{++}$ é o conjunto positivo de números reais. O estimador ML de μ_0 é a média amostral de w_t. O estimador GMM de μ_0 baseado na condição de média zero $\mathrm{E}(w_t - \mu_0) = 0$, também é a média amostral de w_t.

ML Condicional

Na maioria das aplicações o vetor $\boldsymbol{w}_t$ é: y_t e $\boldsymbol{x}_t$. O interesse do econommetrista é saber como $\boldsymbol{x}_t$ influencia a distribuição condicional de y_t dado $\boldsymbol{x}_t$. y_t é a variável dependente e $\boldsymbol{x}_t$ é o regressor.*

Faça $f(y_t \mid \boldsymbol{x}_t; \boldsymbol{\theta}_0)$ ser a densidade condicional de y_t dado $\boldsymbol{x}_t$, e faça $f(\boldsymbol{x}_t; \boldsymbol{\psi}_0)$ ser a densidade marginal de $\boldsymbol{x}_t$. Então, a densidade conjunta de $\boldsymbol{w}_t = (y_t, bmx_t')'$ é:

$$f(y_t, \boldsymbol{x}_t; \boldsymbol{\theta}_0, \boldsymbol{\psi}_0) = f(y_t \mid \boldsymbol{x}_t; \boldsymbol{\theta}_0) f(\boldsymbol{x}_t; \boldsymbol{\psi}_0) \tag{9}$$

Suponha por enquanto que $\boldsymbol{\theta}_0$ e $\boldsymbol{\psi}_0$ não são funcionalmente relacionados. A log verossimilhança média dos dados $(w_1, ..., w_n)$

*Os resultados valem para y_t sendo escalar ou vetor.

é

$$\frac{1}{n}\sum_{t=1}^{n}\log f(\boldsymbol{w}_t;\boldsymbol{\theta},\psi) = \underbrace{\frac{1}{n}\sum_{t=1}^{n}\log f(y_t \mid \boldsymbol{x}_t;\boldsymbol{\theta})}_{\text{log verossimilhança condicional}} + \frac{1}{n}\sum_{t=1}^{n}\log f(\boldsymbol{x}_t;\boldsymbol{\psi}) \quad (10)$$

O estimador ML condicional de $\boldsymbol{\theta}_0$ maximiza o primeiro termo de (10), ignorando o segundo termo. Este é um estimador-M com:

$$m(\boldsymbol{w}_t;\boldsymbol{\theta}) = \log f(y_t \mid \boldsymbol{x}_t;\boldsymbol{\theta}_0), \text{isto é} \quad (11)$$

$$Q_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{t=1}^{n}\log f(y_t \mid \boldsymbol{x}_t)$$

O segundo termo de (11) é a log verossimilhança marginal média.

Este termo não depende de $\boldsymbol{\theta}$, então a estimativa ML condicional de $\boldsymbol{\theta}_0$ é numericamente a mesma se o segundo termo fosse incluido.

Relação funcional entre $\boldsymbol{\theta}_0$ e $\boldsymbol{\psi}_0$: por exemplo, $\boldsymbol{\theta}_0$ e $\boldsymbol{\psi}_0$ podem ser dividos como:

- $\boldsymbol{\theta}_0 = (\boldsymbol{\alpha}_0', \boldsymbol{\beta}_0')'$

- $\boldsymbol{\psi}_0 = (\boldsymbol{\beta}_0', \boldsymbol{\gamma}_0')'$

Neste caso o estimador completo e o condicional não são mais numericamente iguais. Neste caso, o estimador ML condicional de $\boldsymbol{\theta}_0$ 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 parâmetrica para $f(\boldsymbol{x}_t; \boldsymbol{\psi})$.

Exemplo 7.2

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

$$\{y_t, \boldsymbol{x}_t\} \text{ é iid.}$$

$$y_t = \boldsymbol{x}_t' \boldsymbol{\beta}_0 + \varepsilon_t \tag{12}$$

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

A versossimilhança de $y_t \mid \boldsymbol{x}_t$, $f(y_t \mid \boldsymbol{x}_t; \boldsymbol{\theta})$, é a função de densidade de $N(\boldsymbol{x}_t' \boldsymbol{\beta}, \sigma_0^2)$. Dado $\boldsymbol{\theta} = (\boldsymbol{\beta}', \sigma^2)'$ e $\boldsymbol{w}_t = (y_t, \boldsymbol{x}_t')'$, então a função m é

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

$$= \log\left\{\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{(y_t - \boldsymbol{x}_t'\boldsymbol{\beta})^2}{2\sigma^2}\right]\right\}$$

$$= -\frac{1}{2}\log(2\pi) - \log(\sigma) - \frac{1}{2}\left(\frac{y_t - \boldsymbol{x}_t'\boldsymbol{\beta}}{\sigma}\right)^2 \quad (14)$$

O espaço dos parâmetros Θ é $\Re^K \times \Re_{++}$, tal que K é a dimensão de $\boldsymbol{\beta}$ e $\Re_{++}$ 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. $\boldsymbol{x}_t$ pode incluir a renda do marido. A probabilidade condicional de y_t dado um vetor de regressores $\boldsymbol{x}_t$ é dado por

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

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

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

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

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

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

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 $\boldsymbol{\lambda} = \boldsymbol{\tau}(\boldsymbol{\theta})$ definido em Θ. Faça Λ ser o intervalo do mapeamento:

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

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

$$\boldsymbol{\tau} : \Theta \to \Lambda$$

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

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

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

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

Para ver isto faça $\hat{\boldsymbol{\lambda}} = \boldsymbol{\tau}(\widehat{\boldsymbol{\theta}})$. Para qualquer $\boldsymbol{\lambda} \in \Lambda$, temos $\tilde{Q}_n(\boldsymbol{\lambda}) = Q_n(\boldsymbol{\tau}^{-1}(\boldsymbol{\lambda})) \leq Q_n(\widehat{\boldsymbol{\theta}})$, porque $\widehat{\boldsymbol{\theta}}$ maximiza $Q_n(\widehat{\boldsymbol{\theta}})$ sobre Θ e $\boldsymbol{\tau}^{-1}$

está em Θ. Mas $Q_n(\hat{\boldsymbol{\theta}}) = Q_n(\boldsymbol{\tau}^{-1}(\boldsymbol{\lambda})) = \tilde{Q}_n(\hat{\boldsymbol{\lambda}})$. Então $\tilde{Q}_n(\boldsymbol{\lambda}) \leq \tilde{Q}_n(\hat{\boldsymbol{\lambda}})$ para todo $\boldsymbol{\lambda} \in \Lambda$.

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

NLS

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

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

tal que $\mathrm{E}(\varepsilon_t \mid \boldsymbol{x}_t) = 0$ 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:

$$m(\boldsymbol{w}_t; \boldsymbol{w}_t) = -[y_t - \mathsf{E}(y_t \mid \boldsymbol{x}_t)]^2 \tag{22}$$

i.e.

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

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 adtivos. Considere a função de produção CES:

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

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

Esse é um modelo de média condicional (21) com $y_t = Q_t$, $\boldsymbol{x}_t = (K_t, L_t)'$, $\boldsymbol{\theta}_0 = (A_0, \delta_0, \rho_0)'$ e $\varphi(\boldsymbol{x}_t; \boldsymbol{\theta}) = A_0 \left[\delta_0 K_t^{-\rho_0} + (1 - \delta_0) L_t^{-\rho_0}\right]^{-1/\rho_0}$. 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 Θ.

GMM linear e não-linear

O modelo GMM linear é:

$$y_t = \boldsymbol{z}_t'\boldsymbol{\theta}_0 + \varepsilon_t \tag{24}$$

com $\boldsymbol{x}_t$ sendo o vetor de instrumentos, as condições de ortogonalidade são:

$$\mathsf{E}[\boldsymbol{x}_t.(y_t - \boldsymbol{z}_t'\boldsymbol{\theta}_0)] = \mathbf{0} \tag{25}$$

O modelo corretamente especificado aqui é um conjunto de processos ergódicos estacionários $\boldsymbol{w}_t = (y_t, \boldsymbol{z}_t', \boldsymbol{x}_t')'$, tal que estas condições de média-zero valem para $\boldsymbol{\theta}_0 \in \Theta$. O estimador GMM linear de $\boldsymbol{\theta}_0$ é um estimador GMM com a função $\boldsymbol{g}$ da função objetivo GMM (3) dada por:

$$g(\boldsymbol{w}_t; \boldsymbol{\theta}) = \boldsymbol{x}_t.(y_t - \boldsymbol{z}_t'\boldsymbol{\theta}) = \boldsymbol{x}_t.y_t - \boldsymbol{x}_t.\boldsymbol{z}_t'\boldsymbol{\theta} \tag{26}$$

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

$$a(y_t, \boldsymbol{z}_t; \boldsymbol{\theta}_0) = \varepsilon_t \tag{27}$$

Apenas faça

$$g(\boldsymbol{w}_t; \boldsymbol{\theta}) = \boldsymbol{x}_t . a(y_t, \boldsymbol{z}_t; \boldsymbol{\theta}) \tag{28}$$

Este estimador é denominado *estimador de variáveis instrumentais generalizado*. Este ainda é um caso especial de GMM porque a função $g(\boldsymbol{w}_t; \boldsymbol{\theta})$ 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:

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

R_{t+1} é a taxa de retorno bruta (1 + taxa de retorno), c_t é o consumo de não-duráveis, β_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:

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

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

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

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

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

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

com

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

Consistência

A função objetivo $Q_n(\cdot)$ é uma função aleatória porque para cada $\boldsymbol{\theta}$ o valor $Q_n(\boldsymbol{\theta})$ é uma variável aleatória, pois $Q_n(\boldsymbol{\theta})$ depende de $(\boldsymbol{w}_1, ..., \boldsymbol{w}_n)$.

