Simulação

(rascunho de notas de aula)

(Train, 2009: cap.)

Victor Gomes

Universidade de Brasilia

07/06/2019

Maximização Numérica

Maximização

Função log-verossimilhança:

$$LL(\beta) = \sum_{n=1}^{N} \ln P_n(\beta)/N$$

tal que $P_n(\beta)$ é a probabilidade do resultado observado da escolha individual. N é o tamanho da amostra e β é um vetor $K \times 1$ dos parâmetros. A função log-verossimilhança é dividida por N, então LL é a log-verossimilhança média na amostra. Isto não afeta a localização do máximo e facilita a interpretação.

O objetivo é encontrar o valor de β que maximiza $LL(\beta)$: $\widehat{\beta}$. O econometrista especifica os valores iniciais β_0. Represente β_t o valor de β para cada passo t da simulação. A questão é qual o

melhor passo que se pode tomar em seguida, i.e. qual o melhor valor de β_{t+1}?

O gradiente em β_t é:

$$g_t = \left(\frac{\partial LL(\beta)}{\partial \beta} \right)_{\beta_t} \tag{1}$$

Este vetor nos diz de que forma o passo "sobe" (aproxima o máximo) a função de verossimilhança.

A matriz de segunda derivadas (Hessiano) é:

$$H_t = \left(\frac{\partial g_t}{\partial \beta \prime} \right)_{\beta_t} = \left(\frac{\partial^2 LL(\beta)}{\partial \beta \partial \beta \prime} \right)_{\beta_t} \tag{2}$$

O gradiente tem dimensão $K \times 1$ e o Hessiano é $K \times K$. O Hessiano ajuda a saber quão o tamanho do passo enquanto o gradiente diz qual a direção do passo.

Algoritmos

Newton-Raphson

Para determinar o melhor valor de β_{t+1}, tome a uma aproximação de Taylor de segunda ordem de $LL(\beta_{t+1})$ em torno de $LL(\beta_t)$:

$$LL(\beta_{t+1}) = LL(\beta_t) + (\beta_{t+1} - \beta_t)\prime g_t + \frac{1}{2}(\beta_{t+1} - \beta_t)\prime H_t(\beta_{t+1} - \beta_t) \quad (3)$$

Agora encontre o valor de β_{t+1} que maximiza esta aproximação para $LL(\beta_{t+1})$:

$$\frac{\partial LL(\beta_{t+1})}{\partial \beta_{t+1}} = g_t + H_t(\beta_{t+1} - \beta_t) = 0 \quad (4)$$

$$H_t(\beta_{t+1} - \beta_t) = -g_t \quad (5)$$

$$\beta_{t+1} = \beta_t + (-H_t^{-1})g_t \quad (6)$$

Newton-Raphson usa esta fórmula. O passo seguinte é $(-H_t^{-1})g_t$. Cada passo é a inclinação da log-verossimilhança divido pela sua curvatura.

Quadrática

Se a $LL(\beta)$ é quadrática em β, então o procedimento Newton-Raphson deve alcançar o máximo em um passo a partir do valor inicial.

Isto pode ser verificado para o caso com $K = 1$. Se $LL(\beta)$ é quadrática então ela pode ser escrita como:

$$LL(\beta) = a + b\beta + c\beta^2. \tag{7}$$

O máximo é

$$\frac{\partial LL(\beta_{t+1})}{\partial \beta_{t+1}} = b + 2c\beta = 0 \tag{8}$$

$$\widehat{\beta} = -\frac{b}{2c} \tag{9}$$

O gradiente e o Hessiano são $g_t = b + 2c\beta_t$ e $H_t = 2c$, respectivamente. Então o algoritmo Newton-Raphson resulta em:

$$\beta_{t+1} = \beta_t - H_t^{-1} g_t \tag{10}$$

$$= \beta_t - \frac{1}{2c}(b + 2c\beta_t) \tag{11}$$

$$= -\frac{b}{2c} = \widehat{\beta} \tag{12}$$

Tamanho do passo. Devido ao tamanho do passo, é possível no procedimento NR passar do ponto de máximo da função verdadeira.

