

Ec241a

Econometrics

Spring 2004

UC Berkeley Department of Economics

MAXIMUM LIKELIHOOD ESTIMATION II:
CONSISTENCY, NORMALITY AND EFFICIENCY (FERGUSON)

Lancaster estimates an exponential model where the conditional density of unemployment duration Y given covariates X is assumed to have an exponential distribution with density

$$f(y|x, \beta_0) = e^{x'\beta_0} \exp(-ye^{x'\beta_0}),$$

for positive y . He estimates β_0 by maximizing the log likelihood function

$$L(\beta) = \sum_{i=1}^N x_i'\beta - y_i \cdot \exp(x_i'\beta),$$

over β . How does one ascertain the properties of the mle in a case like this? In particular, how does one derive the following results that Lancaster relies on

$$\hat{\beta}_{mle} \xrightarrow{p} \beta_0, \quad \text{or} \quad \lim_{N \rightarrow \infty} Pr[\|\hat{\beta}_{mle} - \beta_0\| > \varepsilon] = 0, \quad \forall \varepsilon > 0,$$

and

$$\sqrt{N}(\hat{\beta}_{mle} - \beta_0) \xrightarrow{d} \mathcal{N}\left(0, \left(\mathbb{E}\left[Y_i X_i X_i' \exp(X_i' \beta_0)\right]\right)^{-1}\right),$$

where he estimates $\mathbb{E}[Y_i X_i X_i' \exp(X_i' \beta_0)]$ with $\sum_i y_i x_i x_i' \exp(x_i' \hat{\beta}_{mle})/N$? In addition, the choice of mle is at least partially motivated by a large sample efficiency argument. We shall now look at each of these three results.

Consistency

The basic consistency result for maximum likelihood estimators, as well as for some of the other estimators we shall look at later, is the following:

Result 1 CONSISTENCY

Suppose that there is a sequence of functions $Q_N(\theta)$ from $\Theta \subset \mathbb{R}^K$ to \mathbb{R} and a function $Q_0(\theta)$

also from Θ to \mathbb{R} such that

(i), $Q_N(\theta)$ converges to $Q_0(\theta)$ uniformly in θ , for $\theta \in \Theta$.

(ii), $Q_N(\theta)$ is continuous in θ .

(iii), $Q_0(\theta)$ is uniquely minimized at θ_0 ,

(iv), Θ is compact.

Then the solution $\hat{\theta}$ to $\min_{\theta \in \Theta} Q_N(\theta)$ converges to θ_0 in probability.

Note: by uniform convergence we mean that

$$Pr \left(\sup_{\theta \in \Theta} |Q_N(\theta) - Q_0(\theta)| > \varepsilon \right) \longrightarrow 0,$$

for all $\varepsilon > 0$.

Here is a simple proof: Let Ω be an arbitrary open neighbourhood in \mathbb{R}^K containing θ_0 . We shall show that the probability that $\hat{\theta}_{mle}$ is in Ω goes to one as the sample size gets large. Because Ω is open, the complement Ω^c is closed and the intersection of Ω^c with Θ , $\Omega^c \cap \Theta$, is compact. Therefore the solution to

$$\min_{\theta \in (\Omega^c \cap \Theta)} Q_0(\theta),$$

exists. Let

$$\varepsilon = \min_{\theta \in (\Omega^c \cap \Theta)} Q_0(\theta) - Q_0(\theta_0).$$

Because $\theta_0 \notin \Omega^c$, and θ_0 is the unique minimand of $Q_0(\theta)$, it follows that $\varepsilon > 0$.

Consider the event A_N , defined as:

$$\sup_{\theta \in \Theta} |Q_N(\theta) - Q_0(\theta)| < \varepsilon/2.$$

By the uniform convergence, the probability of the event A_N converges to one for all $\varepsilon > 0$, so also for the ε defined above.