Idéia para consistência: $Q_n(\boldsymbol{\theta})$ converge em probabilidade para $Q_0(\boldsymbol{\theta})$, e o parâmetro verdadeiro $\boldsymbol{\theta}$ soluciona o "problema limite" de maximização da função limite $Q_0(\boldsymbol{\theta})$, então o limite do máximo $\widehat{\boldsymbol{\theta}}$ deve ser $\boldsymbol{\theta}_0$.

Necessário: (i) convergência em probabilidade uniforme; (ii) Θ 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 $\text{plim}_{n\to\infty} Q_n(\boldsymbol{\theta}) = Q_0(\boldsymbol{\theta})$ para todo $\boldsymbol{\theta}$ – a sequência de variáveis aleatórias $|Q_n(\boldsymbol{\theta}) - Q_0(\boldsymbol{\theta})| (n = 1, 2, ...)$ 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 Θ no seguinte sentido:

$$\sup_{\boldsymbol{\theta}\in\Theta} |Q_n(\boldsymbol{\theta}) - Q_0(\boldsymbol{\theta})| \to_p 0 \text{ enquanto } n \to \infty \tag{33}$$

- 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 $\{\boldsymbol{h}_n(\cdot)\}$ converge uniformemente em probabilidade para uma função não aleatória $\boldsymbol{h}_0(\cdot)$ se cada elemento converge uniformente. Esta convergência elemento por elemento é equivalente a convergência na norma:

$$\sup_{\boldsymbol{\theta}\in\Theta} \|\boldsymbol{h}_n(\boldsymbol{\theta}) - \boldsymbol{h}_0(\boldsymbol{\theta})\| \to_p 0 \text{ enquanto } n \to \infty \tag{34}$$

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

Consistência com espaço dos parâmetros compacto

Proposição 7.1 (Consistência): Suponha que (i) Θ é um subconjunto compato de $\Re^p$, (ii) $Q_n(\boldsymbol{\theta})$ é contínua em $\boldsymbol{\theta}$ para qualquer dado $(\boldsymbol{w}_1, ..., \boldsymbol{w}_n)$, e (iii) $Q_n(\boldsymbol{\theta})$ é função mensurável do dado para todo $\boldsymbol{\theta}$ em Θ. Se existe uma função $Q_0(\boldsymbol{\theta})$ tal que

1. (identificação) $Q_0(\boldsymbol{\theta})$ é unicamente maximizado sobre Θ em $\boldsymbol{\theta}_0 \in \Theta$.

2. (convergência uniforme) $Q_n(\cdot)$ converge uniformemente em probabilidade para $Q_0(\cdot)$, então $\widehat{\boldsymbol{\theta}} \rightarrow_p \boldsymbol{\theta}_0$.

Consistência sem espaço dos parâmetros compacto

Proposição 7.2 (Consistência): Suponha que (i) o vetor verdadeiro dos parâmetros $\boldsymbol{\theta}_0$ é um elemento do interior de espaço de parâmetros convexo $\Theta(\subset \Re^p)$, (ii) $Q_n(\boldsymbol{\theta})$ é côncava sobre o espaço dos parâmetros para qualquer dado $(\boldsymbol{w}_1, ..., \boldsymbol{w}_n)$, e (iii) $Q_n(\boldsymbol{\theta})$ é função mensurável do dado para todo $\boldsymbol{\theta}$ em Θ. Faça $\widehat{\boldsymbol{\theta}}$ ser o estimador extremo definido anteriormente (7.1). Se existe uma função $Q_0(\boldsymbol{\theta})$ tal que

1. (identificação) $Q_0(\boldsymbol{\theta})$ é unicamente maximizado sobre Θ em $\boldsymbol{\theta}_0 \in \Theta$.

2. (convergência pontual) $Q_n(\cdot)$ converge em probabilidade para $Q_0(\cdot)$ para todo $\boldsymbol{\theta}_0 \in \Theta$, então a medida que $n \to \infty$, $\widehat{\boldsymbol{\theta}}$ existe com probabilidade aproximando 1 e $\widehat{\boldsymbol{\theta}} \to_p \boldsymbol{\theta}_0$.

Na maioria das aplicações Θ é um conjunto convexo aberto, então a condição (i) é satisfeita. Esta proposição requer que $Q_n(\boldsymbol{\theta})$ 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

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

Se $\{\boldsymbol{w}_t\}$ 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

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

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

Lema 7.2 (Lei uniforme dos grandes números): Faça $\{\boldsymbol{w}_t\}$ ser um processo estacionário e ergódico. Suponha que (i) o conjunto Θ é compacto, (ii) $m(\boldsymbol{w}_t;\boldsymbol{\theta})$ é contínua em $\boldsymbol{\theta}$ e (iii) $m(\boldsymbol{w}_t;\boldsymbol{\theta})$ é mensurável em $\boldsymbol{w}_t$ para todo $\boldsymbol{\theta}$ em Θ. Além disso suponha:

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

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

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

Condição de dominância: Faça $\{\boldsymbol{w}_t\}$ ser um processo estacionário ergódico. Suponha que (i) o espaço de parâmetros Θ é compacto, (ii) $\boldsymbol{h}(\boldsymbol{w}_t;\boldsymbol{\theta})$ é contínuo em $\boldsymbol{\theta}$ para todo $\boldsymbol{w}_t$ e (iii) $\boldsymbol{h}(\boldsymbol{w}_t;\boldsymbol{\theta})$ é mensurável em $\boldsymbol{w}_t$ para todo $\boldsymbol{\theta}$ em Θ. Além disso, suponha

$$\textbf{condição de dominância: } \mathrm{E}[\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \| \boldsymbol{h}(\boldsymbol{w}_t;\boldsymbol{\theta}) \|] < \infty$$

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

Consistência para estimadores-M com espaço dos parâmetros compacto

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

1. (identificação) $\mathsf{E}[m(\boldsymbol{w}_t;\boldsymbol{\theta})]$ é unicamente maximizado sobre Θ em $\boldsymbol{\theta}_0 \in \Theta$.

2. (dominância) $\mathsf{E}[\sup_{\boldsymbol{\theta}\in\Theta}|\boldsymbol{h}(\boldsymbol{w}_t;\boldsymbol{\theta})|] < \infty$

Então $\widehat{\boldsymbol{\theta}} \to_p \boldsymbol{\theta}_0$

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 $\Theta(\subset \Re^p)$, (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 Θ. Faça $\hat{\boldsymbol{\theta}}$ ser o estimador-M definido por (7.1.1) e (7.1.2). Suponha além disso:

1. (identificação) $\mathrm{E}[m(\boldsymbol{w}_t; \boldsymbol{\theta})]$ é unicamente maximizado sobre Θ em $\boldsymbol{\theta}_0 \in \Theta$.

2. (convergência pontual) $\mathrm{E}[|m(\boldsymbol{w}_t;\boldsymbol{\theta})|] < \infty$ para todo $\boldsymbol{\theta}$ em Θ.

Então, a medida que $n \to \infty$, $\widehat{\boldsymbol{\theta}}$ existe com probabilidade aproximando 1 e $\widehat{\boldsymbol{\theta}} \to_p \boldsymbol{\theta}_0$.

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 $\boldsymbol{\tau}(\boldsymbol{\theta}) : \Theta \rightarrow \Lambda \equiv \boldsymbol{\tau}(\Theta)$ tal que

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

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

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

seja a função objetivo após a reparametrização.

O estimador é consistente porque

$$\underset{n\to\infty}{\mathrm{plim}}\widehat{\boldsymbol{\theta}} = \underset{n\to\infty}{\mathrm{plim}}\boldsymbol{\tau}^{-1}(\widehat{\boldsymbol{\lambda}}) =$$

$$= \boldsymbol{\tau}^{-1}(\underset{n\to\infty}{\mathrm{plim}}\widehat{\boldsymbol{\lambda}}) = \boldsymbol{\tau}^{-1}(\boldsymbol{\lambda}_0) = \boldsymbol{\lambda}_0 \qquad (36)$$

Identificação NLS

Mean square error:

$$\mathsf{E}[\{y_t - h(\boldsymbol{x}_t)\}^2] =$$

$$= \mathsf{E}[\{y_t - \mathsf{E}(y_t \mid \boldsymbol{x}_t)\}^2] + \mathsf{E}[\{\mathsf{E}(y_t \mid \boldsymbol{x}_t) - h(\boldsymbol{x}_t)\}^2] \geq$$
$$\geq \mathsf{E}[\{y_t - \mathsf{E}(y_t \mid \boldsymbol{x}_t)\}^2] \quad (37)$$

