

Asymptotic Properties of Maximum Likelihood Estimation

Following Greene, pp. 476–483. Let θ_0 be the true value of θ , and simply use θ as notation for ‘generic values’ of θ (i.e. typically, values of θ that are not θ_0). Consider the random variable $L(\theta)/L(\theta_0)$. What is its expectation?

$$E_0 \left[\frac{L(\theta)}{L(\theta_0)} \right] = \int \left(\frac{L(\theta)}{L(\theta_0)} \right) L(\theta_0) dy = 1,$$

because the integral of $L(\theta)$ is the integral of $f(y|\theta)$, which is the integral of a density. Now according to Jensen’s Inequality (discussion)

$$E_0 \left[\log \frac{L(\theta)}{L(\theta_0)} \right] < \log E_0 \left[\frac{L(\theta)}{L(\theta_0)} \right], \quad \text{so}$$

$$E_0 \left[\log \frac{L(\theta)}{L(\theta_0)} \right] < 0, \quad \text{or}$$

$$E_0[\log L(\theta) - \log L(\theta_0)] < 0, \quad \text{thus} \quad E_0[\log L(\theta)] < E_0[\log L(\theta_0)]$$

Consistency of the MLE

We have $E_0[\log L(\theta)] < E_0[\log L(\theta_0)]$ so $E_0[\frac{1}{n} \log L(\theta)] < E_0[\frac{1}{n} \log L(\theta_0)]$.

Now $\frac{1}{n} \log L(\theta) = \frac{1}{n} \sum_{i=1}^n \ln f(y_i|\theta)$, i.e. $\frac{1}{n} \log L(\theta)$ is an average of n *iid* random variables; and we have assumed enough so that sample means converge to population means.

Let's think about the MLE $\hat{\theta}$: in any sample $L(\hat{\theta}) \geq L(\theta)$, so for every n , $\frac{1}{n} \log L(\hat{\theta}) \geq \frac{1}{n} \log L(\theta)$.

Summarizing, in the limit, $\frac{1}{n} \log L(\theta_0) > \frac{1}{n} \log L(\theta)$ and $\frac{1}{n} \log L(\hat{\theta}) \geq \frac{1}{n} \log L(\theta)$; so in the limit $\frac{1}{n} \log L(\hat{\theta}) = \frac{1}{n} \log L(\theta_0)$. Consequently, if we assume that in the limit $L(\theta) \neq L(\theta_0)$ (cf. Greene p. 469), we have $\text{plim } \hat{\theta}_{MLE} = \theta_0$.

Asymptotic Normality of the MLE

Here we use a trick that is used over and over again in asymptotic statistical theory: Taylor series expansion. In MLE,

$$g(\hat{\theta}) = 0;$$

expand the function $g(\cdot)$ in Taylor series around the point $\theta = \theta_0$:

$$g(\hat{\theta}) = g(\theta_0) + H(\bar{\theta})(\hat{\theta} - \theta_0) = 0,$$

where $\bar{\theta}$ is a point between $\hat{\theta}$ and θ_0 (brief discussion). Proceeding rather schematically,

$$\begin{aligned} H(\bar{\theta})(\hat{\theta} - \theta_0) &= -g(\theta_0) \\ \sqrt{n}(\hat{\theta} - \theta_0) &= -[H(\bar{\theta})]^{-1}[\sqrt{n}g(\theta_0)] \\ \sqrt{n}(\hat{\theta} - \theta_0) &\rightarrow^d -[H(\theta_0)]^{-1}[\sqrt{n}g(\theta_0)], \end{aligned}$$

where the last step 'follows' because $\text{plim } \hat{\theta} = \theta_0$ and $H(\cdot)$ is continuous.

Asymptotic Normality of the MLE (2)

We have $\sqrt{n}(\hat{\theta} - \theta_0) \rightsquigarrow -[H(\theta_0)]^{-1}[\sqrt{n}g(\theta_0)]$ or $\sqrt{n}(\hat{\theta} - \theta_0) \rightsquigarrow -[\frac{1}{n}H(\theta_0)]^{-1}[\sqrt{n}g(\theta_0)]$
i.e. $\sqrt{n}(\hat{\theta} - \theta_0) \rightsquigarrow -[\frac{1}{n}H(\theta_0)]^{-1}[\sqrt{n}\bar{g}(\theta_0)]$.

