UMA NOTA SOBRE AS PROPRIEDADES DOS ESTIMADORES

Paula Pereda

Denisard Alves

X é uma varável aleatória caracterizada, entre outros aspectos de sua distribuição, por um parâmetro $\theta.$ A distribuição de X pode ser contínua, discreta e até mesmo ser um atributo, caracterizado por uma variável binária.

Exemplo:

X é a renda familiar, portanto contínua. Uma característica importante é a sua média, que pode ser um parâmetro a ser estimado. Para estimarmos um parâmetro, combinamos informações *a priori* que dispomos e a informação da amostra. As informações *a priori* também são conhecidas por hipótese mantida e, também, por modelo.

As informações *a priori* dizem respeito às características de X na população. Essas informações podem ser sobre a distribuição de X, sobre os parâmetros da distribuição de X, em adição a θ ou uma especificação para o próprio θ. Além do modelo, tem-se a disposição as informações da amostra: $\{x_1 . x_2, \dots, x_n\}$.

Vamos dar início à analise com um modelo simples, que se tornará mais elaborado ao longo do curso.

Um estimador de θ, uma característica da distribuição de X, pode ser chamado de $\hat{\theta}$. Como $\hat{\theta}$ é uma função de valores amostrais, podemos escrever:

$$\hat{\theta} = \hat{\theta}\ (\mathrm{x}_1, \mathrm{x}_2, \dots, \mathrm{x}_\mathrm{n}).$$

A única restrição é que a função de valores da amostra, ou seja, a equação do estimador, não pode ter parâmetros desconhecidos. Assim, o estimador é uma função de variáveis aleatória, ou seja, também é uma variável aleatória.

As características básicas da distribuição de $\hat{\theta}$ são dadas pelos momentos:

- Média: E($\hat{\theta}$)
- Variância: Var($\hat{\theta}$)=E[$\hat{\theta} -$ E($\hat{\theta}$)]^2
- Desvio Padrão: $\sqrt{\mathrm{Var}\widehat{(\theta)}}$
- Erro Padrão: $\sqrt{\widehat{\mathrm{Var}(\hat{\theta}})}$
- Erro Amostral: $X - \theta$
- Viés: $\mathrm{E}\big(\hat{\theta}\big) - \theta$
- Erro Quadrático Médio (sigla em inglês: MSE): $E[\hat{\theta} - \theta\ ]^2$

O MSE mede a dispersão entre o estimador e o verdadeiro valor do parâmetro. Se o viés for nulo, o estimador não é viesado, então o MSE é igual à variância. Para melhor entender essa afirmação vamos decompor o MSE:

$$\text{MSE}(\hat{\theta}) = E[\hat{\theta} - \text{E}(\hat{\theta}) + \text{E}(\hat{\theta}) - \theta]^2$$

$$= E\{[\hat{\theta} - \text{E}(\hat{\theta})] + [\text{E}(\hat{\theta}) - \theta]\}^2$$

$$= E[\hat{\theta} - \text{E}(\hat{\theta})]^2 + \text{E}\{\text{E}[(\hat{\theta}) - \theta]\}^2 + 2\text{E}[\hat{\theta} - \text{E}(\hat{\theta})]E[\text{E}(\hat{\theta}) - \theta]$$

O último termo é nulo, pois

$2\text{E}[\hat{\theta} - \text{E}(\hat{\theta})]E[\text{E}(\hat{\theta}) - \theta] = 2\{[\text{E}(\hat{\theta})^2 - \text{E}(\hat{\theta})^2] - \theta\text{E}(\hat{\theta}) + \theta\text{E}(\hat{\theta})\} = 0$

Portanto,

$\text{MSE}(\hat{\theta}) = \text{E}[\hat{\theta} - [\text{E}(\hat{\theta})]^2 + [\text{E}(\hat{\theta}) - \theta]^2 =$ variância + viés^2

Ou seja, o valor do MSE nunca será menor do que a variância. MSE será igual a variância quando o estimador for não viesado.

