Statistical Asymptotics
Part II: First-Order Theory

Andrew Wood

School of Mathematical Sciences
University of Nottingham

APTS, March 23-27, 2015

Structure of the Chapter

This chapter covers asymptotic normality and related results.

Topics: MLEs, log-likelihood ratio statistics and their asymptotic distributions; M-estimators and their first-order asymptotic theory.

Initially we focus on the case of the MLE of a scalar parameter θ.

Then we study the case of the MLE of a vector θ, first without and then with nuisance parameters.

Finally, we consider the more general setting of M-estimators.

Motivation

Statistical inference typically requires approximations because exact answers are usually not available.

Asymptotic theory provides useful approximations to densities or distribution functions.

These approximations are based on results of probability theory.

The theory underlying these approximation techniques is valid as some quantity, typically the sample size n [or more generally some 'amount of information'], goes to infinity, but the approximations obtained are often accurate even for small sample sizes.

Test statistics

To test the null hypothesis $H_0 : \theta = \theta_0$, where θ_0 is an arbitrary, specified, point in Ω_θ. If desired, we may think of θ_0 as the 'true' value of the parameter, but this is not necessary.

Three statistics that typically differ by $O_p(n^{-1/2})$ are:

(1) the likelihood ratio statistic (also known as the Wilks statistic)

$$w(\theta_0) = 2\{l(\hat{\theta}) - l(\theta_0)\},$$

(2) the score statistic

$$w_U(\theta_0) = U(\theta_0)^\top i(\theta_0)^{-1} U(\theta_0),$$

(3) the Wald statistic

$$w_p(\theta_0) = (\hat{\theta} - \theta_0)^T i(\theta_0)(\hat{\theta} - \theta_0).$$

Scalar case

For a scalar θ, (1) may be replaced by

$$r(\theta_0) = \mathrm{sgn}(\hat{\theta} - \theta_0)\sqrt{w(\theta_0)},$$

the signed root likelihood ratio statistic.

Also (2) and (3) may be replaced by

$$r_U(\theta_0) = U(\theta_0)/\sqrt{i(\theta_0)}$$

and

$$r_p(\theta_0) = (\hat{\theta} - \theta_0)\sqrt{i(\theta_0)}$$

respectively.

Asymptotic normality of MLE: scalar case

We stick with the scalar θ case for the moment, and also assume that the sample is IID.

Our immediate goal is to give a careful explanation of why, in broad generality,

$$n^{1/2}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, \bar{i}(\theta_0)^{-1}), \tag{1}$$

where

- $\hat{\theta}$ and θ_0 are, respectively, the MLE and the 'true' value of θ;
- $\bar{i}(\theta)$ is the Fisher information matrix for a single observation

Taylor expansion: scalar case

In regular settings, $\hat{\theta}$ solves $U(\hat{\theta}) = 0$. Assuming $l(\theta)$ has a continuous 3rd derivative, we may use Taylor's theorem to obtain

$$0 = U(\hat{\theta}) = U(\theta_0) + (\hat{\theta} - \theta_0)\frac{\partial^2 l}{\partial \theta^2}(\theta_0) + \frac{1}{2}(\hat{\theta} - \theta_0)^2 \frac{\partial^3 l}{\partial \theta^3}(\theta^*)$$

$$= n^{-1/2} U(\theta_0) - \bar{j}(\theta_0) n^{1/2}(\hat{\theta} - \theta_0) + n^{-1/2} R^*, \qquad (2)$$

where θ^* lies between $\hat{\theta}$ and θ_0, $\bar{j}(\theta_0) = n^{-1} j(\theta)$ is the (sample) mean observed information at θ_0, and the remainder term R^* is given by

$$R^* = \frac{1}{2}(\hat{\theta} - \theta_0)^2 \frac{\partial^3 l}{\partial \theta^3}(\theta^*). \qquad (3)$$

Taylor expansion: scalar case (continued)

Assuming $j(\theta_0)$ is non-zero, we may rearrange (2) to obtain

$$n^{1/2}(\hat{\theta} - \theta_0) = \bar{j}(\theta_0)^{-1} n^{-1/2} U(\theta_0) + \bar{j}(\theta_0)^{-1} n^{-1/2} R^*. \qquad (4)$$

We will show that

- the first term on the RHS of (4) converges in distribution to $N(0, \bar{i}(\theta_0)^{-1})$;
- the second term is $O_p(n^{-1/2})$.

Then, applying Slutsky's theorem, we obtain (1).

CLT for Score Statistic

When the observations $X_1, \ldots, X_n$ are IID, the score statistic at θ_0,

$$U(\theta_0) = \sum_{i=1}^{n} \frac{\partial \log f}{\partial \theta}(X_i|\theta_0),$$

is the sum of IID random variables.

