Advanced Econometrics I

Jürgen Meinecke

Lecture 10 of 12

Research School of Economics, Australian National University

Roadmap

- Maximum Likelihood (ML) Estimation
  - Maximum Likelihood Estimator
  - Invariance Property
  - Fisher Information, Cramér Rao Bound, Information Equality
  - Minimum Variance Unbiased Estimators
  - Asymptotic Efficiency of ML

Let's take a step back into the univariate world

You have a random sample $Y_1, \ldots, Y_N$ that is generated from either a

- probability mass function (discrete) or
- probability density function (continuous)

For simplicity I will only write pdf (and omit pmf), and denote it by $f(y|\theta)$

Crucial here

- f is known
- θ is unknown and we want to estimate it

Example: f is the pdf of a normal distribution with variance 1 and we are after the expected value θ

This foreshadows one of the biggest drawbacks of ML estimation: you need to know f

Definition (Likelihood Functions)

Given a random sample $Y_1, \dots, Y_N$ where each Y_i has pdf $f(y|\theta)$

- **likelihood function**
  $L(\theta) := f_{Y_1,\dots,Y_N}(y_1, \cdots, y_N|\theta) = \prod_{i=1}^{N} f(y_i|\theta)$
- **log likelihood function**
  $\ln L(\theta) := \ln(L(\theta)) = \sum_{i=1}^{N} \ln f(y_i|\theta)$

We view both as function in the parameter θ; given a random sample the N different values y_i are known

Most generally, the likelihood function is the joint pdf, but here it reduces to the product of pdf's because of iid

The log likelihood function therefore is a sum

Definition (Maximum Likelihood (ML) Estimator)

Given a random sample $Y_1, \ldots, Y_N$ where each Y_i has pdf $f(y|\theta)$ the **maximum likelihood estimator** of θ is defined by

$$\hat{\theta}^{\mathsf{ML}} := \operatorname{argmax}_\theta L(\theta).$$

Corollary

$\operatorname{argmax}_\theta L(\theta) = \operatorname{argmax}_\theta \ln L(\theta)$.

Because the log is strictly monotone

It is usually easier to maximize $\ln L$ analytically and numerically

Example: exponential distribution

Example: Y_i is exponentially distributed

Recall, pdf of Y_i is $f(y|\mu) = (1/\mu)\exp(-y/\mu)$

$1/\mu$ is called the arrival rate

You may remember that $\mathrm{E}Y_i = \mu$ and $\mathrm{Var}\ Y_i = \mu^2$

For a sample of size $N = 1$, the likelihood functions are

$$L(\mu) = (1/\mu)\exp(-y_1/\mu)$$
$$\ln L(\mu) = -(\ln \mu + y_1/\mu)$$

The ML estimator is $\hat{\mu}^{\mathrm{ML}} := \operatorname{argmax}_\mu -(\ln \mu + y_1/\mu)$

First derivative is $-\left((1/\mu) - y_1/\mu^2\right)$

Setting equal to zero results in $\hat{\mu}^{\mathrm{ML}} = y_1$

(The second order condition could also be checked to confirm a local maximum)

Example: exponential distribution (continued)

What if your sample size is $N > 1$?

The likelihood functions are

$$L(\mu) = \mu^{-N} \exp(-\textstyle\sum_{i=1}^{N} y_i/\mu)$$

$$\ln L(\mu) = -N \ln \mu - \textstyle\sum_{i=1}^{N} y_i/\mu = -N \left(\ln \mu + \bar{Y}/\mu\right)$$

The ML estimator is $\hat{\mu}^{\mathsf{ML}} := \operatorname{argmax}_{\mu} -N \left(\ln \mu + \bar{Y}/\mu\right)$

First derivative is $-N \left((1/\mu) - \bar{Y}/\mu^2\right)$

Setting equal to zero results in $\hat{\mu}^{\mathsf{ML}} = \bar{Y}$

This may convince you that ML estimation results in sensible estimators

But why should it be a sensible estimation approach?