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

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

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

Identificação média condicional: a condição de identificação funciona se

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

Identificação para ML Condicional

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

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

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

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

$$\mathsf{E}[\log f(y_t \mid \boldsymbol{x}_t; \boldsymbol{\theta})] > \mathsf{E}[\log f(y_t \mid \boldsymbol{x}_t; \boldsymbol{\theta}_0)] \quad \text{se } \log f(y_t \mid \boldsymbol{x}_t; \boldsymbol{\theta}) \neq \log f(y_t \mid \boldsymbol{x}_t; \boldsymbol{\theta}_0)\} \tag{42}$$

$$\mathsf{E}[\log f(y_t \mid \boldsymbol{x}_t; \boldsymbol{\theta})] = \mathsf{E}[\log f(y_t \mid \boldsymbol{x}_t; \boldsymbol{\theta}_0)] \quad \text{se } \log f(y_t \mid \boldsymbol{x}_t; \boldsymbol{\theta}) = \log f(y_t \mid \boldsymbol{x}_t; \boldsymbol{\theta}_0) \tag{43}$$

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

$$f(y_t \mid \boldsymbol{x}_t; \boldsymbol{\theta}) \neq f(y_t \mid \boldsymbol{x}_t; \boldsymbol{\theta}_0) \tag{44}$$

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 $\{y_t, \boldsymbol{w}_t\}$ serem processos estacionários e ergódicos com densidade condicional $f(y_t \mid \boldsymbol{x}_t; \boldsymbol{\theta})$ e faça $\widehat{\boldsymbol{\theta}}$ ser o estimador ML condicional, que maximiza a média do log da versossimilhança condicional:

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\text{argmax}} \frac{1}{n} \sum_{i=1}^{n} \log f(y_t \mid \boldsymbol{x}_t; \boldsymbol{\theta})$$

Suponha modelo corretamente especificado, então $\boldsymbol{\theta}_0 \in \Theta$. Suponha também que (i) Θ é um subconjunto compato de $\Re^p$, (ii) $f(y_t \mid \boldsymbol{x}_t; \boldsymbol{\theta})$ é contínua em $\boldsymbol{\theta}$ para qualquer $\{y_t, \boldsymbol{w}_t\}$, e (iii) $f(y_t \mid \boldsymbol{x}_t; \boldsymbol{\theta})$ é mensurável em $\{y_t, \boldsymbol{w}_t\}$ para todo para todo $\boldsymbol{\theta}$ em Θ. Suponha além disso:

1. (identificação) $\text{Prob}[f(y_t \mid \boldsymbol{x}_t; \boldsymbol{\theta}) \neq f(y_t \mid \boldsymbol{x}_t; \boldsymbol{\theta}_0)] > 0$ para todo $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0 \in \Theta$.

2. (dominância) $\text{E}[\sup_{\boldsymbol{\theta} \in \Theta} |\log f(y_t \mid \boldsymbol{x}_t; \boldsymbol{\theta})|] < \infty$

Então $\widehat{\boldsymbol{\theta}} \to_p \boldsymbol{\theta}_0$

Consistência para ML condicional sem espaço dos parâmetros compacto

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

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\text{argmax}} \frac{1}{n} \sum_{i=1}^{n} \log f(y_t \mid \boldsymbol{x}_t; \boldsymbol{\theta})$$

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

$f(y_t \mid \boldsymbol{x}_t; \boldsymbol{\theta})$ é mensurável em $\{y_t, \boldsymbol{w}_t\}$ para todo para todo $\boldsymbol{\theta}$ em Θ. Suponha além disso:

1. (identificação) Prob$[f(y_t \mid \boldsymbol{x}_t; \boldsymbol{\theta}) \neq f(y_t \mid \boldsymbol{x}_t; \boldsymbol{\theta}_0)] > 0$ para todo $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0 \in \Theta$.

2. (dominância) E$[|\log f(y_t \mid \boldsymbol{x}_t; \boldsymbol{\theta})|] < \infty$ para todo $\boldsymbol{\theta}_0 \in \Theta$

Então, a medida que $n \to \infty$, $\widehat{\boldsymbol{\theta}}$ existe com probabilidade aproximando 1 e $\widehat{\boldsymbol{\theta}} \to_p \boldsymbol{\theta}_0$

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

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

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

Esta condição é satisfeita se $\mathrm{E}(\boldsymbol{x}_t\boldsymbol{x}_t')$ 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, $\log f(y_t \mid \boldsymbol{x}_t; \boldsymbol{\theta}) \neq \log f(y_t \mid \boldsymbol{x}_t; \boldsymbol{\theta}_0)$ com probabilidade positiva se $\sigma^2 \neq \sigma_0^2$. Então, considerando o caso $\sigma^2 = \sigma_0^2$ mas com $\boldsymbol{\beta} \neq \boldsymbol{\beta}_0$. Então

$$\begin{aligned} \mathsf{E}[(\boldsymbol{x}_t'\boldsymbol{\beta} - \boldsymbol{x}_t'\boldsymbol{\beta}_0)^2] = \mathsf{E}[\{\boldsymbol{x}_t'(\boldsymbol{\beta} - \boldsymbol{\beta}_0)\}^2] = \\ = (\boldsymbol{\beta} - \boldsymbol{\beta}_0)'\mathsf{E}[\boldsymbol{x}_t\boldsymbol{x}_t'](\boldsymbol{\beta} - \boldsymbol{\beta}_0) > 0 \quad (46) \end{aligned}$$

Consequentemente, $\boldsymbol{x}_t'\boldsymbol{\beta} \neq \boldsymbol{x}_t'\boldsymbol{\beta}_0$ com probabilidade positiva; se $\boldsymbol{x}_t'\boldsymbol{\beta} = \boldsymbol{x}_t'\boldsymbol{\beta}_0$ com probabilidade 1, então $\mathsf{E}[(\boldsymbol{x}_t'\boldsymbol{\beta} - \boldsymbol{x}_t'\boldsymbol{\beta}_0)^2]$ será zero. Portanto, $y_t - \boldsymbol{x}_t'\boldsymbol{\beta} \neq y_t - \boldsymbol{x}_t'\boldsymbol{\beta}_0$ com probabilidade positiva, implicando que a condição de identificação seja satisfeita. A mesma condição sobre $\mathsf{E}(\boldsymbol{x}_t\boldsymbol{x}_t')$ também implica condição (b) da Proposição 7.6 porque

$$\mathsf{E}[(\boldsymbol{x}_t'\boldsymbol{\beta} - \boldsymbol{x}_t'\boldsymbol{\beta}_0)^2] = \mathsf{E}[\{\varepsilon_t + \boldsymbol{x}_t'(\boldsymbol{\beta}_0 - \boldsymbol{\beta})\}^2] =$$

$$= \mathsf{E}(\varepsilon_t)^2 + (\boldsymbol{\beta}_0 - \boldsymbol{\beta})'\mathsf{E}(\boldsymbol{x}_t\boldsymbol{x}_t')(\boldsymbol{\beta}_0 - \boldsymbol{\beta})$$

O último termo é finito se $\mathrm{E}(\boldsymbol{x}_t \boldsymbol{x}_t')$ é finito. Estes resultados, junto com o fato observado anteriormente de que a log verossimilhança é côncava após a reparemetrização, implica que o estimador ML condicional do modelo de regressão linear com erro normal é consistente se $\{y_t \boldsymbol{x}_t\}$ é estacionário ergódico e $\mathrm{E}(\boldsymbol{x}_t \boldsymbol{x}_t')$ 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

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

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

A não singularidade de $\mathrm{E}(\boldsymbol{x}_t\boldsymbol{x}_t')$ também implica que a condição (b) da Proposição 7.6 é satisfeita para o modelo probit. É fácil verificar que

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

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 $\mathrm{E}[|\log f(y_t \mid \boldsymbol{x}_t; \boldsymbol{\theta})|] < \infty$ se $\mathrm{E}(\boldsymbol{x}_t\boldsymbol{x}_t')$ existe e é finito (condição implicada pela não-singularidade). Se conclui que o estimador ML probit é consistente se $\{y_t, \boldsymbol{x}_t\}$ é estacionária ergódica e $\mathrm{E}(\boldsymbol{x}_t\boldsymbol{x}_t')$ 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 $\boldsymbol{g}(\boldsymbol{w}_t;\boldsymbol{\theta})$ é contínua em $\boldsymbol{\theta}$ para todo $\boldsymbol{w}_t$. Se $\{\boldsymbol{w}_t\}$ é estacionária ergódica, então $g_n(\boldsymbol{\theta})(\equiv \frac{1}{n}\sum_{t=1}^{n} g(\boldsymbol{w}_t;\boldsymbol{\theta})) \to_p \mathsf{E}[g(\boldsymbol{w}_t;\boldsymbol{\theta})]$. Então a função limite de $Q_0(\boldsymbol{\theta})$ é

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

Esta função é nãopositiva se $\boldsymbol{W}$ é positiva definida. Ela possui um máximo de zero em $\boldsymbol{\theta}_0$ porque $\mathsf{E}[g(\boldsymbol{w}_t;\boldsymbol{\theta}_0)] = \mathbf{0}$ 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 $\mathsf{E}[g(\boldsymbol{w}_t;\boldsymbol{\theta})] \neq \mathbf{0}$ 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 $\boldsymbol{g}_n(\cdot)$ converge uniformemente para $\mathrm{E}[g(\boldsymbol{w}_t;\cdot)]$.