Note that the event A_N implies

$$Q_0(\hat{\theta}) - Q_N(\hat{\theta}) < \varepsilon/2,$$

which, because $\hat{\theta}$ minimizes $Q_N(\theta)$, implies that A_N implies

$$Q_0(\hat{\theta}) < Q_N(\hat{\theta}) + \varepsilon/2 \leq Q_N(\theta_0) + \varepsilon/2.$$

Another implication of A_N is that

$$Q_N(\theta_0) - Q_0(\theta_0) < \varepsilon/2,$$

which combined with the earlier result implies that if A_N is true, then

$$Q_0(\hat{\theta}) < Q_0(\theta_0) + \varepsilon.$$

But if $Q_0(\hat{\theta})$ is less than ε away from $Q_0(\theta)$, it must be that $\hat{\theta}$ is in Ω . Thus A_N is true implies $\hat{\theta} \in \Omega$. Then the fact that the probability of A_N converges to one implies that the probability of $\hat{\theta} \in \Omega$ converges to one. This is true for any open set containing θ_0 , so $\hat{\theta}$ converges to θ_0 . \square .

The most difficult condition is condition (i), the uniform convergence requirement. To see what the role is of this condition, consider the following example.

Example: Suppose

$$Q_N(\theta) = (\theta - 1)^2 + \theta^{1/N} \ln \theta / \sqrt{N},$$

for $\theta \in (0, 1]$, and $Q_N(0) = 0$. For fixed θ ,

$$Q_0(\theta) = \lim_{N \rightarrow \infty} Q_N(\theta) = (\theta - 1)^2,$$

which is minimized at $\theta = 1$. However, consider evaluating $Q_N(\theta)$ at

$$\operatorname{argmin}_{\theta} -\theta^{1/N} \ln \theta / \sqrt{N} = \exp(-N).$$

Then

$$Q_N(\theta) = (\exp(-N) - 1)^2 + \sqrt{N} \cdot \exp(-1).$$

As N increases $Q_N(\exp(-N)) \ll Q_N(1) = 0$, and so the minimand converges to zero, not one. Figure 1 shows the functions $Q_N(\theta)$ for $N = 1, 2, 4, 64$, with the limiting function $Q_0(\theta)$ in dashes. In each graph the point $(\exp(-N), Q_N(\exp(-N)))$ is marked with an $*$. \square

More easily interpretable conditions for uniform convergence are given in the following result:

Result 2 UNIFORM CONVERGENCE

Let Z_1, Z_2, \dots be a sequence of independent and identically distributed random variables, and let $\psi(z, \theta)$ for each z be a function from Θ to R , satisfying

- (i), $\psi(z, \theta)$ is continuous in θ for each z .
- (ii), Θ is compact,
- (iii), $\|\psi(z, \theta)\| \leq K(z)$ with $\mathbb{E}[|K(z)|] < \infty$.

Then

- (i) $\mathbb{E}[\psi(Z, \theta)]$ is continuous in θ , and,
- (ii) $\sum_{i=1}^N \psi(Z_i, \theta)/N$ converges to $\mathbb{E}[\psi(Z, \theta)]$ uniformly in θ .

For proofs see, for example, Ferguson (1996) or Newey and McFadden (1994).

How does this work for the exponential example? We cannot apply this directly because we did not specify the full joint distribution of (Y, X) , only the conditional distribution of Y given X . Let us complete this specification in the following way. Let us assume that the covariates X have a discrete distribution with $Pr(X = \lambda_k) = p_k$, with both λ_k and p_k known. We also assume that $\Theta = \{\beta \mid |\beta_j| \leq c_\beta\}$ for some $c_\beta > 0$. This implies that $x'\beta$ is bounded by some constant, denoted by c . Then the conditions for Result 2 are satisfied for

$$\psi(Y, X, \beta) = - \left(X'\beta - Y \exp(X'\beta) + \ln \left(\sum_{j=1}^K 1\{X = \lambda_j\} \cdot p_j \right) \right).$$