Vamos agora ver as propriedades desejáveis de um estimador. Essas propriedades se dividem em dois grupos:

Propriedades de *pequenas amostras* ou de *amostras finitas*: válidas para qualquer tamanho de amostra;

Propriedades *de grandes amostras,* ou *assintóticas:* válidas para grandes amostras ($n \to \infty$).

Propriedades para Amostras Finitas

A primeira propriedade, e mais conhecida, é a propriedade de **não viés**.

$\hat{\theta}$ é um estimador não viesado de θ se $\text{E}(\hat{\theta}) = \theta$.

A média é o valor central da distribuição. Quando se diz que a esperança matemática (E) é igual ao verdadeiro valor do parâmetro, o que se quer afirmar é que a distribuição amostral do estimador $\hat{\theta}$, $f(\hat{\theta})$, gera estimativas de θ, acima e abaixo de θ, mas a média desses valores será igual a θ. Um exemplo de um estimador não viesado é a média amostral como um estimador da média populacional, pois

$$E(\bar{X}) = \frac{1}{n}\sum_{i}^{n} X_i = \mu$$

Em que μ é média de X na população.

Mas saber que o estimador produz estimativas, cuja média é igual ao verdadeiro valor do parâmetro na população, não é suficiente para caracteriza o estimador como um bom estimador. Se as estimativas forem muito dispersas entorno do verdadeiro parâmetro, talvez não o qualifique com um bom estimador. É bom lembrar que um estimador que tenha pequena dispersão ou variância, mas é viesado (e não conhecemos nem direção nem magnitude do viés) é muito menos útil. Isso pode ser visto, se tomarmos o caso extremo de um estimador que tenha variância nula. Qualquer constante se qualifica para ser este “estimador”. Vamos supor que seja escolhido o valor 10, ou seja, todos os valores gerados por $f(\hat{\theta})$ estarão concentrados no ponto $\hat{\theta} = 10$. Neste caso seria um estimador sem sentido, pois nenhuma atenção foi dada para a amostra, as informações amostrais sequer foram usadas. Em vista desses argumentos seria desejável que um estimador minimizasse o Erro Quadrático Médio (MSE), definido acima. O problema com este estimador ótimo é que ele não poderia ser definido como um estimador, pois frequentemente na sua fórmula apareceria o próprio valor do parâmetro θ, o qual desejamos estimar. Ou seja, a fórmula seria inútil e incalculável.

A discussão anterior nos leva ao conceito de eficiência, embora ainda exista controvérsia sobre o conceito. Alguns autores usam o conceito de eficiência com base no MSE, apesar das dificuldades que apontamos. Outros só consideram apropriado seu uso no seu conceito assintótico (excluindo os casos em que temos amostras finitas). A visão de eficiência que faz parte do senso comum entre os econometristas caracteriza um estimador como eficiente se (e somente se) ele for não viesado e ao mesmo tempo tenha variância mínima. Formalmente,

$\hat{\theta}$ é um estimador eficiente de θ se ele satisfizer as seguintes condições: (i) $\hat{\theta}$ é não viesado; (ii) $Var(\hat{\theta}) \leq Var(\tilde{\theta})$, onde $\tilde{\theta}$ é qualquer outro estimador não viesado de θ.

Um estimador assim definido por vezes também é chamado de Melhor Estimador Não Viesado-MVUE do inglês, “minimum variance unbiased estimator”. Observe que, por esta definição, um estimador viesado jamais será eficiente, por menor que seja o viés.

Matematicamente o não viés é simples de ser constatado, pois basta tomar a esperança matemática do estimador. Eficiência , entretanto, é mais complicado, pois quando dizemos que um estimador é eficiente estamos comparando sua variância com a variância de todos os estimadores não viesados (e essa quantidade de estimadores pode ser imensa). Na prática, o que ocorre é que desejamos comparar um estimador com algum outro estimador. Tal fato simplifica o conceito de eficiência, pois a eficiência se torna relativa. Ou seja, simplesmente dizemos que um estimador é eficiente com relação a um outro estimador.