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 $\{\boldsymbol{w}_t\}$ ser estacionária ergódica e faça $\hat{\boldsymbol{\theta}}$ ser o estimador GMM definido por

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}\in\Theta}{\text{argmin}} \left[\frac{1}{n}\sum_{t=1}^{n} g(\boldsymbol{w}_t;\boldsymbol{\theta})\right]' \hat{\boldsymbol{W}} \left[\frac{1}{n}\sum_{t=1}^{n} g(\boldsymbol{w}_t;\boldsymbol{\theta})\right] \qquad (50)$$

tal que a matriz simétrica $\hat{\boldsymbol{W}}$ é assumida convergir em probabilidade para alguma matriz simétrica positiva definida $\boldsymbol{W}$. Suponha que o modelo é corretamente especificado tal que $\mathsf{E}[g(\boldsymbol{w}_t;\boldsymbol{\theta}_0)] = \boldsymbol{0}$ (i.e. valem as condições de ortogonalidade) para $\boldsymbol{\theta}_0) \in \Theta$. Suponha que (i) o espaço dos parâmetros Θ é um subconjunto compacto de $\Re^p$, (ii) $g(\boldsymbol{w}_t;\boldsymbol{\theta})$ é contínuo em $\boldsymbol{\theta}$ para todo $\boldsymbol{w}_t$ e (iii) $g(\boldsymbol{w}_t; \boldsymbol{\theta})$ é mensurável em $\boldsymbol{w}_t$ para todo $\boldsymbol{\theta}$ em Θ (então $\widehat{\boldsymbol{\theta}}$ é uma variável aleatória bem definida pelo Lema 7.1). Além disso, suponha que

1. (identificação) $\mathrm{E}[g(\boldsymbol{w}_t; \boldsymbol{\theta})] \neq \mathbf{0}$ para todo $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$ em Θ

2. (dominância) $\mathrm{E}[\sup_{\boldsymbol{\theta} \in \Theta} \parallel g(\boldsymbol{w}_t; \boldsymbol{\theta}) \parallel] < \infty$

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

Quando $g(\boldsymbol{w}_t; \boldsymbol{\theta})$ é não-linear em $\boldsymbol{\theta}$, especifar 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

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

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

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

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

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

$\boldsymbol{s}(\boldsymbol{w}_t;\boldsymbol{\theta})$ é denominado *vetor score para observação* t. $\boldsymbol{H}(\boldsymbol{w}_t;\boldsymbol{\theta})$ será definido como o *Hessiano para a observação* t. (obs: não confundir 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 $\widehat{\boldsymbol{\theta}}$ está no interior de $\boldsymbol{\Theta}$. Então, se $\widehat{\boldsymbol{\theta}}$ satisfaz as condições de primeira ordem:

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

Usar o seguinte resultado de cálculo:

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

$$\underset{q\times 1}{\boldsymbol{h}(\boldsymbol{x})} = \underset{q\times 1}{\boldsymbol{h}(\boldsymbol{x}_0)} + \underset{q\times q}{\frac{\partial \boldsymbol{h}(\bar{\boldsymbol{x}})}{\partial \boldsymbol{x}'}} \underset{p\times 1}{(\boldsymbol{x} - \boldsymbol{x}_0)} \tag{55}$$

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

Fazendo $q = p$, $\boldsymbol{x} = \widehat{\boldsymbol{\theta}}$, $\boldsymbol{x}_0 = \boldsymbol{\theta}_0$ e $\boldsymbol{h}(\cdot) = \partial Q_n(\cdot)/\partial \boldsymbol{\theta}$ no Teorema do Valor Médio, obtemos as seguintes expansões do valor médio:

$$\underset{p\times 1}{\frac{\partial Q_n(\widehat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}}} = \underset{p\times 1}{\frac{\partial Q_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}}} + \underset{p\times p}{\frac{\partial^2 Q_n(\bar{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}} \underset{p\times p}{\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right)} = \tag{56}$$

$$= \underset{p\times 1}{\frac{1}{n}\sum_{t=1}^{n} \boldsymbol{s}(\boldsymbol{w}_t; \boldsymbol{\theta}_0)} + \underset{p\times p}{\left[\frac{1}{n}\sum_{t=1}^{n} \boldsymbol{H}(\boldsymbol{w}_t; \bar{\boldsymbol{\theta}})\right]} \underset{p\times p}{\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right)}$$

tal que $\bar{\boldsymbol{\theta}}$ é o valor médio que está entre $\hat{\boldsymbol{\theta}}$ e $\boldsymbol{\theta}_0$.[†] Combinando esta equação com a CPO:

$$\underset{p\times 1}{\mathbf{0}} = \underbrace{\frac{1}{n}\sum_{t=1}^{n} \boldsymbol{s}(\boldsymbol{w}_t;\boldsymbol{\theta}_0)}_{p\times 1} + \underbrace{\left[\frac{1}{n}\sum_{t=1}^{n} \boldsymbol{H}(\boldsymbol{w}_t;\bar{\boldsymbol{\theta}})\right]}_{p\times p} \underset{p\times p}{\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right)} \tag{57}$$

assumindo que $\frac{1}{n}\sum_{t=1}^{n} \boldsymbol{H}(\boldsymbol{w}_t;\bar{\boldsymbol{\theta}})$ é não-singular, esta equação pode ser solucionada para $\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0$ para resultar

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) = -\left[\frac{1}{n}\sum_{t=1}^{n} \boldsymbol{H}(\boldsymbol{w}_t;\bar{\boldsymbol{\theta}})\right]^{-1} \sqrt{n}\frac{1}{n}\sum_{t=1}^{n} \boldsymbol{s}(\boldsymbol{w}_t;\boldsymbol{\theta}_0) \tag{58}$$

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 $\boldsymbol{\theta}_0$.

[†]O requerimento de diferenciabilidade contínua é satisfeito se $m(\boldsymbol{w}_t;\boldsymbol{\theta})$ é continuamente diferenciável duplamente com respeito a $\boldsymbol{\theta}$.

Como $\bar{\boldsymbol{\theta}}$ permanece entre $\hat{\boldsymbol{\theta}}$ e $\boldsymbol{\theta}_0$, $\bar{\boldsymbol{\theta}}$ é consistente para $\boldsymbol{\theta}_0$ se $\hat{\boldsymbol{\theta}}$ é. Se $\{\boldsymbol{w}_t\}$ é estacionária ergódica é natural conjecturar que

$$\frac{1}{n}\sum_{t=1}^{n} \boldsymbol{H}(\boldsymbol{w}_t;\bar{\boldsymbol{\theta}}) \rightarrow_p \mathsf{E}[\boldsymbol{H}(\boldsymbol{w}_t;\boldsymbol{\theta}_0)] \tag{59}$$

A estacionariedade ergódica de $\boldsymbol{w}_t$ e a consistência de $\bar{\boldsymbol{\theta}}$ não são suficientes para garantir o resultado acima. Também é necessário assumir convergência uniforme em (59) em uma vizinhança de $\boldsymbol{\theta}_0$). Mas pelo Teorema da Convergência Uniforme, a convergência uniforme de $\frac{1}{n}\sum_{t=1}^{n} \boldsymbol{H}(\boldsymbol{w}_t;\bar{\boldsymbol{\theta}})$ é satisfeita se a seguinte condição de dominância é satisfeita para o Hessiano: para alguma vizinhança $\aleph$ de $\boldsymbol{\theta}_0$), $\mathsf{E}[\sup_{\boldsymbol{\theta}\in\aleph} \|\boldsymbol{H}(\boldsymbol{w}_t;\boldsymbol{\theta}_0)\|] < \infty$. Esta é uma condição suficiente para (59).

Finalmente, se (59) se sustenta e se

$$\frac{1}{\sqrt{n}}\frac{1}{n}\sum_{t=1}^{n} \boldsymbol{s}(\boldsymbol{w}_t;\boldsymbol{\theta}_0) \to_d N(\mathbf{0},\boldsymbol{\Sigma}) \tag{60}$$

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

$$\sqrt{n}\left(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\right) \to_d N\left(\mathbf{0},(\mathsf{E}[\boldsymbol{H}(\boldsymbol{w}_t;\boldsymbol{\theta}_0)])^{-1}\,\boldsymbol{\Sigma}\,(\mathsf{E}[\boldsymbol{H}(\boldsymbol{w}_t;\boldsymbol{\theta}_0)])^{-1}\right) \tag{61}$$

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 $\{\boldsymbol{w}_t\}$ é estacionário ergódico e o estimador M $\widehat{\boldsymbol{\theta}}$ definido por (7.1.1) e (7.1.2) é consistente. Além disso, suponha que

1. $\boldsymbol{\theta}_0$ é interior de Θ;

2. $m(\boldsymbol{w}_t; \boldsymbol{\theta})$ é continuamente diferenciável duplamente em $\boldsymbol{\theta}$ para qualquer $\boldsymbol{w}_t$,

3. $\frac{1}{n}\sum_{t=1}^{n} \boldsymbol{s}(\boldsymbol{w}_t; \boldsymbol{\theta}_0) \to_d N(\mathbf{0}, \boldsymbol{\Sigma})$, $\boldsymbol{\Sigma}$ é positiva definida, tal que $\boldsymbol{s}(\boldsymbol{w}_t; \boldsymbol{\theta})$ é definida em (7.3.2),

4. (condição de dominância local sobre o Hessiano) para alguma vizinhança de $\aleph$ de $\boldsymbol{\theta}_0$),

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

tal que para qualquer estimador consistente de $\tilde{\boldsymbol{\theta}}$, $\frac{1}{n}\sum_{t=1}^{n} \boldsymbol{H}(\boldsymbol{w}_t;\bar{\boldsymbol{\theta}})$ $\to_p \mathsf{E}[\boldsymbol{H}(\boldsymbol{w}_t;\boldsymbol{\theta}_0)]$, tal que $\boldsymbol{H}(\boldsymbol{w}_t;\boldsymbol{\theta}_0)$ é definido em (53).

