Estimação e Inferência do Método dos Extremos

Bibliografia: Capítulo 7 de Econometric Foundations

Claudio Lucinda

FEA/USP

Overview

Introduction

Estimadores de Extremo são métodos estatísticos que usam a maximização ou minimização de uma função para estimar os parâmetros de um modelo estatístico. Esses estimadores são amplamente utilizados em várias áreas da estatística e econometria, como por exemplo, na regressão linear, no método dos momentos e na máxima verossimilhança. A ideia é encontrar o valor dos parâmetros que maximizam ou minimizam a função objetivo, e essa escolha pode ser justificada por propriedades estatísticas como consistência e eficiência.

Definição Formal - Estimadores de Extremo

- De uma forma geral, os estimadores de extremo englobam qualquer estimador que pode ser definido a partir da otimização de uma função objetivo de estimação, definida a partir de parâmetros θ, e dados $(\mathbf{Y}, \mathbf{X})$.
- Mais especificamente, eles podem ser definidos como

$$\hat{\Theta} = \arg\max_{\theta \in \Omega}[m(\theta, \mathbf{Y}, \mathbf{X})]$$

- Evidentemente, as propriedades deste estimador dependem das propriedades da função objetivo sendo estimada.
- Neste curso, conseguimos estabelecer as propriedades de consistência e normalidade assintótica para os principais estimadores sem condições exageradamente restritivas sobre o modelo probabilístico subjacente.

Consistência do Estimador de Extremo - I

Theorem

O estimador $\hat{\Theta} = \arg\max_{\theta\in\Omega}[m(\theta, \mathbf{Y}, \mathbf{X})]$ *converge em probabilidade para o valor* θ_0 *do parâmetro* θ *se:*

1. $m(\theta, \mathbf{Y}, \mathbf{X})$ *converge em probabilidade para uma função de* θ*, por exemplo* $m_0(\theta)$
2. $m_0(\theta)$ *é contínua em* θ
3. $m_0(\theta)$ *tem um máximo único em* θ_0
4. Ω *é compacto*

Consistência do Estimador de Extremo - II

Theorem

O estimador $\hat{\Theta} = \arg\max_{\theta\in\Omega}[m(\theta, \mathbf{Y}, \mathbf{X})]$ *converge em probabilidade para o valor* θ_0 *do parâmetro* θ *se:*

1. $m(\theta, \mathbf{Y}, \mathbf{X}) \xrightarrow{p} m_0(\theta), \forall\theta \in \Omega$, *em que* $m_0(\theta)$ *representa uma função de* θ
2. $m_0(\theta)$ *é côncava em* θ
3. $m_0(\theta)$ *tem um máximo único em* θ_0
4. Ω *é um conjunto convexo, com* θ_0 *no interior de* Ω

Observações

- Estamos assumindo que θ_0 é o "valor verdadeiro de θ" como uma versão compacta da ideia que existe um modelo de probabilidade que é congruente com o Processo Gerador dos Dados, e esse modelo é indexado e identificado pelo valor θ_0 dos parâmetros.
- A diferença entre os dois teoremas é que o segundo relaxa duas hipóteses do primeiro:
  - O espaço de parâmetros não precisa ser compacto – pode ser simplesmente convexo com o θ_0 interior
  - A convergência em probabilidade da métrica de estimação, ela só precisa ser côncava em θ

Normalidade Assintótica

Theorem

Seja $\hat{\Theta} = \arg\max_{\theta \in \Omega}[m(\theta, \mathbf{Y}, \mathbf{X})]$, *em que*

1. $\hat{\Theta} \xrightarrow{p} \theta_0$, *com* θ_0 *sendo o valor verdadeiro de* θ
2. θ_0 *é um ponto interior de* Ω
3. $m(\theta, \mathbf{Y}, \mathbf{X})$ *é duplamente diferenciável e contínua em* θ, *pelo menos em uma vizinhança* $\kappa(\theta_0)$ *de* θ_0
4. $\sqrt{n}\frac{\partial m(\theta, \mathbf{Y}, \mathbf{X})}{\partial \theta}|_{\theta_0} \xrightarrow{d} \mathbb{N}(\mathbf{0}, \mathbf{\Sigma})$
5. $\lim_{n \to \infty} P\left(\sup_{\theta \in \kappa(\theta_0)} \| \frac{\partial^2 m(\theta, \mathbf{Y}, \mathbf{X})}{\partial \theta \partial \theta'} - \mathbf{h}(\theta) \| < \xi\right) = 1, \quad \forall \xi > 0$, *em que* $\mathbf{h}(\theta)$ *é uma função de* θ
6. $\mathbf{h}(\theta)$ *é uma função contínua e não singular de* θ_0

Então $\sqrt{n}(\hat{\Theta} - \theta_0) \xrightarrow{d} \mathbb{N}(\mathbf{0}, \mathbf{h}(\theta_0)^{-1}\mathbf{\Sigma}\mathbf{h}(\theta_0)^{-1})$

Normalidade Assintótica (II)

Proof.

