Stat 200- Feb. 26, 2001, Monday

Summary of preceding lecture:Maximising log likelihood, with and without constraints, can be an unsolvable problem in closed form, then we have to use iterative procedures. The parametric bootstrap was often the only way to study the sampling distribution of the mle.

8.5.2 Large Sample Theory for MLE (p 261)

MLE estimates are consistent under reasonable conditions. We will give ideas about how this is proved without the technical subtleties.

***Theorem* 8.1** *Under appropriate smoothness conditions on* f *(the density), the mle from an iid sample is consistent.*

Proof:
The estimate maximises $l(\theta\ )$ so it also maximises

$$\frac{1}{n}l(\theta\ ) = \frac{1}{n}\sum_{i=1}^{n} logf(X_i|\theta\ ) \longrightarrow Elogf(X|\theta\ ) = \int logf(X|\theta\ )f(X|\theta_0\ )dx$$

Here we will have to admit that the θ that maximises $l(\theta\ )$ is close to the one that maximises $Elogf(X|\theta\ )$, so if we differentiate with regards to θ we get :

$$\frac{\partial}{\partial\theta}\int logf(X|\theta\ )f(X|\theta_0\ )dx = \int \frac{\frac{\partial f}{\partial\theta}(x|\theta\ )}{f(x|\theta\ )} f(x|\theta_0\ )dx$$

and we see that for the particular value $\theta\ = \theta_0$ this is

$$\frac{\partial}{\partial\theta}\int f(X|\theta_0\ )dx = \frac{\partial}{\partial\theta}(1) = 0$$

so θ_0 IS a stationary point.

We have interchanged derivationa and integration, to be allowed to do this there are smoothness conditions to impose on f.

We are going to use a quantity called Fisher's information, in some sense it's the expected value of the square of the relative rate of change of the density.

As we will see a large information is equivalent to having a bigger sample size and thus a smaller variance of the estimate.

Lemma 1 *Define the quantity*

$$I(\theta\ ) = E[\frac{\partial}{\partial\theta} logf(x|\theta\ )]^2$$

under appropraite smoothness conditions on f, $I(\theta\ )$ *can be written :*

$$I(\theta\ ) = -E[\frac{\partial^2}{\partial\theta^2} logf(x|\theta\ )]$$

Proof:
Double derivation gives the result.(with exchanges between order of diff and integration).

Here are some details:

$$\begin{aligned}
\frac{\partial}{\partial\theta}\int f(x|\theta\ )dx &= \int \left[\frac{\partial}{\partial\theta} logf(x|\theta\ )\right] f(x|\theta\ )dx = 0 \\
&\Longrightarrow E\left[\frac{\partial}{\partial\theta} logf(x|\theta\ )\right] = 0 \\
\frac{\partial^2}{\partial\theta^2}\int f(x|\theta\ )dx &= 0 \\
&= \int \frac{\partial^2}{\partial\theta^2} logf(x|\theta\ )f(x|\theta\ )dx + \int \left[\frac{\partial}{\partial\theta} logf(x|\theta\ )\right]^2 f(x|\theta\ )dx \\
&\Longrightarrow I(\theta\ ) = -E\left[\frac{\partial^2}{\partial\theta^2} logf(x|\theta\ )\right]
\end{aligned}$$

***Theorem* 8.2** *Under smoothness conditions on* f*, the probability distribution of*

$$\sqrt{nI(\theta_0)}(\hat{\theta}_n - \theta_0) \Longrightarrow \mathcal{N}(0,1)$$

This says that the mle is asymptotically unbiased and that it's variance is inversely proportional to n and $I(\theta_0)$.

Proof: (Just formal heuristics)

Taylor expansion of $l'(\hat{\theta})$ which is itself 0:

$$\begin{aligned} 0 = l'(\hat{\theta}) &\approx l'(\theta_0) + (\hat{\theta} - \theta_0) l''(\theta_0) \\ (\hat{\theta} - \theta_0) &\approx -\frac{l'(\theta_0)}{l''(\theta_0)} (*) \\ n^{\frac{1}{2}}(\hat{\theta} - \theta_0) &\approx -\frac{n^{-\frac{1}{2}} l'(\theta_0)}{n^{-1} l''(\theta_0)} \end{aligned}$$

This is an important lemma defining a good way of computing what is known as Fisher's information:

$$I(\theta_0) = E(\ell'(\theta_0)^2) = -E(\ell''(\theta_0))$$

The first theorem said that mle is consistent, the second that we have convergence in distribution of

$$\sqrt{nI(\theta_0)}(\hat{\theta}_n - \theta_0) \Longrightarrow \mathcal{N}(0,1)$$

In particular the "asymptotic variance" of the mle $\hat{\theta}$ is $\frac{1}{nI(\theta_0)}$.