5. $\mathsf{E}[\boldsymbol{H}(\boldsymbol{w}_t;\boldsymbol{\theta}_0)]$ é nãosingular. Então $\hat{\boldsymbol{\theta}}$ é assintoticamente normal com

$$\mathsf{Avar}(\hat{\boldsymbol{\theta}}) = (\mathsf{E}[\boldsymbol{H}(\boldsymbol{w}_t;\boldsymbol{\theta}_0)])^{-1}\,\boldsymbol{\Sigma}\,(\mathsf{E}[\boldsymbol{H}(\boldsymbol{w}_t;\boldsymbol{\theta}_0)])^{-1}$$

Estimação da Variância Assintótica Consistente

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

$$\text{Avar}(\hat{\boldsymbol{\theta}}) = (\text{E}[\boldsymbol{H}(\boldsymbol{w}_t;\boldsymbol{\theta}_0)])^{-1}\,\boldsymbol{\Sigma}\,(\text{E}[\boldsymbol{H}(\boldsymbol{w}_t;\boldsymbol{\theta}_0)])^{-1} \tag{62}$$

Uma vez que $\hat{\boldsymbol{\theta}} \to_p \boldsymbol{\theta}_0$, condição 4 da Proposição 7.8 implica que

$$\frac{1}{n}\sum_{t=1}^{n} \boldsymbol{H}(\boldsymbol{w}_t;\bar{\boldsymbol{\theta}}) \to_p \text{E}[\boldsymbol{H}(\boldsymbol{w}_t;\boldsymbol{\theta}_0)]$$

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

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

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

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

tal que

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

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

$$\text{Avar}(\widehat{\boldsymbol{\theta}}) = -\{\text{E}[\boldsymbol{H}(\boldsymbol{w}_t;\boldsymbol{\theta}_0)]\}^{-1} = \tag{65}$$

$$= \{\text{E}[\boldsymbol{s}(\boldsymbol{w}_t;\boldsymbol{\theta}_0)\boldsymbol{s}(\boldsymbol{w}_t;\boldsymbol{\theta}_0)']\}^{-1}$$

Esta é a prova da Proposição 7.9.

Normalidade assitótica para ML condicional

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

1. $\boldsymbol{\theta}_0$ é um ponto interior de Θ;

2. $f(y_t \mid \boldsymbol{x}_t; \boldsymbol{\theta})$ é continuamente diferenciável duplamente em $\boldsymbol{\theta}$ para todo $(y_t, \boldsymbol{x}_t)$;

3. $\mathsf{E}[\boldsymbol{s}(\boldsymbol{w}_t; \boldsymbol{\theta}_0)] = \boldsymbol{0}$ e $-\mathsf{E}[\boldsymbol{H}(\boldsymbol{w}_t; \boldsymbol{\theta}_0)] = \mathsf{E}[\boldsymbol{s}(\boldsymbol{w}_t; \boldsymbol{\theta}_0)\boldsymbol{s}(\boldsymbol{w}_t; \boldsymbol{\theta}_0)']$, tal que as funções $\boldsymbol{s}$ e $\boldsymbol{H}$ são definidas em (52) e (53);

4. (condição de dominância local para o Hessiano) para alguma vizinhança de $\aleph$ de $\boldsymbol{\theta}_0$)

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

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

5. $\mathsf{E}[\boldsymbol{H}(\boldsymbol{w}_t;\boldsymbol{\theta}_0)]$ é nãosingular.

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

$$\mathsf{Avar}(\hat{\boldsymbol{\theta}}) = -\{\mathsf{E}[\boldsymbol{H}(\boldsymbol{w}_t;\boldsymbol{\theta}_0)]\}^{-1} =$$

$$= \{\mathsf{E}[\boldsymbol{s}(\boldsymbol{w}_t;\boldsymbol{\theta}_0)\boldsymbol{s}(\boldsymbol{w}_t;\boldsymbol{\theta}_0)']\}^{-1}$$

Estimativa de Avar($\widehat{\boldsymbol{\theta}}$)

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

$$\text{1o estimador de Avar}(\widehat{\boldsymbol{\theta}}) = \left(\frac{1}{n}\sum_{t=1}^{n} \boldsymbol{H}(\boldsymbol{w}_t; \widehat{\boldsymbol{\theta}})\right)^{-1} \tag{66}$$

Como $\widehat{\boldsymbol{\theta}} \to_p \boldsymbol{\theta}_0$, este estimador é consistente com a Proposição 7.9.

Outro estimador baseado na relação Avar($\widehat{\boldsymbol{\theta}}$) $= \{\mathrm{E}[\boldsymbol{s}(\boldsymbol{w}_t; \boldsymbol{\theta}_0)\boldsymbol{s}(\boldsymbol{w}_t; \boldsymbol{\theta}_0)']\}^{-1}$ é

$$\text{2o estimador daAvar}(\widehat{\boldsymbol{\theta}}) = \{\frac{1}{n}\sum_{t=1}^{n} \boldsymbol{s}(\boldsymbol{w}_t; \widehat{\boldsymbol{\theta}})\boldsymbol{s}(\boldsymbol{w}_t; \widehat{\boldsymbol{\theta}})'\}^{-1} \tag{67}$$

Para garantir a consistência para este estimador é preciso uma hipótese adicional na Proposição 7.9 que não precisa ser demostranda pois é observada diretamente. A vantagem do segundo estimador é que se usa apenas a primeira derivada e para métodos de aproximação númerica 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:

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

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:

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

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

$$\boldsymbol{s}(\boldsymbol{w}_t;\widehat{\boldsymbol{\theta}})\boldsymbol{s}(\boldsymbol{w}_t;\widehat{\boldsymbol{\theta}})' = \begin{pmatrix} -\frac{1}{\sigma^4}\boldsymbol{x}_t\boldsymbol{x}_t'\widehat{\varepsilon}_t^2 & -\frac{1}{2\sigma^4}\boldsymbol{x}_{t}.\widehat{\varepsilon}_t + \frac{1}{2\sigma^6}\boldsymbol{x}_{t}.\widehat{\varepsilon}_t^3 \\ -\frac{1}{2\sigma^4}\boldsymbol{x}_{t}.\widehat{\varepsilon}_t + \frac{1}{2\sigma^6}\boldsymbol{x}_{t}.\widehat{\varepsilon}_t^3 & \frac{1}{4\sigma^4} - \frac{1}{2\sigma^6}\widehat{\varepsilon}_t^2 + \frac{1}{4\sigma^8}\widehat{\varepsilon}_t^4 \end{pmatrix}$$

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

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

$$-\mathsf{E}[\boldsymbol{H}(\boldsymbol{w}_t;\widehat{\boldsymbol{\theta}}_0)] = \mathsf{E}[\boldsymbol{s}(\boldsymbol{w}_t;\widehat{\boldsymbol{\theta}})\boldsymbol{s}(\boldsymbol{w}_t;\widehat{\boldsymbol{\theta}})'] = \tag{70}$$

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

Se $\mathsf{E}[\boldsymbol{x}_t\boldsymbol{x}_t']$ é nãosingular, então $\mathsf{E}[\boldsymbol{H}(\boldsymbol{w}_t;\widehat{\boldsymbol{\theta}}_0)]$ é nao singular e o item 5 é satisfeito.

Considerando a condição 4, faça $\tilde{\varepsilon}_t = y_t - \boldsymbol{x}_t'\tilde{\boldsymbol{\beta}}$ para algum estimador consistente $\tilde{\boldsymbol{\beta}}$ e $\tilde{\sigma}^2$. Condição (59) neste exemplo é

$$\begin{pmatrix} -\frac{1}{\sigma^2}\frac{1}{n}\sum_{t=1}^{n}\boldsymbol{x}_t.\boldsymbol{x}_t' & -\frac{1}{\sigma^4}\frac{1}{n}\sum_{t=1}^{n}\boldsymbol{x}_t.\widehat{\varepsilon}_t \\ -\frac{1}{\sigma^4}\frac{1}{n}\sum_{t=1}^{n}\boldsymbol{x}_t.\widehat{\varepsilon}_t & -\frac{1}{2\sigma^4}+\frac{1}{2\sigma^6}\frac{1}{n}\sum_{t=1}^{n}\widehat{\varepsilon}_t^2 \end{pmatrix} \tag{71}$$

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

é direto mostrar que $\tilde{\varepsilon}_t = \varepsilon_t - \boldsymbol{x}_t'(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$. Concluímos que todas as condições da Proposição 7.9 são satisfeitas se $\mathsf{E}[\boldsymbol{x}_t\boldsymbol{x}_t']$ é nãosingular.

