STA2212: Inference and Likelihood

A. Notation

One random variable: Given a model for X which assumes X has a density $f(x;\theta)$, $\quad \theta \in \Theta \subset \mathbb{R}^p$, we have the following definitions:

likelihood function	$L(\theta;x) = c(x)f(x;\theta)$	$\mathcal{L}(\theta)$
log-likelihood function	$\ell(\theta;x) = \log L(\theta;x) = \log f(x;\theta) + a(x)$	
score function	$u(\theta) = \partial\ell(\theta;x)/\partial\theta$	$\ell'(x;\theta)$
observed information function	$j(\theta) = -\partial^2\ell(\theta;x)/\partial\theta\partial\theta^T$	$J(\theta) = \mathrm{E}_\theta\{j(\theta)\}$
expected information (in one observation)	$i(\theta) = \mathrm{E}_\theta\{U(\theta)U(\theta)^T\}$[1]	$I(\theta)$ (p.245)

Independent observations: When we have X_i independent, identically distributed from $f(x_i;\theta)$, then, denoting the observed sample $\boldsymbol{x} = (x_1,\ldots,x_n)$ we have:

likelihood function	$L(\theta;\boldsymbol{x}) = \prod_{i=1}^n f(x_i;\theta)$	$\mathcal{L}(\theta)$
log-likelihood function	$\ell(\theta) = \ell(\theta;\boldsymbol{x}) = \sum_{i=1}^n \ell(\theta;x_i)$	$\ell(\theta)$
maximum likelihood estimate	$\hat\theta = \hat\theta(\boldsymbol{x}) = \arg\sup_\theta \ell(\theta)$	$S(\boldsymbol{X})$
score function	$U(\theta) = \ell'(\theta) = \sum U_i(\theta)$	$\boldsymbol{S}(\theta)$ (p.273)
observed information function	$j(\theta) = -\ell''(\theta) = -\ell''(\theta;\boldsymbol{x})$	$nJ(\theta) = \mathrm{E}_\theta\{-\ell''(x;\theta)\}$
observed (Fisher) information	$j(\hat\theta)$	$n\widehat{I(\theta)}$ (p.254)
expected (Fisher) information	$i(\theta) = \mathrm{E}_\theta\{U(\theta)U(\theta)^T\} = ni_1(\theta)$	$I_n(\theta) = nI(\theta)$

Comments:

1. the maximum likelihood estimate $\hat\theta_n$ is usually obtained by solving the *score equation* $\ell'(\theta;\boldsymbol{x}) = 0$. Lazy notation is $\hat\theta$, but for asymptotics $\hat\theta_n$ is preferred.

2. It doesn't really matter for the definitions above if the observations are independent and identically distributed (i.i.d.), or only independent, but the theorems that are proved in MS Ch. 5 and AoS Ch. 9 assume i.i.d.. What really. matters is that we have a central limit theorem for the first derivative; everything follows from that.

3. There are important distinctions to be careful about in the notation for likelihood and its quantities:

   (a) Are we working with a single observation x, X or n observations $\boldsymbol{x}, \boldsymbol{X}$?

   (b) Do we want to find the distribution of something; so $\ell(\theta;\boldsymbol{X})$ or calculate data summaries; $\ell(\theta;\boldsymbol{x})$?

[1] $U(\theta) = u(\theta;X)$

B. First order asymptotic theory MS §5.4

1. θ is a scalar

If the components of $\boldsymbol{X}$ are i.i.d., then the score function $U(\theta; \boldsymbol{X})$ is a sum of i.i.d. random variables, and we can show that it has expected value 0 and variance $I_n(\theta)$ (or $i(\theta)$ in my notation). Under some regularity conditions on the density $f(x_i; \theta)$ (MS A1-A6, p.245), the central limit theorem gives

$$\frac{U(\theta)}{I_n^{1/2}(\theta)} \xrightarrow{d} N(0,1), \text{ equivalently } \frac{1}{\sqrt{n}} U(\theta) \xrightarrow{d} N\{0, I(\theta)\}. \tag{1}$$

