Econometrics – Maximum Likelihood Estimation

William M Volckmann II

January 4, 2017

This set of notes is... not great. It's based on material that was presented in lecture that itself was... not great. So there are some things missing, it's lacking in exposition, and generally is just... not great.

1 MLE Redux

We already know a little bit about maximum likelihood estimation (mle). We've seen the **likelihood function**,

$$L(\theta; x) = \prod_{i=1}^{n} f(x_i; \theta),$$

where we treat L as a function of θ, hence often writing $L(\theta)$. We've used the **log likelihood function**,

$$l(\theta) = \ln\big(L(\theta)\big) = \sum_{i=1}^{n} \ln\big(f(x_i; \theta)\big),$$

which is often easier to work with than the likelihood function itself. The maximizer of $L(\theta)$, which is also the maximizer of $l(\theta)$, is our point estimator for θ, which we denote $\widehat{\theta}$ and call it the **maximum likelihood estimator** of θ. In particular, the mle solves the **estimating equation**

$$\frac{\partial l(\theta)}{\partial \theta} = 0.$$

We will often rely on the following **regularity conditions**.

(R1) The pdfs are distinct for each θ. That is, if $\theta \neq \theta'$, then $f(x_i;\theta) \neq f(x_i;\theta')$.

(R2) The pdfs have common support for all θ.

(R3) The **true value** of θ, denoted θ_0, is an interior point of the sample space.

We will also be assuming that the true data generating process is $f(x;\theta_0)$, that is, assuming that the density is correctly specified.

Theorem 1. *Suppose that $X_1, \ldots, X_n$ satisfy the regularity conditions R1-R3, where θ_0 is the true parameter. Furthermore, suppose that $f(x;\theta)$ is differentiable with respect to θ. Then the likelihood equations,*

$$\frac{\partial L(\theta)}{\partial \theta} = 0 \quad \textit{and} \quad \frac{\partial l(\theta)}{\partial \theta} = 0,$$

have a solution $\widehat{\theta}_n$ such that $\widehat{\theta}_n \xrightarrow{P} \theta_0$.

Theorem 2. *Suppose that $X_1, \ldots, X_n$ satisfy the regularity conditions R1-R3, where θ_0 is the true parameter. Furthermore, suppose that $f(x;\theta)$ is differentiable with respect to θ. Suppose the likelihiid equation has the unique solution $\widehat{\theta}_n$. Then $\widehat{\theta}_n$ is a consistent estimator of θ_0.*

Theorem 3. *Let $X_1, \ldots, X_n$ be iid with the pdf $f(x;\theta)$. For a specified function g, let $\eta = g(\theta)$ be a parameter of interest. Suppose $\widehat{\theta}$ is the mle of θ. Then $\widehat{\theta}$ is the mle of $\eta = g(\theta)$.*

2 Newton's Method of Iteration

Sometimes there might not actually be an analytical solution to the mle. In such a case, we can use numerical analysis by way of **Newton's Method of Iteration** to approximate the solution. We make an initial guess at the solution, $\widehat{\theta}_0$. Take the second-order Taylor expansion of $l(\theta)$ around $\widehat{\theta}_0$,

$$l(\theta) \approx l(\widehat{\theta}_0) + l'(\widehat{\theta}_0)(\theta - \widehat{\theta}_0) + \frac{1}{2} l''(\widehat{\theta}_0)(\theta - \widehat{\theta}_0)^2.$$

Then take the derivative with respect to θ, set it equal to zero, and solve for θ to get the subsequent iteration,

$$l'(\widehat{\theta}_0) + l''(\widehat{\theta}_0)(\theta - \widehat{\theta}_0) := 0 \quad \Longrightarrow \quad \widehat{\theta}_1 = \widehat{\theta}_0 - \frac{l'(\widehat{\theta}_0)}{l''(\widehat{\theta}_0)}.$$

Proceed with the same chain of logic but with the new guess $\widehat{\theta}_1$. Continue with the series of iterative guesses

$$\widehat{\theta}_1 = \widehat{\theta}_0 - \frac{l'(\widehat{\theta}_0)}{l''(\widehat{\theta}_0)} \quad \Longrightarrow \quad \widehat{\theta}_2 = \widehat{\theta}_1 - \frac{l'(\widehat{\theta}_1)}{l''(\widehat{\theta}_1)} \quad \Longrightarrow \quad \ldots$$

and so on. In general, the sth iteration will be

$$\widehat{\theta}_s = \widehat{\theta}_{s-1} - \frac{l'(\widehat{\theta}_{s-1})}{l''(\widehat{\theta}_{s-1})}.$$

Iteration continues until the change in each iteration and the change in $l(\widehat{\theta}_s)$ is deemed sufficiently small.