A intuição é que o estimador de extremo é uma função assintoticamente linear de variáveis aleatórias distribuídas de forma normal multivariada e, portanto, o estimador é assintoticamente normalmente distribuído. Vamos mostrar isso fazendo uma expansão de Taylor em torno do ponto θ_0:

$$\frac{\partial m(\theta, \mathbf{Y}, \mathbf{X})}{\partial \theta} = \left.\frac{\partial m(\theta, \mathbf{Y}, \mathbf{X})}{\partial \theta}\right|_{\theta_0} + \hat{\mathbf{h}}(\theta_*)(\theta - \theta_0)$$

Neste caso, o $\hat{\mathbf{h}}(\theta_*)$ é o hessiano avaliado em um ponto θ_*. Avaliando essa equação em $\theta = \hat{\Theta}$ faz com que o lado direito da igualdade seja igual a zero. Reorganizando e multiplicando por $\sqrt{n}$, temos □

Normalidade Assintótica (III)

Proof.

$$\sqrt{n}(\hat{\Theta} - \theta_0) \xrightarrow{d} \hat{\mathbf{h}}(\theta_*)^{-1}\sqrt{n}\,\frac{\partial m(\theta, \mathbf{Y}, \mathbf{X})}{\partial \theta}\bigg|_{\theta_0}$$

Esse resultado de convergência em distribuição leva ao que tá no teorema pelo Teorema de Slutsky, pelo fato que a derivada parcial possui uma distribuição $\mathbb{N}(\mathbf{0}, \mathbf{\Sigma})$ e pelo fato que combinações lineares de variáveis multivariadamente normalmente distribuídas possuem uma distribuição normal.

$\square$

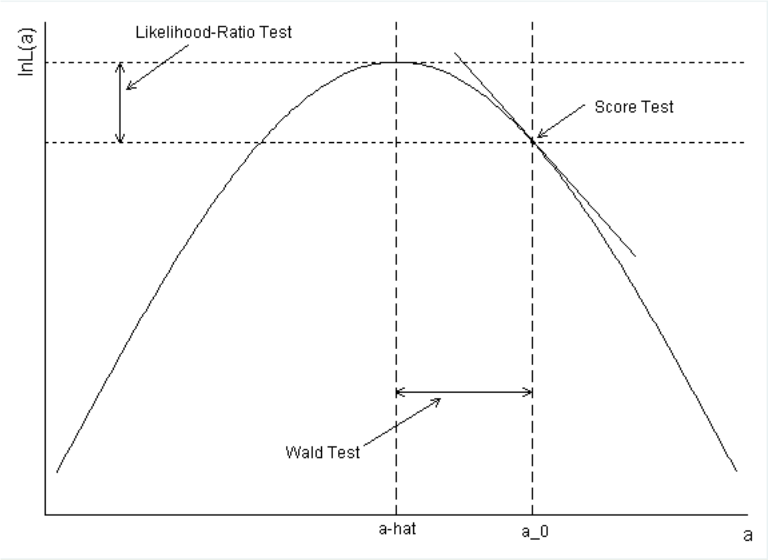
Testes Clássicos de Hipóteses

- A partir da distribuição dos estimadores, iremos agora passar a apresentação dos testes de hipóteses, envolvendo combinações lineares dos coeficientes.
- Entendemos combinações lineares da seguinte forma:
  - Sendo θ o vetor de parâmetros, as combinações lineares envolvem a construção de uma matriz $\mathbf{C}$ e um vetor $\mathbf{r}$ tal que possamos escrever as restrições como sendo $\mathbf{C}\theta = \mathbf{r}$.
  - Supondo que K seja a dimensão de θ e J, o número de restrições, $\mathbf{C}$ é $J \times K$ e r é $K \times 1$
  - Com estas restrições, na prática temos $K - J$ parâmetros "livres" (ver seção 6.2 do Greene).
  - Este é um caso especial de "Testes de Hipóteses Aninhadas" – ou seja, o modelo implícito na hipótese nula pode ser considerado como um caso especial (Aninhado) no modelo implícito na hipótese alternativa.