Almost everything else follows from this result and Slutsky's theorem. For example, we can show that

$$(\hat{\theta} - \theta) I_n^{1/2}(\theta) = U(\theta)/I_n^{1/2}(\theta) + o_p(1),$$

where $o_p(1)$ means a remainder term that goes to 0 in probability as $n \to \infty$, so we have the second result

$$(\hat{\theta} - \theta) I_n^{1/2}(\theta) \xrightarrow{d} N(0,1). \tag{2}$$

These limit theorems give us two corresponding approximations to use with n fixed:

$$U(\theta) \dot{\sim} N\left(0, I_n(\theta)\right), \tag{3}$$

and

$$\hat{\theta} - \theta \dot{\sim} N\left(0, 1/I_n(\theta)\right). \tag{4}$$

The notation $\dot{\sim}$ is read as "is approximately distributed as".

The proof of MS Theorem 5.3 allows that $I(\theta) = \text{var}_\theta\{\ell'(\theta; X_i)\}$ and $J(\theta) = \text{E}_\theta\{-\ell''(\theta); X_i\}$ might be different, which is handy later for the study of misspecified models.

Having the unknown quantity θ in the variance in (3) and (4) is inconvenient, but to the same order of approximation, we can replace $I_n(\theta)$ by $I_n(\hat{\theta})$ or by $j(\hat{\theta})$; the latter is denoted $n\widehat{I(\theta)}$ in MS, p. 254. In AoS, $I_n^{-1/2}(\theta)$ is called `se` and $I_n^{-1/2}(\hat{\theta})$ is called $\widehat{\texttt{se}}$, but the use of $j(\hat{\theta}) = -\ell''(\hat{\theta}; \boldsymbol{x})$ is not mentioned. The approximation with the Hessian at the maximum (in my notation $j(\hat{\theta}) = -\ell''(\hat{\theta}; \boldsymbol{x})$ is then

$$\hat{\theta} - \theta \dot{\sim} N\{0, j^{-1}(\hat{\theta})\}, \tag{5}$$

Proof of (2)

The score equation $\ell'(\hat{\theta}) = 0$ is expanded using Taylor series with remainder:

$$0 = \ell'(\hat{\theta}) = \ell'(\theta) + (\hat{\theta} - \theta)\ell''(\theta) + \frac{1}{2}(\hat{\theta} - \theta)^2 \ell'''(\theta_n^*), \tag{6}$$

where $|\theta_n^* - \theta| \leq |\hat{\theta} - \theta|$ in the remainder term. Re-arranging terms gives

$$-\frac{\ell'(\theta)}{\ell''(\theta)} = (\hat{\theta} - \theta)\left\{1 + \frac{1}{2}(\hat{\theta} - \theta)\frac{\ell'''(\theta_n^*)}{\ell''(\hat{\theta})}\right\}, \tag{7}$$

and introducing dependence on n explicitly gives

$$\begin{aligned}\frac{\ell'(\theta)/\sqrt{n}}{-\ell''(\theta)/n}\cdot\sqrt{n}\frac{I(\theta)}{I^{1/2}(\theta)} &= \sqrt{n}I^{1/2}(\theta)(\hat\theta-\theta)\left\{1-\frac{1}{2}(\hat\theta-\theta)\frac{\ell'''(\theta_n^*)/n}{-\ell''(\hat\theta)/n}\right\},\\ \frac{\ell'(\theta)}{\{nI(\theta)\}^{1/2}}\left\{\frac{I(\theta)}{-\ell''(\theta)/n}\right\} &= \{nI(\theta)\}^{1/2}(\hat\theta-\theta)(1+Z_n), \qquad (8)\end{aligned}$$

where Z_n is shorthand for the remainder term.

