Statistics

Week 5: Optimal Estimators

Etienne Wijler

Econometrics and Data Science
Econometrics and Operations Research
Bachelor Program

VRIJE UNIVERSITEIT AMSTERDAM
SCHOOL OF BUSINESS AND ECONOMICS

Course overview: MLE and estimator evaluation

P4: Estimation

- Week 1 Probability Recap
- Week 2 Statistical Models
- Week 3 Data Reduction and MME
- Week 4 MLE and Evaluation
- Week 5 Estimator Optimality
- Week 6 Consistency

P5: Inference

- Week 7 Hypothesis testing
- Week 8 Mean and Variance testing
- Week 9 Finding test statistics
- Week 10 Evaluating tests
- Week 11 Interval estimation
- Week 12 Asymptotic tests

Optimal estimators: from uniformly better to uniformly best?

Recall, that we call $W_1(\boldsymbol{X})$ uniformly better than $W_2(\boldsymbol{X})$ when $\mathrm{MSE}(\boldsymbol{\theta}, W_1) \leq \mathrm{MSE}(\boldsymbol{\theta}, W_2)$ for all $\boldsymbol{\theta} \in \Theta$.

Question: does there exist an estimator that is uniformly better than all other estimators?

Optimal estimators: from uniformly better to uniformly best?

Recall, that we call $W_1(\boldsymbol{X})$ uniformly better than $W_2(\boldsymbol{X})$ when $\text{MSE}(\boldsymbol{\theta}, W_1) \leq \text{MSE}(\boldsymbol{\theta}, W_2)$ for all $\boldsymbol{\theta} \in \Theta$.

Question: does there exist an estimator that is uniformly better than all other estimators?

Answer: No. Consider the silly estimator $W^*(\boldsymbol{X}) = 3$. This will in general be a terrible estimator, but it's a perfect estimator when $\theta_0 = 3$!

Moreover, $\text{MSE}(3, W^*(\boldsymbol{X})) = 0$, implying that this "estimator" cannot be beaten when evaluated at $\theta = 3$.

Best unbiased estimators

Idea: what if we restrict our attention to the class of unbiased estimators.

Clearly, this excludes the silly estimator from before (why?). Does an optimal unbiased estimator exist?

Definition (7.3.7)

An estimator W^* is a uniform minimum variance unbiased estimator (UMVUE) for $\tau(\theta)$ if $\mathbb{E}_\theta\left(W^*\right) = \tau(\theta)$ and, for any other estimator W with $\mathbb{E}_\theta(W) = \tau(\theta)$, we have $\operatorname{var}_\theta W^* \leq \operatorname{var}_\theta W$ for all θ.

Recall, for unbiased estimators we have $\operatorname{MSE}(\theta, W) = \operatorname{var}_\theta(W)$.

Hence, an UMVUE is therefore an estimator that is uniformly best out of all possible unbiased estimators!

Searching for UMVUEs

Problem: finding an UMVUE is not an easy task.

Challenge: when multiple unbiased estimators exist, infinitely many new unbiased estimators can be constructed!

Example

In the $\{\text{Uniform}(0, \theta) \mid \theta > 0\}$ model, we found two unbiased estimators: $\hat{\theta}_{MM} = 2\bar{X}$ and the bias-corrected maximum likelihood estimator given by $\hat{\theta} = \frac{n+1}{n} X_{(n)}$. Clearly, $\hat{\theta}_a = a\hat{\theta}_{MM} + (1-a)\hat{\theta}$ for $a \in (0, 1)$ is also unbiased.

Stuck? We cannot check infinitely many values for a, nor can we be certain that we have included all unbiased estimators in our comparison. Should we give up? Go home? Give up econometrics all together? No!!!

A theoretical lower bound

Idea: What if we can derive a lower bound $B(\theta)$ on the variance of any unbiased estimator of $\tau(\theta_0)$? That is $\text{MSE}(\theta, W) = \mathbb{V}\text{ar}_\theta W \geq B(\theta)$ for all unbiased estimators and $\theta \in \Theta$.

Implication: An unbiased estimator W^* with $\mathbb{V}\text{ar}_\theta W^* = B(\theta)$ must inevitably be an UMVUE!

Cramér - Rao: Such a theoretical lower bound has been derived by the famous statisticians H. Cramér and C.R. Rao. We will derive this bound by ourselves today!

But first, some preliminaries...