Consider a random variable Y generated according to $f(y|\theta)$

Take any $\tilde{\theta} \neq \theta$ and look at the likelihood ratio $LR = f(Y|\tilde{\theta})/f(Y|\theta)$

Here θ is the *true* parameter and $\tilde{\theta}$ is some other value

By Jensen's inequality $\mathrm{E}(-\ln LR) > -\ln \mathrm{E}(LR)$
(because $-\ln(\cdot)$ is strictly convex)

Substituting back

$$\mathrm{E}\left(-\ln \frac{f(Y|\tilde{\theta})}{f(Y|\theta)}\right) > -\ln \mathrm{E}\left(\frac{f(Y|\tilde{\theta})}{f(Y|\theta)}\right)$$

Notice the expectations are taken over the *true* pdf $f(y|\theta)$
(and not $f(y|\tilde{\theta})$)

The expectation on the rhs simplifies

$$\mathrm{E}\left(\frac{f(Y|\tilde{\theta})}{f(Y|\theta)}\right) = \int \frac{f(y|\tilde{\theta})}{f(y|\theta)} f(y|\theta)dy = \int f(y|\tilde{\theta})dy = 1$$

Therefore

$$\mathrm{E}\left(-\ln\frac{f(Y|\tilde{\theta})}{f(Y|\theta)}\right) > -\ln \mathrm{E}\left(\frac{f(Y|\tilde{\theta})}{f(Y|\theta)}\right) = -\ln 1 = 0$$

Implying

$$\mathrm{E}\left(\ln f(Y|\theta)\right) > \mathrm{E}\left(\ln f(Y|\tilde{\theta})\right)$$

This means that the expected value of the log likelihood is maximized at the *true* value of the parameter

So there's some hope that the sample analog will mimic this feature

Example: exponential distribution (continued)

Writing the log likelihood using the generic placeholder $\tilde{\mu}$

$$\ln L(\tilde{\mu}) = -N\left(\ln \tilde{\mu} + \bar{Y}/\tilde{\mu}\right)$$

Consider its expected value

$$\mathrm{E}\left(\ln L(\tilde{\mu})\right) = -N\left(\ln \tilde{\mu} + \mu/\tilde{\mu}\right)$$

By the analogy principle, this is maximized at $\tilde{\mu} = \mu$

Roadmap

- Maximum Likelihood (ML) Estimation
  - Maximum Likelihood Estimator
  - Invariance Property
  - Fisher Information, Cramér Rao Bound, Information Equality
  - Minimum Variance Unbiased Estimators
  - Asymptotic Efficiency of ML

Here's a simple looking result that is useful

Proposition (Invariance Property of ML estimators)

If $\hat{\theta}^{ML}$ is the ML estimator of θ, then for any function g, the ML estimator of $g(\theta)$ is $g(\hat{\theta}^{ML})$.

Note: I didn't specify the domain and range of g because that arises from the context of θ

Also, g doesn't have to be one-to-one

It's easiest to explain by example...

Example: exponential distribution (continued)

Let $Y_1, \ldots, Y_N$ be iid with pdf $f(y|\mu) = (1/\mu)\exp(-y/\mu)$

Let's take the case $N > 1$

We saw earlier that $\hat{\mu}^{\mathsf{ML}} = \bar{Y}$

Now, define $\lambda := 1/\mu$, which is called the *arrival rate*

Rewrite the pdf: $f(y|\lambda) = \lambda \exp(-\lambda y)$

How would we estimate λ using ML?

Easy, the invariance property suggests: $\hat{\lambda}^{\mathsf{ML}} = 1/\bar{Y}$

Roadmap

- Maximum Likelihood (ML) Estimation
  - Maximum Likelihood Estimator
  - Invariance Property
  - Fisher Information, Cramér Rao Bound, Information Equality
  - Minimum Variance Unbiased Estimators
  - Asymptotic Efficiency of ML

Definition (Score Function)

Let Y be a random variable with pdf $f(y|\theta)$. The **score function** is defined by

$$S(y|\theta) := \frac{\partial \ln f}{\partial \theta}(y|\theta)$$

The score plays an important role in ML theory

Obviously, by definition $(1/N)\sum_{i=1}^{N} S(y_i|\hat{\theta}^{\text{ML}}) = 0$
(necessary condition for a maximum)

This is interesting bc it is true that $\mathrm{E}\left(S(Y_i|\theta)\right) = 0$ (see below)

The ML estimator therefore can be motivated as an analog estimator: it is the value for θ that makes the average score equal zero

Example: exponential distribution (continued)

Going back to the example of the exponential distribution

$S(y|\mu) := -\left((1/\mu) - y/\mu^2\right)$

In the case $N = 1$ we get $S(y_1|\hat{\theta}^{\mathrm{ML}}) = 0$

In the case $N > 1$ we get $(1/N)\sum_{i=1}^{N} S(y_i|\hat{\theta}^{\mathrm{ML}}) = 0$

Showing that the expected value of the score is zero:

Proof.

