Notas Introdutórias Sobre o Princípio de Máxima Verossimilhança: Estimação e Teste de Hipóteses

Marcelo S. Portugal[1]

1. Introdução

O objetivo central destas notas é fornecer uma breve introdução ao método de máxima verossimilhança.[2] Este procedimento, assim como o método de mínimos quadrados, permite a estimação dos parâmetros de modelos econométricos e a realização de testes de hipóteses relativos a restrições lineares e não lineares ao vetor de parâmetros.

Embora seja bastante antigo[3], foi apenas a partir dos anos oitenta, em função do desenvolvimento dos computadores pessoais de grande potência, que o método de máxima verossimilhança começou a ser utilizado extensivamente em econometria. Como veremos a seguir, o grande obstáculo à utilização prática do método de máxima verossimilhança consiste na freqüente incapacidade de obter-se uma solução explícita para a maioria dos problemas em questão. Neste sentido, existe a necessidade de utilizar-se algum método de otimização numérica para a obtenção dos parâmetros de interesse.

A grande importância do método de máxima verossimilhança consiste nas boas propriedades assintóticas dos estimadores, que são consistentes e assintoticamente eficientes.

---

[1] Gostaria de agradecer a colaboração de Suzana Menna Barreto Coccaro e dos bolsistas de iniciação científica Frederico Pinto (CNPq/UFRGS) e Leandro Milititisky (PRUNI/UFRGS).

[2] Uma abordagem detalhada e bastante completa do método de máxima verossimilhança pode ser encontrada em Cramer (1986).

[3] A formulação original foi feita por Fisher (1929).

2. O Método de Máxima Verossimilhança

Uma amostra aleatória $(y_1, y_2, \ldots, y_n)$, retirada de uma população com uma função de densidade de probabilidade $f(y, \theta)$, a qual depende do vetor de parâmetros θ, tem uma função de densidade de probabilidade (pdf) conjunta dada por

$$\prod_{i=1}^{n} f(y_i, \theta).$$

Isto é, a função de densidade de probabilidade conjunta é simplesmente o produto das densidades de cada uma das observações,

$$f(y_1, \theta) \; x \; f(y_2, \theta) \; x \; \ldots \; x \; f(y_n, \theta)$$

onde θ é um vetor de parâmetros (fixo) e y_i é uma variável aleatória (variável).

Note que, antes da retirada da amostra, cada observação é uma variável aleatória cuja função de densidade de probabilidade é igual a função de densidade de probabilidade da população. A média e a variância de cada observação a ser retirada são iguais à média e variância da população em questão. É neste sentido que dizemos que na função de densidade conjunta, antes de retirada a amostra, θ é fixo e y_i é variável.

Contudo, uma vez que tenha sido obtida uma amostra específiva, y_i torna-se fixo e a função de densidade de probabilidade conjunta pode então ser reinterpretada como sendo uma função do vetor de parâmetros θ, que se tornam variáveis. Para uma dada amostra $(y_1, y_2, \ldots, y_n)$ a função

de densidade de probabilidade conjunta vista como função do vetor de parâmetros desconhecidos θ, é denominada de função de verossimilhança.

Em econometria o problema que se coloca é o de, dada uma amostra, obter-se uma estimativa dos valores dos parâmetros populacionais desconhecidos. Uma possibilidade para a resolução do problema de estimação é escolher o vetor $\hat{\theta}$ que maximize a probabilidade de obtenção da amostra específica $(y_1, y_2, \ldots, y_n)$ que se tem em mãos. Em outras palavras, queremos o vetor $\hat{\theta}$ que faz a probabilidade de obter-se a amostra já obtida a maior possível, ou seja, temos que achar o $\hat{\theta}$ que maximize a função de verossimilhança.

Temos portanto a função de verossimilhança $L(\theta, y)$, onde y é fixo e θ é a variável, e o problema consiste em obter-se o vetor $\hat{\theta}$ que maximiza esta função. O estimador de máxima verossimilhança $\hat{\theta}$ é o vetor que faz

$$L(\hat{\theta}, y) > L(\hat{\hat{\theta}}, y)$$

onde $\hat{\hat{\theta}}$ é qualquer outro estimador de θ.

Do ponto de vista matemático a implementação deste procedimento parece ser simples, pois tudo que temos a fazer é maximizar a função de verossimilhança com respeito a $\hat{\theta}$. Para tanto, basta igualar a zero as derivadas parciais da função de verossimilhança e achar o vetor $\hat{\theta}$ que resolve este conjunto de equações. Na maioria dos casos nós trabalharemos com o logaritmo natural da função de verossimilhança (*ln L*), pois maximizar o logaritmo natural de uma função é em geral mais simples e produz os mesmos resultados da maximização da função original.