Mas em muitas situações é possível fazer afirmação com relação à eficiência em termos absolutos. Nesses casos, não é necessário se preocupar em comparar a variância de estimadores não viesados, ou verificar a eficiência relativa. A razão é a existência de um teorema conhecido por *Teorema de Cramer-Rao*.

Teorema de Cramer-Rao: Considere X uma v.a. com função de probabilidade $f(X)$, caracterizada por parâmetros $\theta_1, \theta_2, \ldots . \theta_k$. Considere $\hat{\theta}_1$ um estimador não viesado de θ_1, derivado de uma amostra $\{x_1, x_2, \ldots, x_n\}$ de X. Defina $l = lnL = lnf(x_1, x_2, \ldots, x_n)$, conhecida como função log-verossimilhança (logaritmo neperiano da função de verossimilhança L da amostra). Escreva a seguinte matriz:

$$H_{kk} = \begin{bmatrix} -E[\frac{\partial^2 l}{\partial \hat{\theta}_1^2}] & \cdots & -E[\frac{\partial^2 l}{\partial \hat{\theta}_1 \partial \hat{\theta}_K}] \\ \vdots & \ddots & \vdots \\ -E[\frac{\partial^2 l}{\partial \hat{\theta}_K \partial \hat{\theta}_1}] & \cdots & -E[\frac{\partial^2 l}{\partial \hat{\theta}_K^2}] \end{bmatrix}$$

Essa matriz é conhecida como matriz de informação. Considere agora a inversa da matriz de informação, $I^{kk} = (H^{kk})^{-1}$. Então a desigualdade de Cramer-Rao[1] é definida por

$$Var(\hat{\theta}_k) \geq I^{kk}, k = 1, \dots, K$$

Em que onde I^{kk} é a inversa da matriz de informação. Este Teorema nos permite construir um limite inferior (maior que zero) para a variância de que qualquer estimador não viesado, desde que conheçamos a função de distribuição da população a partir da qual os dados foram gerados. (Ou seja, precisamos saber o "DNA" do gerador dos dados). Esse limite inferior é chamado de *Limite Inferior de Cramer-Rao*, que nos garante que nenhum estimador não viesado terá variância menor que este limite, o máximo que pode ocorrer é da variância do estimador ser igual ao limite de Cramer-Rao, jamais será inferior.

Se isso ocorre, sabemos que nenhum outro estimador não viesado terá variância menor que a deste estimador e podemos afirmar que ele é eficiente, no sentido absoluto do conceito. Sabemos que $Var(\bar{X}) = \sigma^2/n$. Se a distribuição geradora da amostra for Normal pode-se mostrar que (σ^2/n) é o limite inferior de Cramer-Rao para estimadores não viesados da média populacional, portanto, a média amostral é um estimador eficiente da média populacional para uma população normal.

É importante observar que nem sempre o limite inferior de Cramer-Rao funciona para um estimador, isso porque em sempre ele é atingido por um estimador. Por exemplo o limite inferior que precisa ser atingido para que o estimador da variância de uma população Normal é dado por $(2\sigma^4/n)$, mas não existe nenhum estimador não viesado de σ^2 que tenha variância igual a este limite[2].

Esta discussão mostrou que o conceito de estimador eficiente entre todos os estimadores não viesados é bastante complicado. Então, para que se obtenha um conceito de eficiência de mais fácil aplicação é melhor que nos restrinjamos a uma classe específica de estimador. No nosso caso, aqueles que são os mais frequentemente utilizados em análise empírica em economia: estimadores que sejam funções lineares das observações da amostra (estimadores lineares). Isso nos leva a um conceito mais especializado de eficiência que pode ser colocado da seguinte forma:

Definição de Eficiência para Estimadores Lineares

$\hat{\theta}$ é o melhor estimador linear não viesado (da sigla em inglês: BLUE)[3] de θ se as seguintes condições forem satisfeitas:

a. $\hat{\theta}$ é uma função linear das observações da amostra;
b. c é um estimador não viesado de θ; e
c. $\mathrm{Var}(\hat{\theta}) \leq \mathrm{Var}(\tilde{\theta})$, onde $\tilde{\theta}$ é qualquer outro estimador linear não viesado de θ.