For all θ,

$1 = \int f(y|\theta)dy$ and therefore $0 = \frac{\partial}{\partial\theta}\int f(y|\theta)dy$

Swapping differentiation and integration, we get

$$\begin{aligned} 0 &= \int \frac{\partial}{\partial\theta} f(y|\theta)dy \\ &= \int \frac{\partial \ln f}{\partial\theta}(y|\theta) \cdot f(y|\theta)dy \\ &= \int S(y|\theta) \cdot f(y|\theta)dy \\ &= \mathrm{E}\left(S(Y|\theta)\right) \end{aligned}$$

□

How can we use this result?

Definition (Fisher Information)

Let Y be a random variable with pdf $f(y|\theta)$. The **Fisher Information** is defined as $I(\theta) := \mathrm{E}\left(S(Y|\theta)^2\right)$.

Notice that the Fisher Information is also equal to the variance of the score bc Var $S(Y|\theta) = \mathrm{E}\left(S(Y|\theta)^2\right) - \mathrm{E}\left(S(Y|\theta)\right)^2 = \mathrm{E}\left(S(Y|\theta)^2\right)$

Here an important result for unbiased estimators $T(Y_1, \dots, Y_N)$

Proposition (Cramér Rao Bound (CRB))

Let $Y_1, \dots, Y_N$ be iid with pdf $f(y|\theta)$ and let $T(Y_1, \dots, Y_N)$ be an unbiased estimator for θ. Then

$$\mathrm{Var}\, T(Y_1, \dots, Y_N) \geq \frac{1}{N \cdot I(\theta)}$$

Proof.

Let's only check the case $N = 1$. First notice that

$$\begin{aligned}
\text{Cov}\left(S(Y|\theta), T(Y)\right) &= \mathsf{E}(S(Y|\theta) \cdot T(Y)) \\
&= \int S(y|\theta) \cdot T(y) \cdot f(y|\theta) dy \\
&= \int T(y) \frac{\partial \ln f}{\partial \theta}(y|\theta) f(y|\theta) dy \\
&= \int T(y) \cdot \frac{\partial f}{\partial \theta}(y|\theta) dy \\
&= \frac{\partial}{\partial \theta} \int T(y) \cdot f(y|\theta) dy \\
&= \frac{\partial}{\partial \theta} \mathsf{E}\left(T(Y)\right) = \frac{\partial}{\partial \theta} \theta = 1
\end{aligned}$$

Use Cauchy Schwarz inequality: $\text{Cov}(A, B)^2 \leq \text{Var}\, A \cdot \text{Var}\, B$

$$\text{Var}\, T(Y) \geq \frac{\text{Cov}\left(S(Y|\theta), T(Y)\right)^2}{\text{Var}\, S(Y|\theta)} = \frac{1}{\text{Var}\, S(Y|\theta)} = \frac{1}{I(\theta)}$$

□

What's the point of the CRB?

It gives us a lower bound for the variance of any unbiased estimator (not only ML estimators)

Aside:
Finding a lower bound for the variance is, in principle, not hard at all: zero is always a lower bound

The nice thing about the CRB, as we will see, is that it can actually be attained by some estimators

And ML estimators have a special relationship with the CRB as we will see

Example: exponential distribution (continued)

Again going back to the example of the exponential distribution

The score was $S(y|\mu) := -\left((1/\mu) - y/\mu^2\right)$

Notice that $\mathrm{E}\left(S(Y|\mu)\right) = -1/\mu + \mathrm{E}Y/\mu^2 = -1/\mu + \mu/\mu^2 = 0$

(This is a quick cross check that we haven't made any mistakes)

Therefore $\mathrm{E}\left(S(Y|\mu)^2\right) = \mathrm{Var}\, S(Y|\mu) = VY/\mu^4 = \mu^{-2}$