Para lidar com essa possibilidade é possível multiplicar o termo do passo por um escalar λ na fórmula NR:

$$\beta_{t+1} = \beta_t + \lambda(-H_t^{-1})g_t \tag{13}$$

Então o vetor $(-H_t^{-1})g_t$ é chamdo de *direção* e λ o tamanho do passo. λ é reduzido para garantir que cada passo do procedimento produza um aumento na $LL(\beta)$. Este ajustamento é realizado em cada iteração.

Começe com $\lambda = 1$.

- Se $LL(\beta t + 1) > LL(\beta_t)$, então se muda para β_{t+1} e começa nova iteração.

- Se $LL(\beta t + 1) < LL(\beta_t)$, então faça $\lambda = 1/2$ e tente de novo. Se com $\lambda = 1/2$, $LL(\beta t + 1)$ ainda é $< LL(\beta_t)$, então faça $\lambda = 1/4$ e tente novamente. Continue com este processo até $LL(\beta t + 1) > LL(\beta_t)$.

- Se este processo resultar em um λ muito pequeno, pouco progresso será feito para encontrar o máximo. Isto pode ser um sinal de que é preciso outro procedimento de maximização.

- Um procedimento análogo pode ser feito em outra direção. A vantagem desta idéia e que usualmente se reduz o número de iterações. Novos valores de λ podem ser tentados sem recalcular g_t e H_t. Ajustar apenas λ é mais rápido.

Convavidade. Se a log-verossimilhança é globalmente côncava, então é garantido que o procedimento NR fornece um aumento na função de verossimilança em cada iteração.

O procedimento NR tem dois problemas. O cálculo do Hessiano é usualmente intensivo em cálculo. Procedimentos que evitam o Hessiano são mais rápidos. O segundo problema é que o NR não garante aumento em cada passo se a função verossimilhança não for globalmente côncava (quando $-H^{-1}$ não for positiva definida o aumento não é garantido).

BHHH

Partir do fato de que a função a ser maximizada é a soma dos termos na amostra. Notação para mostrar que a log-verossimilhança é uma soma das observações.

O *score* de uma função é a derivada da log verossimilança da observação com respeito aos parâmetros: $s_n(\beta_t) = \partial P_n(\beta)/\partial\beta$ solucionada em β_t. O gradiente é o *score* médio: $g_n = \sum_n s_n(\beta_t)/N$. O produto da score da observação n é a matriz $(K \times K)$:

$$s_n(\beta_t)s_n(\beta_t)' = \begin{pmatrix} s_n^1 s_n^1 & \dots & s_n^1 s_n^K \\ \vdots & \ddots & \vdots \\ s_n^K s_n^1 & \dots & s_n^K s_n^K \end{pmatrix}$$

tal que s_n^k é o elemento k de $s_n(\beta_t)$ com a dependência de β_t omitido por conveniência. A matriz de produto na amostra é

$$B_t = \sum_n \frac{s_n(\beta_t)s_n(\beta_t)'}{N}$$

Esta média é relacionada com a matriz de covariância: se o score médio for zero, então B deve ser a matriz de covariância dos scores na amostra.

Nos parâmetros que maximizam a função de verossimilhança, o score médio é zero. O máximo ocorre quando a inclinação é zero. Isso significa que o grandiente, i.e. o score médio, é zero. Então no máximo B é a variância dos scores na amostra. A variância dos scores fornece informação importante para localizar o máximo de uma função. Esta variância fornece *uma medida da curvatura da função de log-verossimilhança.*

A curvatura é pequena quando a variância dos scores é pequena (caso de scores parecidos). Quando existe bastante diferença na amostra a variância é maior. No segundo caso a função log-verossimilhança possui um pico claro, refletindo o fato de qua a amostra possui boa informação para os valores de β.

Essas idéias sobre variância dos scores e sua relação com a curvatura são formalizadas na *indentidade de informação*. A iden-

tidade afirma que a covariância dos scores nos parâmetros verdadeiros é igual a negativa do Hessiano esperado.

