ApEc 8213: Econometric Analysis III -- Lecture #1

Maximum Likelihood Estimation
Hansen, *Probability and Statistics for Economists* Chap. 10

Maximum likelihood estimation (MLE) is a commonly used estimation method. For linear models, it is rare to use MLE because efficient estimators can be obtained without assuming that the error term follows a particular distribution. This is harder to show for nonlinear models, but **MLE is efficient if the distributional assumptions** (e.g. that the error term follows a specific statistical distribution) **are correct**. If the distributional assumptions are **not correct**, MLE could lead to **inconsistent estimates**.

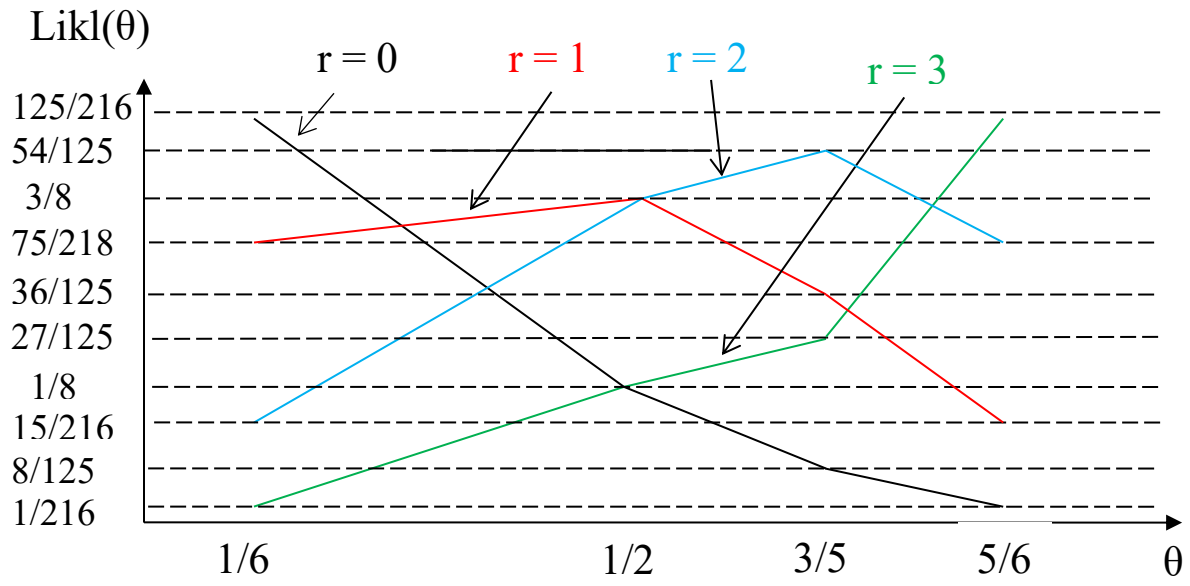
I. Intuition for Maximum Likelihood Estimation

So far, the methods you have seen in class estimate the parameters by minimizing the sum of the squared error terms. **MLE finds the parameters that are most likely to produce the actual data you have**. Here is a simple example to provide the intuition. You have a "population" of red balls and white balls. Let θ be the proportion of red balls; θ has only 4 possible values, 1/6, ½, 3/5 and 5/6.

You have a sample of 3 balls. Let r = number of red balls. The relationship between θ and your data (r) is:

	r = 0	r = 1	r = 2	r = 3
θ = 1/6	$(5/6)^3 = 125/216$	$(1/6)(5/6)^2\times3 = 75/218$	$(1/6)^2(5/6)\times3 = 15/216$	$(1/6)^3 = 1/216$
θ = ½	$(1/2)^3 = 1/8$	$(1/2)(1/2)^2\times3 = 3/8$	$(1/2)^2(1/2)\times3 = 3/8$	$(1/2)^3 = 1/8$
θ = 3/5	$(2/5)^3 = 8/125$	$(2/5)^2(3/5)\times3 = 36/125$	$(2/5)(3/5)^2\times3 = 54/125$	$(3/5)^3 = 27/125$
θ = 5/6	$(1/6)^3 = 1/216$	$(1/6)^2(5/6)\times3 = 15/216$	$(1/6)(5/6)^2\times3 = 75/218$	$(5/6)^3 = 125/216$

This gives the following **likelihoods** as **functions of θ**:



If your sample (data) contains 0 red balls (r = 0) then the θ mostly likely to generate it is 1/6, so that is your MLE estimate of θ. If your sample has 1 red ball, the θ most likely to generate it is 1/2, so that is your estimate of θ.