Let $T(Y_1, \ldots, Y_N)$ be an unbiased estimator for μ

It follows that $\mathrm{Var}\, T(Y_1, \ldots, Y_N) \geq \mu^2/N$

The CRB for the variance of $T(Y_1, \ldots, Y_N)$ is μ^2/N

Another important result in ML theory

Proposition (Information Equality)

Let Y be a random variable with pdf $f(y|\theta)$. Then

$$I(\theta) = -E\left(\frac{\partial S}{\partial \theta}(Y, \theta)\right).$$

This gives us 3 ways to obtain the Fisher information; via the:

- variance of the score
- second moment of the score
- first derivative of the score
  (which is the second derivative of the log likelihood)

Writing the information equality in terms of the original pdf

$$\mathrm{E}\left(\left(\frac{\partial \ln f}{\partial \theta}(Y|\theta)\right)^2\right) = -\mathrm{E}\left(\frac{\partial^2 \ln f}{\partial \theta^2}(Y|\theta)\right)$$

Proof.

When we derived the expected value of the score, we obtained

$$\int \frac{\partial \ln f}{\partial \theta}(y|\theta) \cdot f(y|\theta) dy = 0$$

Differentiating both sides

$$\begin{aligned}
&\int \frac{\partial^2 \ln f}{\partial \theta^2}(y|\theta) \cdot f(y|\theta) dy + \int \frac{\partial \ln f}{\partial \theta}(y|\theta) \cdot \frac{\partial f}{\partial \theta}(y|\theta) dy \\
&= \mathrm{E}\left(\frac{\partial^2 \ln f}{\partial \theta^2}(Y|\theta) \right) + \int \frac{\partial \ln f}{\partial \theta}(y|\theta) \frac{\partial \ln f}{\partial \theta}(y|\theta) f(y|\theta) dy \\
&= \mathrm{E}\left(\frac{\partial^2 \ln f}{\partial \theta^2}(Y|\theta) \right) + \int \left(\frac{\partial \ln f}{\partial \theta}(y|\theta) \right)^2 \cdot f(y|\theta) dy \\
&= \mathrm{E}\left(\frac{\partial^2 \ln f}{\partial \theta^2}(Y|\theta) \right) + \mathrm{E}\left(\left(\frac{\partial \ln f}{\partial \theta}(Y|\theta) \right)^2 \right) = 0
\end{aligned}$$

□

Example: exponential distribution (continued)

Yet again using the example of the exponential distribution

The score was $S(y|\mu) := -\left((1/\mu) - y/\mu^2\right)$

We saw earlier that $\mathrm{E}\left(S(Y|\mu)^2\right) = \mu^{-2}$

Let's look at the derivative of the score

$$\frac{\partial S}{\partial \mu}(y|\mu) = -(-1/\mu^2 + 2y/\mu^3)$$

Recalling $\mathrm{E}Y = \mu$, you see quickly that $\mathrm{E}\left(\frac{\partial S}{\partial \mu}(y|\mu)\right) = -\mu^{-2}$

This confirms the information equality

Example: exponential distribution (continued)

Another thing that's interesting here is that we do know the actual variance of $\hat{\mu}^{\mathrm{ML}}$ because it is easy to derive

Recall that $\hat{\mu}^{\mathrm{ML}} = \bar{Y}$
(the case in which $N > 1$)

Therefore Var $\hat{\mu}^{\mathrm{ML}} =$ Var $\bar{Y} = \mu^2/N$

So when Y_i has an exponential distribution with parameter μ then the ML estimator $\hat{\mu}^{\mathrm{ML}}$ has a variance that *attains* the CRB

We've also seen that the ML estimator is unbiased

So $\hat{\mu}^{\mathrm{ML}}$ is the unbiased estimator with minimum variance

Is this true for all maximum likelihood estimators?

Roadmap

- Maximum Likelihood (ML) Estimation
  - Maximum Likelihood Estimator
  - Invariance Property
  - Fisher Information, Cramér Rao Bound, Information Equality
  - Minimum Variance Unbiased Estimators
  - Asymptotic Efficiency of ML

What have we discovered so far?