BHHH* usam essa relação, B_t, no algoritmo de maximização no lugar de $-H^{-1}$. Cada iteração é definida como:

$$\beta_{t+1} = \beta_t + \lambda B_t^{-1} g_t \tag{14}$$

Vantagens do BHHH sobre NR:

- B_t é mais rápido de calcular do que H_t. Os scores devem ser calculados para obter o gradiente para o procedimento NR e, portanto, calculando B_t como média do produto "externo"

*Berndt, Hall, Hall e Hausman, 1974.

dos scores não toma tempo extra. Em contraste, calculando H_t requer o cálculo das segunda derivadas do função log-verossimilhança.

- B_t é necessariamente positiva definda. O procedimento BHHH garante em cada iteração um aumento em $LL(\beta)$, mesmo na parte convexa da função. Usando a prova fornecida previamente para o NR quando $-H_t$ é positivo definida, o passo BHHH λB_t^{-1} aumenta a $LL(\beta)$ para λ suficientemente pequeno.

Comentário. $B \to -H$ a medida que $N \to \infty$. Esta relação entre as duas matrizes é a implicação da identidade de informação.

Se espera que B_t seja uma melhor aproximação mais perto do máximo de $LL(\beta)$.

Problema:

- O procedimento pode fornecer pequenos passos que aumentem muito pouco a $LL(\beta)$, especialmente quando o procedimento está longe do máximo – especialmente se B_t estiver muito longe do valor verdadeiro $-H_t$ ou porque $LL(\beta)$ é altamente não quadrática na área que o problema está ocorrendo.

BHHH-2

Uma variante do BHHH é obtida pela subtração do score médio antes de calcular o produto "externo." Para qualquer nível de score médio, a covariância dos scores sobre a média dos indivíduis é

$$W_t = \sum_n \frac{(s_n(\beta_t) - g_t)(s_n(\beta_t) - g_t)'}{N} \tag{15}$$

tal que o gradiente g_t é o score médio. W_t é a covariância dos scores em torno de sua média e B_t é a média do produto "externo" dos scores. W_t e B_t são os mesmos quando o gradiente médio é 0.

O estimador BHHH-2 usa W_t no lugar de B_t:

$$\beta_{t+1} = \beta_t + \lambda W_t^{-1} g_t \tag{16}$$

W_t é necessariamente positiva definda, dado que é uma matriz de covariância. Então o procedimento é garantido um aumento em $LL(\beta)$ em cada iteração.[†]

A principal vantagem de BHHH-2 é deixar clara a relação entre a covariância dos scores e o produto externo dos scores.

Steepest Ascent

Este procedimento é definido como:

$$\beta_{t+1} = \beta_t + \lambda g_t \tag{17}$$

A matriz definidora para este procedimento é a identidade I. Como ela é positiva definida ele garante um aumento em cada

[†]Aqui também $W \to -H$ a medida que $N \to \infty$.

passo da iteração – no caso, o maior aumento possível (*steepst ascent*).

DFP e BFGS

Os métodos DFP (Davidson-Fletcher-Powell) e BFGS (Broyden-Fletcher-Goldfard-Shanno) calculam o Hessiano aproximado numa forma que usar informação em mais de um ponto da função de verossimilhança. Estes métodos usam mais de um ponto para aproximar a curvatura da função log-verossimilhança.

O Hessiano é a matriz de segunda derivadas. Ela fornece o montante de mudança na inclinação da curva. O Hessiano é definido por movimentos infinitesimais. Um arco-Hessiano pode ser definido baseado em como o gradiente muda de um ponto para o

outro. Por exemplo, para a função $f(x)$, suponha a inclinação em $x = 3$ como 25 e em $x = 4$ a inclinação é 19. A mudança de inclinação para uma unidade de mudança é -6. Nesse caso, o arco-Hessiano é -6, representando a mudança na inclinação quando o passo muda de 3 para 4. Esse é o conceito usado aqui para aproximar o Hessiano.

No DFP e BFGS, o gradiente é calculado em cada passo da iteração. A diferença no gradiente é usado para calcular o arco-Hessiano. O arco-Hessiano reflete a mudança no gradiente que ocorre para movimentos empíricos na curva. A cada nova iteração esses procedimentos atualizam o arco-Hessiano usando o novo gradiente.