Cosidere agora as seguintes definições:

i) escore eficiente (*efficient score*): $\frac{\partial \ln L}{\partial \theta} = S(\theta)$;

ii) matriz de informação: $E\left(-\frac{\partial^2 \ln L}{\partial \theta \partial \theta'}\right) = I(\theta)$

Note que o estimador de máxima verissimilhança ($\hat{\theta}$) vai ser a solução do conjunto de equações $S(\theta) = 0$. Mais ainda, dadas algumas condições bem gerais, é possível mostar-se que $\hat{\theta}$ é consistente, assintoticamente normalmente distribuído e tem variância $[I(\theta)]^{-1}$. Este valor, $[I(\theta)]^{-1}$, é conhecido como o limite inferior de Cramer-Rao, pois não existe outro estimador consistente do vetor θ que tenha variância menor. Neste sentido, o estimador de máxima verossimilhança ($\hat{\theta}$) é também eficiente assintoticamente.

Vamos agora apresentar dois exemplos para facilitar a visualização do funcionamento do método de máxima verossimilhança e da composição da matriz de informação.

Exemplo 1: Considere uma variável aleatória *y* com distribuição normal, média μ e variância σ^2.

$$y \sim N(\mu, \sigma^2)$$

A função de densidade de probabilidade de cada observação é também normal e dada por

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

e a função de densidade conjunta é dada por

$$\prod_{t=1}^{T} f(y_t;\mu,\sigma^2).$$

Logo a função de verossimilhança é

$$L=\prod_{t=1}^{T} f(\mu,\ \sigma^2;\ y_t)$$

e o logaritmo natural de L é

$$ln\ L(\mu,\ \sigma^2;\ y_t) = \sum_{t=1}^{T} ln\ f(\mu,\ \sigma^2;\ y_t)$$

$$= \sum_{t=1}^{T}\left[-\ln\sqrt{2\pi\sigma^2} - \frac{1}{2\sigma^2}\sum_{t=1}^{T}(y_t-\mu)^2\right]$$

$$\ln L = -\frac{T}{2}\ln 2\pi - \frac{T}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{t=1}^{T}(y_t-\mu)^2. \qquad (1)$$

A equação (1) acima é a forma mais usual de apresentação do $ln\ L$. Vamos agora encontrar os estimadores de máxima verossimilhança da média (μ) e da variância (σ^2), isto é vamos obter o vetor (μ, σ^2) que

maximizará a equação (1). Para tanto temos que igualar o escore eficiente a zero($S(\theta) = 0$).[4]

$$\frac{\partial \ln L}{\partial \mu} = \frac{1}{\sigma^2}\sum_{t=1}^{T}\left(y_t - \mu\right)^2 = 0 = \frac{1}{\sigma^2}\sum_{t=1}^{T} y_t - \frac{1}{\sigma^2}(T\mu) \qquad (2)$$

$$\frac{\partial \ln L}{\partial \sigma^2} = -\frac{T}{2\sigma^2} + \frac{1}{2\left(\sigma^2\right)^2}\sum_{t=1}^{T}\left(y_t - \mu\right)^2 = 0 \qquad (3)$$

Resolvendo as equações (2) e (3) para μ e σ^2 temos

$$\hat{\mu} = \frac{1}{T}\sum_{t=1}^{T} y_t = \bar{y} \qquad \text{e} \qquad \hat{\sigma}^2 = \frac{1}{T}\sum_{t=1}^{T}(y_t - \bar{y})^2,$$

que são os estimadores de máxima verossimilhança para a média e a variância. Para obtermos a matriz de informações $I(\theta)$ precisamos encontrar as derivadas segundas de *ln L* com respeito aos parâmetros de interesse.

$$\frac{\partial^2 \ln L}{\partial \mu^2} = -\frac{T}{\sigma^2} \qquad (4)$$

$$\frac{\partial^2 \ln L}{\partial\left(\sigma^2\right)^2} = \frac{T}{2\left(\sigma^2\right)^2} - \frac{1}{\left(\sigma^2\right)^3}\sum_{t=1}^{T}\left(y_t - \mu\right)^2 \qquad (5)$$

[4] Para uma revisão sobre variáveis aleatórias, distribibuição de probabilidade e estatística básica em geral, ver Larson (1982) e Silvey (1975).

$$\frac{\partial^2 \ln L}{\partial \mu \ \partial \sigma^2} = - \frac{1}{(\sigma^2)^2} \sum_{t=1}^{T} (y_t - \mu) \qquad (6)$$