Questions: What is your estimate of θ if r = 2? If r = 3?

II. Parametric Models (10.2)

Consider a variable X. A **parametric model** of X is a **complete probability function** that includes an **unknown parameter vector θ.**

When X is a **discrete** variable (e.g. a binary variable such as being employed) we can write this as the **probabilities for each value that X takes**: $\boldsymbol{\pi(x|\,\theta)}$. When X is a **continuous** variable, we can write this as a **density function $\boldsymbol{f(x|\,\theta)}$**. In **both cases**, the **possible values for θ**, which can be a vector, **belong to a set Θ**, which is called the **parameter space**.

The **parametric model** for X can produce **different distributions** for X by using **different values of θ**. These different possible distributions are called a **parametric family**.

Example 1: $X \sim \mathrm{N}(\mu, \sigma^2)$, thus $f(x|\,\mu, \sigma^2) = \dfrac{\phi(\frac{x-\mu}{\sigma})}{\sigma}$

Example 2: X has the exponential distribution: $f(x|\,\lambda) = e^{-x/\lambda}/\lambda$

MLE assumes some parametric family, and then estimates the associated parameters (θ). **If the parametric family assumption is correct, then the distribution of X in the population is "fully described"**. More formally, we can define a parametric model as follows:

Definition 10.1. A **model** for a random sample is the assumption that X_i, $i = 1, 2, \ldots n$, are i.i.d. and are drawn from a known density function $f(x|\theta)$ or mass function $\pi(x|\theta)$ with unknown parameter $\theta \in \Theta$.

Since one could choose an incorrect model when using MLE, we need to **define** a **correctly specified model**:

Definition 10.2. A model is **correctly specified** when there is a **unique** parameter value $\theta_0 \in \Theta$ such that $f(x|\theta_0) = f(x)$, the "true" distribution of x. The parameter value $\boldsymbol{\theta_0}$ is called the **true parameter value**, and it is unique if there is no other θ_0 such that $f(x|\theta_0) = f(x)$. A model is **misspecified** if there is no $\theta \in \Theta$ such that $f(x|\theta) = f(x)$.

Hansen gives some **examples** on p.193. One is $f(x) = 2e^{-2x}$: both the exponential distribution and the gamma distribution are correctly specified models for it. Another is a "mixture" (weighted average of) two normally distributed models. This can fully "capture" a single normal distribution, but it can do so with 2+ sets of parameters, so it is not unique.

III. Likelihood (joint density) (10.3)

The likelihood of your data is the **joint density** as assumed by your model. Because we **assume** that the **observations are independent** of each other, the joint density is simply the **product of densities of all the observations**. Thus:

$$f(x_1, x_2, \ldots x_n|\,\theta) = f(x_1|\,\theta) \times f(x_2|\,\theta) \times \ldots f(x_n|\,\theta)$$

$$= \prod_{i=1}^{n} f(x_i|\theta)$$

This joint density **evaluated at the observed data**, which is a function of θ, is **called** the **likelihood function**:

Definition 10.3. The **likelihood function** for a sample of n observations, $L_n(\theta)$, is:

$$L_n(\theta) \equiv f(X_1, X_2, \ldots X_n|\theta) = \prod_{i=1}^{n} f(X_i|\theta) \quad (continuous\ X)$$

or

$$L_n(\theta) \equiv \prod_{i=1}^{n} \pi(X_i|\theta) \qquad (discrete\ X)$$

MLE finds the value of θ that best describes the data, which is θ_0. The likelihood function, $L_n(\theta)$, shows, for different values of θ, the probability (likelihood) that we observe the data that we actually have. **The value of θ that is most likely to give the data that we have** is the value that **maximizes the likelihood function**.