Moreover, we know from standard likelihood theory that

$$E_{\theta_0}[U(\theta_0)] = 0 \quad \text{and} \quad \mathrm{Var}_{\theta_0}[U(\theta_0)] = i(\theta_0) = n\bar{i}(\theta_0).$$

Hence, by an application of the CLT, we conclude that

$$n^{-1/2} U(\theta_0) \xrightarrow{d} N(0, \bar{i}(\theta_0)).$$

Law of large numbers for observed information

When the sample $X_1, \ldots, X_n$ is IID from f, the observed information is given by

$$j(\theta_0) = -\sum_{i=1}^{n} \frac{\partial^2 \log f}{\partial \theta^2}(X_i|\theta_0),$$

and so is a sum of IID random variables. Moreover, we know that, from the definition of Fisher information,

$$E_{\theta_0}[j(\theta_0)] = i(\theta_0) = n\bar{i}(\theta_0).$$

Hence, by the Weak Law of Large Numbers,

$$n^{-1} j(\theta_0) \xrightarrow{p} \bar{i}(\theta_0).$$

Remainder term

In broad generality, R^* in (3) is $O_p(1)$, i.e. bounded in probability.

To see that this is plausible, note that when

$$n^{1/2}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, \bar{i}(\theta_0)^{-1}),$$

we have $(\hat{\theta} - \theta_0)^2 = O_p(n^{-1})$.

Also, $\frac{\partial^3 l}{\partial \theta^3}$ is a sum of n terms.

Hence under reasonable conditions we can hope that

$$R^* = O_p(n^{-1})O_p(n) = O_p(1).$$

However, providing a general proof is quite a challenging problem. Below, we outline an approach which works in many cases.

Final step

Assuming that $R^* = O_p(1)$ and putting the last three slides together, we see that

$$n^{1/2}(\hat{\theta} - \theta_0) = \bar{j}(\theta_0)^{-1} n^{-1/2} U(\theta_0) + \bar{j}(\theta_0)^{-1} n^{-1/2} R^*$$

$$\xrightarrow{d} \bar{i}(\theta_0)^{-1} N(0, \bar{i}(\theta_0)) + O_p(n^{-1/2}) + O_p(1).n^{-1/2}.O_p(1)$$

$$\xrightarrow{d} N(0, \bar{i}(\theta_0)^{-1})$$

where Slutsky's theorem has been used in the final step.

Similar reasoning is used in the multivariate case, the main differences being that the multivariate versions of the CLT, WLLN and Taylor's theorem are used.

Exponential example

Here we have

$$l(\lambda) = n\log(\lambda) - \lambda\sum_{k=1}^{n} x_k, \qquad U(\lambda) = n\lambda^{-1} - \sum_{k=1}^{n} X_k,$$

$j(\lambda) = i(\lambda) = n\lambda^{-2}$, $\bar{j}(\lambda) = \bar{i}(\lambda) = \lambda^{-2}$, and

$$R^* = \frac{1}{2}(\hat{\lambda} - \lambda_0)^2 \frac{2n}{(\lambda^*)^3},$$

where λ^* lies between $\hat{\lambda}$ and λ_0.

Since $n^{1/2}(\hat{\lambda} - \lambda_0) \xrightarrow{d} N(0, \bar{i}(\lambda_0)^{-1})$, it follows that $\lambda^* \xrightarrow{p} \lambda_0$, and so we may conclude that $R^* = O_p(1)$.

Proving that $R^* = O_p(1)$

A useful general strategy for proving that $R^* = O_p(1)$, and hence asymptotic normality of the MLE, is as follows.

Step 1. Show that there there exists a consistent solution $\hat{\theta}$ of the score equation $U(\hat{\theta}) = 0$.

Step 2. Use a *uniform* law of large numbers on a bounded neighbourhood of θ_0 to prove that $R^* = O_p(1)$.

For results relating to Step 1, see for example van der Vaart (1998, Chapter 5).

For Step 2, a version of the uniform law of large numbers (ULLN) is now given. See, for example, Jennrich (1969, Annals of Mathematical Statistics).

Note: stronger versions of the result that we state here are available.

Uniform law of large numbers

Let F denote a distribution function on $\mathcal{X} \subseteq \mathbb{R}^p$, and let $X_1, \ldots, X_n$ denote an IID sample from F.

Consider a vector-valued function $g(x, \theta)$ which is continuous as a function of $\theta \in \Theta \subseteq \mathbb{R}^d$ for each fixed x, and measureable as a function of $x \in \mathcal{X} \subseteq \mathbb{R}^p$ for each fixed θ.

Suppose that, for some bounded open set $\Delta \subset \Theta$, the following statements hold:

$$\text{(i)} \quad \sup_{\theta \in \Delta} ||g(x, \theta)|| \leq h(x), \quad \text{and} \quad \text{(ii)} \quad \int_{x \in \mathcal{X}} h(x) dF(x) < \infty.$$

Write $\tau(\theta) = \int_{x \in \mathcal{X}} g(x, \theta) dF(x)$. Then $\tau(\theta)$ is continuous on Δ and

$$\sup_{\theta \in \Delta} \left|\left| n^{-1} \sum_{k=1}^{n} g(X_i, \theta) - \tau(\theta) \right|\right| \xrightarrow{p} 0 \quad \text{as} \quad n \to \infty.$$

Application of the ULLN to the remainder

We now return to the scalar θ case.