A matriz de informação é formada pelas derivadas segundas do logaritmo da função de verossimilhança avaliadas no ponto de máximo, isto é, em $\hat{\mu}$ e $\hat{\sigma}^2$. Se multiplicarmos e dividirmos o lado direito da equação (5) por *T* e lembrando que $E(y_t) = \mu$, temos

$$I(\theta) = \begin{bmatrix} T/\hat{\sigma}^2 & 0 \\ 0 & T/2\hat{\sigma}^4 \end{bmatrix} \qquad (7)$$

As variâncias dos estimadores de máxima verossimilhança podem então ser obtidas através da inversão da matriz de informação.

$$[I(\theta)]^{-1} = \begin{bmatrix} \hat{\sigma}^2/T & 0 \\ 0 & 2\hat{\sigma}^4/T \end{bmatrix} \qquad (8)$$

Exemplo 2: Consideremos agora o modelo de regressão simples que apresenta resíduos com distribuição normal.

$$y_i = \beta x_i + u_i \ ; \ \text{onde} \qquad u_i \sim N(0, \sigma^2)$$

Neste caso, a função de densidade do erro u_i é dada por

$$f(\mu_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-u_i^2 / 2\sigma^2\right) \quad \text{para} \ \ i = 1,2,\ldots,n$$

Como erro u_i tem distribuição normal com média zero e variância σ^2, y_i vai também ser normal e tem uma função de densidade dada por

$$f(y_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(y_i - \beta x_i)^2\right] \quad \text{para } i = 1,2,\ldots,n. \qquad (9)$$

Logo o logaritmo natural da função de verossimilhança é dado por

$$ln\, L\,(\beta,\sigma^2; y_i) = \sum_{i=1}^{n} ln\;\; f(y_i). \qquad (10)$$

Substituindo (9) em (10) temos

$$\ln(\beta,\sigma^2; y_i) = \sum_{i=1}^{n}\left[-\ln\sqrt{2\pi\sigma^2} - \frac{1}{2\sigma^2}(y_i - \beta x_i)^2\right]$$

$$= -\frac{n}{2}\; ln(2\pi) \;-\; \frac{n}{2}\; ln(\sigma^2) \;-\; \frac{1}{2\sigma^2}\sum_{i=1}^{n} (y_i - \beta x_i)^2 \qquad (11)$$

Para obter os estimadores de máxima verossimilhança de β e σ^2 temos que igualar as derivadas primeiras da equação (11) à zero.

$$\frac{\partial ln\, L}{\partial \beta} = \;-\; \frac{1}{2\sigma^2} \sum_{i=1}^{n} 2\left(y_i - \beta x_i\right)\left(-x_i\right) = 0 \qquad (12)$$

$$\frac{\partial \ln L}{\partial \sigma^2} = \;-\; \frac{n}{2}\frac{1}{\sigma^2} \;+\; \frac{1}{2\left(\sigma^2\right)^2} \sum_{i=1}^{n}\left(y_i - bx_i\right)^2 = 0 \qquad (13)$$

A resolução das equações (12) e (13) para β e σ^2 fornece os estimadores de máxima verossimilhança.

$$\hat{\beta} = \sum_{i=1}^{n} (y_i x_i) \Big/ \sum_{i=1}^{n} x_i^2 \quad = \hat{\beta}_{OLS} \qquad (15)$$

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} (y_i - \beta x_i)^2}{n} \qquad (16)$$

Note que o estimador de máxima verossimilhança de β é igual ao estimador de mínimos quadrados ordinários e que o estimador de σ^2 é viezado para amostras pequenas, pois estamos dividindo por *n* e não por *n-1*, como seria o caso do estimador não tendencioso da variância. Para amostras grandes, contudo o viés torna-se irrelevante.

Passando agora para a matriz de informação, temos de calcular as derivadas segundas.

$$\frac{\partial^2 \ln L}{\partial \beta^2} = -\frac{1}{\sigma^2} \sum_{i=1}^{n} x_i^2 \qquad (17)$$

$$\frac{\partial^2 \ln L}{\partial (\sigma^2)^2} = \frac{n}{2} \frac{1}{(\sigma^2)^2} - \frac{1}{(\sigma^2)^3} \sum_{i=1}^{n} (y_i - \beta x_i)^2 \qquad (18)$$

$$\frac{\partial^2 \ln L}{\partial \beta \, \partial \sigma^2} = -\frac{1}{(\sigma^2)^2} \sum_{i=1}^{n} (y_i - \beta x_i) x_i \qquad (19)$$

Tomando-se o valor esperado das equações (17), (18) e (19) e lembrando-se que $E(x_i u_i) = 0$, pois *x* e *u* não são correlacionados, e que $E\left(\sum_{i=1}^{n} u_i^2\right) = n\sigma^2$, obtemos a matriz de informação avaliada no máximo a a sua inversa.