Definition 10.4. The **maximum likelihood estimator** $\hat{\theta}$ of θ is the value of θ that maximizes $L_n(\theta)$:

$$\hat{\theta} = \arg\max_{\theta \in \Theta} L_n(\theta)$$

Example: $f(x|\lambda) = \lambda^{-1}e^{-x/\lambda}$. The likelihood function is:

$$L_n(\lambda) = \prod_{i=1}^{n}\left(\frac{1}{\lambda}e^{-X_i/\lambda}\right) = \frac{1}{\lambda^n}e^{-n\bar{X}_n/\lambda}$$

Differentiate this with respect to λ (first order condition):

$$0 = \partial L_n(\lambda)/\partial\lambda = -n\frac{1}{\lambda^{n+1}}e^{-n\bar{X}_n/\lambda} + \frac{1}{\lambda^n}e^{-n\bar{X}_n/\lambda} \times \frac{n\bar{X}_n}{\lambda^2}$$

A little algebra shows that this holds for $\lambda = \bar{X}_n$, so $\hat{\lambda} = \bar{X}_n$.

In most cases, it is **more convenient to maximize the <u>logarithm</u> of the likelihood function**:

Definition 10.5. The **log-likelihood function** is:

$$\ell_n(\theta) \equiv \log\big(L_n(\theta)\big) = \sum_{i=1}^{n}\log\left(f(X_i|\theta)\right)$$

Since the log function is a monotonic increasing function, **the value of θ that maximizes $\ell_n(\theta)$ also maximizes $L_n(\theta)$**:

Question: For any i, is $\log(f(X_i|\theta)) > 0$ or < 0?

Theorem 10.1. $\hat{\theta} = \arg\max_{\theta\in\Theta}\ell_n(\theta)$.

Example. $f(x|\lambda) = \lambda^{-1}e^{-x/\lambda}$. Recall that $L_n(\lambda) = \lambda^{-n}e^{-n\bar{X}_n/\lambda}$. Thus $\ell_n(\lambda) = \log(L_n(\lambda)) = -n\log(\lambda) - n\bar{X}_n\lambda^{-1}$. Differentiating this with respect to λ gives:

$$0 = -n\lambda^{-1} + n\bar{X}_n\lambda^{-2}$$

This yields $\lambda = \bar{X}_n$, so we set $\hat{\lambda} = \bar{X}_n$, the same $\hat{\lambda}$ as above.

Example. Bernoulli distribution: $\pi(x|p) = p^x(1-p)^{1-x}$, where $x = 0$ or 1, and Prob$[x = 1] = p$. For any observation, the likelihood that x takes a specific value is $\pi(x|p)$, so $L(p) = \pi(x|p) = p^x(1-p)^{1-x}$. The log of this likelihood is $x\log(p) + (1-x)\log(1-p)$. For a sample of n observations, $\ell_n(\lambda) = \sum_{i=1}^{n}\{X_i\log(p) + (1-X_i)\log(1-p)\} = n\bar{X}_n\log(p) + n(1-\bar{X}_n)\log(1-p)$. Differentiate this w.r.t. p:

$$0 = n\bar{X}_n p^{-1} - n(1-\bar{X}_n)(1-p)^{-1}$$

Solving this yields $p = \bar{X}_n$.

IV. Likelihood Analog Principle (10.4)

A general estimation method in statistics and econometrics is to **use the data in our sample to estimate parameters that are functions of the variables in the population**. The **starting point** is the **expected log density** function:

$$\ell(\theta) = \mathrm{E}[\log(f(X|\,\theta))]$$

This is a function of θ. Note also **no *n* subscript on $\ell(\theta)$**.