3 Efficiency

We will now introduce two additional regularity conditions.

(R4) The pdf $f(x;\theta)$ is twice differentiable as a function of θ.

(R5) The integral $\int f(x;\theta)\ dx$ can be differentiated twice under the integral sign as a function of θ.

3.1 Fisher Information

Since $f(x;\theta)$ is a pdf, it follows that

$$\int_{-\infty}^{\infty} f(x;\theta)\ dx = 1.$$

Supposing the integrals and derivatives are interchangeable, we can take the derivative with respect to θ to get

$$\int_{-\infty}^{\infty} \frac{\partial f(x;\theta)}{\partial \theta}\ dx = 0,$$

which, by introducing $f(x;\theta)/f(x;\theta)$, can be written

$$\int_{-\infty}^{\infty} \frac{\partial f(x;\theta)/\partial\theta}{f(x;\theta)} f(x;\theta)\ dx = 0.$$

Note that for any function $g(y)$, we have

$$\frac{\partial \ln\big(g(y)\big)}{\partial y} = \frac{1}{g(y)}\frac{\partial g(y)}{\partial y} \quad \text{and} \quad \frac{\partial \ln\big(g(y)\big)}{\partial y} g(y) = \frac{\partial g(y)}{\partial y}.$$

Keeping in mind that we are treating f as a function of θ, we thus can write

$$\int_{-\infty}^{\infty} \frac{\partial \ln\big(f(x;\theta)\big)}{\partial \theta} f(x;\theta)\ dx = 0. \tag{1}$$

Differentiate one more time, applying the product rule, to get

$$\begin{aligned}
&\int_{-\infty}^{\infty} \frac{\partial^2 \ln\big(f(x;\theta)\big)}{\partial\theta^2} f(x;\theta) + \frac{\partial \ln\big(f(x;\theta)\big)}{\partial\theta}\frac{\partial f(x;\theta)}{\partial\theta}\; dx \\
&= \int_{-\infty}^{\infty} \frac{\partial^2 \ln\big(f(x;\theta)\big)}{\partial\theta^2} f(x;\theta)\; dx + \int_{-\infty}^{\infty} \frac{\partial \ln\big(f(x;\theta)\big)}{\partial\theta}\frac{\partial \ln\big(f(x;\theta)\big)}{\partial\theta} f(x;\theta)\; dx \\
&= \int_{-\infty}^{\infty} \frac{\partial^2 \ln\big(f(x;\theta)\big)}{\partial\theta^2} f(x;\theta)\; dx + \int_{-\infty}^{\infty} \left[\frac{\partial \ln\big(f(x;\theta)\big)}{\partial\theta}\right]^2 f(x;\theta)\; dx \qquad (2) \\
&= 0.
\end{aligned}$$

The second term in equation (2) can be written as an expectation. We refer to it as **Fisher information** and denote it with $I(\theta)$:

$$I(\theta) = E\left[\left(\frac{\partial \ln\big(f(X;\theta)\big)}{\partial\theta}\right)^2\right].$$

Alternatively, from equation (2) we can write

$$I(\theta) = -E\left[\frac{\partial^2 \ln\big(f(X;\theta)\big)}{\partial\theta^2}\right].$$

Usually the second form involving the second derivative is easier to compute.

Notice that equation (1) can also be written as an expectation, in particular, as

$$E\left[\frac{\partial \ln\big(f(X;\theta)\big)}{\partial\theta}\right] = 0.$$

Let $Y = \partial \ln\big(f(X;\theta)\big)/\partial\theta$. Then we can write

$$\operatorname{Var}(Y) = E[Y^2] - E[Y]^2 = I(\theta).$$

Thus, Fisher information is the variance of the random variable $\partial \ln\big(f(X;\theta)\big)/\partial\theta$.