Exemplo 7.11

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

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

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

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

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

Aqui $\mathbf{\Phi}'(\cdot)$ é função de densidade cumulativa de $N(0,1)$ e $\phi(\cdot) = \mathbf{\Phi}'(\cdot)$ é a função de densidade. É fácil provar a condição 3:

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

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

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

$$-\mathsf{E}[\boldsymbol{H}(\boldsymbol{w}_t;\boldsymbol{\theta}_0)] = \mathsf{E}[\boldsymbol{s}(\boldsymbol{w}_t;\boldsymbol{\theta}_0)\boldsymbol{s}(\boldsymbol{w}_t;\boldsymbol{\theta}_0)'] = \tag{76}$$

$$= \mathsf{E}[\boldsymbol{\lambda}(\boldsymbol{x}_t'\boldsymbol{\theta}_0)\boldsymbol{\lambda}(-\boldsymbol{x}_t\boldsymbol{\theta}_0)'\boldsymbol{x}_t\boldsymbol{x}_t']$$

tal que

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

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

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

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

Normalidade assintótica do GMM

A função objetivo do GMM é:

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

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

O Teorema do Valor Médio é aplicado em $g_n(\boldsymbol{\theta})$ e não na primeira derivada. Em GMM, a função objetivo precisa ser diferiável 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:

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

$\boldsymbol{G}_n(\boldsymbol{\theta})$ é o Jacobiano de $\boldsymbol{g}_n(\boldsymbol{\theta})$

$$\boldsymbol{G}_n(\boldsymbol{\theta}) \equiv \frac{\partial \boldsymbol{g}_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \tag{81}$$

Aplicando o Teorema do Valor Médio para $\boldsymbol{g}_n(\boldsymbol{\theta})$ se obtém:

$$\underset{(p\times 1)}{\mathbf{0}} = -\underset{(p\times K)}{\boldsymbol{G}_n(\hat{\boldsymbol{\theta}})'}\ \underset{(K\times K)}{\hat{\boldsymbol{W}}}\ \underset{(K\times 1)}{\boldsymbol{g}_n(\boldsymbol{\theta}_0)} - \underset{(p\times K)}{\boldsymbol{G}_n(\hat{\boldsymbol{\theta}})'}\ \underset{(K\times K)}{\hat{\boldsymbol{W}}}\ \underset{(K\times p)}{\boldsymbol{G}_n(\bar{\boldsymbol{\theta}})}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \tag{82}$$

Solucionando para $(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ e substituindo $\boldsymbol{g}_n(\boldsymbol{\theta})$ por $\frac{1}{n}\sum_{t=1}^{n} \boldsymbol{g}(\boldsymbol{w}_t; \boldsymbol{\theta})$:

$$\underset{(p\times 1)}{(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)} = -\left[\underset{(p\times K)}{\boldsymbol{G}_n(\hat{\boldsymbol{\theta}})'}\underset{(K\times K)}{\hat{\boldsymbol{W}}}\underset{(K\times p)}{\boldsymbol{G}_n(\bar{\boldsymbol{\theta}})}\right]^{-1} - \underset{(p\times K)}{\boldsymbol{G}_n(\hat{\boldsymbol{\theta}})'}\underset{(K\times K)}{\hat{\boldsymbol{W}}}\frac{1}{\sqrt{n}}\frac{1}{n}\underset{(K\times 1)}{\sum_{t=1}^{n}} \boldsymbol{g}(\boldsymbol{w}_t; \boldsymbol{\theta}_0) \tag{83}$$

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

$$\boldsymbol{G}_n(\tilde{\boldsymbol{\theta}}) \equiv \frac{\partial \boldsymbol{g}_n(\boldsymbol{\theta})}{\partial \tilde{\boldsymbol{\theta}}} \tag{84}$$

Se $\boldsymbol{w}_t$ é estacionário ergódico e $\tilde{\boldsymbol{\theta}}$ é consistente, é natural conjecturar que esta expressão converge para $\mathrm{E}[\frac{\partial \boldsymbol{g}_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_0}]$. Como foi verdade para o estimador GMM, esta conjectura é verdade se $\frac{\partial \boldsymbol{g}_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_0}$

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 $\{\boldsymbol{w}_t\}$ é estacionário ergódico, $\hat{\boldsymbol{W}}(K \times K)$ converge em probabilidade para uma matriz $\boldsymbol{W}$ simétrica positiva definida e o estimador GMM de $\hat{\boldsymbol{\theta}}$ com dimensão p é consistente. Além disso, suponha que

1. $\boldsymbol{\theta}_0$ é interior de Θ;
2. $\underset{(K\times 1)}{\boldsymbol{g}}(\boldsymbol{w}_t; \boldsymbol{\theta})$ é continuamente diferenciável duplamente em $\boldsymbol{\theta}$ para qualquer $\boldsymbol{w}_t$,
3. $\frac{1}{n}\sum_{t=1}^{n} \boldsymbol{g}(\boldsymbol{w}_t; \boldsymbol{\theta}_0) \rightarrow_d N(\mathbf{0}, \boldsymbol{S})$, $\boldsymbol{S}$ $(K \times K)$ é positiva definida;

4. (condição de dominância local sobre $\frac{\partial \boldsymbol{g}_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$) para alguma vizinhança de $\aleph$ de $\boldsymbol{\theta}_0$),

$$\mathsf{E}[\sup_{\boldsymbol{\theta}\in\aleph} \|\frac{\partial \boldsymbol{g}_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\|] < \infty$$

tal que para qualquer estimador consistente de $\tilde{\boldsymbol{\theta}}$, $\frac{1}{n}\sum_{t=1}^{n} \frac{\partial \boldsymbol{g}_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$ $\rightarrow_p$ $\mathsf{E}[\frac{\partial \boldsymbol{g}_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}]$.

5. $\mathsf{E}[\frac{\partial \boldsymbol{g}_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}]$ é coluna posto completo (full column rank).

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

   $$\mathsf{Avar}(\widehat{\boldsymbol{\theta}}) = \left(\widehat{\boldsymbol{G}}'\widehat{\boldsymbol{W}}\widehat{\boldsymbol{G}}\right)^{-1}\left(\widehat{\boldsymbol{G}}'\widehat{\boldsymbol{W}}\widehat{\boldsymbol{S}}\widehat{\boldsymbol{W}}\widehat{\boldsymbol{G}}\right)\left(\widehat{\boldsymbol{G}}'\widehat{\boldsymbol{W}}\widehat{\boldsymbol{G}}\right)^{-1}$$

   tal que $\underset{K\times p}{\widehat{\boldsymbol{G}}} \equiv \boldsymbol{G}_n(\widehat{\boldsymbol{\theta}}) = \frac{\partial \boldsymbol{g}_n(\widehat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}}$

A hipótese de que $\boldsymbol{S}$ 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:

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

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

$$\mathrm{E}(\boldsymbol{g}_i(\delta)) = 0 \tag{86}$$

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

$$J(\tilde{\delta}) = n.\bar{\boldsymbol{g}}_n'(\boldsymbol{\delta})\widehat{W}\bar{\boldsymbol{g}}_n(\boldsymbol{\delta}) \tag{87}$$

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

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

O estimador GMM eficiente pode ser construído determinando

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

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

Distribuição do Estimador GMM Não-linear: Sob condições de regularidade gerais,

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

tal que

$$\boldsymbol{V}_\delta = \left(\boldsymbol{Q}' \boldsymbol{W} \boldsymbol{Q} \right)^{-1} \left(\boldsymbol{Q}' \boldsymbol{W} \boldsymbol{S} \boldsymbol{W} \boldsymbol{Q} \right) \left(\boldsymbol{Q}' \boldsymbol{W} \boldsymbol{Q} \right)^{-1} \tag{90}$$

$$\boldsymbol{S} = \mathsf{E}(\boldsymbol{g}_i \boldsymbol{g}_i') \tag{91}$$

e

$$\boldsymbol{Q} = \mathsf{E}\left(\frac{\partial}{\partial \delta'} \boldsymbol{g}_i(\delta)\right) \tag{92}$$

$\boldsymbol{Q} = \boldsymbol{G}$ i.e. compatibilizando notações de Hansen e Hayashi.

Se a matriz eficiente de pesos é utilizada, então:

$$\boldsymbol{V}_\delta = \left(\boldsymbol{Q}' \boldsymbol{S}^{-1} \boldsymbol{Q}\right)^{-1} \tag{93}$$

ML vs GMM

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

Caso iid com densidade de $\boldsymbol{w}_t$ dada por $f(\boldsymbol{w}_t;\boldsymbol{\theta})$. Faça $\widehat{\boldsymbol{\theta}}$ ser o estimador GMM associado com as condições de ortogonalidade $\mathrm{E}[\boldsymbol{g}(\boldsymbol{w}_t;\boldsymbol{\theta}_0)] = \mathbf{0}$. Sua variância assintótica é dada na Proposição 7.10. Pode se mostrar que:

$$\mathrm{Avar}(\widehat{\boldsymbol{\theta}}) \geq \mathrm{E}[\boldsymbol{s}(\boldsymbol{w}_t;\boldsymbol{\theta}_0)\boldsymbol{s}(\boldsymbol{w}_t;\boldsymbol{\theta}_0)']^{-1} \tag{94}$$

tal que

$$\boldsymbol{s}(\boldsymbol{w}_t;\boldsymbol{\theta}_0) = \frac{\partial \boldsymbol{g}(\boldsymbol{w}_t;\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}}$$

Isto é, a inversa da matriz de informação $\mathrm{E}[\boldsymbol{s}\boldsymbol{s}']$ 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 $f(\boldsymbol{w}_t; \boldsymbol{\theta})$ que permite a conexão entre diferenciação e integração.

Eficiência assintótica do ML sobre GMM deepende deste resultado porque o limite inferior $\mathsf{E}[\boldsymbol{s}\boldsymbol{s}']^{-1}$ 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 $f(\boldsymbol{w}_t; \boldsymbol{\theta})$, enquanto o GMM não faz isso.