Greene handles this by saying $\frac{1}{n}H(\theta_0)$ converges to $E_0[\frac{1}{n}H(\theta_0)]$, so

$$\sqrt{n}(\hat{\theta} - \theta_0) \rightsquigarrow \{-E_0[\frac{1}{n}H(\theta_0)]^{-1}\}[\sqrt{n}\bar{g}(\theta_0)]$$

so that we have $\sqrt{n}(\hat{\theta} - \theta_0)$ is distributed as 'something with a plim' times \sqrt{n} times a sample mean; this has a nondegenerate distribution and thus $\sqrt{n}(\hat{\theta} - \theta_0)$ also has a nondegenerate distribution (it is neither collapsing as n grows nor exploding.) Using $\mathcal{H} = E_0[\frac{1}{n}H(\theta_0)]$ to reduce clutter and proceeding mechanically, the variance of $\sqrt{n}(\hat{\theta} - \theta_0)$ is

$$\sqrt{n}(\hat{\theta} - \theta_0)\sqrt{n}(\hat{\theta} - \theta_0)' = -\mathcal{H}^{-1}[\sqrt{n}\bar{g}(\theta_0)][\sqrt{n}\bar{g}(\theta_0)]'(-\mathcal{H}^{-1})$$

Asymptotic Normality of the MLE (3)

We have

$$n(\hat{\theta} - \theta_0)(\hat{\theta} - \theta_0)' = -\mathcal{H}^{-1}[\sqrt{n}\bar{g}(\theta_0)][\sqrt{n}\bar{g}(\theta_0)]'(-\mathcal{H}^{-1}),$$

but the information matrix equality is equivalent to $E_0\{[\sqrt{n}\bar{g}(\theta_0)][\sqrt{n}\bar{g}(\theta_0)]'\} = -\mathcal{H}$, and thus

$$\begin{aligned} V(\sqrt{n}(\hat{\theta} - \theta_0)) &= -\mathcal{H}^{-1}(-\mathcal{H})(-\mathcal{H}^{-1}) \\ &= -\mathcal{H}^{-1} \end{aligned}$$

Suppose instead of $E_0\{[\sqrt{n}\bar{g}(\theta_0)][\sqrt{n}\bar{g}(\theta_0)]'\} = -\mathcal{H}$, we faced a situation where $E_0\{[\sqrt{n}\bar{g}(\theta_0)][\sqrt{n}\bar{g}(\theta_0)]'\} = \mathcal{J}$, where \mathcal{J} is something else that is the variance of the gradient at θ_0 . We would then have

$$\begin{aligned} V(\sqrt{n}(\hat{\theta} - \theta_0)) &= -\mathcal{H}^{-1}(\mathcal{J})(-\mathcal{H}^{-1}) \\ &= \mathcal{H}^{-1}\mathcal{J}\mathcal{H}^{-1} \end{aligned}$$

Summary of the Discussion

1. The MLE is \sqrt{n} consistent and has asymptotic variance $\frac{1}{n}\mathcal{H}^{-1}$, that is $\hat{\theta}_{MLE}$ is ‘approximately’ $N(\theta_0, \frac{1}{n}\mathcal{H}^{-1})$, or $\sqrt{n}(\hat{\theta} - \theta_0) \rightsquigarrow N(0, \mathcal{H}^{-1})$.
2. The rate of \sqrt{n} and the normality derive from the central limit theorem being applied to sample averages.
3. Less obvious, (or perhaps less emphasized), the consistency is being driven by (1) maximizing an objective function that (2) is uniquely maximized asymptotically at θ_0 .
4. We did not discuss efficiency. The Cramer–Rao bound says that if $\hat{\theta}$ is a \sqrt{n} consistent estimator of θ , then $\sqrt{n}(\hat{\theta} - \theta_0)$ must have a variance of at least \mathcal{H}^{-1} . (Greene doesn’t give a proof of this.) The ‘reason’ the MLE reaches this bound is that $\mathcal{J} = -\mathcal{H}$.

Roadmap of Topics to Follow

1. QMLE—quasi maximum likelihood—as an estimation procedure. QMLE is using a ‘nearly correct’ likelihood function, on purpose. The limiting properties of QMLE depend on understanding the KLIC (Kullback–Leibler information criterion), so we’ll do that.
2. Estimation of $V(\hat{\theta})$.
3. OLS is MLE classical regression model under normality, etc.
4. Three testing principles in MLE/QMLE. (LR, Wald, score/LM).
5. Extremum or M–estimation.
6. Nonlinear models in estimation and testing; the delta method (testing nonlinear restrictions on θ .)

The Kullback–Leibler Information Criterion (KLIC)

The KLIC is a measure of the discrepancy between two distributions. It is not a *metric*—that is, the KLIC of F from (or relative to) G is not the same as that from G to F .

Let F and G be two distribution functions and f and g the corresponding densities. In addition, if the underlying random variable X is discrete let p be the vector of probabilities associated with each of the support points x_i , $i = 1, \dots, s$ under F and w those probabilities or weights under G , i.e. $p_i = f_X(x_i)$, $w_i = G_X(x_i)$. To compare F and G we require them to have the same support.