Clearly $\psi(Y, X, \beta)$ is continuous in β . It is also bounded in absolute value by

$$K(Y) = c + Y \cdot \exp(c) - \min_j \ln p_j.$$

This function $K(Y)$ has finite expectation

$$\mathbb{E}[K(Y)] = c + \sum_{j=1}^K p_j \exp(-\lambda_j'\beta) \cdot \exp(c) - \min_j \ln p_j.$$

Next we shall apply Result 1. The objective function is

$$Q_N(\beta) = -L(\beta)/N = -\frac{1}{N} \sum_{i=1}^N X'_i \beta - Y_i \exp(X'_i \beta) + \ln \left(\sum_{j=1}^K 1\{X_i = \lambda_j\} \cdot p_j \right).$$

By Result 2 this converges uniformly to

$$\begin{aligned} Q_0(\beta) &= \mathbb{E}[Q_N(\beta)] = \mathbb{E}[\mathbb{E}[Q_N(\beta)|X]] = -\mathbb{E}[X'_i \beta - \exp(X'_i(\beta - \beta_0))] \\ &= -\sum_{j=1}^K p_j \lambda'_j \beta - p_j \exp(\lambda'_j(\beta - \beta_0)), \end{aligned}$$

which, also by Result 2, is continuous in β . Take the derivative with respect to β to get

$$\frac{\partial Q_0}{\partial \beta}(\beta) = -\sum_{j=1}^K p_j \lambda_j (1 - \exp(\lambda'_j(\beta - \beta_0))),$$

with second derivative

$$\frac{\partial^2 Q_0}{\partial \beta \partial \beta'}(\beta) = \sum_{j=1}^K p_j \lambda_j \lambda'_j \exp(\lambda'_j(\beta - \beta_0)).$$

The first derivative is zero at $\beta = \beta_0$ and because the second derivative is positive definite, this is the global minimum. Therefore Result 1 applies and the mle is consistent for β_0 .

Remark: The assumption that the covariates X_i have a discrete distribution with known support is just a device to apply full parametric maximum likelihood methods. We never actually use the knowledge of the support points or the probabilities in the computation. We can clearly make weaker assumptions there but at the same time there is no harm in assuming discreteness: in the end the world is discrete and continuity is just an approximation that can often make life easier (but not here). The same technique will be used in some of the discussion of GMM methods. \square

The fact that $\mathbb{E}[\ln f_Z(Z; \theta)]$ is maximized at the true value θ_0 as shown above for the exponential example, is true much more generally. Here are two arguments. The first is based on Jensen's inequality:

Result 3 JENSEN'S INEQUALITY

Let $g(z)$ be a convex function. Then $\mathbb{E}[g(Z)] \geq g(\mathbb{E}[Z])$, with equality only in the case of a linear function.

Proof: This is for the continuous case, with $g(z)$ twice continuously differentiable, so that $g''(z) \geq 0$. We can expand $g(z)$ around $\mathbb{E}[Z]$:

$$g(z) = g(\mathbb{E}[Z]) + g'(\mathbb{E}[Z]) \cdot (z - \mathbb{E}[Z]) + g''(\tilde{z}) \cdot (z - \mathbb{E}[Z])^2,$$

for some \tilde{z} in between z and $\mathbb{E}[Z]$. Because of the convexity, the last term is non-negative, so

$$g(z) \geq g(\mathbb{E}[Z]) + g'(\mathbb{E}[Z]) \cdot (z - \mathbb{E}[Z]).$$

Then

$$\begin{aligned} \mathbb{E}[g(Z)] &\geq \mathbb{E}\left[g(\mathbb{E}[Z]) + g'(\mathbb{E}[Z]) \cdot (Z - \mathbb{E}[Z])\right] \\ &= g(\mathbb{E}[Z]) + g'(\mathbb{E}[Z]) \cdot \mathbb{E}\left[(Z - \mathbb{E}[Z])\right] = g(\mathbb{E}[Z]), \end{aligned}$$

