LTCC/Reid: Likelihood and derived quantities   November 5, 2012

Given a model for Y which assumes Y has a density $f(y;\theta), \quad \theta \in \Theta \subset \mathbb{R}^d$, we have the following definitions:

observed likelihood function	$L(\theta; y) = c(y) f(y;\theta)$
log-likelihood function	$\ell(\theta; y) = \log L(\theta; y) = \log f(y;\theta) + a(y)$
score function	$U(\theta) = \partial \ell(\theta; y)/\partial\theta$
observed information function	$j(\theta) = -\partial^2 \ell(\theta; y)/\partial\theta\partial\theta^T$
expected information (in one observation)	$i(\theta) = \mathrm{E}_\theta U(\theta)U(\theta)^T$ (called $i_1(\theta)$ in CH)

When we have Y_i independent, identically distributed from $f(y_i;\theta)$, then, denoting the observed sample $y = (y_1, \ldots, y_n)$ we have:

log-likelihood function	$\ell(\theta) = \ell(\theta; y) + a(y)$	$O_p(n)$
maximum likelihood estimate	$\hat\theta = \hat\theta(y) = \arg\sup_\theta \ell(\theta)$	$\theta + O_p(n^{-1/2})$
score function	$U(\theta) = \ell'(\theta) = \sum U_i(\theta) = U_+(\theta)$	$O_p(n^{1/2})$
observed information function	$j(\theta) = -\ell''(\theta) = -\ell(\theta; Y)$	$O_p(n)$
observed (Fisher) information	$j(\hat\theta)$	
expected (Fisher) information	$i(\theta) = E_\theta\{U(\theta)U(\theta)^T\} = n i_1(\theta)$	$O(n)$,

where with the risk of some confusion we use the same notation. Sometimes the expected Fisher information is defined instead as $i(\theta) = \mathrm{E}_\theta\{-\partial U(\theta; Y)/\partial\theta^T\}$ (e.g. in BNC). In models for which we can interchange differentiation and integration in $\int f(y;\theta)dy = 1$, these are the same due to the Bartlett identities:

$$\begin{aligned}
\mathrm{E}_\theta\{U(\theta)\} &= 0,\\
\mathrm{E}_\theta\{U'(\theta)\} + \mathrm{E}_\theta\{U^2(\theta)\} &= 0,\\
\mathrm{E}_\theta\{U''(\theta)\} + 3\mathrm{E}_\theta\{U(\theta)U'(\theta)\} + \mathrm{E}_\theta\{U^3(\theta)\} &= 0,
\end{aligned}$$

and so on, where the result applies to vector θ, but as presented here is for scalar θ. (In the vector setting the second derivative of U is a $d \times d \times d$ array.)

First order asymptotic theory

The following results are used for approximate inference based on the likelihood function:

1. θ is a scalar

$\frac{1}{\sqrt{n}}U(\theta)/i_1^{1/2}(\theta) \xrightarrow{d} N(0,1)$	by the central limit theorem
standardized score statistic	$r_u = U(\theta)/j^{1/2}(\hat\theta) \xrightarrow{d} N(0,1)$
$\sqrt{n}(\hat\theta-\theta)i_1^{1/2}(\theta) =$	$\dfrac{1}{\sqrt{n}}\dfrac{U(\theta)}{i_1^{1/2}(\theta)}\{1+o_p(1)\}$
standardized m.l.e.	$r_e = (\hat\theta-\theta)j^{1/2}(\hat\theta) \xrightarrow{d} N(0,1)$
(log) likelihood ratio statistic	$w(\theta) = 2\{\ell(\hat\theta)-\ell(\theta)\} = (\hat\theta-\theta)^2 i(\theta)\{1+o_p(1)\}$
	$w(\theta) \xrightarrow{d} \chi^2_1$
likelihood root	$r(\theta) = \text{sign}(\theta-\hat\theta)\{w(\theta)\}^{1/2}$
	$r(\theta) \xrightarrow{d} N(0,1)$

2. θ a vector of length d

$\frac{1}{\sqrt{n}}\{U(\theta)\} \xrightarrow{d} N_d\{0, i_1(\theta)\}$	by the central limit theorem
standardized score statistic	$w_u = U(\theta)^T\{i(\theta)\}^{-1}U(\theta)$
$\sqrt{n}(\hat\theta-\theta) =$	$\frac{1}{\sqrt{n}}i_1^{-1}(\theta)U(\theta)\{1+o_p(1)\}$
standardized m.l.e.	$w_e = (\hat\theta-\theta)^T i(\theta)(\hat\theta-\theta)$
likelihood ratio statistic	$w = 2\{\ell(\hat\theta)-\ell(\theta)\} = (\hat\theta-\theta)^T i(\theta)(\hat\theta-\theta)\{1+o_p(1)\}$
	$w(\theta) \xrightarrow{d} \chi^2_d$