The function $\partial \ln\big(f(x;\theta)\big)/\partial\theta$ is called the **score function**. Recall that it determines the estimating equations for the mle – recall that the mle $\widehat{\theta}$ solves

$$\sum_{i=1}^{n} \frac{\partial \ln\big(f(x_i;\theta)\big)}{\partial\theta} = 0.$$

3.2 Rao-Cramer Lower Bound

When we have only one sample, say X_1, Fisher information is the variance of the random variable $\partial \ln\big(f(X_1;\theta)\big)/\partial\theta$. For sample size n, the likelihood $L(\theta)$ is the pdf of the random sample, and the random variable whose variance is the information in the sample is given by

$$\frac{\partial \ln\big(L(\theta,\mathbf{X})\big)}{\partial\theta} = \sum_{i=1}^{n} \frac{\partial \ln\big(f(X_i;\theta)\big)}{\partial\theta}.$$

Each summand is iid with common variance $I(\theta)$. Recall that the variance of the sum of independent random variables is the sum of the variance of those variables. Thus, we simply have

$$\mathrm{Var}\left(\frac{\partial \ln\big(f(\theta,\mathbf{X})\big)}{\partial\theta}\right) = nI(\theta).$$

Theorem 4. *Let $X_1,\dots,X_n$ be iid with common pdf $f(x;\theta)$. Assume that the regularity conditions R1-R5 hold. Let $Y = u(X_1,\dots,X_n)$ be a statistic with mean $E[Y] = k(\theta)$. Then*

$$\mathrm{Var}(Y) \geq \frac{[k'(\theta)]^2}{nI(\theta)}.$$

Corollary 1. *Suppose $Y = u(X_1,\dots,X_n)$ is an unbiased estimator of θ, so* *that $k(\theta) = \theta$. Then the Rao-Cramer inequality becomes*

$$\operatorname{Var}(Y) \geq \frac{1}{nI(\theta)}.$$

Definition 1. Let Y be an unbiased estimator of a parameter θ in the case of point estimation. The statistic Y is called an **efficient estimator** of θ if and only if the variance of Y attains the Rao-Cramer lower bound.

Definition 2. In cases in which we can differentiate with respect to a parameter under an integral or summation symbol, the ratio of the Rao-Cramer lower bound to the actual variance of an any unbiased estimator of a parameter is called the **efficiency** of that estimator.

4 Asymptotics

Let's impose one more regularity condition.

(R6) The pdf $f(x;\theta)$ is three times differentiable as a function of θ. Further, for all θ, there exists a constant c and a function $M(x)$ such that

$$\left| \frac{\partial^3}{\partial\theta^3} \ln\big(f(x;\theta)\big) \right| \leq M(x),$$

which $E_{\theta_0}[M(X)] \leq \infty$ for all $\theta_0 - c < \theta < \theta_0 + c$ and all x in the support of X.

Yeah, that's a hell of a condition. But the proceeding theorem, which R6 justifies, is very nice.

Theorem 5. *Assume $X_1, \dots, X_n$ are iid with pdf $f(x;\theta_0)$ such that the regularity conditions R1-R6 are satisfied. Suppose further than the Fisher information satisfies $0 < I(\theta_0) < \infty$. Then any consistent sequence of solutions* *of the mle equations satisfies*

$$\sqrt{n}(\widehat{\theta} - \theta_0) \overset{D}{\to} \mathcal{N}\left(0, \frac{1}{I(\theta_0)}\right).$$

As in the previous notes, we can write the result of the previous theorem with regard to asymptotic distribution, i.e. for large (but not limiting to infinity) n, as

$$\widehat{\theta} \overset{a}{\sim} \mathcal{N}\left(\theta_0, \frac{1}{nI(\theta_0)}\right).$$

This theorem also implies that $I(\widehat{\theta}) \overset{P}{\to} I(\theta_0)$. As a consequence, the interval

$$\left(\widehat{\theta}_n - z_{\alpha/2}\frac{1}{\sqrt{nI(\widehat{\theta}_n)}}, \widehat{\theta}_n + z_{\alpha/2}\frac{1}{\sqrt{nI(\widehat{\theta}_n)}}\right)$$

is an approximate $(1-\alpha)100\%$ confidence interval for θ.

Corollary 2. *Assume $X_1, \ldots, X_n$ are iid with pdf $f(x;\theta_0)$ such that the regularity conditions R1-R6 are satisfied. Suppose further than the Fisher information satisfies $0 < I(\theta_0) < \infty$. Finally, suppose $g(x)$ is a continuous function of x which is differentiable at θ_0 such that $g'(\theta_0) \neq 0$. Then*

$$\sqrt{n}\left[g(\widehat{\theta}_n) - g(\theta_0)\right] \overset{D}{\to} \mathcal{N}\left(0, \frac{g'(\theta_0)^2}{I(\theta_0)}\right).$$

This should remind you of the Δ-method, and indeed, it is an application of the Δ-method to the previous theorem.