Theorem 10.2. When the model is correctly specified the "true" parameter θ_0 maximizes the expected log density $\ell(\theta)$:

$$\theta_0 = \arg\max_{\theta \in \Theta} \ell(\theta).$$

Note again no *n* subscript on $\ell(\theta)$.

The **sample analog** of $\ell(\theta)$, $\bar{\ell}_n(\theta)$, is the **average over the data of the log-likelihood** (note the *n* subscript on $\bar{\ell}_n(\theta)$):

$$\bar{\ell}_n(\theta) = (1/n)\ell_n(\theta) = (1/\mathrm{n})\sum_{i=1}^{n} \log\left(f(X_i|\theta)\right)$$

Thus, **the $\widehat{\boldsymbol{\theta}}$ that maximizes $\bar{\boldsymbol{\ell}}_{\boldsymbol{n}}(\boldsymbol{\theta})$** (see p.6) **is the analog estimator of** the $\boldsymbol{\theta_0}$ (the value of θ that maximizes $\ell(\theta)$).

Example: $f(x|\,\lambda) = \lambda^{-1}e^{-x/\lambda}$. The log of this density is $-\log(\lambda) - x/\lambda$. The *expected log density* is:

$$\ell(\lambda) = \mathrm{E}[\log(f(X|\,\lambda))] = \mathrm{E}[-\log(\lambda) - X/\lambda] = -\log(\lambda) - \mathrm{E}[X]/\lambda$$

We want to find the "true" value of λ, which we call λ_0. For this function, which is the exponential function (see pp.59-60 of Hansen's *Probability and Statistics for Economists*), E[*X*] equals "λ", which we can denote as λ_0. Thus we have

$\ell(\lambda) = -\log(\lambda) - \lambda_0/\lambda$. Differentiating this with respect to λ yields $0 = -\lambda^{-1} + \lambda_0/\lambda^{-2}$, which implies that $\lambda = \lambda_0$. Thus **maximization of the expected log density yields the "true" value of λ**, which we denote by $\boldsymbol{\lambda_0}$.

Hansen gives two other examples on pp.196-97, including the normal density function with mean μ and variance σ^2.

He also points out (Section 10.5) that **MLE is "invariant" to transformations** ("invariance property"):

Theorem 10.3. If $\hat{\theta}$ is the MLE estimate of $\theta \in R^m$, then for any transformation $\beta = h(\theta) \in R^{\ell}$, the MLE estimator of β is $\hat{\beta} = h(\hat{\theta})$.

That is, we can "reparameterize" the model by replacing θ with $h^{-1}(\beta)$ and maximize the model with respect to β. The $\hat{\beta}$ obtained by MLE will equal $h(\hat{\theta})$. **This is important because sometimes we need transformations of $\widehat{\boldsymbol{\theta}}$, and we want these transformation to be MLE estimators**.

In Section 10.6, Hansen gives **six steps for MLE Estimation** (and some examples of applying these 6 steps):

1. Construct $f(x|\ \theta)$ as a function of x and θ.
2. Take the logarithm of $f(x|\ \theta)$: $\log[f(x|\ \theta)]$.
3. Evaluate at $x = X_i$ and sum: $\ell_n(\theta) = \sum_{i=1}^{n} \log\left(f\left(X_i|\theta\right)\right)$
4. If possible, solve the F.O.C. to find the maximum.

5. Check the S.O.C. to verify that it is a maximum.
6. If solving the F.O.C. is not possible, use numerical methods to maximize $\ell_n(\theta)$.

V. Score, Hessian and Information (10.7)

The log-likelihood function, $\ell_n(\theta)$ has some useful properties:

$$\ell_n(\theta) = \sum_{i=1}^{n} \log\left(f(X_i|\theta)\right)$$

Assume that $f(x|\,\theta)$ is differentiable with respect to θ. The **likelihood score** is the derivative of the likelihood function:

$$S_n(\theta) = \partial\ell_n(\theta)/\partial\theta = \sum_{i=1}^{n} \partial \log\left(f(X_i|\theta)\right)/\partial\theta$$

Since $\boldsymbol{\theta}$ **is often a vector**, $\boldsymbol{S_n(\theta)}$ is a **vector of the same dimension**. The score measures the "sensitivity" of $\ell_n(\theta)$ to θ, **When $\boldsymbol{\hat{\theta}}$ is an interior solution** to maximizing $\ell_n(\theta)$, then $\boldsymbol{S_n(\hat{\theta}) = 0}$ (derivatives for each element of $\hat{\theta}$ equal 0).