Assume that $\hat{\theta}$ is a consistent estimator of θ_0 and define

$$g(x, \theta) = \frac{\partial^3 \log f}{\partial \theta^3}(x|\theta).$$

Suppose we can find a small (in particular, bounded) open neighbourhood Δ of θ_0 such that

$$\sup_{\theta \in \Delta} |g(x, \theta)| \leq h(x),$$

where $E_{\theta_0}[h(X)] < \infty$.

Application of ULLN (continued)

Then, writing $\bar{g}(\theta) = n^{-1}\sum_{k=1}^{n} g(X_k, \theta)$, we have the identity

$$\frac{\partial^3 l}{\partial \theta^3}(\theta^*) = \tau(\theta^*) + \{\bar{g}(\theta^*) - \tau(\theta^*)\}.$$

Consequently,

$$\left| n^{-1}\frac{\partial^3 l}{\partial \theta^3}(\theta^*) \right| \leq |\tau(\theta^*)| + \sup_{\theta \in \Delta} |\bar{g}(\theta^*) - \tau(\theta^*)|.$$

If $\hat{\theta}$ is consistent, then the first term on the RHS converges to $\tau(\theta_0)$ by the continuous mapping theorem, and the second term is $o_p(1)$ by the ULLN.

Therefore, in this case, $n^{-1}\partial^3 l(\theta^*)/\partial\theta^3 = O_p(1)$ and, consequently, $R^* = O_p(1)$ as required.

Application of ULLN to information

In many situations, we would like to be able to conclude that

$$\bar{j}(\hat{\theta}) \equiv n^{-1} j(\hat{\theta}) \xrightarrow{p} \bar{i}(\theta_0). \tag{5}$$

This can be established provided we can show that

- $\hat{\theta}$ is a consistent estimator of θ_0; and
- we can find a suitable bounding function which enables an application of the ULLN.

Moreover, a stronger result typically holds: that the differences between each of the quantities $\bar{i}(\theta_0)$, $\bar{i}(\hat{\theta})$, $\bar{j}(\hat{\theta})$, $\bar{j}(\theta_0)$ is $O_p(n^{-1/2})$.

Orders

In the IID case, and often more generally, the following order statements are valid as $n \to \infty$:

$$\begin{aligned} U(\theta_0) &\equiv n^{1/2}\bar{U}(\theta_0) = O_p(n^{1/2}), \\ i(\theta_0) &\equiv n\bar{i}(\theta_0) = O(n), \\ \hat{\theta} - \theta_0 &= O_p(n^{-1/2}), \end{aligned}$$

where $\bar{i}(\theta_0)$ is the average information per observation and $\bar{U}(\theta_0) = n^{-1/2}U(\theta_0)$ is a normalised score function. If the observations are IID, $\bar{i}$ is the information for a single observation.

The three likelihood statistics

We shall now investigate the first-order behaviour of the 3 likelihood statistics under the assumption that

$$n^{1/2}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, \bar{i}(\theta_0)^{-1}).$$

First let us consider Wilks's statistic $w(\theta_0) = 2\{l(\hat{\theta}) - l(\theta_0)\}$.

Taylor expanding $l(\theta_0)$ about $l(\hat{\theta})$ this time, we obtain

$$l(\theta_0) = l(\hat{\theta}) + (\theta_0 - \hat{\theta})\frac{\partial l}{\partial \theta}(\hat{\theta}) + \frac{1}{2!}(\theta_0 - \hat{\theta})^2 \frac{\partial^2 l}{\partial \theta^2}(\hat{\theta}) + \frac{1}{3!}(\theta_0 - \hat{\theta})^3 \frac{\partial^3 l}{\partial \theta^3}(\theta^{**}), \quad (6)$$

where θ^{**} lies between $\hat{\theta}$ and θ_0.

The three likelihood statistics (continued)

Noting that, in regular models, $\partial l(\theta)/\partial\theta = 0$ at $\theta = \hat{\theta}$, and substituting $j(\hat{\theta}) = -\partial^2 l(\hat{\theta})/\partial\theta^2$, we may rearrange (6) to obtain

$$2\{l(\hat{\theta}) - l(\theta_0)\} = (\hat{\theta} - \theta_0)^2 j(\hat{\theta}) + R^{**},$$

where

$$R^{**} = \frac{1}{3}(\hat{\theta} - \theta_0)^3 \frac{\partial^3 l}{\partial\theta^3}(\theta^{**}).$$

Using the strategy outlined before, i.e. prove consistency and then apply a suitable ULLN, it can be shown that in broad generality, $R^{**} = O_p(n^{-1/2})$.

The three likelihood statistics (continued)

Similar arguments, broadly applicable, show that

$$n^{-1} j(\hat{\theta}) \xrightarrow{p} \bar{i}(\theta_0),$$

which implies that Wilks's statistics $w(\theta_0)$ differs from the Wald statistic $W_p(\theta_0)$ by $O_p(n^{-1/2})$.