which completes the proof. \square

Now consider the random variable Y defined as the ratio of the density function at some arbitrary value of θ to the density function at θ_0 , both evaluated at the random variable Z :

$$Y = f_Z(Z; \theta) / f_Z(Z; \theta_0).$$

Take $g(\cdot)$ to be minus the logarithmic function: $g(a) = -\ln(a)$, so $g'(a) = -1/a$, and $g''(a) = 1/a^2 > 0$ and $g(\cdot)$ is convex. Then, by Jensen's inequality

$$\mathbb{E}[-\ln Y] \geq -\ln \mathbb{E}[Y],$$

implying

$$\mathbb{E}\left[-\ln\left(\frac{f_Z(Z; \theta)}{f_Z(Z; \theta_0)}\right)\right] \geq -\ln\left(\mathbb{E}\left[\frac{f_Z(Z; \theta)}{f_Z(Z; \theta_0)}\right]\right),$$

where the expectation is over the distribution of Z , that is the density $f_Z(z; \theta_0)$. The expectation on the right therefore simplifies:

$$\mathbb{E}\left[\frac{f_Z(Z; \theta)}{f_Z(Z; \theta_0)}\right] = \int \frac{f_Z(z; \theta)}{f_Z(z; \theta_0)} \cdot f_Z(z; \theta_0) dz = \int f_Z(z; \theta) dz = 1,$$

for all values of θ , so that after taking the log, the righthand side is equal to zero, and thus

$$\mathbb{E} \left[-\ln \left(\frac{f_Z(Z; \theta)}{f_Z(Z; \theta_0)} \right) \right] \geq 0$$

implying

$$-\mathbb{E} \left[\ln f_Z(Z; \theta) \right] + \mathbb{E} \left[\ln f_Z(Z; \theta_0) \right] \geq 0,$$

and thus

$$\mathbb{E} \left[\ln f_Z(Z; \theta_0) \right] \geq \mathbb{E} \left[\ln f_Z(Z; \theta) \right],$$

for all θ . This implies that the expected value of the log likelihood is maximized at the true value of θ , and therefore there is some hope that the actual log likelihood function is maximized at a value close to θ_0 , and therefore at a value that is a good estimate of θ_0 .

Note that we always take the expectation over the true distribution, $f(z; \theta_0)$, even if we are taking expectations of functions of θ , evaluated at all possible values of θ , and in particular evaluated at values other than the true value θ_0 .

A second argument why $\mathbb{E}[\ln f_Z(Z; \theta)]$ is maximized at the true value θ_0 relies on the information matrix equality

$$-\mathbb{E} \left[\frac{\partial^2 \ln f}{\partial \theta^2}(z; \theta_0) \right] = \mathbb{E} \left[\frac{\partial \ln f}{\partial \theta}(z; \theta_0)^2 \right].$$

To see why this holds recall that for all θ

$$1 = \int_z f_Z(z; \theta) dz,$$

implying,

$$0 = \frac{\partial}{\partial \theta} \int_z f_Z(z; \theta) dz.$$

and thus, assuming we can change the order of differentiation and integration, we get

$$0 = \int_z \frac{\partial f_Z}{\partial \theta}(z; \theta) dz = \int_z \frac{\partial \ln f_Z}{\partial \theta}(z; \theta) \cdot f_Z(z; \theta) dz. \quad (1)$$

Thus:

$$\mathbb{E} \left[\frac{\partial \ln f_Z}{\partial \theta}(z; \theta_0) \right] = 0.$$