The **likelihood Hessian**, denoted by $\mathscr{H}_n(\theta)$, is the negative of the **second derivative** of the likelihood function:

$$\mathscr{H}_n(\theta) = -\frac{\partial^2 \ell_n(\theta)}{\partial\theta\partial\theta'} = -\sum_{i=1}^{n} \frac{\partial^2 \log(f(X_i|\theta))}{\partial\theta\partial\theta'}$$

The Hessian **shows the degree of curvature** in the log-likelihood: smaller values indicate a flatter likelihood.

The **efficient score** is the derivative of $\ell_n(\theta)$ **for a single observation**, **evaluated at** a random vector X and **the "true" θ (θ_0)**:

$$S = \partial \log(f(X|\,\theta_0)/\partial\theta$$

Theorem 10.4. Assume that the model is correctly specified, the support of X does not depend on θ, and θ_0 is in the interior of Θ. Then the efficient score S satisfies $\mathrm{E}[S] = 0$

Definition 10.6. The **Fisher information matrix** is the **variance of the efficient score**:

$$\mathscr{I}_\theta = \mathrm{E}[SS']$$

Definition 10.7. The **expected Hessian** (recall that $\ell(\theta) = \mathrm{E}[\log(f(X|\,\theta))]$) is:

$$\mathscr{H}_\theta = -\frac{\partial^2 \ell(\theta_0)}{\partial\theta\partial\theta'}$$

If $f(X|\,\theta)$ is twice differentiable in θ, and the support of X does not depend on θ, $\mathscr{H}_\theta$ equals the expectation of the likelihood Hessian for a single observation:

$$\mathscr{H}_\theta = -\mathrm{E}\left[\frac{\partial^2 \log(f(X|\theta_0))}{\partial\theta\partial\theta'}\right]$$

This leads to …

Theorem 10.5 Information Matrix Equality. Assume that the model is correctly specified, and that the support of X does not depend on θ. Then the Fisher information matrix equals the expected Hessian:

$$\mathscr{I}_\theta = \mathscr{H}_\theta$$

This says that **the curvature in the likelihood function and the variance of the score are equal to each other**. Hansen says that there is **no intuition for this**, and that it is useful to simplify the asymptotic variance of the MLE.

Hansen gives some examples of this on pp.204-206.

VI. Cramér-Rao Lower Bound (10.9)

A useful property of MLE is that there is **a lower bound of the covariance matrix for any unbiased estimate of θ**:

Theorem 10.6: Cramér-Rao Lower Bound. Assume that the model is correctly specified, that the support of X does not depend on θ, and that θ_0 is in the interior of Θ. **If $\tilde{\boldsymbol{\theta}}$ is an unbiased estimator of $\boldsymbol{\theta}$**, then:

$$\text{var}[\tilde{\theta}] \geq (n\mathscr{I}_\theta)^{-1}$$

In other words, the smallest possible covariance matrix for any unbiased estimator of θ is $(n\mathscr{I}_\theta)^{-1}$, so **any unbiased estimator of θ that has a covariance matrix of $(n\mathscr{I}_\theta)^{-1}$**

has achieved maximum statistical efficiency (there exists no other unbiased estimator that is more efficient).

Definition 10.8. The **Cramér-Rao Lower Bound** is $(n\mathcal{I}_\theta)^{-1}$.

Definition 10.9. An estimator $\tilde{\theta}$ is **Cramér-Rao efficient** if it is unbiased for θ and $\text{var}[\tilde{\theta}] = (n\mathcal{I}_\theta)^{-1}$.