The term in brackets on the left hand side of (8) converges in probability to 1, by the weak law of large numbers. To see that the remainder term Z_n converges in probability to 0, we assume $\hat\theta$ converges in probability to θ, which implies that θ_n^* does as well. The limit in probability of $\ell'''(\theta_n^*)/n$ is $\mathrm{E}_\theta\{\ell'''(\theta;Y)\}$and that of $-\ell''(\hat\theta)/n$ is $I_1(\theta)$, which gives the desired result. Thus

$$\sqrt{n}(\hat\theta-\theta)I^{1/2}(\theta)=\frac{1}{\sqrt{n}}\frac{\ell'(\theta)}{I^{1/2}(\theta)}\{1+o_p(1)\}, \qquad (9)$$

showing that the standardized maximum likelihood estimator converges to a standard normal distribution.

An equivalent result to (2) is

$$\sqrt{n}(\hat\theta-\theta)\xrightarrow{d}N\{0,I(\theta)\},$$

always remembering that $I(\theta)$ is the expected information with one observation, and $I_n(\theta)=nI(\theta)$ is the total information in the sample $(X_1,\ldots,X_n)$. This follows from (7) using Slutsky's theorem, or from (8) multiplying both sides by $I(\theta)$.

It wouldn't be correct though to write $\hat\theta\xrightarrow{d}N\{\theta,I_n^{-1}(\theta)\}$, even though it sort of looks right, because both the LHS and RHS depend on n. In asymptotic theory, you have to get all the n-dependence on one side of the limit.

2. θ is a vector of length p MS 5.4 pp256ff

The results above all generalize directly to a vector θ of unknown parameters. The notation on p.1 already includes this case. The score function is a $p\times 1$ vector and the observed and expected Fisher information are $p\times p$ matrices. The limit theorems corresponding to (1) and (2) are

$$I_n^{-1/2}(\theta)U_n(\theta)\xrightarrow{d}N_p(0,\mathcal{I}_p),\quad I_n^{1/2}(\theta)(\hat\theta-\theta)\xrightarrow{d}N_p(0,\mathcal{I}_p), \qquad (10)$$

where $N_p(0,\mathcal{I}_p)$ is the multivariate standard normal distribution and $\mathcal{I}_p$ is the $p\times p$ identity matrix. Because this limit statement involves taking the square root of the matrix I_n, the results in (5) are rarely used in this form. So Theorem 5.4 in MS is stated as

$$\sqrt{n}(\hat\theta-\theta)\xrightarrow{d}N_p\{0,J^{-1}(\theta)I(\theta)J^{-1}(\theta)\}; \qquad (11)$$

where $J(\theta)=\mathrm{E}_\theta\{\ell''(\theta;X_1)\}$ and $I(\theta)=\mathrm{var}\{\ell'(\theta;X_1)\}$, i.e. the information about θ in a single observation. The notation ℓ' and ℓ'' is lazy shorthand for the vector of

first derivatives and matrix of second derivatives, and "var" is short for variance-covariance matrix. Under our regularity conditions on the density $f(x;\theta)$, $I(\theta) = J(\theta)$ and (11) becomes

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N_p\{0, I^{-1}(\theta)\} \tag{12}$$

The approximation obtained from this result is

$$\hat{\theta} \dot{\sim} N\{\theta, j^{-1}(\hat{\theta})\}, \tag{13}$$

or in MS notation (see (5) above)

$$\hat{\theta} \dot{\sim} N\left(\theta, \{n\widehat{I(\theta)}\}^{-1}\right). \tag{14}$$

This approximation is for the whole vector $\hat{\theta}$ but that's not so useful in practice. However we can specialize the result to a single component, giving, for example,

$$\hat{\theta}_j - \theta_j \dot{\sim} N\left(0, \{j^{-1}(\hat{\theta})\}_{jj}\right), \tag{15}$$

i.e. the jth diagonal element of the inverse matrix is the approximate variance of the jth component of the vector $\hat{\theta}$. We also have that $\{j^{-1}(\hat{\theta})\}_{jk}$ is the asymptotic covariance of $\hat{\theta}_j, \hat{\theta}_k$.