The CRB gives a lower bound for the variance of unbiased estimators

Let's say you have obtained an unbiased estimator

You calculate its variance and compare that to the CRB

The CRB is most useful when we can show that our estimator actually attains it

If its variance is equal to the CRB then we know it must be the minimum variance unbiased estimator
(as was the case for $\hat{\mu}^{\mathrm{ML}}$ in exponential distribution example)

If its variance is not equal to the CRB then there are two possibilities: the estimator

- is not a minimum variance unbiased estimator
  (there is another unbiased estimator with smaller variance)
- is a minimum variance unbiased estimator with a variance greater than the CRB (the bound cannot be attained)

Under what conditions does a minimum variance unbiased estimator with variance equal to the CRB exist?

And how do we find such an estimator?

This is where ML estimation is so powerful:

Theorem

Let $Y_1, \dots, Y_N$ be iid with pdf $f(y|\theta)$. An unbiased estimator $T(Y_1, \dots, Y_N)$ with variance equal to the Cramér Rao bound exists if and only if the score function can be written

$$\frac{1}{N} \sum_{i=1}^{N} S(y_i|\theta) = a(\theta) \cdot (T(Y_1, \dots, Y_N) - \theta),$$

for some function $a(\theta)$.

The minimum variance unbiased estimator is then equal to the maximum likelihood estimator. Its variance is $1/a(\theta)$.

Proof.

Let's only check the case $N = 1$.

(only if)

Suppose $T(Y)$ is an unbiased estimator with variance equal to the CRB. Then it must be the case that

$$\operatorname{Var} T(Y) = \frac{1}{I(\theta)} = \frac{1}{\operatorname{Var} S(Y|\theta)} = \frac{\operatorname{Cov}\left(S(Y|\theta), T(Y)\right)^2}{\operatorname{Var} S(Y|\theta)}$$

which implies $|\operatorname{Corr}\left(T(Y), S(Y|\theta)\right)| = 1$.

It follows that the score is a linear function of $T(Y)$, with coefficients possibly depending on θ:

$$S(Y|\theta) = b(\theta) + a(\theta) \cdot T(Y)$$

Since $T(Y)$ is unbiased, and because $\mathrm{E}(S(Y|\theta)) = 0$, it follows that $b(\theta) = -a(\theta) \cdot \theta$, therefore

$$S(Y|\theta) = a(\theta) \cdot (T(Y) - \theta)$$

□

Proof.

(if)

Suppose the score can be written $S(Y|\theta) = a(\theta) \cdot (T(Y) - \theta)$.

Because $\mathrm{E}S(Y|\theta) = 0$, $\mathrm{E}T(Y) = \theta$, so it is unbiased. Furthermore

$$\mathrm{Var}\ S(Y|\theta) = a(\theta)^2 \mathrm{Var}\ T(Y)$$

which implies

$$\mathrm{Var}\ T(Y) = a(\theta)^{-2} \mathrm{Var}\ S(Y|\theta) = a(\theta)^{-2} \cdot I(\theta)$$

At the same time

$$\mathrm{E}\left(\frac{\partial S}{\partial \theta}(Y|\theta)\right) = \mathrm{E}\left(\frac{\partial a}{\partial \theta}(\theta)\,(T(Y) - \theta) - a(\theta)\right) = -a(\theta)$$

So $I(\theta) = a(\theta)$ by the information equality, and therefore

$$\mathrm{Var}\ T(Y) = a(\theta)^{-1} = I(\theta)^{-1}.$$

Lastly, that estimator would be found by $S(y|\hat{\theta}) = 0$ which is the ML estimator via the analogy principle $\square$

Example: exponential distribution (continued)

We already know that $\hat{\mu}^{\mathrm{ML}}$ attains the CRB

Let's show that the score can be written like in the theorem

Also let's understand the role of $a(\theta)$ in that theorem

We know from earlier that

$$\begin{aligned} S(y|\mu) &= -\left((1/\mu) - y/\mu^2\right) \\ &= \frac{y-\mu}{\mu^2} \end{aligned}$$

Therefore

$$(1/N)\sum_{i=1}^{N} S(y_i|\mu) = \frac{\bar{Y}-\mu}{\mu^2} = \frac{\hat{\mu}^{\mathrm{ML}}-\mu}{\mu^2} = \mu^{-2}\left(\hat{\mu}^{\mathrm{ML}}-\mu\right),$$

implying $a(\mu) = \mu^{-2}$

So $a(\mu)$ is the Fisher Information!