Equivalence up to $O_p(n^{-1/2})$ with the score statistic,

$$w_U(\theta_0) = U(\theta_0)^2 / i(\theta_0),$$

follows from the result established earlier that

$$n^{1/2}(\hat{\theta} - \theta_0) = \bar{i}(\theta_0)^{-1} n^{-1/2} U(\theta_0) + O_p(n^{-1/2}).$$

The three likelihood statistics (continued)

Finally, we show that the 3 likelihood statistics have χ^2_1 distributions when θ is a scalar.

In view of the equivalence up to $O_p(n^{-1/2})$ of the 3 statistics, it is sufficient just to consider the score statistic.

We have already seen that, in the IID case,
$n^{-1/2}U(\theta_0) \xrightarrow{d} N(0, \bar{i}(\theta_0))$.

Therefore, by the continuous mapping theorem,

$$\{n^{-1/2}U(\theta_0)\}^2/\bar{i}(\theta_0) = U(\theta_0)^2/i(\theta_0) \xrightarrow{d} \chi^2_1.$$

Signed root statistic

When θ is a scalar, the signed root likelihood ratio statistic

$$r = \mathrm{sgn}(\hat{\theta} - \theta)\sqrt{w(\theta)}$$

satisfies

$$r = \hat{j}^{-1/2} U + o_p(n^{-1/2})$$

so that $r \xrightarrow{d} N(0,1)$.

A Confidence Interval

For scalar θ, we have $i(\hat{\theta})^{1/2}(\hat{\theta} - \theta)$ asymptotically $N(0, 1)$, so an approximate $100(1 - \alpha)\%$ confidence interval for θ is

$$\hat{\theta} \mp i(\hat{\theta})^{-1/2}\Phi^{-1}(1 - \alpha/2).$$

Exponential example (continued)

Let us return to the exponential example, where $x_1, \ldots, x_n$ is a sample, assumed IID, from the pdf $\lambda e^{-\lambda x}$, where $\lambda > 0$ and $x > 0$.

Writing $\bar{x} = n^{-1} \sum_{k=1}^{n} x_k$ and solving $U(\hat{\lambda}) = 0$, we find that $\hat{\lambda} = \bar{x}^{-1}$. Recall that $i(\lambda_0) = n\lambda_0^{-2}$.

Some elementary calculations show that

$$w(\lambda_0) = 2n\{\lambda_0 \bar{x} - 1 - \log(\lambda_0 \bar{x})\}, \qquad w_U(\lambda_0) = n(\lambda_0 \bar{x} - 1)^2$$

and

$$w_P(\lambda_0) = n\{(\lambda_0 \bar{x})^{-1} - 1\}^2.$$

Exponential example (continued)

Note that all 3 statistics are different. In fact, $w(\lambda_0)$ and $w_U(\lambda_0)$ are closer than they first appear.

If we approximate the log in $w(\lambda_0)$ by

$$\log(\lambda_0\bar{x}) = \log\{1 + (\lambda_0\bar{(}x) - 1)\} \approx (\lambda_0\bar{x} - 1) - \frac{1}{2}(\lambda_0\bar{x} - 1)^2,$$

then the resulting approximation to $w(\lambda_0)$ is identical to $w_U(\lambda_0)$.

Some practical comments

Typically, in practical work, the performance of the Wilks statistic $w(\theta_0)$ is superior to that of the other two statistics.

However, $w_U(\theta_0)$ has the advantage that the MLE is not required, and $w_P(\theta_0)$ is often convenient to use because it depends on quantities which are routinely available from a model fit.

A further point is that $w(\theta_0)$ and $w_U(\theta_0)$ are both invariant with respect to a $1:1$ transformation of θ, whereas the Wald statistic does not possess this invariance.

Furthermore, the choice of parametrisation of the Wald statistic can have a big effect on the value of the statistic, especially when approaching a boundary of the parameter space.

Some practical comments (continued)

When the sample size is moderate, the normal approximation to the distribution of the signed likelihood root r is, in practice, typically at least as good as or better than the normal approximation to the distribution of the MLE.

Non-uniqueness of MLEs

A serious problem, which we have not considered, is that in general multiple solutions of the score equation $U(\hat{\theta}) = 0$ may exist, even asymptotically.

A class of models where such non-uniqueness typically occurs is the curved exponential family, which was briefly introduced in Chapter 1.

Non-uniqueness complicates the asymptotic theory of MLEs, especially statements about consistency.

In these lectures we shall just acknowledge the existence of the problem and not investigate it further.

Vector parameter θ: preliminary comments

We now move on to vector θ, first without and then with nuisance parameters.

The intuitions and techniques underlying the proofs of asymptotic normality of the MLE and the χ^2 distribution of the 3 likelihood statistics in the case of vector θ are very similar to those used in the case of scalar θ.

All that changes is that we work with multivariate versions of Taylor's expansion, the CLT and the WLLN, and we are now dealing with vectors and matrices.

However, it is inevitable that the notation becomes more complex in the multivariate case. We shall make use of a summation convention from time to time in order to avoid over-long formulae.

Vector θ, no nuisance parameters

Denote by l_r the rth component of $U(\theta)$, by l_{rs} the (r,s)th component of $\nabla_\theta \nabla_\theta^T l$. Let $[l_{rs}]^{-1} = [l^{rs}]$.