Some useful definitions

Definition (Score and more)

- The *score* random variable is defined as

$$\mathcal{S} = \mathcal{S}(\theta, \boldsymbol{X}) = \frac{\partial}{\partial\theta}\ell(\theta \mid \boldsymbol{X}).$$

- The *Hessian* random variable is defined as

$$\mathcal{H} = \mathcal{H}(\theta, \boldsymbol{X}) = \frac{\partial^2}{\partial\theta^2}\ell(\theta \mid \boldsymbol{X}).$$

- The *Fisher information* is defined as

$$I_\theta = \mathbb{E}_\theta \left(\frac{\partial}{\partial\theta} \log f(X \mid \theta) \right)^2 = \mathbb{E}_\theta S(\theta, \boldsymbol{X})^2.$$

The Cauchy-Schwarz inequality

Lemma (Cauchy-Schwarz)

Let Y, Z be two random variables, then

$$(\mathbb{E}YZ)^2 \leq \mathbb{E}(Y^2)\mathbb{E}(Z^2).$$

Alternatively: it follows that $\mathbb{E}(YZ) \leq \sqrt{\mathbb{E}Y^2\mathbb{E}Z^2}$.

Correlation: a direct corollary of the Cauchy-Schwarz inequality is that the correlation between two random variables lies between -1 and 1. Why?

Extension: More flexible inequalities exist, such as Hölder's inequality. You will learn about these later in your econometric careers.

Proof of Cauchy-Schwarz

Proof of the Cauchy-Schwarz inequality.

Let $a, b \geq 0$ be two numbers, then

$$\mathbb{E}(aY + bZ)^2 = a^2\mathbb{E}(Y^2) + b^2\mathbb{E}(Z^2) + 2ab\mathbb{E}(YZ) \geq 0,$$
$$\mathbb{E}(aY - bZ)^2 = a^2\mathbb{E}(Y^2) + b^2\mathbb{E}(Z^2) - 2ab\mathbb{E}(YZ) \geq 0.$$

Now choose $a^2 = \mathbb{E}(Z^2)$ and $b^2 = \mathbb{E}(Y^2)$ to obtain

$$2a^2b^2 + 2ab\mathbb{E}(YZ) \geq 0,$$
$$2a^2b^2 - 2ab\mathbb{E}(YZ) \geq 0.$$

and therefore

$$-\sqrt{\mathbb{E}(Y^2)\mathbb{E}(Z^2)} = -ab = -\frac{2a^2b^2}{2ab} \leq \mathbb{E}(YZ) \leq \frac{2a^2b^2}{2ab} = ab = \sqrt{\mathbb{E}(Y^2)\mathbb{E}(Z^2)}.$$

Taking squares finishes the proof. $\square$

Cramér-Rao lower bound

Theorem (7.3.9, Cramér-Rao inequality)

Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ *be a random vector from the statistical model* $\{f(\boldsymbol{x} \mid \theta) \mid \theta \in \Theta\}$ *and let* $W(\boldsymbol{X})$ *be an unbiased estimator of the univariate* $\tau(\theta_0)$. *Suppose that* τ *is differentiable with respect to* θ *and that some regularity conditions hold, i.e. that we can swap differentiation and integration. Then*

$$\mathbb{V}ar_\theta W(\boldsymbol{X}) \geq \frac{\tau'(\theta)^2}{I_\theta}.$$

Note: in many applications $\tau(\theta) = \theta$, such that we have $\tau'(\theta) = 1$ and the Cramèr-Rao lower bound reads as $\mathbb{V}\mathrm{ar}_\theta W(\boldsymbol{X}) \geq I_\theta^{-1}$.

Proof: blackboard.

Information equality

Recall that the Fisher information number is given by $I_\theta = \mathbb{E}_\theta S(\theta, \boldsymbol{X})^2$.

Problem: The squared score often yields unwieldy expressions, making the Fisher Information number difficult to derive.

Solution: under some regularity conditions, we may work with the Hessian instead.

Lemma (Information equality)

If

$$\frac{d}{d\theta}\int_{\mathbb{R}^n}\left(\frac{\partial}{\partial\theta}\ell(\theta \mid \boldsymbol{x})\right) f(\boldsymbol{x} \mid \theta) d\boldsymbol{x} = \int_{\mathbb{R}^n}\frac{\partial}{\partial\theta}\left[\left(\frac{\partial}{\partial\theta}\ell(\theta \mid \boldsymbol{x})\right) f(\boldsymbol{x} \mid \theta)\right] d\boldsymbol{x},$$

(which is true for an exponential family), then

$$\mathbb{E}_\theta \mathcal{S}(\theta, \boldsymbol{X})^2 = -\mathbb{E}_\theta \mathcal{H}(\theta, \boldsymbol{X})$$

Proof of information equality

Proof.