3. $\theta = (\psi,\lambda) = (\psi_1,\dots,\psi_q,\lambda_1,\dots,\lambda_{d-q})$ We partition the information matrices compatibly and write

$$U(\theta) = \begin{pmatrix} U_\psi(\theta) \\ U_\lambda(\theta) \end{pmatrix},$$

$$i(\theta) = \begin{pmatrix} i_{\psi\psi} & i_{\psi\lambda} \\ i_{\lambda\psi} & i_{\lambda\lambda} \end{pmatrix} \quad j(\theta) = \begin{pmatrix} j_{\psi\psi} & j_{\psi\lambda} \\ j_{\lambda\psi} & j_{\lambda\lambda} \end{pmatrix}$$

and

$$i^{-1}(\theta) = \begin{pmatrix} i^{\psi\psi} & i^{\psi\lambda} \\ i^{\lambda\psi} & i^{\lambda\lambda} \end{pmatrix} \quad j^{-1}(\theta) = \begin{pmatrix} j^{\psi\psi} & j^{\psi\lambda} \\ j^{\lambda\psi} & j^{\lambda\lambda} \end{pmatrix}.$$

The constrained maximum likelihood estimator of λ is denoted by $\hat\lambda_\psi$, which in regular models satisfies $U_\lambda(\psi,\hat\lambda_\psi) = 0$.

Note that

$$i^{\psi\psi}(\theta) = \{i_{\psi\psi}(\theta) - i_{\psi\lambda}(\theta)i_{\lambda\lambda}^{-1}(\theta)i_{\lambda\psi}(\theta)\}^{-1}, \quad (1)$$

using the formula for the determinant of a partitioned matrix. A similar result holds for j.

The profile log-likelihood function is $\ell_{\rm P}(\psi) = \ell(\psi, \hat\lambda_\psi)$, and the (observed) profile information function is $j_{\rm P}(\psi) = -\ell''_{\rm P}(\psi)$, a $q \times q$ matrix.

The limiting results above can be used to derive the following

$$\begin{aligned} w_u(\psi) = U_\psi(\psi, \hat\lambda_\psi)^T \{i^{\psi\psi}(\psi, \hat\lambda_\psi)\} U_\psi(\psi, \hat\lambda_\psi) &\ \dot\sim\ \chi^2_q \\ w_e(\psi) = (\hat\psi - \psi)\{i^{\psi\psi}(\hat\psi, \hat\lambda)\}^{-1}(\hat\psi - \psi) &\ \dot\sim\ \chi^2_q \\ w(\psi) = 2\{\ell(\hat\psi, \hat\lambda) - \ell(\psi, \hat\lambda_\psi)\} = 2\{\ell_{\rm P}(\hat\psi) - \ell_{\rm P}(\psi)\} &\ \dot\sim\ \chi^2_q; \end{aligned}$$

see (52), (54) and (56) in CH §9.3.

This determines the following first-order pivotal quantities, for scalar ψ:

$$\begin{aligned} r_e(\psi) &= (\hat\psi - \psi) j_{\rm P}^{1/2}(\hat\psi) \dot\sim N(0,1), \\ r_u(\psi) &= \ell'_{\rm P}(\psi) j_{\rm P}^{-1/2}(\hat\psi) \dot\sim N(0,1), \\ r(\psi) &= \text{sign}(\hat\psi - \psi)\sqrt{2\{\ell_{\rm P}(\hat\psi) - \ell_{\rm P}(\psi)\}} \dot\sim N(0,1) \\ w(\psi) &= 2\{\ell_{\rm P}(\hat\psi) - \ell_{\rm P}(\psi)\} \dot\sim \chi^2_1, \end{aligned}$$

where the third form follows from the fourth.

Exercises

1. *Orthogonal nuisance parameters.* In a model $f(y;\theta)$ with $\theta = (\psi, \lambda)$, the component parameter ψ and λ are orthogonal (with respect to Fisher information) if $i_{\psi\lambda}(\theta) = 0$.