The maximum likelihood estimate for given observations y is, for regular problems, defined as the solution, assumed unique, of the likelihood equation

$$U(\hat{\theta}) = 0.$$

where now $U(\theta)$ is a d-dimensional vector.

Test statistics

To test the null hypothesis $H_0 : \theta = \theta_0$, where θ_0 is an arbitrary, specified, point in Ω_θ. If desired, we may think of θ_0 as the 'true' value of the parameter, but this is not necessary.

Three statistics that typically differ by $O_p(n^{-1/2})$ are:

(1) the likelihood ratio statistic (also known as the Wilks statistic)

$$w(\theta_0) = 2\{l(\hat{\theta}) - l(\theta_0)\},$$

(2) the score statistic

$$w_U(\theta_0) = U(\theta_0)^\top i(\theta_0)^{-1} U(\theta_0),$$

(3) the Wald statistic

$$w_p(\theta_0) = (\hat{\theta} - \theta_0)^T i(\theta_0)(\hat{\theta} - \theta_0).$$

Distributions

In a first-order asymptotic theory, the likelihood statistics (1)–(3) have, asymptotically, the chi-squared distribution with $d_\theta = \dim(\Omega_\theta)$ degrees of freedom.

Confidence regions at level $1 - \alpha$ are formed approximately as, for example,

$$\{\theta : w(\theta) \leq \chi^2_{d_\theta,\alpha}\},$$

where $\chi^2_{d_\theta,\alpha}$ is the upper α point of the relevant chi-squared distribution.

In considerable generality U is asymptotically multivariate normal with zero mean and variance $i(\theta)$.

In the IID case, i.e. when $Y = (Y_1, \ldots, Y_n)^\top$ and $Y_1, \ldots, Y_n$ are IID random vectors, then $U(\theta)$ is a sum of n IID vectors, each of which has mean 0 and covariance matrix $\bar{i}(\theta)$.

Therefore we may apply the multivariate CLT to obtain

$$\{n\bar{i}(\theta)\}^{-1/2} U(\theta) \xrightarrow{d} N_d\big(0, I_d\big),$$

I_d is identity matrix, '$1/2$' indicates matrix square root.

Note: If Σ is symmetric and positive definite with spectral decomposition $\Sigma = Q\Lambda Q^\top$, where $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_d)$, then we define $\Sigma^{1/2} = Q\Lambda^{1/2}Q^\top$ where $\Lambda^{1/2} = \mathrm{diag}(\lambda_1^{1/2}, \ldots, \lambda_d^{1/2})$, and $\Sigma^{-1/2} = Q\Lambda^{-1/2}Q^\top$ where $\Lambda^{-1/2} = (\Lambda^{1/2})^{-1}$.

An aside: summation convention

Whenever an index occurs both as a subscript and as a superscript in an expression, summation over possible values of that index is to be assumed.

Distribution of $\hat{\theta}$

Expand the score $l_r(\theta)$ in a Taylor series around θ, writing

$$\begin{aligned} l_r(\theta) &= U_r(\theta) = \sqrt{n}\bar{l}_r(\theta) = \sqrt{n}\bar{U}_r(\theta), \\ l_{rs}(\theta) &= n\bar{l}_{rs}(\theta) = -j_{rs}(\theta) = -n\bar{j}_{rs}(\theta), \\ \bar{\delta}^r &= \sqrt{n}(\hat{\theta}^r - \theta^r), l_{rst}(\theta) = n\bar{l}_{rst}(\theta), \\ i(\theta) &= n\bar{i}(\theta), \text{ etc.} \end{aligned}$$

Then, $l_r(\hat{\theta}) = 0$, so

$$\begin{aligned} \sqrt{n}\bar{l}_r(\theta) &+ n\bar{l}_{rs}(\theta)\bar{\delta}^s/\sqrt{n} \\ &+ \tfrac{1}{2}n\bar{l}_{rst}(\theta)\bar{\delta}^s\bar{\delta}^t/n + \cdots = 0. \end{aligned}$$

To a first-order approximation, ignoring the third term, we have

$$\begin{aligned}\bar{\delta}^r &= -\bar{I}^{rs}(\theta)\bar{l}_s(\theta) + O_p(n^{-1/2}) \\ &= \bar{j}^{rs}(\theta)\bar{l}_s(\theta) + O_p(n^{-1/2}).\end{aligned}$$

Now $j^{rs}/i^{rs} \xrightarrow{p} 1$, so

$$\bar{\delta}^r = \bar{i}^{rs}(\theta)\bar{l}_s(\theta) + O_p(n^{-1/2}),$$

a linear function of asymptotically normal variables of zero mean. It follows that $[\bar{\delta}^r]$ is asymptotically normal with zero mean and covariance matrix $[\bar{i}^{rs}]$. We have

$$\{n\bar{i}(\theta)\}^{1/2}(\hat{\theta} - \theta) \xrightarrow{d} N_d\big(0, I_d\big).$$

Other quantities

By direct expansion of log-likelihood in θ around $\hat{\theta}$ we obtain, writing $\hat{j}_{rs} = j_{rs}(\hat{\theta})$,

$$w(\theta) = \hat{j}_{rs}(\hat{\theta} - \theta)^r(\hat{\theta} - \theta)^s + o_p(1)$$

or equivalently

$$w(\theta) = i^{rs} l_r l_s + o_p(1),$$

so $w(\theta) \xrightarrow{d} \chi^2_d$.