Recall that $\mathbb{E}_\theta(\mathcal{S}(\theta, \boldsymbol{X})) = 0$. Then,

$$\begin{aligned}
0 &= \frac{d}{d\theta}\mathbb{E}_\theta(\mathcal{S}(\theta, \boldsymbol{X})) = \frac{d}{d\theta}\int_{\mathbb{R}^n}\left(\frac{\partial}{\partial\theta}\ell(\theta \mid \boldsymbol{x})\right) f(\boldsymbol{x} \mid \theta)d\boldsymbol{x} \\
&\overset{(*)}{=} \int_{\mathbb{R}^n}\frac{\partial}{\partial\theta}\left(\frac{\partial}{\partial\theta}\ell(\theta \mid \boldsymbol{x})\right) f(\boldsymbol{x} \mid \theta)d\boldsymbol{x} \\
&\overset{(**)}{=} \int_{\mathbb{R}^n}\left(\frac{\partial^2}{\partial\theta^2}\ell(\theta \mid \boldsymbol{x})\right) f(\boldsymbol{x} \mid \theta) + \left(\frac{\partial}{\partial\theta}\ell(\theta \mid \boldsymbol{x})\right)\frac{\partial}{\partial\theta}f(\boldsymbol{x} \mid \theta)d\boldsymbol{x} \\
&\overset{(***)}{=} \int_{\mathbb{R}^n}\left(\frac{\partial^2}{\partial\theta^2}\ell(\theta \mid \boldsymbol{x})\right) f(\boldsymbol{x} \mid \theta) + \left(\frac{\partial}{\partial\theta}\ell(\theta \mid \boldsymbol{x})\right)^2 f(\boldsymbol{x} \mid \theta)d\boldsymbol{x} \\
&= \mathbb{E}_\theta(\mathcal{H}(\theta, \boldsymbol{X})) + \mathbb{E}_\theta(\mathcal{S}(\theta, \boldsymbol{X})^2).
\end{aligned}$$

$\square$

Cramér-Rao lower bound: example

Example

Let $X_1, \ldots, X_n$ be a random sample from an Exponential(θ) distribution with pdf given by $f(x \mid \theta) = \frac{1}{\theta} e^{-x/\theta}$. Recall that the MLE of θ is $\hat{\theta} = \bar{X}$. Derive the Cramér - Rao lower bound and the variance of $\hat{\theta}$. What does this tell us about the MLE?

Hint: Think about which simplifications we are able to use.

Note: an alternative parameterization of the exponential distribution is $f(x \mid \lambda) = \lambda e^{-\lambda x}$. Do not confuse these two cases.

Fisher information for iid random variables

Simplification: As usual, we are able to simplify our derivations further by exploiting the assumption of iid random variables.

Definition

We define the Fisher information in the iid case for an individual observation as

$$i_\theta = \mathbb{E}_\theta \left(\frac{\partial}{\partial \theta} \log g(X_1 \mid \theta) \right)^2 .$$

Under some regularity conditions, we may equivalently define the individual information number as

$$i_\theta = -\mathbb{E}_\theta \left(\frac{\partial^2}{\partial \theta^2} \log g(X_1 \mid \theta) \right) .$$

Cramér-Rao for iid random variables

Corollary (7.3.10)

Suppose the assumptions of the Cramér-Rao lower bound theorem hold and that additionally $\boldsymbol{X} = (X_1, \dots, X_n)$ *is a vector of iid random variables, each with pdf* $g(x \mid \theta)$. *Then*

$$\mathbb{V}ar_\theta W(\boldsymbol{X}) \geq \frac{\tau'(\theta)^2}{n i_\theta}.$$

Example (7.3.12)

Consider the statistical model $\{\text{Poisson}(\lambda) \mid \lambda > 0\}$. Let $X_1, \dots, X_n$ denote a random sample with a population pdf in this model. Show that the moment and ML estimator $\hat{\lambda} = \bar{X}$ is an UMVUE.

The return of sufficiency!

Problem: We now know how to verify whether an estimator is UMVUE, but we still need to find the estimator first. How to do this?

Recall: a sufficient statistic incorporates all relevant information about the parameter of interest.

Implication: estimators that do not depend on sufficient statistics miss out on some of this information!

Idea: estimators that do not depend on a sufficient statistic can be improved upon by "making them a function" of a sufficient statistic.

Question: How?

The power of averaging

Suppose that we have an unbiased estimator $W(\boldsymbol{X})$ of $\tau(\theta_0)$ and a sufficient statistic $T(\boldsymbol{X})$.