Now differentiate the righthand side of (1) again to get

$$\begin{aligned} 0 &= \int_z \frac{\partial^2 \ln f_Z}{\partial \theta^2}(z; \theta) \cdot f_Z(z; \theta) dz + \int_z \frac{\partial \ln f_Z}{\partial \theta}(z; \theta) \cdot \frac{\partial f_Z}{\partial \theta}(z; \theta) dz \\ &= \int_z \frac{\partial^2 \ln f_Z}{\partial \theta^2}(z; \theta) \cdot f_Z(z; \theta) dz + \int_z \frac{\partial \ln f_Z}{\partial \theta}(z; \theta) \cdot \frac{\partial \ln f_Z}{\partial \theta}(z; \theta) f_Z(z; \theta) dz \\ &= \int_z \frac{\partial^2 \ln f_Z}{\partial \theta^2}(z; \theta) \cdot f_Z(z; \theta) dz + \int_z \left(\frac{\partial \ln f_Z}{\partial \theta}(z; \theta) \right)^2 f_Z(z; \theta) dz \end{aligned}$$

Hence

$$\mathbb{E} \left[\frac{\partial^2 \ln f}{\partial \theta^2}(z; \theta_0) \right] + \mathbb{E} \left[\frac{\partial \ln f}{\partial \theta}(z; \theta_0)^2 \right] = 0,$$

and $\mathbb{E} \left[\frac{\partial^2 \ln f_Z}{\partial \theta^2}(z; \theta_0) \right]$ is negative definite, and θ_0 must be the maximand of $\mathbb{E}[\ln f_Z(Z; \theta)]$.

The next result gives general conditions under which the mle has an approximate normal distribution in large samples. First we establish some additional notation. Let $\mathcal{S}(z, \theta) = \frac{\partial \ln f}{\partial \theta}(z, \theta)$ be the score function, or the vector of derivatives of the log of the density function, and $\mathcal{H}(z, \theta) = \frac{\partial^2 \ln f}{\partial \theta \partial \theta'}(z, \theta)$ be the Hessian, or the matrix of second derivatives of the log of the density function.

Result 4 *Let Z_1, Z_2, \dots be a sequence of independent and identically distributed random variables with density $f(z|\theta)$, and let $\theta_0 \in \text{int}(\Theta)$ be the true value of the parameter θ .*

Suppose

(i) Θ is compact,

(ii) *The second derivative of $f(z|\theta)$ with respect to θ exists and may be passed under the integral sign in $\int f(z|\theta) dz$,*

(iii), *There exists a function $K(z)$ such that $\mathbb{E}[K(Z)] < \infty$ and each component of $\mathcal{H}(Z, \theta)$ is bounded in absolute value by $K(Z)$ uniformly in some neighbourhood of θ_0 ,*

(iv), $\mathcal{I}(\theta_0) = -\mathbb{E}[\mathcal{H}(Z, \theta_0)]$ is positive definite.

(v), If $f(z, \theta) = f(z|\theta_0)$ almost everywhere, then $\theta = \theta_0$.

Then, there exists a strongly consistent sequence of roots of the likelihood equation $\hat{\theta}_n$ such that

$$\sqrt{N}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \mathcal{N}(0, \mathcal{I}(\theta_0)^{-1}).$$

Proof: see Ferguson.

Consider our exponential example. The score, hessian and information matrix are, respectively,

$$\mathcal{S}(z, \beta) = x'(1 - y \exp(x'\beta)),$$

$$\mathcal{H}(z, \beta) = -yxx' \exp(x'\beta),$$

and

$$\mathcal{I}(\beta_0) = \mathbb{E}[YXX' \exp(X'\beta_0)] = \mathbb{E}[XX'],$$

respectively. Assuming again that X is discrete it is easy to bound each element of $\mathcal{H}(y, x, \beta)$. The hessian is negative definite as long as $\mathbb{E}[XX']$ is positive definite. Condition (v) is also satisfied. Finally, the solution to the likelihood equations is unique and equal to the mle, so in large samples the mle has the normal approximation given above.