These theorems are very useful and in particular they allow the construction of:

8.5.3 Confidence Intervals , p. 266

- exact methods
- approximate methods based on the theorems above
- bootstrap methods

Example:

Poisson λ, $I(\lambda) = -E\ell''(\lambda)$ and $E(x) = \lambda$.

$$\begin{aligned} \ell(\lambda) &= x log\lambda - \lambda - log x! \\ \ell'(\lambda) &= \frac{x}{\lambda} - 1 \\ \ell''(\lambda) &= -\frac{x}{\lambda^2} \end{aligned}$$

$$I(\lambda) = \frac{1}{\lambda}$$

We plug in the estimate $\bar{X}$ for λ and then we can compute the confidence interval which will simply be:

$$\bar{X} +/- z_{\frac{\alpha}{2}} \sqrt{\frac{\bar{X}}{n}}$$

This is only an approximate CI, since $\bar{X}$ is only asymptotically Normal.

8.6 Efficiency and Cramer-Rao Lower Bound, p. 273

We will often be looking for the best possible estimators, then the question arises :

Is there a lower bound for the Mean Square Error of any estimators? (we can't do better than this lower bound.....)

Such a bound would provide a benchmark, an estimator achieving the lower bound would be impossible to improve upon.

When the estimator is UNBIASED we DO have such a bound due to the following 'information inequality bound':

Theorem 8.3 *If* $\boldsymbol{X_1, \ldots, X_n}$ *are iid with density* $\boldsymbol{f(x|\theta\ )}$ *and* $\boldsymbol{T = t(x1, , x_n)}$ *is a statistic that provides an unbiased estimate of* $\boldsymbol{\theta}$.

Then under smoothness conditions needed for inversion of integration and differentiation :

$$\boldsymbol{var(T) \geq \frac{1}{nI(\theta\ )}}$$

Proof:

Call $\boldsymbol{Z}$ the random variable

$$\boldsymbol{Z = \sum_{i=1}^{n} \frac{\partial}{\partial\theta}[logf(X_i|\theta\ )]}$$

$\boldsymbol{E(Z) = 0}$ and $\boldsymbol{Var(Z) = nI(\theta\ )}$

If we can prove $\boldsymbol{Cov(Z, T) = 1}$ we are home, because :

$\boldsymbol{\frac{Cov^2(Z,T)}{Var(T)Var(Z)} \leq 1}$ and $\boldsymbol{Cov(Z, T) = 1}$ would imply that $\boldsymbol{Var(T) \geq \frac{1}{Var(Z)} = \frac{1}{nI(\theta\ )}}$

Here is how we do that :

$$\begin{aligned}
\boldsymbol{Cov(ZT)} &= \boldsymbol{E(ZT) - 0 = \int\int \ldots \int t(x_1..t_n) \left[\sum_{i=1}^{n} \frac{\frac{\partial}{\partial\theta} f(x_i|\theta\ )}{f(x_i|\theta\ )}\right] \prod_{j=1}^{n} f(x_j|\theta) dx_j} \\
&\text{and } \boldsymbol{\sum_{i=1}^{n} \frac{\partial}{\partial\theta} f(x_i|\theta\ ) \prod_{j\neq i} f(x_j|\theta\ ) dx_j = \frac{\partial}{\partial\theta} \prod_{j=1}^{n} f(x_i|\theta\ )} \\
&= \boldsymbol{\int\int \ldots \int t(x_1..t_n) \frac{\partial}{\partial\theta} \prod_{j=1}^{n} f(x_j|\theta) dx_j} \\
&= \boldsymbol{\frac{\partial}{\partial\theta} E(T) = 1}
\end{aligned}$$

8.6.1 Efficiency

Given this lower bound, we have a reference to which we compare variances of all estimators.

Definition:

$$\text{Efficiency of } \boldsymbol{\hat{\theta}} \text{ relative to } \boldsymbol{\tilde{\theta}} = \boldsymbol{eff(\hat{\theta}\ , \tilde{\theta}\ ) = \frac{Var(\tilde{\theta}\ )}{Var(\hat{\theta}\ )}}$$

Most meaningful when both estimators are unbiased or have the same bias.

When the Cramer-Rao lower bound is attained, the estimator is said to be efficient.

Example:

Poisson: $\boldsymbol{I(\lambda) = \frac{1}{\lambda}}$. For any unbiased estiamte of $\boldsymbol{\lambda}$, $\boldsymbol{var(T) \geq \frac{\lambda}{n}}$.

$$\boldsymbol{\lambda_{MLE} = \bar{X} = \frac{S}{n}}, \text{ with } \boldsymbol{var(S) = n\lambda}, \text{ so } \boldsymbol{var(\bar{X}\ ) = \frac{\lambda}{n}}$$

Most often, for instance for the MLE, we have asyptotic efficiency, or asymptotic relative efficency (ARE).

8.7 Sufficiency, p. 280

If an estimator is sufficient for a parameter $\boldsymbol{\theta}$ we can compute just that estimate and throw away all the other data.

The rigourous definition of a sufficient statistic is that it is suff. iff the conditional distribution (density or frequency) of the vector $\underline{\boldsymbol{X}}$ does not depend on $\boldsymbol{\theta}$ for any value of $\boldsymbol{T} = \boldsymbol{t}$.