5 Failure of Assumptions

Sometimes we might specify the wrong density function $f(x;\theta)$ and do the typical MLE routine. The the MLE is called a **quasi-MLE** or **pseudo-MLE**. In general, the quasi-MLE is *inconsistent* for θ_0, and the information equality does *not* hold. In other words, we should not use the MLE. This can't be all that surprising, can it?

However, if the specified density $f(x;\theta)$ is in the exponential family, then the quasi-MLE remains consistent for θ_0.

6 Maximum Likelihood Tests

We will be considering a two-sided test of $H_0 : \theta = \theta^*$ vs $H_1 : \theta \neq \theta^*$ at level α. There are three maximum likelihood tests we can do concerning whether or not to reject the null.

(a) They are asymptotically equivalent under H_0.

(b) They are asymptotically equivalent under local alternatives, i.e. $H_1 : \theta = \theta^* + c/\sqrt{n}$.

(c) They are all $\chi^2(1)$ distributed.

Thus, we can use whichever test is simplest given the context.

6.1 The Likelihood Ratio Test

Consider the ratio of two likelihood functions. In particular, define

$$\Lambda = \frac{L(\theta^*)}{L(\widehat{\theta})}.$$

If we assume regularity conditions R1-R6 all hold, then under the null hypothesis $H_0 : \theta = \theta^*$, we have

$$-2\ln(\Lambda) \overset{D}{\to} \chi^2(1).$$

Thus, we reject H_0 if $-2\ln(\Lambda) \geq \chi^2_\alpha(1)$. Equivalently, reject H_0 if

$$2\ln[L(\widehat{\theta}) - L(\theta^*)] \geq \chi^2_\alpha(1).$$

We call this the **likelihood ratio test**. It is favored by statisticians due to the Neyman-Pearson lemma, which we will cover later. It relies entirely on correct specification of the density, however.

6.2 The Wald Test

The **Wald test** relies on the statistic

$$\chi^2_W = \left[\sqrt{nI(\widehat{\theta})}(\widehat{\theta} - \theta^*)\right]^2 .$$

We reject H_0 if $\chi^2_W \geq \chi^2_\alpha(1)$. Note that for a one-sided test, we can use $\chi_W \overset{D}{\to} \mathcal{N}(0,1)$.

The Wald test is favored by econometricians because it is easy to robustify.

6.3 The Score Test

Define the statistic

$$\chi^2_R = \left[\frac{l'(\theta^*)}{\sqrt{nI(\theta^*)}}\right]^2 .$$

We reject H_0 in favor of H_1 if $\chi^2_R \geq \chi^2_\alpha(1)$. This is called the **score test**. Note that for a one-sided test, we can use $\chi_R \overset{D}{\to} \mathcal{N}(0,1)$.

The score test is advantageous when it is much simpler to estimate the model under H_0.

7 Optimality of Maximum Likelihood Tests

Recall that the size of a test is the probability of rejecting H_0 given that H_0 is actually true; and the power of a test is the probability of rejecting H_0 given that H_1 is true. A test a fixed size α will be unbiased if power is always greater than or equal to the size; and it will be consistent if power approaches 1 as n approaches infinity. The idea is that if we have a lot of data, we should have more ability to reject false nulls.

A **most powerful test** is a test that has equal or higher power than any other test. A most powerful test exists for testing one simple hypothesis against another, i.e. $H_0 : \theta = \theta'$ vs $H_1 : \theta = \theta''$. By the Neyman-Pearson lemma, this is a function of the likelihood ratio. Which is ti say, a most powerful test reject H_0 if

$$\frac{L(\theta'; x)}{L(\theta''; x)} \leq k \quad \iff \quad \ln\big(L(\theta'; x)\big) - \ln\big(L(\theta''; x)\big) \leq \ln(k),$$

where k is determined by the test size α.

Now consider a simple null vs a composite alternative, e.g. $H_0 : \theta = \theta'$ vs $H_1 : \theta > \theta'$. A **uniformly most powerful test** of size α is one that has the most power against every value of θ'' under H_1. A uniformly most powerful test does not always exist, but it does for sure when there is a monotone likelihood ratio.