The Cramér-Rao Lower Bound (CRLB) is a **famous result**. It **provides a general bound on the precision of estimation**. When θ is a vector (which is usually the case) then CRLB states that the covariance matrix is bounded from below by the matrix inverse of the Fisher information matrix, which means that the difference between the two matrices is a positive semi-definite matrix.

Example. Suppose that X is normally distributed $N(\mu, \sigma^2)$, with unknown μ and σ^2. Then $f(x|\,\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-(x-\mu)^2/2\sigma^2}$. The **log density** is:

$$\log(f(x|\,\mu, \sigma^2)) = -(1/2)\log(2\pi) - (1/2)\log(\sigma^2) - \frac{(x-\mu)^2}{2\sigma^2}$$

The **log likelihood is**:

$$\ell_n(\mu, \sigma^2) = -(n/2)\log(2\pi) - (n/2)\log(\sigma^2) - \frac{\sum_{i=1}^{n}(X_i-\mu)^2}{2\sigma^2}$$

Differentiate this with respect to μ and σ^2:

$$\partial \ell_n(\mu, \sigma^2)/\partial\mu = (1/(2\sigma^2))\sum_{i=1}^{n} 2(X_i - \mu) = 0$$

$$\sum_{i=1}^{n} X_i - n\mu = 0$$

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

$$\partial \ell_n(\mu, \sigma^2)/\partial\sigma^2 = -(n/2)(1/\sigma^2) + \frac{\sum_{i=1}^{n}(X_i-\mu)^2}{2(\sigma^2)^2} = 0$$

$$-n\sigma^2 + \sum_{i=1}^{n}(X_i - \mu)^2 = 0$$

$$\sigma^2 = (1/n)\sum_{i=1}^{n}(X_i - \mu)^2$$

The F.O.C. for μ implies that the MLE is $\hat{\mu} = \bar{X}_n$. The F.O.C. for σ^2 implies that the MLE is $\hat{\sigma}^2 = (1/n)\sum_{i=1}^{n}(X_i - \mu)^2$, which can be calculated by replacing μ with $\bar{X}_n$.

Continuing with this example, return to the **log density**:

$$\log(f(x|\, \mu, \sigma^2)) = -(1/2)\log(2\pi) - (1/2)\log(\sigma^2) - \frac{(x-\mu)^2}{2\sigma^2}$$

To **get the score**, **differentiate** this **with respect to μ and σ^2**:

$$\partial\log(f(x|\, \mu, \sigma^2))/\partial\mu = \frac{x-\mu}{\sigma^2} = 0$$

$$\partial\log(f(x|\, \mu, \sigma^2))/\partial\sigma^2 = -\frac{1}{2\sigma^2} + \frac{(x-\mu)^2}{2\sigma^4} = 0$$

The second derivatives are:

$$\partial^2 \log(f(x|\,\mu, \sigma^2)) / \partial \mu^2 = -\frac{1}{\sigma^2}$$

$$\partial^2 \log(f(x|\,\mu, \sigma^2)) / \partial (\sigma^2)^2 = \frac{1}{2\sigma^4} - \frac{(x-\mu)^2}{\sigma^6}$$

$$\partial^2 \log(f(x|\,\mu, \sigma^2)) / \partial \mu \partial \sigma^2 = -\frac{x-\mu}{2\sigma^4}$$

The expected Fisher information matrix is:

$$\mathcal{I}_\theta = \mathrm{E}\begin{bmatrix} \frac{1}{\sigma^2} & \frac{X-\mu}{2\sigma^4} \\ \frac{X-\mu}{2\sigma^4} & \frac{(X-\mu)^2}{\sigma^6} - \frac{1}{2\sigma^4} \end{bmatrix} = \begin{bmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{bmatrix}$$

Finally, the lower bound is:

$$\mathrm{CRLB} = (n\mathcal{I}_\theta)^{-1} = \begin{bmatrix} \frac{\sigma^2}{n} & 0 \\ 0 & \frac{2\sigma^4}{n} \end{bmatrix}$$

The **MLE estimate of the sample mean**, that is $\hat{\mu} = \bar{X}_n$, **attains the CRLB** because the variance of the sample mean $\bar{X}_n$ is σ^2/n (see p.137). In contrast, the MLE estimate of the variance of X, $\hat{\sigma}^2 = (1/n)\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$, is (slightly) biased (see p.139). **The estimate, $s^2 = (1/(n\text{-}1))\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$ is unbiased but (slightly) misses the CRLB**. See p.208.