$$E\left(-\frac{\partial^2 \ln L}{\partial \theta \, \partial \theta'}\right) = \quad I(\theta) = \begin{bmatrix} \sum_{i=1}^{n} x_i^2 \Big/ \sigma^2 & 0 \\ 0 & \frac{n}{2}\frac{1}{(\sigma^2)^2} \end{bmatrix}$$

$$[I(\theta)]^{-1} = \begin{bmatrix} \hat{\sigma}^2 \Big/ \sum_{i=1}^{n} x_i^2 & 0 \\ 0 & 2\hat{\sigma}^4 / n \end{bmatrix} = \begin{bmatrix} Var(\hat{\beta}) & 0 \\ 0 & Var(\hat{\sigma}^2) \end{bmatrix}$$

Podemos ainda escrever a função de verossimilhança no ponto de máximo como uma função da soma do quadrado dos resíduos (RSS), destacando assim a relação entre os métodos de máxima verossimilhança e mínimos quadrados ordinários.

$$\ln L(\hat{\theta}) = const. - \frac{n}{2} \ln\left(\frac{RSS}{n}\right)$$

3. Função de Verossimilhança Concentrada

Em alguns situações, nem todos os parâmetros são de interesse. Aqueles parâmetros que não são de interesse, por serem conhecidos ou estimados de alguma outra forma, são chamados de parâmetros de

perturbação (*nuisance*). Neste caso, podemos reduzir o espaço dos parâmetros e trabalhar apenas com aqueles que são de nosso interesse. Para tanto pode-se trabalhar com a função de verossimilhança concentrada.

Considere a possibilidade de de estabelecermos uma partição do espaço paramétrico θ, de forma que tenhamos

$$\theta = \theta_1 \times \theta_2$$

A função de verossimilhança pode então ser reescrita como

$$L\left(\theta; y_1 \ldots y_n\right) = L\left(\theta_1, \theta_2; y_1, \ldots, y_n\right)$$

Considere agora a condição de primeira ordem para a obtenção do estimador de θ_2

$$\frac{\partial L}{\partial \theta_2} = 0$$

e escreva esta condição em função de θ_1,

$$\frac{\partial L}{\partial \theta_2} = 0 = g\left(\theta_1\right)$$

Se substituirmos $g\left(\theta_1\right)$ na função L acima teremos a função de verossimilhança concentrada em relação a θ_2

$$L_c\left(\theta_1, g(\theta_1)\,; y_1, \ldots, y_n\right)$$

que é a função apenas de θ_1. Para facilitar a visualização da função de verossimilhança concentrada vejamos um novo exemplo.

Exemplo 3: Como havíamos visto no exemplo 2 o estimador de máxima verossimilhança da variância do erro na equação de regressão é uma função do parâmetro β.

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \beta x_i\right)^2 = g(\beta) \qquad (20)$$

Para obtermos a função de verossimilhança concentrada (*ln* L_c) em relação a σ^2, basta substituir a equação (20) em (11).

$$\ln L_c\left(\beta; y_1, \ldots, y_n, \hat{\sigma}^2\right) = -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln\left[\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \beta x_i\right)^2\right] - \frac{\sum_{i=1}^{n}\left(y_i - \beta x_i\right)^2}{2\left[\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \beta x_i\right)^2\right]}$$

$$\ln L_c = -\frac{n}{2}(\ln 2\pi + 1) - \frac{n}{2}\ln\left(\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \beta x_i\right)^2\right) \qquad (21)$$

4. Decomposição em Função do Erro de Previsão

Até agora nós consideramos um conjunto de *T* observaçãoes independentes. Mas para o caso de variáveis dependentes, pode-se construir a função de verossimilhança de maneira análoga. Este é o caso, por exemplo, quando os resíduos da equação de regressão são heterocedásticos ou autocorrelacionados serialmente.

Considere um conjunto de observações com média μ e variância $\sigma^2 V$. O problema da dependência entre as observações está em V. Se $V=I$, estamos de volta ao caso anterior. Se a diagonal principal de V não for composta de constantes temos heterocedasticidade e se os elementos fora da diagonal principal forem diferentes de zero temos autocorrelação serial. Para facilitar a visualização do problema podemos escrever a matriz de covariâncias como

$$E\left(u_i, u_j\right)^2 = \begin{bmatrix} E\left(u_1^2\right) & E\left(u_1 u_2\right) & E\left(u_1 u_3\right) & \dots & E\left(u_1 u_n\right) \\ E\left(u_2 u_1\right) & E\left(u_2^2\right) & E\left(u_2 u_3\right) & \dots & E\left(u_2 u_n\right) \\ E\left(u_3 u_1\right) & E\left(u_3 u_2\right) & E\left(u_3^2\right) & \dots & E\left(u_3 u_n\right) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ E\left(u_n u_1\right) & E\left(u_n u_2\right) & E\left(u_n u_3\right) & \dots & E\left(u_n^2\right) \end{bmatrix}$$