[1] Essa desigualdade se mantém sob condições bem gerais. Veja Rao, C.R., *Linear Statistical Inference and Its Applications*, 2nd Edition, N.Y. Wiley, 1973

[2] Veja Rao, C.R. *op. cit*.

[3] BLUE do inglês "Best Linear Umbiased Estimator"

A condição de linearidade significa que para uma amostra $x_1, x_2, \ldots, x_n$, onde o estimador tem que ter a forma $a_1x_1 + a_2x_2 + \cdots + a_n\, x_n$, que é exatamente aquela apresentada, por exemplo, para a média amostral:

$$\bar{X} = \frac{1}{n}x_1 + \frac{1}{n}x_2 + \cdots + \frac{1}{n}x_n$$
$$em\ que\ a_1 = a_2 = \cdots = a_n = \frac{1}{n}$$

Ou seja, o estimador da média amostral é linear nas observações da amostra. Como já vimos, ele é também não viesado e sua variância é igual ao limite mínimo de Cramer-Rao, quando a distribuição geradora da amostra for normal. Como consequência, a média amostral é um estimador BLUE (eficiente) da média populacional.

Alguns livros de estatística e de econometria também definem a propriedade de suficiência. Um estimador é dito ser suficiente se ele usar todas as informações da amostra. Portanto, um estimador para ser eficiente, ou BLUE, precisa usar todas as informações da amostra. Embora pouco mencionada é óbvio que se um estimador não usar todas as informações disponíveis na amostra ele será um estimador ineficiente quando comparado a um que usa todas. Suficiência é uma condição necessária para eficiência[4]

As três propriedades – Não Viés, Eficiência e BLUE – representam as três propriedades de pequenas amostras de um estimador, sendo que BLUE equivale a eficiência somente quando o estimador for linear. Todas as três são definidas em termos de média e de variância. Essas propriedades representam para os econometristas todas as propriedades desejáveis para um estimador. Mas, uma palavra de cautela é sempre necessária, pois elas não podem ser definidas para estimadores que não tenham média ou variância (não definidas, ou inexistentes).

Propriedades Assintóticas

Propriedades assintóticas são aquelas válidas apenas para grandes amostras, ou para amostras com tamanho *n* se aproximando do infinito. A distribuição amostral de um estimador é diferente para tamanhos de amostras diferentes. A distribuição da média amostral para amostras de qualquer população é definida pelo **Teorema do Limite Central**.

Teorema do Limite Central

Para qualquer que seja a distribuição de X, com média μ e variância σ^2, a distribuição de

$$(\bar{X} - \mu)/\sigma_{\bar{X}}$$

converge para a N(0,1) (normal padronizada) quando o tamanho da amostra, n, converge para infinito. Portanto a distribuição de $\bar{X}$ em grandes amostras se aproxima da $N(\mu, \sigma^2/n)$.

A Teorema nos garante que quando o tamanho da amostra cresce, a distribuição assintótica será aproximadamente normal. Então, para qualquer que seja a distribuição que tenha gerado os dados

[4] Veja Lindgreen, B. W, *Statistical Theory*, 3rd Edition, N.Y. MacMillan, 1976, p.264.

amostrais, sabemos que o estimador convergirá para a distribuição normal, sendo esta a distribuição assintótica do estimador.

Quando falamos em distribuição assintótica, é importante saber que ela não é a forma final da distribuição do estimador assume quando o tamanho da amostra se aproxima do infinito. O que vai ocorrer, quando n vai para infinito, é que a distribuição do estimador colapsa (melhor ainda quando isso ocorre sobre o verdadeiro valor do parâmetro). Uma distribuição que colapsa em um ponto é chamada de degenerada.