Critério de convergência

Na teoria o máximo é quando o vetor gradiente é zero. Na prática, o gradiente nunca é exatamente zero.

A estatística $m_t = g_t'(-H_t^{-1})$ as vezes é usada para julgar a convegência. Pode se especificar um valor pequeno, como 0.0000001 e determinar que cada iteração ocorra enquanto $M_t < 0.0000001$. Se esta desigualdade é satisfeita então o algoritmo para é os parâmetros são considerados como os valores que convergem.

Max local vs global

Todos estes métodos podem convergir para um máximo local e não o global. Quando a função log-verossimilhança é globalmente côncava (e.e. modelo logit linear nos parâmetros) existe

apenas um máximo. Entretanto, vários modelos de escolha discreta não são globalmente côncavos.

Uma forma de investigar é usar diversos valores iniciais para observar a convergência do modelo. Ver Andrews, Gentzkow, e Shapiro, QJE, 2017.

Variância. Ver Hayashi, cap. 7.

Simulação e Densidades

Sorteios

Normal e uniforme padronizada

Sorteio com distribuição normal e uniforme padronizadas usando gerador de números aleatórios. Notação:

- η sorteio com normal padronizada $N(0, 1)$;
- μ sorteio com uniforme padronizada .

Algumas variáveis aleatórias são transformações de uma normal padronizada. Exemplos:

- Sorteio de uma normal $N(b, s^2)$ é obtida como $\varepsilon = b + s\eta$;

- Sorteio de uma densidade lognormal (exponencial do sorteio de uma normal): $\varepsilon = e^{b+s\eta}$. Nesse caso a média é $\exp\{b + (s^2/2)\}$ e a variância é $\exp(2b + s^2)(\exp(s^2) - 1)$

Cumulativa inversa para densidades univariadas

Considere uma variável aleatória* com densidade $f(\varepsilon)$ e a distribuição cumulativa como $F(\varepsilon)$. Se F é invertível, então sorteios de ε podem ser obtidos de uma uniforme padronizada.

*Apenas funciona com distribuições univariadas.

Por definição, $F(\varepsilon) = k$ sgnifica que a probabilidade de obter um sorteio igual ou abaixo de ε é k, tal que k está entre 0 e 1. Um sorteio μ da uniforme padronizada fornece um número entre 0 e 1. Se pode fazer $F(\varepsilon) = \mu$ e solucionar para o ε correspondente: $\varepsilon = F^{-1}(\mu)$. Quando ε é sorteado desta forma, a distribuição cumulativa dos sorteios é igual a F, tal que é equivalente a sorteios diretos de F.

Exemplo da valor extremo. A densidade é $f(\varepsilon) = \exp(-\varepsilon).\exp(-\exp(-\varepsilon)$ com distribuição cumulativa $F(\varepsilon) = \exp(-\exp(-\varepsilon))$. Um sorteio dessa densidade é obtido como $\varepsilon = -\ln(-\ln \mu)$.

Densidade univariada truncada

Considere uma variável aleatória entre a e b com densidade proporcional a $f(\varepsilon)$ neste intervalo. I.e. a densidade é $(1/k)f(\varepsilon)$

para $a \leq \varepsilon \leq b$ e 0 caso contrário. k é a constante normalizadora que garante que a densidade integra para 1: $k = \int_a^b f(\varepsilon)d\varepsilon = F(b) - F(a)$. Um sorteio dessa densidade pode ser obtido usando o procedimento da cumulatia inversa (acima).

Cacule a média ponderada de $F(a)$ e $F(b)$ como $\bar{\mu} = (1-\mu)F(a) + \mu F(b)$. Então calcule $\varepsilon = F^{-1}(\bar{\mu})$. Como $\bar{\mu}$ está entre $F(a)$ e $F(b)$, ε está entre a e b. O sorteio μ determina o qual longe ir entre a e b (observe que a constante k não precisa ser calculada).