   (a) Suppose we have a sample $y_1, \ldots, y_n$ from the density $f(y;\theta)$. Show that
   $$\hat\lambda_\psi = \hat\lambda + O_p(n^{-1/2}),$$
   whereas if ψ and λ are orthogonal that
   $$\hat\lambda_\psi = \hat\lambda + O_p(n^{-1}).$$

   (b) Assume y_i follows an exponential distribution with mean $\lambda e^{-\psi x_i}$, where x_i is known. Find conditions on the sequence $\{x_i, i = 1, \ldots, n\}$ in order that λ and ψ are orthogonal with respect to expected Fisher information. Find an expression for the constrained maximum likelihood estimate $\hat\lambda_\psi$ and show the effect of parameter orthogonality on the form of the estimate.

2. *Sufficient statistics (CH Exercise 2.2).* Find the log-likelihood function for a sample of size n from an $AR(1)$ process:

$$y_t = \mu + \rho(y_{t-1} - \mu) + \epsilon_t, \quad \epsilon_t(i.i.d.) \sim N(0, \sigma^2), \quad t = 1, ..., n,$$

where $|\rho| < 1$, as a function of $\theta = (\mu, \sigma^2, \rho)$ and y_0. Write down the likelihood for data $y_1, \ldots, y_n$ in the cases where the initial value y_0 is

   (a) a given constant;

   (b) normally distributed with mean μ and variance $\sigma^2/(1-\rho^2)$;

   (c) assumed equal to y_n,

and give the sufficient statistic for each case.

Measure theory

The likelihood function is defined as (proportional to) the density function, and this is a density with respect to some dominating measure. Since θ varies in Θ, we need f to be a density function with respect to the same dominating measure for each value of θ. Schervish (p.13) states it this way:

Let $(S, \mathcal{A}, \mu)$ be a probability space, and let $(\mathcal{X}, \mathcal{B})$ be a Borel space. Let $X : S \longrightarrow \mathcal{X}$ be measurable. The parametric family of distributions for X is the set

$$\{P_\theta : \forall A \in \mathcal{B}, P_\theta(A) = \Pr(X \in A), \theta \in \Theta\}.$$

Assume that each P_θ, considered as a measure on $(\mathcal{X}, \mathcal{B})$ is absolutely continuous with respect to a measure ν on $(\mathcal{X}, \mathcal{B})$. We write

$$f(x;\theta) = \frac{dP_\theta}{d\nu}(x);$$

this is the likelihood function for θ.

Some books describe the likelihood function as the Radon-Nikodym derivative of the probability measure with respect to a dominating measure. Sometimes the dominating measure is taken to be P_{θ_0} for a fixed value $\theta_0 \in \Theta$. When we consider probability spaces and/or parameter spaces that are infinite dimensional, it is not obvious what to use as a dominating measure. For counting processes, this is done rigorously in Ch.II of Andersen et al. The result is Jacod's formula for the likelihood ratio:

Suppose we have a counting process $N(\cdot)$ on $[0, \tau]$, and a filtration $\mathcal{F}_t = \mathcal{F}_0 \cup \sigma\{N(s); s \leq t\}$, with $\mathcal{F} = \mathcal{F}_\tau$. A counting process is a piecewise constant, non-decreasing, stochastic process with jumps of size $+1$. It can be shown to be a local submartingale, with compensator Λ. Suppose P and $\tilde{P}$ are two probability measures on $\mathcal{F}$, for which the two compensators are Λ and $\tilde{\Lambda}$. Suppose $\tilde{P}$ is absolutely continuous with respect to P. If Λ and $\tilde{\Lambda}$ are absolutely continuous a.s. P, then

$$\frac{d\tilde{P}}{dP} = \frac{d\tilde{P}}{dP}\bigg|_{\mathcal{F}_0} \frac{\prod_t \tilde{\lambda}(t)^{\triangle N(t)} \exp\{-\tilde{\Lambda}(\tau)\}}{\prod_t \lambda(t)^{\triangle N(t)} \exp\{-\Lambda(\tau)\}}.$$

Except for the somewhat unfamiliar notation, this is identical to the likelihood function for the non-homogeneous Poisson process (SM, Ex.6.5),

$$\prod_{j=1}^{n} \lambda(t_j) \exp\{-\int_0^\tau \lambda(u)du\}, \quad 0 < t_1 < \cdots < t_n < \tau.$$

References

Andersen, P.K., Borgan, O., Gill, R.D. and Keiding, N. (1993). *Statistical Models Based on Counting Processes.* Springer, New York.
[BNC] Barndorff-Nielsen, O.E. and Cox, D.R. (1994). *Inference and Asymptotics.* Chapman & Hall, London.
[SM] Davison, A.C. (2003). *Statistical Models.* Cambridge University Press, Cambridge.
Schervish, M.J. (1995). *Theory of Statistics.* Springer, New York.