Fact: If $T(\boldsymbol{x}) = T(\boldsymbol{y})$, we should obtain the same conclusion about $\tau(\theta_0)$.

However, what if $W(\boldsymbol{x}) \neq W(\boldsymbol{y})$? The difference must come from noise in the data.

Note: For any collection of iid random variables $X_1, \dots, X_n$, we have that $\mathbb{E}\bar{X} = \mathbb{E}X_1$ and $\mathbb{V}\mathrm{ar}\bar{X} = \frac{\mathbb{V}\mathrm{ar}X_1}{n}$. Hence, averaging reduces the variance.

Idea: take an average over $\{W(\boldsymbol{x}) \mid \boldsymbol{x} \text{ s.t. } T(\boldsymbol{x}) = t\}$ to reduce the variance!

Rao-Blackwell

Better idea: Perhaps we should not give equal weight to all outcomes $\boldsymbol{x}$, but rather weigh them by their "likeliness".

Discrete case:

$$\phi(t) = \sum_{\boldsymbol{x}\colon T(\boldsymbol{x})=t} W(\boldsymbol{x}) P(\boldsymbol{X} = \boldsymbol{x} \mid T = t).$$

General case: what we have just defined is actually equal to $\phi(T) = \mathbb{E}(W \mid T)$.

Theorem (7.3.17, Rao-Blackwell)

Let W be an unbiased estimator of $\tau(\theta_0)$ and let T be a sufficient statistic for θ_0. Then $\phi(T) = \mathbb{E}(W \mid T)$ is also an unbiased estimator of $\tau(\theta_0)$ and $\mathbb{V}ar_\theta \phi(T) \leq \mathbb{V}ar_\theta W$. That is, $\phi(T)$ is uniformly better than W.

Proof of Rao-Blackwell

Proof.

The statistic T has to be sufficient to ensure that $\phi(T)$ is well defined and does not depend on θ_0. Recall results 4.4.3 and 4.4.7 in C&B:

$$\mathbb{E}_\theta W = \mathbb{E}_\theta(\mathbb{E}(W \mid T)) \qquad \text{and} \qquad \mathbb{V}\text{ar}_\theta W = \mathbb{E}_\theta(\mathbb{V}\text{ar}(W \mid T)) + \mathbb{V}\text{ar}_\theta(\mathbb{E}(W \mid T)).$$

From the first equality we get that $\mathbb{E}_\theta(\mathbb{E}(W \mid T)) = \tau(\theta)$. From the second equality we obtain

$$\mathbb{V}\text{ar}_\theta(\mathbb{E}(W \mid T)) = \mathbb{V}\text{ar}_\theta W - \mathbb{E}_\theta(\mathbb{V}\text{ar}(W \mid T)) \leq \mathbb{V}\text{ar}_\theta W.$$

$\square$

The importance of sufficiency in Rao-Blackwell

Example (7.3.18)

Suppose we have the statistical model $\{\text{Normal}(\mu, 1) \mid \mu \in \mathbb{R}\}$, let $W(\boldsymbol{X}) = \overline{X}$ and $T(\boldsymbol{X}) = X_1$. Clearly W is unbiased and therefore $\phi(T) = \mathbb{E}(\overline{X} \mid X_1)$ is also unbiased and has smaller variance than $W(\boldsymbol{X})$. However, T is not a sufficient statistics. Derive an expression of $\phi(T)$ to understand what is going wrong here.

Example

We have the statistical model $\{\text{Poisson}(\lambda) \mid \lambda > 0\}$ and are interested in estimating λ_0 by using the poor estimator $W(\boldsymbol{X}) = X_1$. The set of Poisson distributions is an exponential family, so it is straightforward to show that $T(\boldsymbol{X}) = \sum_{i=1}^{n} X_i$ is a sufficient statistic. Use "Rao-Blackwellization" to improve upon this silly estimator. How good is our improvement?

Extra UMVUE example (if time permits)

Example

You are testing the strength of iron bars used in construction by applying a downward force in kiloton (kt) on one end of the bar. Let X_i, $i = 1, \ldots, n$, denote the amount of kt it takes to bend the i-th bar more than 1 cm. Consider the statistical model $\{g(x \mid \beta) \mid \beta > 0\}$ with

$$g(x \mid \beta) = \frac{4}{\beta} x^3 e^{-x^4/\beta}, \qquad x > 0,\ \beta > 0$$

Derive the MLE of β. Is this an UMVUE?

Hint: You may use that $\mathbb{E}X^n = \beta^{n/4}\Gamma(1 + n/4)$.