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

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

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

Erro amostral em formato comum

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

Notando que $\frac{\partial Q_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} = \frac{1}{n}\sum_{t=1}^{n} \boldsymbol{s}(\boldsymbol{w}_t; \boldsymbol{\theta})$, a expansão do valor médio em (58) pode ser escrita como

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) = -\left[\frac{1}{n}\sum_{t=1}^{n} \boldsymbol{H}(\boldsymbol{w}_t; \bar{\boldsymbol{\theta}})\right]^{-1} \sqrt{n}\frac{\partial Q_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} \tag{96}$$

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

$$\boldsymbol{\Psi} = \mathsf{E}[\boldsymbol{H}(\boldsymbol{w}_t; \boldsymbol{\theta}_0)] \tag{97}$$

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

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

Por construção o termo entre chaves converge em probabilidade para zero. Pela condição 3 da Proposição 7.8 $\sqrt{n}\frac{\partial Q_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}}(= 1/n\sum \boldsymbol{s}(\boldsymbol{w}_t;\hat{\boldsymbol{\theta}}))$ 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:

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

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 $\sqrt{n}\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right)$ e $-\left[\boldsymbol{\Psi}\right]^{-1}\sqrt{n}\frac{\partial Q_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}}$ desaparecem, a distribuição assintótica de $\sqrt{n}\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right)$ é a mesma que $-\left[\boldsymbol{\Psi}\right]^{-1}\sqrt{n}\frac{\partial Q_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}}$ pelo Lema 2.4(a). Então $\sqrt{n}\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right)$ converge para uma distribuição normal com média zero e variância assintótica dado por:

$$\mathsf{Avar}(\widehat{\boldsymbol{\theta}}) = \left[\boldsymbol{\Psi}\right]^{-1} \boldsymbol{\Sigma} \left[\boldsymbol{\Psi}\right]^{-1} \tag{100}$$

tal que

$$\underset{(p\times p)}{\boldsymbol{\Sigma}} \equiv \mathsf{Avar}\left(\frac{\partial Q_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}}\right)$$

Para estimadores-M, $\sqrt{n}\frac{\partial Q_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} = \frac{1}{n}\sum \boldsymbol{s}(\boldsymbol{w}_t; \widehat{\boldsymbol{\theta}})$ e $\boldsymbol{\Sigma}$ é a variância de longo prazo de $\boldsymbol{s}(\boldsymbol{w}_t; \boldsymbol{\theta}_0)$. Fazendo $\boldsymbol{\Psi} = \mathsf{E}[\boldsymbol{H}(\boldsymbol{w}_t; \boldsymbol{\theta}_0)]$ fornece uma proposição para a variância na Proposição 7.8.

Agora calculamos para a variância para o GMM. No caso GMM:

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

Uma vez que $\boldsymbol{G}_n(\boldsymbol{\theta}_0) = \frac{\partial \boldsymbol{g}(\boldsymbol{w}_t;\hat{\boldsymbol{\theta}}_0)}{\partial \boldsymbol{\theta}'} \to_p \boldsymbol{G}$, $\widehat{\boldsymbol{W}} \to_p \boldsymbol{W}$ e $\frac{1}{n}\sum_{t=1}^{n}\boldsymbol{g}(\boldsymbol{w}_t;\hat{\boldsymbol{\theta}}_0) \to_p N(0,\boldsymbol{S})$ 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

$$\underset{(p\times p)}{\boldsymbol{\Sigma}} \equiv \text{Avar}\left(\frac{\partial Q_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}}\right) = \\ = \underset{(p\times K)}{\boldsymbol{G}'}\underset{(K\times K)}{\boldsymbol{W}}\underset{(K\times K)}{\boldsymbol{S}}\underset{(K\times K)}{\boldsymbol{W}}\underset{(K\times p)}{\boldsymbol{G}} \tag{102}$$

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

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) = -\boldsymbol{B}^{-1}\boldsymbol{c} \tag{103}$$

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

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

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

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

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

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

Teste de hipóteses

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

Como considerado anteriormente desejamos testar um conjunto de r restrições possivelmente nãolineares. Faça $\boldsymbol{\theta}_0$ ser o parâmetro do modelo com dimensão p. A hipótese nula pode ser expressa como

$$H_0 : \underset{(r\times 1)}{\boldsymbol{a}(\boldsymbol{\theta}_0)} = \mathbf{0} \tag{105}$$

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

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

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

$$\boldsymbol{A}_0 = \boldsymbol{A}(\boldsymbol{\theta}_0) \text{ é posto completo.} \tag{107}$$

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

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

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

O parâmetro verdadeiro $\boldsymbol{\theta}_0$ 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 $\boldsymbol{\theta}_0$ satisfaz a restrição. A convergência uniforme em probabilidade de $Q_n(\cdot)$ em $Q_0(\cdot)$ garante que o limite do estimador restrito $\tilde{\boldsymbol{\theta}}$ é a solução limite do problema restrito, que é $\boldsymbol{\theta}_0$.

A trindade

Proposição 7.11 (A trindade): Faça $\widehat{\boldsymbol{\theta}}$ o estimador extremo definido em (1). Considere a hipótese nula em (105) tal que $\boldsymbol{a}(\cdot)$ é continuamente diferenciável (então o Jacobiano é contínuo) e a condição de posto (107) é satisfeita. Faça $\tilde{\boldsymbol{\theta}}$ 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 $\boldsymbol{W} = \boldsymbol{S}^{-1}$ e dado que é disponível estimador consistente $\widehat{\boldsymbol{S}}$ de $\boldsymbol{S}$ construído

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

Defina as estatísticas Wald, LM e LR como:

Wald: $\underset{(1\times r)}{n\boldsymbol{a}(\hat{\boldsymbol{\theta}})'}\left[\underset{(r\times p)}{\boldsymbol{A}(\hat{\boldsymbol{\theta}})}\underset{(p\times p)}{\hat{\boldsymbol{\Sigma}}^{-1}}\underset{(p\times r)}{\boldsymbol{A}(\hat{\boldsymbol{\theta}})'}\right]^{-1}\underset{(r\times 1)}{\boldsymbol{a}(\hat{\boldsymbol{\theta}})}$

LM: $n\underset{(1\times p)}{\left(\frac{\partial Q_n(\tilde{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}}\right)'}\underset{(p\times p)}{\hat{\boldsymbol{\Sigma}}^{-1}}\underset{(p\times 1)}{\left(\frac{\partial Q_n(\tilde{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}}\right)}$

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

Os termos para substituição são:

$Q_n(\boldsymbol{\theta})$**:** • ML: $\frac{1}{n}\sum_{t=1}^{n}\log f(y_t \mid \boldsymbol{x}_t;\boldsymbol{\theta})$;

- GMM eficiente: $-\frac{1}{2}\boldsymbol{g}_n(\boldsymbol{\theta})'\widehat{\boldsymbol{S}}^{-1}\boldsymbol{g}_n(\boldsymbol{\theta})$

$\widehat{\boldsymbol{\Sigma}}$**:** • ML: $-\frac{\partial^2 Q_n(\widehat{\boldsymbol{\theta}})}{\partial\boldsymbol{\theta}\boldsymbol{\theta}'}$ ou $\frac{1}{n}\sum \boldsymbol{s}(\boldsymbol{w}_t;\widehat{\boldsymbol{\theta}})\boldsymbol{s}(\boldsymbol{w}_t;\widehat{\boldsymbol{\theta}})'$

- GMM: $\widehat{\boldsymbol{G}}'\widehat{\boldsymbol{S}}^{-1}\widehat{\boldsymbol{G}}$, $\underset{(K\times p)}{\widehat{\boldsymbol{G}}} = \boldsymbol{G}_n(\widehat{\boldsymbol{\theta}})$

$\tilde{\boldsymbol{\Sigma}}$**:** • ML: substituir $\widehat{\boldsymbol{\theta}}$ por $\tilde{\boldsymbol{\theta}}$ em $\widehat{\boldsymbol{\Sigma}}$

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

Portanto:

1. O estimador extremo restrito $\tilde{\boldsymbol{\theta}}$ 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:

$$\mathsf{E}(\boldsymbol{e}_i(\beta) \mid \boldsymbol{z}_i) = \mathbf{0} \tag{109}$$

aqui $\boldsymbol{e}$ é uma possível função não-linear $(s \times 1)$ dos parâmetros β (em Hayshi era θ). A variável $\boldsymbol{z}_i$ 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 $y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + e_i$ com instrumentos $\boldsymbol{z}_i$ se aplica a esta classe de modelos com $\mathsf{E}(e_i \mid \boldsymbol{z}_i) = 0$. Nesse caso $\boldsymbol{e}_i(\boldsymbol{\beta}) = y_i - \boldsymbol{x}_i'\boldsymbol{\beta}$. No modelo não-linear $\boldsymbol{e}_i(\boldsymbol{\beta}) = y_i - \boldsymbol{g}(\boldsymbol{x}_i'; \boldsymbol{\beta})$. Em um modelo conjunto de média condicional $\mathsf{E}(y_i \mid \boldsymbol{x}_i) = \boldsymbol{x}_i\boldsymbol{\beta}$ e $\mathsf{Var}(y_i \mid \boldsymbol{x}_i) = \boldsymbol{f}(\boldsymbol{x}_i)'\gamma$, então $(s = 2)$

$$\boldsymbol{e}_i(\boldsymbol{\beta}, \gamma) = \left\{ \begin{array}{c} y_i - \boldsymbol{x}_i'\boldsymbol{\beta} \\ (y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2 - \boldsymbol{f}(\boldsymbol{x}_i)'\boldsymbol{\gamma} \end{array} \right.$$