Vamos usar a média amostral como exemplo. Sabemos que, para qualquer tamanho da amostra:

$E(\bar{X}) = \mu$, em que μ é a média de X na população; e

Var$(\bar{X}) = (\sigma^2/n)$, em que σ^2 é a variância de X na população e n é o tamanho da amostra.

Neste caso, é óbvio que a distribuição final da média amostral se degenera na média populacional. A distribuição se degenera quando n vai para infinito, pois a variância se torna zero (conforme n cresce, a distribuição vai ficando cada vez mais magra, com a diminuição da variância com o crescimento da amostra). Mas, ficando mais próxima da normal, mesmo no caso da distribuição que gerou os dados não ser normal, isso garantido pelo Teorema do Limite Central, como vimos acima. É a distribuição antes do colapso final.

Tendo discutido a distribuição assintótica, vamos verificar as possíveis formas que ela pode assumir. Basicamente 3 formas englobam todas as possibilidades:

> 1º. A distribuição assintótica do estimador é conhecida e é a mesma para qualquer tamanho da amostra. Um exemplo é a média amostral como estimador da média de uma população normal. Neste caso,.a distribuição da média amostral é normal para qualquer tamanho da amostra, com média μ e variância σ^2/n.
>
> 2º . A distribuição do estimador não é da mesma forma para diferentes tamanho da amostra. Este é o caso da distribuição da proporção de sucessos. Que tem distribuição binomial, mas ela tem distribuição assintótica normal, e isso é garantido pelo TLC.
>
> 3º . Por vezes, a forma da distribuição do estimador é desconhecida para diferentes tamanhos da amostra, a menos daquela que é assintótica (com tamanho da amostra convergindo para o infinito). É o caso da média amostral de uma população não normal. Podemos desconhecer a sua distribuição para amostras finitas, mas sabemos que a distribuição assintótica é normal, garantido pelo TLC.

Essas três categorias de estimadores cobrem praticamente todos os problemas de estimação em econometria. De forma similar à distribuição para amostras finitas, as distribuições assintóticas são também caracterizadas por dois momentos: média e variância assintóticas.

A média da distribuição assintótica do estimador é o valor assumido pela média quando o tamanho da amostra converge para infinito:

Média Assintótica é dada por $\lim_{n\to\infty} E(\hat{\theta})$, desde que a média de $\hat{\theta}$ exista.

No entanto, a variância não é dada por $\lim_{n\to\infty} Var(\hat{\theta})$. Para verificar este fato, basta olhar para $\lim_{n\to\infty} Var(\bar{X})$= $\lim_{n\to\infty} (\sigma^2/n)$=0. Para essa definição de variância assintótica, temos que a variância quase sempre é nula. Mas o que queremos de fato é a variância da distribuição assintótica, que à força de hábito chamamos, equivocadamente, de "variância assintótica". De forma correta, definimos a variância da distribuição assintótica da seguinte forma:

$$VarAss\ \frac{1}{n}\lim_{n\to\infty}[\sqrt{n}\,E(\hat{\theta}-E(\hat{\theta})]^2$$

Para evitar a multiplicação do limite pelo reciproco de n, a definição mais limpa é, em geral, preferida:

$$VarAss[\sqrt{n}(\hat{\theta}-\theta)] = \lim_{n\to\infty} E[\sqrt{n}(\hat{\theta}-\theta)]^2$$

Agora podemos definir as propriedades assintóticas do estimador, sendo elas: não viés assintótico; consistência; e eficiência assintótica.

$\hat{\theta}$ é um estimador assintoticamente não viesado de θ se $\lim_{n\to\infty} E(\hat{\theta}) = \theta$.