Se $E\left(u_i^2\right) = \sigma^2$ e $E\left(u_i, u_j\right) = \sigma^2 I$ para *i, j = 1, 2, ..., n* temos $E\left(u_i, u_j\right) = \sigma^2 I$. Se temos $\sigma^2 V$, com $V \neq I$, nós temos problemas de autocorrelação e/ou heterocedasticidade. A presença destes problemas no caso do método de mínimos quadrados, implica na necessidade de ajustes que levam ao "método" de mínimos quadrados generalizados. Os estimadores de mínimos quadrados ordinários e mínimos quadrados generalizados são dados por

$$\hat{\beta}_{OLS} = (X'X)^{-1} X'y \qquad \text{e} \qquad \hat{\beta}_{OLS} = \left(X'V^{-1}X\right)^{-1} X'V^{-1}y$$

Neste caso, a densidade conjunta das observações é dada pela equação (22) apresentada abaixo.

$$\ln L(y_1,\ldots,y_T;\mu,\sigma^2 V) = -\frac{T}{2}\ln 2\pi - \frac{T}{2}\ln\sigma^2 - \frac{1}{2}\ln|V| - \frac{1}{2\sigma^2}(y_t-\mu)'V^{-1}(y_t-\mu)$$

Note que se *V=I* temos a equação (1), pois $|I|=1$ e *ln 1 =0*. A equação (22) é a forma usual de apresentação da função de verossimilhança quando as observações não forem independentes.

O principal objetivo desta seção, contudo, é obter a função de verossimilhança em função do error de previsão. Para tanto devemos lembrar que podemos fatorar a densidade conjunta como

$$\ln l\left(y_1,\ldots,y_T\right) = \ln l\left(y_1,\ldots,y_{T-1}\right) + \ln l\left(y_T/y_1,\ldots,y_{T-1}\right). \qquad (23)$$

Esta fatoração é realizada utilizando-se um resultado básico de probabilidade condicional que sugere que dados dois eventos *A* e *B*, temos que *P(A) = P(A/B)P(B)*. Neste sentido, o segundo termo do lado direito da equação (23) é a distribuição de y_T dada toda a informação até *T-1*.

Considere agora o problema de estimar y_T usando-se toda a informação disponível até *T-1*.Para tanto vamos utilizar o estimador que minimiza o erro médio quadrado é ($\hat{y}_{T/T-1}$), onde

$$\hat{y}_{T/T-1} = E\left(y_T/y_1,\ldots,y_{T-1}\right)$$

A variância do erro de previsão associada a $\hat{y}_{T/T-1}$ é dada por

$$Var\left(y_T/y_{T-1},\ldots,y_1\right) = \sigma^2 f_T$$

As propriedades básicas da distribuição normal, nos garantem que as duas distribuições na fatoração de $ln\,l\left(y_1,\ldots,y_T\right)$ são normais. Portanto, a distribuição condicional pode ser reescrita na forma da equação (24) abaixo.

$$\ln l\left(y_T/y_1,\ldots,y_{T-1}\right) = -\frac{1}{2}\ln 2\pi - \frac{1}{2}\ln\sigma^2 - \frac{1}{2}\ln f_T - \frac{1}{2\sigma^2}\frac{\left(y_T-\hat{y}_{T/T-1}\right)^2}{f_T}$$

A equação (24) foi obtida a partir da equação (22), fazendo-se $T = 1$ e $V = f_T$ (note que não precisamos do módulo pois temos apenas a última observação). Vamos agora repetir a fatoração feita acima para todas as demais observações de forma a obter,

$$\ln l\left(y_1,\ldots,y_T\right) = \sum_{t=2}^{T}\ln l\left(y_t/y_{t-1},\ldots,y_1\right) + \ln l\left(y_1\right). \qquad (25)$$