The asymptotic χ^2 distribution of the Wald and score statistics follows similarly.

Profile likelihood

Consider the multiparameter problem in which $\theta = (\theta^1, \ldots, \theta^d) \in \Omega_\theta$, an open subset of $\mathbb{R}^d$.

Interest lies in inference for a subparameter or parameter function $\psi = \psi(\theta)$.

The profile likelihood $L_{\mathrm{p}}(\psi)$ for ψ is

$$L_{\mathrm{p}}(\psi) = \sup_{\theta:\psi(\theta)=\psi} L(\theta),$$

the supremum of $L(\theta)$ over all θ that are consistent with the given value of ψ.

The profile log-likelihood is $l_{\mathrm{p}} = \log L_{\mathrm{p}}$.

The usual case

Usually ψ is a component of a given partition $\theta = (\psi, \chi)$ of θ into sub-vectors ψ and χ of dimension $d_\psi = d - d_\chi$ and d_χ respectively.

Then

$$L_{\mathrm{p}}(\psi) = L(\psi, \hat{\chi}_\psi),$$

where $\hat{\chi}_\psi$ denotes the maximum likelihood estimate of χ for a given value of ψ.

Properties of profile likelihood

The maximum profile likelihood estimate of ψ equals $\hat{\psi}$.

The profile log-likelihood ratio statistic $2\{l_{\mathrm{p}}(\hat{\psi}) - l_{\mathrm{p}}(\psi_0)\}$ equals the log-likelihood ratio statistic for $H_0 : \psi = \psi_0$,

$$2\{l_{\mathrm{p}}(\hat{\psi}) - l_{\mathrm{p}}(\psi_0)\} \equiv 2\{l(\hat{\psi}, \hat{\chi}) - l(\psi_0, \hat{\chi}_0)\} \equiv w(\psi_0),$$

where l is the log-likelihood and $\hat{\chi}_0 \equiv \hat{\chi}_{\psi_0}$.

The asymptotic null distribution of the profile log-likelihood ratio statistic is $\chi^2_{d_\psi}$.

This result follows from multivariate Taylor expansion, the continuous mapping theorem and the 'nested quadratic forms' result given in Chapter 1.

Multiparameter problems: further statistics

To test $H_0 : \psi = \psi_0$, in the presence of nuisance parameter χ.

Partition the maximum likelihood estimate, the score vector, the information matrix and its inverse:

$$
\begin{aligned}
U(\theta) &= \begin{pmatrix} U_\psi(\psi,\chi) \\ U_\chi(\psi,\chi) \end{pmatrix}, \\
i(\theta) &= \begin{bmatrix} i_{\psi\psi}(\psi,\chi) & i_{\psi\chi}(\psi,\chi) \\ i_{\chi\psi}(\psi,\chi) & i_{\chi\chi}(\psi,\chi) \end{bmatrix}, \\
i(\theta)^{-1} &= \begin{bmatrix} i^{\psi\psi}(\psi,\chi) & i^{\psi\chi}(\psi,\chi) \\ i^{\chi\psi}(\psi,\chi) & i^{\chi\chi}(\psi,\chi) \end{bmatrix}.
\end{aligned}
$$

Wald statistic

We have $\hat{\psi}$ asymptotically normally distributed with mean ψ_0 and covariance matrix $i^{\psi\psi}(\psi_0, \chi_0)$, which can be replaced by $i^{\psi\psi}(\psi_0, \hat{\chi}_0)$.

So a version of the Wald test statistic for the nuisance parameter case is:

$$w_p(\psi_0) = (\hat{\psi} - \psi_0)^T [i^{\psi\psi}(\psi_0, \hat{\chi}_0)]^{-1} (\hat{\psi} - \psi_0).$$

Score statistic

A version of the score statistic for testing $H_0 : \psi = \psi_0$ is:

$$w_u(\psi_0) = U_\psi(\psi_0, \hat{\chi}_0)^T i^{\psi\psi}(\psi_0, \hat{\chi}_0) U_\psi(\psi_0, \hat{\chi}_0).$$

This test has the advantage that MLE has to be obtained only under H_0, and is derived from the asymptotic normality of U.

Asymptotic distributions

Both $w_p(\psi_0)$ and $w_u(\psi_0)$ have asymptotically a chi-squared distribution with d_ψ degrees of freedom.

Effects of parameter orthogonality

Assume that it is possible to make the parameter of interest ψ and the nuisance parameter, now denoted by λ, orthogonal.

Any transformation from, say, (ψ, χ) to (ψ, λ) necessary to achieve this leaves the profile log-likelihood unchanged.

The matrices $i(\psi, \lambda)$ and $i(\psi, \lambda)^{-1}$ are block diagonal. Therefore, $\hat{\psi}$ and $\hat{\lambda}$ are asymptotically independent.