Por exemplo o estimador da variância de $X \sim N(\mu, \sigma^2)$, então

$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$, que é a contrapartida amostral do momento populacional e é um estimador viesado de σ^2 . A variância de *X*, assintoticamente não viesada, é

$$E(\hat{\sigma}^2) = \frac{1}{n}\sum E(X_i - \bar{X})^2$$

mas, não sabemos o que é $\sum E(X_i - \bar{X})^2$, mas sabemos que

$$E(X_i - \mu)^2 = Var(X_i) = \sigma^2$$

Sabemos também que

$$E(\bar{X} - \mu)^2 = Var(\bar{X}) = \sigma^2/n,$$

Vamos tomar $\frac{1}{n}\sum E(X_i - \bar{X})^2$ e vamos adicionar e subtrair μ dentro do parênteses para termos:

$$E(\hat{\sigma}^2) = \frac{1}{n}\sum E[(X_i - \mu) - E(\bar{X} - \mu)]^2$$

$$= \frac{1}{n}\sum[E(X_i - \mu)^2 + E(\bar{X} - \mu)^2 - 2E(X_i - \mu)(\bar{X} - \mu)]$$

$$= \frac{1}{n}\sum\sigma^2 + \frac{1}{n}\sum\frac{\sigma^2}{n} - 2E(\bar{\bar{X}} - \mu)\frac{1}{n}\sum(X_i - \mu)]$$

$=$

Mas,

$$2E(\bar{\bar{X}} - \mu)\frac{1}{n}\sum(X_i - \mu)] = 2E(\bar{X} - \mu)^2 = 2\frac{\sigma^2}{n}$$

Portanto,

$$E(\hat{\sigma}^2) = \frac{1}{n}\sum\sigma^2 + \frac{1}{n}\sum\frac{\sigma^2}{n} - 2\frac{\sigma^2}{n}$$

$$= \sigma^2 + \frac{\sigma^2}{n} - 2\frac{\sigma^2}{n} = \sigma^2 - \frac{\sigma^2}{n} = \left(\frac{n-1}{n}\right)\sigma^2$$

De onde se conclui que $E(\hat{\sigma}^2) \neq \sigma^2$. Logo $\hat{\sigma}^2$ é um estimador viesado de σ^2. Neste caso, o viés assintotico é nulo (desaparece quando *n* tende para infinito).

Consistência

$\hat{\theta}$ é um estimador consistente de θ se $plim(\hat{\theta})= \theta.$

O limite de probabilidade (plim) é dado por:

$$\lim_{n\to\infty}(\ \theta - \varepsilon \leq \hat{\theta} \leq \theta + \varepsilon) = 1$$

Em que ε é um número positivo, tão pequeno quanto se queira.

A consistência é a única propriedade que se baseia na distribuição degenerada, ou seja, ela ocorre quando a distribuição de $\hat{\theta}$ colapsa sobre o verdadeiro valor do parâmetro, $\theta.$

Uma outra forma de se pensar na consistência do estimador é através do MSE$\left(\hat{\theta}\right)$,. quando *n* tende para infinito. Se $\lim_{n\to\infty} MSE(\hat{\theta}) = 0$, o $\hat{\theta}$ é um estimador consistente, pois a distribuição amostral colapsou em cima do verdadeiro valor do parâmetro com viés e variância nulos.

Eficiência Assintótica

$\hat{\theta}$ é estimador assintoticamente eficiente de θ

se as seguinte condições forem satisfeitas:

a $\hat{\theta}$ tem distribuição assintótica com média e variância finitas;
b $\hat{\theta}$ é consistente; e
c nenhum outro estimador consistente de θ tem variância assintótica menor que a variância assintótica de $\hat{\theta}$.

A eficiência assintótica só pode ser definida quando se conhece a forma da distribuição que gerou os dados. Mas, nestes casos basta comparar a variância assintótica do estimador com o limite mínimo de Cramer-Rao. Se a variância assintótica atingir o limite mínimo o estimador é assintoticamente eficiente. O que se observa, então é que eficiência implica em eficiência assintótica, mas o reverso não é verdade.