Agora o estimador de erro quadratico médio mínimo (y_t) é

$$\hat{y}_{t/t-1} = E\left(y_t/y_1,\ldots,y_{t-1}\right).$$

de forma que a equação (25) pode, então, ser escrita como

$$\ln L\left(y_1,\ldots,y_T\right) = -\frac{T}{2}\ln 2\pi - \frac{T}{2}\ln\sigma^2 - \frac{1}{2}\sum_{t=1}^{T}\ln f_t - \frac{1}{2\sigma^2}\sum_{t=1}^{T}\frac{v_t^2}{f_t} \qquad (26)$$

onde $v_t = y_t - \hat{y}_{t/t-1}$ é o erro de previsãoum passo a frente. Chegamos assim a expressão (26), que é uma outra forma de apresentação da função

de verossimilhança, agora com base no erro de previsão. A vantagem em se escrever a função de verossimilhança desta forma, é que v_t e f_t podem ser facilmente calculados de forma sequencial pelo filtro de Kalman.[5]

Em modelos mais gerais onde um vetor de *N x 1* é observado a cada ponto no tempo o argumento é o mesmo de antes com v_t sendo agora também um vetor de dimensão *N x 1*, com os erros de previsão com média zero e matriz de variância F_t. O *ln L* pode então ser escrito como

$$\ln L = -\frac{TN}{2}\ln 2\pi - \frac{1}{2}\sum_{t=1}^{T}\ln|F_t| - \frac{1}{2}\sum_{t=1}^{T} v_t' F_t^{-1} v_t$$

5. Maximização da Verrosimilhança

Na maioria das vezes, contudo, ao contrário do que ocorreu nos exemplos 1 e 2, as condições de primeira ordem para o problema de maximização da função de verossimilhança, não permitem a obtenção de uma solução explícita para os estimadores em questão. O sistema de equações gerado pelas condições de primeira ordem é quase sempre não-linear, obrigando que a maximização seja feita por algum processo numérico. Os procedimentos de otimização numéricos funcionam de forma recursiva, sendo o valor dos parâmetros no período *t+1* uma função do valor destes no período *t*. O algoritmo numérico consiste em tentar um valor para o parâmetro, e depois corrigi-lo continuamente até que algum critério de convergência seja atendido, quando então tem-se um máximo para a função de verossimilhança. Em alguns casos, quando não ocorre convergência, o processo de iteração tem de ser interrompido depois de um número específico de iterações.

[5] Para maiores detalhes a este respeito ver Harvey (1993), capítulo 4.

As recursões são em geral da forma:

$$\hat{\theta}_{i+1} = \hat{\theta}_i + \lambda_i d_i$$

onde

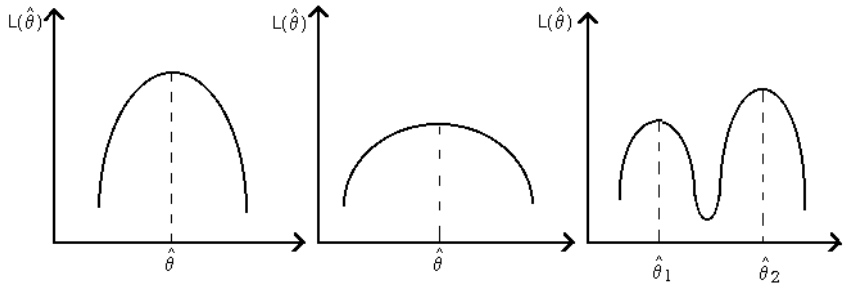
$$\lambda_i d_i = \lambda_i \left[I^*\left(\hat{\theta}\right) \right]^{-1} \left(\left. \frac{\partial L}{\partial \theta} \right|_{\theta=\hat{\theta}} \right)$$

Os diferentes métodos otimização numérica variam quanto a forma de $I^*\left(\hat{\theta}\right)$, que pode ser a matriz de informação, como é o caso no método de *score*, ou alguma aproximação dela. A expressão acima dá algumas pistas no que diz respeito à forma de correção aplicada em cada interação. Ela vai depender da variância do estimador e do gradiente da função de verossimilhança no ponto considerado. Quando o gradiente for bem inclinado e a variância alta, a mudança no parâmetro de uma recursão para a outra é maior, e vice-versa. Neste caso, segue-se a superfície de erro na direção indicada pelo gradiente mais inclinado.

Muitos problemas podem surgir dependendo da forma da função de verossimilhança. A figura (1a) representa o formato ideal para a função de verossimilhança. Contudo, na prática podemos encontrar funções que sejam "achatadas" ao redor do máximo, como é o caso da figura (1b), fazendo com que o algoritmo numérico seja interrompido longe do verdadeiro máximo. Por fim, se existirem máximos locais o algoritmo pode ficar preso no máximo local, uma vez que o algoritmo segue a superfície de erro na direção indicada pelo gradiente mais inclinado, e o valor obtido vai depender do ponto escolhido para iniciar as iterações.[6]

[6] Para maiores detalhes sobre diferentes métodos de otimização numérica, ver Harvey (1990), capítulo 4.

