STA3000: Asymptotic theory for likelihood

Assume we have a sample $Y = (Y_1, \ldots, Y_n)$, where the Y_i are independent, identically distributed with density $f(y_i; \theta)$. Refer to the handout of October 4 for the definitions and orders of magnitude of the score function, maximum likelihood estimate, observed and expected Fisher information. Also there we give the first order theory for θ in the case that θ is a vector of length k, as well as the special case $k = 1$. The vector version results are repeated here:

$$\begin{aligned}
\frac{1}{\sqrt{n}}\{U(\theta)\} &\overset{d}{\to} N_k(0, i_1(\theta)) && (1)\\
\sqrt{n}(\hat\theta - \theta) &= \frac{1}{\sqrt{n}} i_1^{-1}(\theta)U(\theta)\{1 + o_p(1)\}, && (2)\\
2\{\ell(\hat\theta) - \ell(\theta)\} &= (\hat\theta - \theta)^T i(\theta)(\hat\theta - \theta)\{1 + o_p(1)\} && (3)
\end{aligned}$$

from which we have the approximations

$$\begin{aligned}
w_u(\theta) = U(\theta)^T\{i(\theta)\}^{-1}U(\theta) &\mathrel{\dot\sim} \chi^2_k, && (4)\\
w_e(\theta) = (\hat\theta - \theta)^T i(\theta)(\hat\theta - \theta) &\mathrel{\dot\sim} \chi^2_k, && (5)\\
w(\theta) = 2\{\ell(\hat\theta) - \ell(\theta)\} &\mathrel{\dot\sim} \chi^2_k. && (6)
\end{aligned}$$

A moderately rigorous proof of (2) and (3) follows, for scalar θ. The vector case is unchanged, except for tedious notational changes in the Taylor series, although of course we need the dimension of θ fixed as $n \to \infty$.

For (2), we have

$$\begin{aligned}
\ell'(\hat\theta) &= \ell'(\theta) + (\hat\theta - \theta)\ell''(\theta) + \frac{1}{2}(\hat\theta - \theta)^2\ell'''(\theta_n^*),\\
-\frac{\ell'(\theta)}{\ell''(\theta)} &= (\hat\theta - \theta)\{1 + \frac{1}{2}(\hat\theta - \theta)\frac{\ell'''(\theta_n^*)}{\ell''(\hat\theta)}\},\\
\frac{\frac{1}{\sqrt{n}}\ell'(\theta)}{-\ell''(\theta)/n} \cdot \frac{i_1(\theta)}{i_1(\theta)} &= \sqrt{n}(\hat\theta - \theta)\{1 - \frac{1}{2}(\hat\theta - \theta)\frac{\ell'''(\theta_n^*)/n}{-\ell''(\theta)/n}\},\\
\frac{\frac{1}{\sqrt{n}}\ell'(\theta)}{i_1(\theta)}\left(\frac{i_1(\theta)}{-\ell''(\theta)/n}\right) &= \sqrt{n}(\hat\theta - \theta)\{1 + Z_n\}.
\end{aligned}$$

The term in brackets on the LHS of the last line converges in probability to 1, by the WLLN, so can be written $1 + o_p(1)$. The remainder term Z_n converges in probability to 0, because we assume $\hat\theta \overset{p}{\to} \theta$, so that $\theta_n^* \overset{p}{\to} \theta$, because $|\hat\theta - \theta| < |\theta_n^* - \theta|$. Also $\frac{1}{n}\ell'''(\theta_n^*) \overset{p}{\to} E\{\ell'''(\theta; Y)\}$ which we assume is finite (p.281 of CH, for example); similarly $-\frac{1}{n}\ell''(\theta) \overset{p}{\to} i_1(\theta)$, so $Z_n = o_p(1)O(1) = o_p(1)$. Then we can move over the LHS term as

$$\frac{1}{\sqrt{n}}\frac{\ell'(\theta)}{i_1(\theta)}\{1 + o_p(1)\} = \sqrt{n}(\hat\theta - \theta),$$

because $1 + o_p(1)$ is the same as $1 - o_p(1)$, and $\{1 + o_p(1)\}^{-1} = 1 - o_p(1)$.

For (3), we have

$$\begin{aligned}
\ell(\theta) &= \ell(\hat\theta) + (\theta - \hat\theta)\ell'(\hat\theta) + \frac{1}{2}(\theta - \hat\theta)^2\ell''(\hat\theta) + \frac{1}{6}(\theta - \hat\theta)^3\ell'''(\theta_n^*), \\
\ell(\hat\theta) - \ell(\theta) &= \frac{1}{2}(\hat\theta - \theta)^2\{-\ell''(\hat\theta)\} + \frac{1}{6}(\hat\theta - \theta)^3\ell'''(\theta_n^*), \\
2\{\ell(\hat\theta) - \ell(\theta)\} &= (\hat\theta - \theta)^2 i(\theta)\{-\ell''(\hat\theta)/i(\theta)\}\{1 + \frac{1}{3}(\hat\theta - \theta)\frac{\ell'''(\theta_n^*)}{-\ell''(\hat\theta)}\}, \\
&= (\hat\theta - \theta)^2 i(\theta)(1 + Z_n)
\end{aligned}$$

where again $Z_n \overset{p}{\to} 0$ as above.

This begs the question of whether the maximum likelihood estimator is the root of $\ell'(\hat\theta) = 0$, and whether the maximum likelihood estimator converges in probability to θ. Wald's proof of the consistency of the MLE relies on showing (roughly) that the likelihood function is maximized at the true value, in the limit, so that the parameter point that maximizes the likelihood function will converge to that true value. However the devil is in the details. A good discussion is given in Knight, Ch.4; there is more detail in van der Waart, Ch. 5.

An easier approach is to assume enough about the density to be able to prove that there are consistent solutions to the score equation; then if the likelihood function has its maximum in the interior of the parameter space, and the solution to the score equation is unique, it is the MLE. TPE gives the details for this approach; I also found the encylopedia article by Scholz very helpful.

BNC avoid all these problems by just assuming that the score equation gives the MLE, and 'enough regularity' on the model to ensure consistency. After that asymptotic normality follows if one has a central limit theorem for the score function. This can hold much more generally that in the i.i.d. setting.

References

[BNC] Barndorff-Nielsen & Cox (1994). *Inference and Asymptotics.*

[TPE] Lehmann, E.L. and Casella, G. (2003). *Theory of Point Estimation.*

Knight, K. (2002). *Mathematical Statistics.*

Scholz, F. (2006). *Encyclopedia of Statistical Sciences*: Maximum likelihood estimation.

Van der Waart, A. (1998). *Asymptotic Statistics.*