Transformação de Choleski para normal multivariada

Procedimento similar a $\varepsilon = b + s\eta$ para normal multivariada. Faça ε ser um vetor com K elementos distribuidos como $N(b, \Omega)$.

Um fator Choleski de Ω é definido como uma matriz triangular inferior L tal que $LL' = \Omega$.[†]

Sorteio de ε de $N(b, \Omega)$: Tome K sorteios de uma normal padronizada e denomine os vetores desses sorteios $\eta = (\eta_1, ..., \eta_K)'$. Calcule $\varepsilon = b + L\eta$.

Propriedades: ε é normalmente distribuído uma vez que soma de normais é uma normal. A média é b: $E(\varepsilon) = b + LE(\eta) = b$. A covariância é Ω: $\text{Var}(\varepsilon) = E(L\eta(L\eta)') = LE(\eta\eta')L' = L\text{Var}(\eta)L' = LIL' = LL' = \Omega$.

[†]Isso às vezes é chamado de raiz quadrada generalizada de Ω ou desvio padrão generalizado de ε.

Exemplo com 3 dimensões ε e média 0. Um sorteio de ε é calculado como:

$$\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \end{pmatrix} = \begin{pmatrix} s_{11} & 0 & 0 \\ s_{21} & s_{22} & 0 \\ s_{31} & s_{32} & s_{33} \end{pmatrix} \begin{pmatrix} \eta_1 \\ \eta_2 \\ \eta_3 \end{pmatrix}$$

ou

$$\begin{aligned} \varepsilon_1 &= s_{11}\eta_1, \\ \varepsilon_2 &= s_{21}\eta_1 + s_{22}\eta_2, \\ \varepsilon_3 &= s_{31}\eta_1 + s_{32}\eta_2 + s_{33}\eta_3. \end{aligned}$$

Então $\text{Var}(\varepsilon_1) = s_{11}^2$, $\text{Var}(\varepsilon_2) = s_{11}^2 + s_{22}^2$ e $\text{Var}(\varepsilon_3) = s_{11}^2 + s_{22}^2 + s_{33}^2$. Por exemplo: $\text{Cov}(\varepsilon_1, \varepsilon_2) = s_{11}s_{21}$, etc. Os elementos ε_1 e ε_2 são correlacionados por meio de η_1. Em suma, o fator

Choleski expressa K termos correlacionados que surgem de K componentes independentes.

Aceita-rejeita para densidade multivariada truncada

Densidade truncada para densidade multivariada. Suponha sorteio de uma densidade multivariada $g(\varepsilon)$ no intervalo $a \leq \varepsilon \leq b$. a e b são vetores do mesmo tamanho de ε. I.e., sorteio de $f(\varepsilon) = \frac{1}{k} g(\varepsilon)$ se $a \leq \varepsilon \leq b$ e 0 caso contrário (k é a constante de normalização). Se pode obter sorteios de f pelo simples sorteio de g e reter (aceitar) o valor sorteado que pertence ao intervalo relevante. Sorteio fora do intervalo relevante são rejeitados.

Estimação por Simulação

Máxima verossimilhança simulada

No MSL a função de verossimilhança é:

$$LL(\theta) = \sum_n \ln P_n(\theta) \tag{18}$$

tal que θ é um vetor de parâmetros, $P_n(\theta)$ é a probabilidade (exata) da escolha observada para n, e a soma é sobre a amostra de N observações independentes. O estimador ML é o valor de θ que maximiza $LL(\theta)$. Como o gradiente de $LL(\beta)$ é zero no máximo, o estimador ML pode ser definido como o valor de θ que faz com que:

$$\sum_n s_n(\theta) = 0 \tag{19}$$

tal que $s_n(\theta) = \frac{\partial P_n(\theta)}{\partial \theta}$ é o score da observação n.

Faça $\check{P}_n(\theta)$ ser uma aproximação simulada de $P_n(\theta)$. A função simulada da log-verossimilhança é

$$SLL(\theta) = \sum_n \ln \check{P}_n(\theta)$$

e o estimador MSL é o valor θ que maximiza $SLL(\theta)$. O estimador é valor de θ em que $\sum_n \check{s}_n(\theta) = 0$ tal que $\check{s}_n(\theta) = \partial \ln \check{P}_n(\theta)/\partial\theta$.