FIGURA 1



(1a) (1b) (1c)

6. Testes de Hipótese

Considere a possibilidade de testar-se um vetor de restrições (*h(θ)*) nos parâmetros. A hipótese nula pode ser representada por

$$H_0: h(\theta) = 0$$

Esta notação cobre hipóteses simples tais como $\theta_1 = 0$, restrições lineares, como $\theta_1 + \theta_2 - 1 = 0$, e não lineares, tais como $\theta_3 + \theta_1\theta_2 = 0$.

Existem três testes distintos, embora assintoticamente equivalentes, baseados no princípio da máxima verossimilhança. A escolha entre eles vai depender, em cada caso, do conhecimento de suas propriedades para amostras pequenas, quando disponível, e da conveniência computacional. Neste trabalho discutiremos apenas os princípios básicos para a a execução destes testes. A aplicação destes

princípios a problemas específicos, tais como autocorrelação serial, fatores comuns, etc, tem de ser feita caso a caso.

6.1 Teste da Razão da Verossimilhança (LR)

Este teste requer a estimação do modelo restrito e sem restrição. Vamos denominar o vetor de parâmetros restrito de $\tilde{\theta}$, isto é, a hipótese a ser testada é $h(\tilde{\theta}) = 0$, e o vetor não restrito de $\hat{\theta}$. Logo, podemos calcular o valor da função de verossimilhança no ponto de máximo com e sem a restrição, vale dizer, $L(\tilde{\theta})$ e $L(\hat{\theta})$ respectivamente. Se a restrição for verdadeira, o valor da função de verossimilhança avaliada em $\tilde{\theta}$ e $\hat{\theta}$ devem estar "próximos", indicando que os dados estão dando suporte a restrição. A questão é como definir precisamente o que seja "próximo".

O teste *LR* é baseado no *ln* da razão entre as duas verossimilhanças, isto é, na diferença entre o $\ln L(\tilde{\theta})$ e $\ln L(\hat{\theta})$. Se H_0 é verdadeiro, a estatística é da forma

$$LR = -2\left[\ln L(\tilde{\theta}) - \ln L(\hat{\theta})\right] \sim \chi^2_g$$

onde *g* é o número de restrições. O teste é, portanto, distribuído assintoticamente como uma *chi-quadrado* com *g* graus de liberdade. Se o valor da estatística for maior que o valor crítico ao nível de significância desejado nós rejeitamos H_0.

6.2 Teste de Wald

Este teste depende apenas da estimação do modelo sem restrição, e a idéia básica é investigar se a estimativa sem restrição esta perto de cumprir a restrição. Isto é, nós utilizamos $\hat{\theta}$, o vetor estimado sem

restrição, para testar se $h(\hat{\theta})$ está próximo de zero. Caso tenhamos $h(\hat{\theta}) = 0$ a restrição estará sendo satisfeita pelos dados. A questão é, novamente, definir o que significa "próximo" de zero. Para fazermos qualquer consideração a este respeito precisamos primeiro saber quem é a variância de $h(\hat{\theta})$. É possível mostrar-se que

$$Var\left[h\left(\hat{\theta}\right)\right] = J'Var(\hat{\theta})J$$

onde o vetor J é dado por

$$J = \left[\partial h_i / \partial \hat{\theta}_i\right].$$

O teste de Wald (W) tem uma distribuição *chi-quadrado* com g graus de liberdade, onde g é o número de restrições testadas.

$$W = h(\hat{\theta})'\left[Var\, h\left(\hat{\theta}\right)\right]^{-1} h(\hat{\theta}) \sim \chi^2_g$$

Para facilitar a visualização do teste Wald vejamos sua aplicação prática para o caso de teste para fatores comuns (CONFAC).

Exemplo 4: Considere o modelo de regressão dinâmico abaixo

$$\hat{y}_t = \hat{\alpha}\, y_{t-1} + \hat{\beta}_0\, x_t + \hat{\beta}_1\, x_{t-1} + \hat{\beta}_2\, z_t + \hat{\beta}_3\, z_{t-1}$$

e suponha que as restrições a serem testadas são

$$h_1\left(\hat{\theta}\right) = \hat{\alpha}\,\hat{\beta}_0 + \hat{\beta}_1 = 0$$

$$h_2(\hat{\theta}) = \hat{\alpha}\,\hat{\beta}_2 + \hat{\beta}_3 = 0.$$

Neste sentido temos

$$h(\hat{\theta}) = \begin{bmatrix} \hat{\alpha}\hat{\beta}_0 + \hat{\beta}_1 \\ \hat{\alpha}\hat{\beta}_2 + \hat{\beta}_3 \end{bmatrix}$$