On page 208 Hansen presents Theorem 10.7, which extends the Cramér-Rao Lower Bound to functions of parameters.

VII. Consistent Estimation (10.12)

Statisticians have worked out the assumptions needed for MLE estimates, denoted by $\hat{\theta}$, to be consistent. The general result is:

Theorem 10.8. Assume:

1. X_i are i.i.d (independent and identically distributed).
2. $E[\log(f(X|\theta))] \leq G(X)$ and $E[G(X)] < \infty$.
3. $\log(f(X|\theta_1))$ is continuous in θ with probability one.
4. Θ is compact.
5. For all $\theta \neq \theta_0$, $\ell(\theta) < \ell(\theta_0)$.

Then $\hat{\theta} \underset{p}{\rightarrow} \theta_0$ as $n \rightarrow \infty$.

VIII. Asymptotic Normality (10.13 and 10.14)

Hansen shows (pages 209-211), that under a few additional technical assumptions that the MLE estimator $\hat{\theta}$ is asymptotically normally distributed with a covariance matrix that is the inverse of the information matrix $\mathscr{I}_\theta$:

$$\sqrt{n}(\hat{\theta} - \theta_0) \underset{d}{\rightarrow} N(0, \mathscr{I}_\theta^{-1}).$$

This is Theorem 10.9. He also shows (Theorem 10.10) that **maximum likelihood estimation is asymptotically Cramér-Rao efficient**.

Question: Doesn't Theorem 10.10 contradict the point made above (page 16) that the MLE estimate for σ^2 for $X \sim N(\mu, \sigma^2)$ is biased?

IX. Variance Estimation (10.15)

To conduct statistical tests and construct confidence intervals for our estimator, $\hat{\theta}$, we need to estimate its variance: $\mathcal{I}_\theta^{-1}$. We can denote such estimates by $\widehat{\boldsymbol{V}}$. In fact, there are **three different ways to estimate the covariance matrix**. They are **all asymptotically equivalent**, so in practice researchers use the one that is most convenient. Here they are:

Expected Hessian Estimator.

Recall the expected Hessian matrix (p.12):

$$\mathcal{H}_\theta = -\,\mathrm{E}\left[\frac{\partial^2 \log(f(X|\theta_0))}{\partial\theta\partial\theta'}\right]$$

We can define something slightly different by changing the order of differentiation and expectations:

$$\mathcal{H}_\theta(\theta) = \frac{\partial^2}{\partial\theta\partial\theta'}\mathrm{E}[\log(f(X|\,\theta))]$$

Denote $\hat{\mathcal{H}}_\theta = \mathcal{H}_\theta(\hat{\theta})$ ($\mathcal{H}_\theta(\theta)$ evaluated at $\hat{\theta}$).

The **expected Hessian estimator of the variance is**:

$$\widehat{\boldsymbol{V}}_0 = \hat{\mathcal{H}}_\theta^{-1}$$

This can be computed only when $\mathscr{H}_\theta(\theta)$ can be expressed as an explicit function of θ, which is often not possible.

Sample Hessian Estimator. This is almost the same as the expected Hessian estimator except that we take the (second) derivative of the likelihood function instead of the *expected* likelihood function:

$$\hat{\mathcal{H}}_\theta = \frac{1}{n}\sum_{i=1}^{n} - \frac{\partial^2 \log\left(f(X_i|\hat{\theta})\right)}{\partial\theta\partial\theta'} = -\frac{1}{n}\frac{\partial^2}{\partial\theta\partial\theta'}\ell_n(\hat{\theta})$$

$$\widehat{\boldsymbol{V}}_1 = \hat{\mathcal{H}}_\theta^{-1}$$

This is **easier to compute**, and Hansen says that it is the most commonly used estimator for the variance of $\hat{\theta}$.