$\check{P}_n(\theta)$ é um estimador consistente de $P_n(\theta)$ quando R aumenta com N, tal que $E_r\check{P}_n(\theta) = P_n(\theta)$ e $r \in R$. $\check{P}_n(\theta)$ é consistente e eficiente quando R aumenta mais rápido que $\sqrt{N}$.

Método dos momentos simulados

MOM é o valor dos parâmetros que faz com que os resíduos na amostra sejam não correlacionados com as variáveis exógenas. I.e. MOM é definido como os parâmetros que solucionam a equação:

$$\sum_n \sum_j \left[d_{nj} - P_{nj}(\theta) \right] z_{nj} = 0, \tag{20}$$

tal que d_{nj} é a variável dependente que identifica a alternativa escolhida: $d_{nj} = 1$ se n escolhe j e 0 caso contrário. z_{nj} é vetor de variáveis exógenas chamado instrumentos. Os resíduos são $\left[d_{nj} - P_{nj}(\theta) \right]$.

ML para um modelo de escolha discreta é um caso especial do MOM. Se os instrumentos são os scores, $z_{nj} = \partial \ln P_n(\theta)/\partial\theta$, então MOM é o mesmo que ML.

A versão simulada de MOM é chamada de método dos momentos simulados (MSM). No caso é substituido $P_{nj}(\theta)$ pela sua versão simulada $\check{P}_{nj}(\theta)$. O estimador MSM é o conjunto de parâmetros θ que soluciona

$$\sum_n \sum_j \left[d_{nj} - \check{P}_{nj}(\theta) \right] z_{nj} = 0, \tag{21}$$

O estimador é consistente mesmo quando o número de sorteios R é fixo. MSM é assintóticamente equivalente a MOM se R aumenta com N.

MSM é menos eficiente que MSL a menos que instrumentos ideais sejam usados. Entretanto, instrumentos ideais são funções de $\ln P_{nj}$. MSM é usualmente aplicado sem os pesos ideais, que significa que existe perda de eficiência. MSM quando simulado com pesos ideais se torna MSS.

Método scores simulados

MSS possui consistência sem perda de eficiência. O MSS é definido pela seguinte condição:

$$\sum_n s_n(\theta) = 0 \tag{22}$$

tal que $s_n(\theta) = \partial \ln P_n(\theta)/\partial\theta$ é o score para a observação n. O método de scores é simplesmente o ML.

O método dos scores substiui o score exato pelo simulado: $\breve{s}(\theta)$.

A dificuldade do MSS é devido a encontrar um estimador não-viesado do score. O score pode ser reescrito como:

$$s_n(\theta) = \frac{\partial \ln P_{nj}(\theta)}{\partial\theta} = \frac{1}{P_{nj}(\theta)}\frac{\partial P_{nj}}{\partial\theta} \tag{23}$$

Um simulador não-viesado para $\frac{\partial P_{nj}}{\partial \theta}$ é facilmente obtido. A dificuldade é o encontrar o estimador não-viesado para $\frac{1}{P_{nj}(\theta)}$ (a inversa da probabilidade produz viés).

A inversa de $P_n(\theta)$ pode ser simulada da seguinte forma:

1. Faça um sorteio de termos aleatórios da densidade.

2. Calcule a utilidade de cada alternativa desse sorteio

3. Determine qual alternativa j tem a maior utilidade. Se sim, marque o sorteio como aceito. Caso contrário, descarte.

Defina B^r como o número de sorteios até o primeiro ser aceito.

4. Faça os passos R vezes obtendo B^r para $r = 1, ..., R$. O simulador é $1/P_n(\theta) = 1/R \sum_{r=1}^{B} B^r$

Solução numérica

Os estimadores são definidos como o valor de θ que soluciona $\breve{g}(\theta) = 0$, tal que $\breve{g}(\theta) = \sum_n \breve{g}_n(\theta)/N$ é a média amostral da estatística simulada $\breve{g}(\theta)$.