Dada uma restrição de momento condicional, uma restrição de momento incondicional pode ser sempre construída. Isto é, para qualquer função $\boldsymbol{\phi}(\boldsymbol{z}_i'; \boldsymbol{\beta})$ $(l \times 1)$ podemos determinar $\boldsymbol{g}_i(\boldsymbol{\beta}) = \boldsymbol{\phi}(\boldsymbol{z}_i'; \boldsymbol{\beta})\boldsymbol{e}_i(\boldsymbol{\beta})$ que satisfaz $\mathsf{E}(\boldsymbol{g}_i(\beta)) = \mathbf{0}$ 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 Chamberlein (1987). Assuma o caso com $s = 1$. Faça

$$\boldsymbol{R}_i = \mathsf{E}\left(\frac{\partial \boldsymbol{e}_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \mid \boldsymbol{z}_i\right) \tag{110}$$

e

$$\sigma_i^2 = \mathsf{E}(\boldsymbol{e}_i(\boldsymbol{\beta})^2 \mid \boldsymbol{z}_i) \tag{111}$$

Então o instrumento ótimo é:

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

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

$$\boldsymbol{A}_i = -\sigma_i^{-2}\boldsymbol{R}_i \tag{113}$$

Nesse caso o momento ótimo é:

$$\boldsymbol{g}_i(\boldsymbol{\beta}) = \boldsymbol{A}_i \boldsymbol{e}_i(\boldsymbol{\beta}) \tag{114}$$

fazendo $\boldsymbol{g}_i(\boldsymbol{\beta})$ ser esta escolha isto resulta no melhor estimador GMM possível. Na prática $\boldsymbol{A}_i$ é desconhecido, mas a sua forma pode nos ajudar sobre a construção de intrumentos ótimos. Por exemplo, no modelo linear $\boldsymbol{e}_i(\boldsymbol{\beta}) = y_i - \boldsymbol{x}_i'\boldsymbol{\beta}$, observe que

$$\boldsymbol{R}_i = -\mathsf{E}\left(\boldsymbol{x}_i \mid \boldsymbol{z}_i\right)$$

$$\sigma_i^2 = \mathsf{E}\left(\boldsymbol{e}_i^2 \mid \boldsymbol{z}_i\right)$$

Portanto,

$$\boldsymbol{A}_i = \sigma_i^{-2}\mathsf{E}\left(\boldsymbol{x}_i \mid \boldsymbol{z}_i\right)$$

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

No caso das variáveis endógenas, observe que o instrumento eficiente $\boldsymbol{A}_i$ contém a estimação da média condicional de $\boldsymbol{x}_i$ dado $\boldsymbol{z}_i$. Em outras palavras, para conseguir os melhores instrumentos para $\boldsymbol{x}_i$, é necessário o melhor modelo da média condicional de

$\boldsymbol{x}_i$ dado $\boldsymbol{z}_i$, não apenas uma projeção linear arbitrária. O instrumento eficienteé 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.

Bootstraping 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

$$
\begin{aligned}
Q_n(\boldsymbol{\theta}) \cong Q_n(\widehat{\boldsymbol{\theta}}_j) + \boldsymbol{s}_n(\widehat{\boldsymbol{\theta}}_j)'(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_j) + \\
+ \frac{1}{2}(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_j)'\boldsymbol{H}_n(\widehat{\boldsymbol{\theta}}_j)(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_j) \quad (115)
\end{aligned}
$$

tal que $\widehat{\boldsymbol{\theta}}_j$ é a estimativa na iteração j do processo iterativo a ser descrito e $\boldsymbol{s}_n$ e $\boldsymbol{H}_n$ são o gradiente e o Hessiano da função objetivo:

$$\boldsymbol{s}_n = \frac{\partial Q_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'} \tag{116}$$

$$\boldsymbol{H}_n = \frac{\partial^2 Q_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \tag{117}$$

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

$$\widehat{\boldsymbol{\theta}}_{j+1} = \widehat{\boldsymbol{\theta}}_j - [\boldsymbol{H}_n(\widehat{\boldsymbol{\theta}}_j)]^{-1} \boldsymbol{s}_n(\widehat{\boldsymbol{\theta}}_j) \tag{118}$$

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($\widehat{\boldsymbol{\theta}}$) é obtida como $-[\boldsymbol{H}_n(\widehat{\boldsymbol{\theta}})]^{-1}$.*

Gauss-Newton

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

$$\underset{(K\times 1)}{\boldsymbol{g}_n(\boldsymbol{\theta})} \cong \boldsymbol{g}_n(\widehat{\boldsymbol{\theta}}_j) + \underset{(K\times p)}{\boldsymbol{G}_n(\widehat{\boldsymbol{\theta}}_j)}\underset{(p\times 1)}{(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_j)}$$

$$= [\boldsymbol{g}_n(\widehat{\boldsymbol{\theta}}_j) - \boldsymbol{G}_n(\widehat{\boldsymbol{\theta}}_j)\widehat{\boldsymbol{\theta}}_j] - [-\boldsymbol{G}_n(\widehat{\boldsymbol{\theta}}_j)]\boldsymbol{\theta}$$

*$\boldsymbol{H}_n(\widehat{\boldsymbol{\theta}})^{-1} = \frac{1}{n}\sum \boldsymbol{H}(\boldsymbol{w}_t;\boldsymbol{\theta})$.

$$= \boldsymbol{v}_j - \boldsymbol{G}_j\boldsymbol{\theta} \tag{119}$$

tal que

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

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

Se a função $\boldsymbol{g}_n(\boldsymbol{\theta})$ (produto erro-instrumento) na expressão para a função objetivo do GMM dada por $\boldsymbol{v}_j - \boldsymbol{G}_j\boldsymbol{\theta}$, então a função objetivo deve ser quadrática em $\boldsymbol{\theta}$ e o maximizador (ou o minimizador da distância GMM) deve ser o estimador GMM linear:

$$\widehat{\boldsymbol{\theta}}_{j+1} = (\boldsymbol{G}_j'\widehat{\boldsymbol{W}}\boldsymbol{G}_j)^{-1}\boldsymbol{G}_j'\widehat{\boldsymbol{W}}\boldsymbol{v}_j \tag{121}$$

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 comun

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

$$\hat{\boldsymbol{\theta}}_{j+1} = -[-\boldsymbol{G}_n(\hat{\boldsymbol{\theta}}_j)'\widehat{\boldsymbol{W}}\boldsymbol{G}_n(\hat{\boldsymbol{\theta}}_j)]^{-1}[-\boldsymbol{G}_n(\hat{\boldsymbol{\theta}}_j)'\widehat{\boldsymbol{W}}\boldsymbol{g}_n(\hat{\boldsymbol{\theta}}_j)] \tag{122}$$

O segundo termo entre colchetes não é o gradiante solucionado em $\hat{\boldsymbol{\theta}}_j$ para a função objetivo GMM $Q_n(\boldsymbol{\theta}) = -\frac{1}{2}\boldsymbol{g}_n(\boldsymbol{\theta})'\widehat{\boldsymbol{W}}\boldsymbol{g}_n(\boldsymbol{\theta})$. O papel do Hessiano na função objetivo em (**??**) é realizado aqui pelo termo entre colchetes, que é a estimativa baseada em $\hat{\boldsymbol{\theta}}_j$ de $\boldsymbol{G}'\boldsymbol{W}\boldsymbol{G}$. Se $\boldsymbol{g}_n$ 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 $\boldsymbol{g}_n(\boldsymbol{\theta}) = 1/n \sum \boldsymbol{g}(\boldsymbol{w}_t; \boldsymbol{\theta})$ é não linear em $\boldsymbol{\theta}$ 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 $\boldsymbol{\theta}$).

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 $\boldsymbol{g}(y_t, \boldsymbol{z}_t, \boldsymbol{\theta})$ pode ser escrito como $a(y_t, \boldsymbol{z}_t; \boldsymbol{\theta})\boldsymbol{x}_t$. Suponha que a equação $a(y_t, \boldsymbol{z}_t; \boldsymbol{\theta})$ assuma a forma:

$$a(y_t, \boldsymbol{z}_t; \boldsymbol{\theta}) = a_0(y_t, \boldsymbol{z}_t) + \underset{(1\times q)}{\boldsymbol{a}_1(y_t, \boldsymbol{z}_t)'} \underset{(q\times 1)}{\boldsymbol{\alpha}(\boldsymbol{\theta})} \tag{123}$$

tal que $a_0(\cdot)$ e $\boldsymbol{a}_1(\cdot)$ são funções conhecidas de $(y_t, \boldsymbol{z}_t)$ (mas não são funções de $(\boldsymbol{\theta})$). Uma função dessa forma é dita ser *apenas não linear nos parâmetros* ou *linear nas variáveis*. Isto pode ser visto:

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

$$\boldsymbol{G}_n(\boldsymbol{\theta}) \equiv \frac{\partial \boldsymbol{g}_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'} = \underset{(K\times q)}{\left(\frac{1}{n}\sum_{t=1}^{n} \boldsymbol{x}_t \boldsymbol{a}_1(y_t, \boldsymbol{z}_t)'\right)} \underset{(q\times p)}{\frac{\partial \boldsymbol{\alpha}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}} \quad (125)$$

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.