Also, $\hat{\lambda}_\psi$, the MLE of λ for specified ψ, varies only slowly in ψ in the neighbourhood of $\hat{\psi}$, and there is a corresponding slow variation of $\hat{\psi}_\lambda$ with λ: if $\psi - \hat{\psi} = O_p(n^{-1/2})$, then $\hat{\lambda}_\psi - \hat{\lambda} = O_p(n^{-1})$.

For a nonorthogonal nuisance parameter χ, we would have $\hat{\chi}_\psi - \hat{\chi} = O_p(n^{-1/2})$.

Sketch Proof, Scalar Case

If $\psi - \hat{\psi} = O_p(n^{-1/2}), \chi - \hat{\chi} = O_p(n^{-1/2})$, we have

$$l(\psi, \chi) = l(\hat{\psi}, \hat{\chi}) - \tfrac{1}{2}\{\hat{j}_{\psi\psi}(\psi - \hat{\psi})^2 + 2\hat{j}_{\psi\chi}(\psi - \hat{\psi})(\chi - \hat{\chi}) + \hat{j}_{\chi\chi}(\chi - \hat{\chi})^2\} + O_p(n^{-1/2}).$$

It then follows that

$$\begin{aligned}\hat{\chi}_\psi - \hat{\chi} &= \frac{-\hat{j}_{\psi\chi}}{\hat{j}_{\chi\chi}}(\psi - \hat{\psi}) + O_p(n^{-1}) \\ &= \frac{-i_{\psi\chi}}{i_{\chi\chi}}(\psi - \hat{\psi}) + O_p(n^{-1}).\end{aligned}$$

Then, because $\psi - \hat{\psi} = O_p(n^{-1/2})$, $\hat{\chi}_\psi - \hat{\chi} = O_p(n^{-1/2})$ unless $i_{\psi\chi} = 0$, the orthogonal case, when the difference is $O_p(n^{-1})$.

Further remarks

So far as asymptotic theory is concerned, we can have $\hat{\chi}_\psi = \hat{\chi}$ independently of ψ only if χ and ψ are orthogonal. In this special case we can write $l_{\mathrm{p}}(\psi) = l(\psi, \hat{\chi})$.

In the general orthogonal case, $l_{\mathrm{p}}(\psi) = l(\psi, \hat{\chi}) + o_p(1)$, so that a first-order theory could use $l^*_{\mathrm{p}}(\psi) = l(\psi, \hat{\chi})$ instead of $l_{\mathrm{p}}(\psi) = l(\psi, \hat{\chi}_\psi)$.

Distribution theory

The log-likelihood ratio statistic $w(\psi_0)$ can be written as

$$w(\psi_0) = 2\{l(\hat{\psi}, \hat{\chi}) - l(\psi_0, \chi)\} - 2\{l(\psi_0, \hat{\chi}_0) - l(\psi_0, \chi)\},$$

as the difference of two statistics for testing hypotheses without nuisance parameters.

Taylor expansion about (ψ_0, χ), where χ is the true value of the nuisance parameter, gives, to first-order (i.e. ignoring terms of order $o_p(1)$),

$$\begin{aligned} w(\psi_0) &= \begin{bmatrix} \hat{\psi} - \psi_0 \\ \hat{\chi} - \chi \end{bmatrix}^T i(\psi_0, \chi) \begin{bmatrix} \hat{\psi} - \psi_0 \\ \hat{\chi} - \chi \end{bmatrix} \\ &\quad -(\hat{\chi}_0 - \chi)^T i_{\chi\chi}(\psi_0, \chi)(\hat{\chi}_0 - \chi). \end{aligned}$$

The linearised form of the MLE equations is

$$\begin{bmatrix} i_{\psi\psi} & i_{\psi\chi} \\ i_{\chi\psi} & i_{\chi\chi} \end{bmatrix} \begin{bmatrix} \hat{\psi} - \psi_0 \\ \hat{\chi} - \chi \end{bmatrix} = \begin{bmatrix} U_\psi \\ U_\chi \end{bmatrix},$$

so

$$\begin{bmatrix} \hat{\psi} - \psi_0 \\ \hat{\chi} - \chi \end{bmatrix} = \begin{bmatrix} i^{\psi\psi} & i^{\psi\chi} \\ i^{\chi\psi} & i^{\chi\chi} \end{bmatrix} \begin{bmatrix} U_\psi \\ U_\chi \end{bmatrix}.$$

Also $\hat{\chi}_0 - \chi = i_{\chi\chi}^{-1} U_\chi$, to first-order. Then, to first-order,

$$w(\psi_0) = [U_\psi^T \, U_\chi^T] \begin{bmatrix} i^{\psi\psi} & i^{\psi\chi} \\ i^{\chi\psi} & i^{\chi\chi} \end{bmatrix} \begin{bmatrix} U_\psi \\ U_\chi \end{bmatrix} - U_\chi^T \, i_{\chi\chi}^{-1} U_\chi.$$

Then,

$$w(\psi_0) \sim Q_U - Q_{U_\chi} = Q_{U_\psi . U_\chi},$$

a difference of two nested quadratic forms, and is thus asymptotically $\chi^2_{d_\psi}$, by the result given in Chapter 1.