KLIC (2)

The KLIC of G from F is given by:

$$D(G; F) = \mathbf{E}_F(\log(f/g))$$

so that if X is a continuous random variable:

$$D(G; F) = \int \{\log(f(x)) - \log(g(x))\} f(x) dx$$

and if X is discrete:

$$D(G; F) = \sum_{i=1}^s \{\log p_i - \log w_i\} p_i$$

Fundamental properties of $D(G; F)$ include $D(G; F) \geq 0$ and $D(G; F) = 0$ iff $G = F$.

Quasi-Maximum Likelihood Estimation (QMLE).

Now suppose that the DGP is F but that we assume the DGP is in a parametric family $G(x; \theta)$ not including F . If we estimate by maximum likelihood for $G(x; \theta)$ we maximize

$$\sum_{i=1}^n \log(g(x_i; \theta))$$

in our sample. This corresponds to choosing θ to max an estimate of $n \mathbb{E}_F \log(g(x; \theta))$ and so θ converges to the value that does $\max \mathbb{E}_F \log(g(x; \theta))$.

Rewriting the definition of

$$D(G; F) = \int \{\log(f(x))f(x) - \log(g(x; \theta))f(x)\} dx$$

it is apparent that the first term is unaffected by θ and that maximizing the second term is equivalent to minimizing $D(G; F)$.

The Distribution of the QMLE

When we estimate using G instead of F (or g instead of f , or $g(y|\theta)$ instead of $f(y|\theta)$), our estimate of θ still converges to something, namely the θ that corresponds to the G family member that is $(F) - KLIC$ closest to $F(\theta_0)$ (also called $f(y|\theta_0)$). The limiting value of this θ is sometimes called the 'pseudo-true value' of θ , say θ^* . In a sample, we of course only estimate θ^* . The limiting distribution of $\hat{\theta}^*$ of course (sic) we already know:

$$\sqrt{n}(\hat{\theta}^* - \theta^*) \rightsquigarrow N(0, H_*^{-1} J_* H_*^{-1}),$$

where H_* and J_* are Hessian and variance of quasi-scores matrices, computed at θ^* , with expectations taken according to the true data generating process, i.e. $f(y|\theta)$.

The Relevance of the QMLE

Perhaps the relevance of the QMLE is best seen by thinking of QMLE as ‘mis-specified maximum likelihood.’ If we get the likelihood function wrong, then *if* the pseudo true values have an interpretation, then we are already equipped with an inferential procedure concerning these values. A simple but useful example is the classical linear regression model with heteroscedasticity: the coefficients still converge to the ‘right thing’ AND now we know a correct estimate of their variance (that given by the formula $\sqrt{n}(\hat{\theta}^* - \theta^*) \rightsquigarrow N(0, H_*^{-1} J_* H_*^{-1})$.)

Notice that a big role is played by the incorrect $G(\theta)$ that we specify. In certain cases (linear exponential family models) if G specifies certain conditional means correctly, these means will be correctly estimated by the G model at θ^* (the heteroscedastic CLRM is just one example of this phenomenon.)

Many *specification tests* are based on the information matrix equality, $\mathcal{J} + \mathcal{H} = 0$, which obviously is related to whether $H_*^{-1} J_* H_*^{-1} = \mathcal{H}^{-1}$.

Estimating $V(\hat{\theta})$ 'in Practice'

Let us return to MLE under correct specification. How should estimate $V(\hat{\theta})$? We have many alternatives because (1) the (negative) limiting inverse Hessian and the limiting inverse variance of the score are the same, and (2) we can try to calculate these analytically or by averaging over the sample.

Typically, it's reasonably hard to compute e.g. $E(n^{-1}\partial^2 \log L(\theta)/\partial\theta\partial\theta')$, though sometimes not (the normal linear regression model.) Even when we can, it typically requires averaging over X in some way, and that requires estimation or an assumption. The same calculations can be done with the variance of the score (and it had better give the same answer.)

Estimating $V(\hat{\theta})$ 'in Practice' (2)

In practice it is easier just to take sample averages. Under independence, this is particularly easy (footnote: time series examples without independence are typically built out of independent components which then form the basis of the computations, but not always and not always easily): an estimate of \mathcal{H} is the sample average of the observation contributions, i.e.

$$\hat{H} = n^{-1} \sum_{i=1}^n \frac{\partial^2 \log L_i(\theta)}{\partial \theta \partial \theta'}$$

(I can't get a $\hat{\cdot}$ over \mathcal{H}). Under independence, the sample average of the outer product of score contributions estimates the variance of the score

$$\hat{J} = n^{-1} \sum_{i=1}^n \frac{\partial \log L_i(\theta)}{\partial \theta} \frac{\partial \log L_i(\theta)}{\partial \theta'}$$

Thus you can 'easily' estimate the $V(\hat{\theta})$ in three ways: \hat{H}^{-1} , \hat{J}^{-1} , or $\hat{H}^{-1} \hat{J} \hat{H}^{-1}$, with corresponding standard errors, 'Hessian', 'OPG' or 'score', and 'robust' or 'sandwich'.