A Trindade dos Testes – Intuição

- **Razão de Verossimilhança:** É realizado estimando os dois modelos (com e sem restrição) e comparando o ajuste de um modelo com o outro. Impor a restrição no modelo quase sempre fará com que o modelo tenha um pior desempenho (em termos do valor da função objetivo), mas é necessário testar se a diferença é estatisticamente significante.
- **Teste de Wald:** Ele se aproxima do teste de Razão de Verossimilhança, mas apenas precisa estimar um modelo (sem a restrição). Ele funciona testando a hipótese que um conjunto de parâmetros é igual a um valor.
- **Teste do Multiplicador de Lagrange ou Score:** Como no teste de Wald, o teste LM exige que estimemos apenas um modelo. A diferença é que no teste LM, o modelo estimado **não** inclui os parâmetros de interesse. A estatística de teste é calculada com base na inclinação da função objetivo no modelo restrito.
- No contexto de estimadores de extremo, TODAS as estatísticas de teste possuem uma distribuição $\chi^2(J, 0)$, não central com parâmetro de não centralidade igual a zero. As estatísticas de teste no contexto mais geral são as equações (7.6.15) e (7.6.20) para o teste LM (versão completa e versão score), (7.6.22) para o teste de Wald e (7.6.33) para o de Razão de Verossimilhança.

A Trindade dos Testes – Graficamente



Casos Especiais do Estimador de Extremo

- Todos os estimadores que eu conheço são casos especiais de estimadores de extremo – ou seja, boa parte dos resultados derivados acima se aplicam a eles.
- Exemplos:
  - Mínimos Quadrados Ordinários
  - Máxima Verossimilhança
  - Método Generalizado dos Momentos
  - Método do Mínimo Desvio Médio (e regressão quantílica)
  - Método da Máxima Verossimilhança Empírica
  - Método da Máxima Entropia Generalizada
- Alguns desses estimadores serão abordados nas suas seções específicas do curso; aqui falaremos brevemente dos outros.

Máxima Verossimilhança Empírica

- Suponham que $Y_1, Y_2, \cdots, Y_n$ sejam V.A. iid com uma distribuição de probabilidade comum F. Estamos interessados em alguns (ou todos) os valores dos parâmetros $\theta \in \mathbb{R}^K$, associados com F (mas que não descrevem a distribuição!).
- A relação entre θ e F é resumida em um conjunto de funções $\mathbf{h}(Y,\theta) = (h_1(Y,\theta), \cdots, h_m(Y,\theta)'$
- A esperança de $E_\theta[\mathbf{h}(Y,\theta)] = 0$ quando avaliada em θ_0.
- Na prática, os p_i se tornam pesos de probabilidade empírica a serem estimados junto com os θ
- Com isso, a função objetivo é definida como:

$$ln[E_{EL}(\theta;Y)] = \max_{p}\left[\sum_{i=1}^{n}\ln(p_i) \quad t.q. \sum_{i=1}^{n} p_i\mathbf{h}(y_i,\theta) = \mathbf{0} \quad \sum_{i=1}^{n} p_i = 1\right]$$

Entropia Generalizada

- Esse é parecido com o anterior, mas a função a ser maximizada é a medida de entropia de Shannon.
- Nesse caso, os π que são os pesos de probabilidade a serem estimados juntos com o θ

$$KL(\pi, n^{-1}\mathbf{1}) = \sum_{i=1}^{n} \pi_i \ln(n\pi_i) \quad suj. \quad a \quad \sum_{i=1}^{n} \pi_i \mathbf{h}(y_i, \theta) = \mathbf{0} \quad \sum_{i=1}^{n} \pi_i = 1$$

Mínimo Desvio Absoluto e Regressão Quantílica

- O Mínimo Desvio Absoluto passa por encontrar os coeficientes que minimizam a soma dos valores absolutos dos resíduos:

$$S = \sum_{i=1}^{n} |y_i - f(x_i)|$$

- Regressão Quantílica é uma extensão disso, com a função objetivo definida da seguinte forma para cada observação

$$S_i = \begin{cases} \tau(y_i - X_i\theta), & se \quad y_i - X_i\theta \geq 0 \\ (1-\tau)(y_i - X_i\theta), & se \quad y_i - X_i\theta < 0 \end{cases}$$

A função objetivo nesse caso é $S = \sum_{i=1}^{n} S_i$.