The Wald statistic $w_p(\psi_0)$ is based directly on a quadratic form of $\hat{\psi} - \psi_0$, and so can be seen immediately to be asymptotically $\chi^2_{d_\psi}$, from properties of the multivariate normal distribution.

Note that to first-order we have

$$w_p(\psi_0) = [i^{\psi\psi} U_\psi + i^{\psi\chi} U_\chi]^T (i^{\psi\psi})^{-1} [i^{\psi\psi} U_\psi + i^{\psi\chi} U_\chi].$$

We can express the statistic $w_U(\psi_0)$ in terms of the score vector U. To first-order we have

$$w_U(\psi_0) = (U_\psi - i_{\psi\chi} i_{\chi\chi}^{-1} U_\chi)^T i^{\psi\psi} (U_\psi - i_{\psi\chi} i_{\chi\chi}^{-1} U_\chi).$$

This follows since, to first-order,

$$\begin{aligned} U_\psi(\psi_0, \hat{\chi}_0) &= U_\psi + \frac{\partial U_\psi}{\partial \chi} (\hat{\chi}_0 - \chi) \\ &= U_\psi - i_{\psi\chi} i_{\chi\chi}^{-1} U_\chi. \end{aligned}$$

The equivalence of the three statistics, and therefore the asymptotic distribution of $w_U(\psi_0)$, follows on showing that the three first order quantities are identical.

M-estimators

The asymptotic theory of M-estimators is relevant in several contexts.

1. Misspecified models. What happens when the 'wrong' likelihood is maximised?

2. Estimating functions. Sometimes (e.g. when the full likelihood function is very complex) we may wish to set up an alternative system of equations for estimating θ.

3. Robust estimators. If outliers are a major concern we may wish to use an estimator of θ that is insensitive to outliers, i.e. a robust estimator.

M-estimators (continued)

Consider a sample of IID observations $X_1, \ldots, X_n$ from F_0.

We wish to estimate a d-dimensional vector parameter $\theta_0 = \theta(F_0)$.

Suppose that we can determine a d-dimensional vector valued function $G(x, \theta)$ such that

$$E_{F_0}[G(X, \theta_0)] = \int G(x, \theta_0) dF(x) = 0 \tag{7}$$

Our estimating function is then given by

$$G(\theta) = \sum_{i=1}^{n} G_i(\theta) \equiv \sum_{i=1}^{n} G(X_i, \theta).$$

M-estimators (continued)

The estimator $\hat{\theta}_n \equiv \hat{\theta}_n(X_1, \ldots, X_n)$ is then determined by solving

$$G(\hat{\theta}_n) = \sum_{i=1}^{n} G_i(\hat{\theta}_n) = 0. \tag{8}$$

Theorem. Under mild conditions the following results hold.

(i) There exists a sequence $\{\hat{\theta}_n\}$ of solutions of (8) such that $\hat{\theta}_n \xrightarrow{p} \theta_0$ as $n \to \infty$, where θ_0 solves (7).

(ii) As $n \to \infty$, $n^{1/2}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N_d(0, H(\theta_0)V(\theta_0)H(\theta_0)^\top)$

where

$$V(\theta_0) = \mathrm{Cov}_{F_0}[G(X, \theta_0)] \quad \text{and} \quad H(\theta_0) = \{E_{F_0}[\nabla_\theta^\top G(X, \theta_0)]\}^{-1}.$$

Comments

1. When $\hat{\theta}_n$ is an MLE in a regular model, $V(\theta_0) = \bar{i}(\theta_0)$ and $H(\theta_0) = \bar{i}(\theta_0)^{-1}$, so we recover the standard result

$$n^{1/2}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N_d(0, \bar{i}(\theta_0)^{-1}).$$

2. The estimator $H(\theta_0)V(\theta_0)H(\theta_0)^\top$ is known as the sandwich variance estimator.

3. In broad generality, we can estimate $V(\theta_0)$ and $H(\theta_0)$ consistently by

$$\hat{H} = [n^{-1}\sum_{i=1}^{n} \nabla_\theta^\top G_i(\hat{\theta}_n)]^{-1} \quad \text{and} \quad \hat{V} = n^{-1}\sum_{i=1}^{n} G_i(\hat{\theta}_n)G_i(\hat{\theta}_n)^\top. \tag{9}$$

4. Approximate confidence intervals for individual parameters and confidence regions for subsets of parameters can be obtained using the normal approximation plus the estimators in (9).

Applications

We do not have time for a detailed discussion of applications, but we mention 3 areas in which this theory is important.

1. Composite likelihood approaches. Here, the full likelihood is intractable, but suitable estimating functions can be constructed by combining components of the likelihood, typically low-dimensional marginal or conditional distributions.

2. Partial likelihood. This approach was developed by David Cox and is widely used in survival analysis. The idea is to set up estimating functions which just use information at the failure times, so that one can avoid modelling the times between failures.

3. Robust estimation. M-estimators play an important role in robust approached to estimation. Examples are Huber's M-estimators, and Maronna's family of robust estimators of location and scatter.