Roadmap

Maximum Likelihood (ML) Estimation

- Maximum Likelihood Estimator
- Invariance Property
- Fisher Information, Cramér Rao Bound, Information Equality
- Minimum Variance Unbiased Estimators
- Asymptotic Efficiency of ML

Example: exponential distribution (continued)

The example of the exponential distribution again is instructive

We have $Y_1, \ldots, Y_N$ iid with pdf $f(y|\mu) = (1/\mu)\exp(-y/\mu)$

Now, define $\lambda := 1/\mu$ (recall the *arrival rate*)

Rewrite the pdf: $f(y|\lambda) = \lambda \exp(-\lambda y)$

How would we estimate λ using ML?

Easy, by the invariance property: $\hat{\lambda}^{\mathrm{ML}} = 1/\bar{Y}$

Interestingly, $\hat{\lambda}^{\mathrm{ML}}$ is biased

It can be shown that $\mathrm{E}\hat{\lambda}^{\mathrm{ML}} = \frac{N}{N-1}\lambda \neq \lambda$

So $\hat{\lambda}^{\mathrm{ML}}$ cannot be a minimum variance unbiased estimator

Luckily, we have the following result

Theorem (Asymptotic Normality of ML Estimators)

Let $Y_1, \ldots, Y_N$ be iid with pdf $f(y|\theta)$ Then

$$\sqrt{N}\left(\hat{\theta}^{ML} - \theta\right) \xrightarrow{d} N\left(0, I(\theta)^{-1}\right)$$

Notice that $I(\theta)^{-1}$ is the CRB

Loosely, we take this to mean that $\hat{\theta}^{\text{ML}} \overset{approx.}{\sim} \text{N}\left(\theta, \frac{1}{N \cdot I(\theta)}\right)$

Corollary (Consistency of ML Estimators)

$\hat{\theta}^{ML} = \theta + o_p(1)$

That is, maximum likelihood estimators are consistent and asymptotically normal distributed and attain the CRB

Asymptotically, ML estimators are minimum variance unbiased estimators

Sketch of proof is instructive

Start with the log likelihood $\ln L(\theta) = \sum_{i=1}^{N} \ln f(y_i|\theta)$

By definition, the derivative of $\ln L$ evaluated at $\hat{\theta}^{\mathrm{ML}}$ is zero

$$\frac{\partial \ln L}{\partial \theta}(\hat{\theta}^{\mathrm{ML}}) = 0$$

Apply the mean value theorem around θ (the "true" parameter)

$$0 = \frac{\partial \ln L}{\partial \theta}(\hat{\theta}^{\mathrm{ML}}) = \frac{\partial \ln L}{\partial \theta}(\theta) + \frac{\partial^2 \ln L}{\partial \theta^2}(\tilde{\theta}) \cdot (\hat{\theta}^{\mathrm{ML}} - \theta),$$

for some $\tilde{\theta}$ between θ and $\hat{\theta}^{\mathrm{ML}}$

Multiplying by $\sqrt{N}$ and rearranging results in

$$\begin{aligned}
\sqrt{N}\left(\hat{\theta}^{\mathrm{ML}} - \theta\right) &= \left(-\frac{1}{N}\frac{\partial^2 \ln L}{\partial \theta^2}(\tilde{\theta})\right)^{-1} \cdot \left(\frac{1}{\sqrt{N}}\frac{\partial \ln L}{\partial \theta}(\theta)\right) \\
&\xrightarrow{\mathrm{d}} \left(I(\theta)^{-1} + \mathrm{o}_p(1)\right) \cdot \mathrm{N}\left(0, I(\theta)\right) \\
&= \mathrm{N}\left(0, I(\theta)^{-1}\right)
\end{aligned}$$

Example: exponential distribution (continued)

For $\hat{\lambda}^{\mathrm{ML}} = 1/\bar{Y}$ we have $\hat{\lambda}^{\mathrm{ML}} = \lambda + \mathrm{o}_p(1)$

Asymptotically, the bias goes away: $\mathrm{E}\hat{\lambda}^{\mathrm{ML}} = \frac{N}{N-1}\lambda \to \lambda$

$V\hat{\lambda}^{\mathrm{ML}} \simeq \lambda^2/N$

Life is good