Outer Product Estimator. This is based on the formula for the Fisher information matrix:

$$\hat{\mathcal{J}}_\theta = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial\log\left(f(X_i|\hat{\theta})\right)}{\partial\theta}\right)\left(\frac{\partial\log\left(f(X_i|\hat{\theta})\right)}{\partial\theta}\right)'$$

$$\widehat{\boldsymbol{V}}_2 = \hat{\mathcal{J}}_\theta^{-1}$$

Theorem 10.11 shows that all 3 converge in probability to $\boldsymbol{V}$ (which equals $\mathscr{H}_\theta^{-1}$, which also equals $\mathscr{I}_\theta^{-1}$).

X. Kullback-Leibler Divergence + Approximating Models

A useful way to measure the difference between two densities, $f(\mathrm{x})$ and $g(x)$, is the Kullback-Leibler divergence:

$$\mathrm{KLIC}(f, g) = \int f(\mathrm{x})[\log(f(\mathrm{x})) - \log(g(\mathrm{x}))]dx$$

Theorem 10.12 KLIC has the following three properties:

1. $\mathrm{KLIC}(f, f) = 0$
2. $\mathrm{KLIC}(f, g) \geq 0$ (for any f and g)
3. $f = \underset{g}{\operatorname{argmin}}\ \mathrm{KLIC}(f, g)$

For any model specification, such as **maximum likelihood**, it is **unlikely to be the "true" functional form**. **But it still may be a "good fit"**, and we can use KLIC to measure that.

Definition 10.12 The **pseudo-true parameter θ_0 for a parametric model f_θ** that best fits the "true" density f is the value that minimizes the Kullback-Leibler divergence:

$$\theta_0 = \arg\min_{\theta \in \Theta}\ \mathrm{KLIC}(f, f_\theta)$$

Thus, **for a mis-specified parametric model f_θ, define θ_0** as **the θ that minimizes KLIC** for that model.

Theorem 10.14 Under misspecification, the pseudo-true parameter satisfies $\theta_0 = \arg\max_{\theta \in \Theta} \ell(\theta)$.

So maximum likelihood estimation finds the "best fit" to the true model when a misspecified model is used for estimation.

Thus, go ahead and use **MLE**. **Even if misspecified, it gives the best possible fit** (defined as minimizing KLIC). However, you cannot use $\mathscr{I}_\theta^{-1}$ or $\mathscr{H}_\theta^{-1}$ to estimate the covariance matrix for θ_0. Instead use $\mathscr{H}_\theta^{-1}\mathscr{I}_\theta\,\mathscr{H}_\theta^{-1}$; see Theorem 10.16 and the discussion on pages 216-217. This is "robust" estimation of the MLE covariance matrix, which Hansen recommends.

XI. Likelihood Ratio Test (Wooldridge, 2010, p.481)

It is **very easy to test a set of constraints when using MLE**. Recall the log likelihood function based on the data:

$$\ell_n(\theta) = \sum_{i=1}^{n} \log\left(f(X_i|\theta)\right)$$

Let $\hat{\theta}$ denote the unconstrained estimate of θ, that is the estimate of θ that maximizes $\ell_n(\theta)$ without imposing any constraint. Let $\tilde{\theta}$ denote a constrained estimate of θ imposing q constraints; that is, $\tilde{\theta}$ maximizes $\ell_n(\theta)$ after imposing the q constraints. To test the hypothesis that the constraints are "true", use the likelihood ratio (LR) test:

$$\text{LR} = 2[\ell_n(\hat{\theta}) - \ell_n(\tilde{\theta})]$$

LR is distributed as chi-square with q degrees of freedom.

Question: Is it possible for LR < 0?