OLS is the MLE of the CLRM with Normality

Following Greene pp. 492-494. Considering the classical $y_i = x_i'\beta + \varepsilon$, the likelihood function for a sample of n independent, identically and normally distributed disturbances is

$$L = (2\pi\sigma^2)^{-n/2} e^{-\varepsilon'\varepsilon/(2\sigma^2)}$$

and the transformation from ε_i to y_i is $\varepsilon_i = y_i - x_i'\beta$ with a Jacobian of $|\partial\varepsilon_i/\partial y_i|$ of one. (Discussion.) Thus

$$L = (2\pi\sigma^2)^{-n/2} e^{-(1/(2\sigma^2))(y-X\beta)'(y-X\beta)}$$

$$\log L = -\frac{n}{2} \ln 2\pi - \frac{n}{2} \ln \sigma^2 - \frac{(y-X\beta)'(y-X\beta)}{2\sigma^2} \quad \text{and}$$

$$\begin{bmatrix} \frac{\partial \ln L}{\partial \beta} \\ \frac{\partial \ln L}{\partial \sigma^2} \end{bmatrix} = \begin{bmatrix} \frac{X'(y-X\beta)}{\sigma^2} \\ \frac{-n}{2\sigma^2} + \frac{(y-X\beta)'(y-X\beta)}{2\sigma^4} \end{bmatrix}$$

It would be more useful to have these in 'contribution' form (with observations explicit).

OLS is the MLE of the CLRM with Normality (2)

(N.B. Greene is differentiating wrt σ^2 , not σ , treat σ^2 as a symbol.) Differentiating once more

$$E \begin{bmatrix} \frac{\partial^2 \ln L}{\partial \beta \partial \beta'} & \frac{\partial^2 \ln L}{\partial \beta \partial \sigma^2} \\ \frac{\partial^2 \ln L}{\partial \sigma^2 \partial \beta'} & \frac{\partial^2 \ln L}{\partial (\sigma^2)^2} \end{bmatrix} = \begin{bmatrix} -\frac{X'X}{\sigma^2} & -\frac{X'\varepsilon}{\sigma^4} \\ -\frac{\varepsilon'X}{\sigma^4} & \frac{n}{2\sigma^4} - \frac{\varepsilon'\varepsilon}{\sigma^6} \end{bmatrix}$$

$$E \begin{bmatrix} \frac{\partial^2 \ln L}{\partial \beta \partial \beta'} & \frac{\partial^2 \ln L}{\partial \beta \partial \sigma^2} \\ \frac{\partial^2 \ln L}{\partial \sigma^2 \partial \beta'} & \frac{\partial^2 \ln L}{\partial (\sigma^2)^2} \end{bmatrix} = \begin{bmatrix} -\frac{X'X}{\sigma^2} & 0 \\ 0 & -\frac{n}{2\sigma^4} \end{bmatrix}$$

Inverting and multiplying by -1 :

$$\left\{ -E \begin{bmatrix} \frac{\partial^2 \ln L}{\partial \beta \partial \beta'} & \frac{\partial^2 \ln L}{\partial \beta \partial \sigma^2} \\ \frac{\partial^2 \ln L}{\partial \sigma^2 \partial \beta'} & \frac{\partial^2 \ln L}{\partial (\sigma^2)^2} \end{bmatrix} \right\}^{-1} = \begin{bmatrix} \sigma^2(X'X)^{-1} & 0 \\ 0 & \frac{2\sigma^4}{n} \end{bmatrix}$$

Notice the diagonal structure (discussion).

Three 'Principles' of Testing in the ML Framework

We face the situation of testing the null hypothesis that a parametric restriction of a model is true. Suppose this is expressed as $H_0 : c(\beta) = 0$. We can

1. Estimate the model in both the restricted and unrestricted forms, and test the hypothesis by comparing the log likelihoods of the two models. In particular, twice the difference in the log likelihoods is $\chi^2(J)$, where J is the difference in the dimensions of the parameter space of the two models.
2. Estimate the model in the unrestricted form only and then use the estimated $V(\hat{\beta})$ matrix to see if $c(\beta) = 0$ is plausible. This also gives rise to a $\chi^2(J)$ test. When $c(\beta)$ is nonlinear, this requires the 'delta method'; otherwise we are computing the variance of a linear combination of normally distributed objects, and testing the hypothesis that these are zero.

3. Estimate the model in restricted form only. Heuristically, we then proceed to see whether the relaxing the constraint will increase the log likelihood a lot, basing this computation on the quadratic approximation to likelihood implied by the model estimated under the null hypothesis. This also results in $\chi^2(J)$ test.