e a matriz de derivadas $(\partial h / \partial \hat{\theta}_i)$ é

$$J = \partial h / \partial \hat{\theta}_i = \begin{bmatrix} \partial h_1/\partial\hat{\alpha} & \partial h_2/\partial\hat{\alpha} \\ \partial h_1/\partial\hat{\beta}_0 & \partial h_2/\partial\hat{\beta}_0 \\ \partial h_1/\partial\hat{\beta}_1 & \partial h_2/\partial\hat{\beta}_1 \\ \partial h_1/\partial\hat{\beta}_2 & \partial h_2/\partial\hat{\beta}_2 \\ \partial h_1/\partial\hat{\beta}_3 & \partial h_2/\partial\hat{\beta}_3 \end{bmatrix} = \begin{bmatrix} \hat{\beta}_0 & \hat{\beta}_2 \\ \hat{\alpha} & 0 \\ 1 & 0 \\ 0 & \hat{\alpha} \\ 0 & 1 \end{bmatrix}.$$

Tendo o vetor J e a $Var(\hat{\theta})$, que é um resultado simples sempre é fornecido pelos programas de computador, podemos podemos facilmente calcular $Var[h(\hat{\theta})]$ e consequentemente o teste W.

$$Var(\hat{\theta}) = \begin{bmatrix} Var(\hat{\alpha}) & & & cov & \\ & Var(\hat{\beta}_0) & & & \\ & & Var(\hat{\beta}_1) & & \\ cov & & & Var(\hat{\beta}_2) & \\ & & & & Var(\hat{\beta}_3) \end{bmatrix}$$

6.3 Multiplicador de Lagrange ou Escore Eficiente (LM)

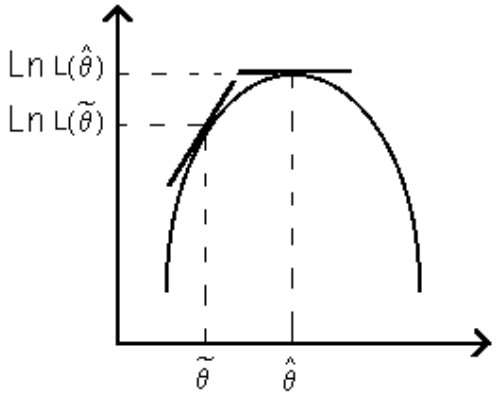
Este teste requer apenas a estimação do modelo com restrição. O primeiro nome sugere que precisamos solucionar um problema de maximização condicionada, enquanto que o segundo nome dá uma idéia mais intuitiva do teste. Se a restrição for válida, isto é H_0 for verdadeira, nós podemos esperar que $S(\tilde{\theta})$, o escore eficiente avaliado em $\tilde{\theta}$, seja "próximo" de zero. Lembre-se que no ponto de máximo temos $S(\theta) = 0$.

Neste sentido, o valor do score eficiente avaliado em $\tilde{\theta}$ é que determina a aceitação ou não da restrição. A estatística é dada por

$$LM = S(\tilde{\theta})'\left[I(\tilde{\theta})\right]^{-1} S(\tilde{\theta}) \sim \chi^2_g.$$

Este é um teste muito usado, e existe uma grande literatura colocando o teste LM em uma forma mais simples para o teste de hipóteses específicas.

FIGURA 2



Como foi dito no início desta seção os três testes são assintoticamente equivalentes e a relação existente entre os três testes podem ser mais facilmente visualizada através da figura 2. O teste de Wald procura medir a distância entre $\tilde{\theta}$ e $\hat{\theta}$, o teste da razão da verossimilhança ocupa-se da distância entre $\ln L(\tilde{\theta})$ e $\ln L(\hat{\theta})$ e, por fim, o teste do multiplicador de Lagrange compara as tangentes nos pontos $\tilde{\theta}$ e $\hat{\theta}$. É fácil ver que quando um dos testes aceita a restrição os demais também o farão

Referências

CRAMER, J. S. (1986), *Econometric Applications of Maximum Likelihood Methods*, Cambridge University Press, Cambridge.

FISHER, R. A. (1921), "On Mathematical Foundations of Theoretical Statistics", *Philosophical Transactions of the Royal Society of London*, série A 222, p309-368.

HARVEY, A. C . (1990), *The Econometric Analysis of Time Series*, segunda edição, Philip Allan, Londres.

HARVEY, A. C. (1993), *Time Series Models*, segunda edição, Harvester Wheatsheaf, Londres.

LARSON, H. J. (1982) *Introduction to Probability Theory and Statistical Inference*, terceira edição, John Wiley & Sons, Singapura.

SILVEY, S. D. (1975), *Statistical Inference*, Chapman and Hall, Londres.