Finally, we consider efficiency. The most important result is the Cramér–Rao bound:

Result 5 *Let Z have density $f(z|\theta_0)$ with nonsingular information matrix $\mathcal{I}(\theta_0)$. Assume that $\partial f(z|\theta)/\partial \theta$ exists and that the derivative with respect to θ can be passed under the integral sign in $\int f(z|\theta) dz$ and $\int \hat{\theta}_{mle} f(z|\theta) dz$. Then:*

$$\text{Var}(\hat{\theta}) \geq \mathcal{I}(\theta_0)^{-1},$$

for any unbiased estimator $\hat{\theta}$.

All conditions are satisfied in our example, implying there is no unbiased estimator with a variance smaller than the large sample variance of $\hat{\theta}_{mle}$.

Example: The relevance of the efficiency result is that we do not have to look for other estimators. In the exponential example such an alternative, but inefficient, estimator is available in the least squares estimator for the regression of $\ln Y$ on X . To see this, first consider for an exponential distribution with parameter λ the moment generating function of $\ln Y$:

$$M_{\ln(Y)}(t) = \mathbb{E}[\exp(t \ln Y)] = \mathbb{E}[Y^t].$$

The moment generating function of an exponentially distributed Y is

$$M_Y(t) = \frac{\lambda}{\lambda - t},$$

and so the expectation of Y^t is

$$M_{\ln(Y)}(t) = \left. \frac{\partial^t M_Y}{\partial z^t}(t) \right|_{z=0} = \mathbb{E}[Y^t] = \frac{\Gamma(t+1)}{\lambda^t},$$

where

$$\Gamma(x) = \int_0^\infty y^{x-1} \exp(-y) dy,$$

the Gamma function, which for integer x is equal to $\Gamma(x) = (x-1)!$. Hence for integer t ,

$$M_{\ln(Y)}(t) = \mathbb{E}[Y^t] = \frac{\Gamma(t+1)}{\lambda^t} = \frac{t!}{\lambda^t},$$

Taking the log to get the cumulant generating function for $\ln Y$ gives:

$$K_{\ln Y}(t) = \ln \Gamma(t+1) - t \cdot \ln \lambda.$$

Taking the first and second derivatives to get the mean and variance of $\ln Y$ gives:

$$\mathbb{E}[\ln Y] = \ln \lambda + \psi(1),$$

and

$$V(\ln Y) = \psi'(1),$$

where $\psi(\cdot)$ and $\psi'(\cdot)$ are the first and second derivatives of the logarithm of the complete Gamma function:

$$\psi(x) = \frac{\partial \ln \Gamma}{\partial x}(x), \quad \text{and} \quad \psi'(x) = \frac{\partial^2 \ln \Gamma}{\partial x^2}(x).$$

Approximately,

$$\psi(1) \approx -0.5772, \quad \psi'(1) = 1.6449.$$

So, for the exponential distribution with hazard rate $\exp(x'\beta)$, we have

$$\mathbb{E}[\ln Y_i | X_i] = X_i' \beta + \psi(1) \approx X_i' \beta - 0.5772, \quad (2)$$

$$V[\ln Y_i | X_i] = \psi'(1) \approx 1.6449,$$

Because of (2) one might consider estimating β by least squares regression of $\ln Y_i - \psi(1)$ on X_i . This is a perfectly decent and consistent estimator. Its variance, $\psi'(1) \cdot \mathbb{E}[XX']^{-1} \approx 1.6449 \cdot \mathbb{E}[XX']^{-1}$, however, compares unfavourably with that of the mle, which is $\mathbb{E}[XX']^{-1}$, as already implied by the above efficiency argument.

Another argument for preferring ml to ols is that the latter is more difficult to extend to allow for right censoring. \square

REFERENCES

AMEMIYA, T., *Advanced Econometrics*, Harvard University Press, Cambridge.

FERGUSON, T., (1996), *A Course in Large Sample Theory*, Chapman and Hall, London.

NEWBY, W., AND D. MCFADDEN, "Large Sample Estimation and Hypothesis Testing", Chapter 36 *Handbook of Econometrics*, Vol 4, McFadden and Engle, (eds.), Elsevier, North Holland.

