ECON 7330, Fall 2025.

Bent E. Sørensen

Notes on Maximum Likelihood.

Consider a sample of independent (later generalized to dependent) variables $x_1, ..., x_N$ with density $f(x;\theta)$, where θ is a $k-$vector of parameters. (E.g., for the linear regression model, θ is the coefficients to the regressors and the variance.) The maximum likelihood estimator is

$$\hat{\theta} = \text{argmax}_\theta \pounds(\theta) = \text{argmax}_\theta f(x_1, ..., x_N; \theta)$$

or, equivalently,

$$\hat{\theta} = \text{argmax}_\theta l(\theta) = \text{argmax}_\theta \log f(x_1, ..., x_N; \theta)$$

where $l(\theta) = \log(\pounds(\theta))$. (In this note, when we find the maximum likelihood estimator, θ is the variable we vary for given x-vector. When we want to take expectations, the "x-part" is the varying variable. This is why this is hard, even if there are not many lines of proof. That cannot be helped, sorry. For some people, even successful academic economists, this will always be a bit mysterious. But you have to know how to use the results below.) We assume that the likelihood function is well-behaved; in particular, that it has a unique maximum (we say that model is "identified"). In practical applications, especially with small datasets/too complicated models, the unique maximum may be hard to find, we will talk about those issue later.) For i.i.d. observations the maximization problem (using $f(x_1, ..., x_N; \theta) = f(x_1; \theta) \times ... \times f(x_N; \theta)$) is

$$\hat{\theta} = \text{argmax}_\theta \; l(\theta; x_1) + ... + l(\theta; x_N) \; .$$

Note: $l(\theta)$ is always the full log-likelihood function. $l(\theta; x_i)$ is log of the density for the i'th observation, which we also may write as $l(\theta; x_i)$ or as $\log f(x_i; \theta)$. Mathematically, it is the exact same, but it is often suggestive to write the argument that is not kept fixed first. So when we find the estimator, θ is the variable written first.

The maximum likelihood (ML) estimator has very good properties, in particular it satisfies the Cramer-Rao lower bound, which states that among consistent estimators the ML estimator has the lowest asymptotic variance. Consistency means that when the sample size goes to infinity the estimator converges to the true value θ_0, which underlies the Date Generating Process (DGP) (you will sometimes encounter that jargon). In a lot of the homeworks, the DGP is one you construct in Matlab.

The ML estimator is consistent by the following argument (which is a sketch—a rigorous proof takes the same form but verifies that each step is valid). We assume that the likelihood function is concave, with the optimum inside some compact interval and that the data are generated by the corresponding density for some true parameter value θ_0. That the density that generated the data is the same as the function used for ML, gives all the nice properties of ML-estimation. We use a little bit of a trick by comparing the likelihood function at any value to the likelihood function at the true value by considering the log of the ratio of the former to the latter:

$$\log(\frac{\pounds(\theta;x)}{\pounds(\theta_0;x)}) \tag{1}$$

The log is strictly concave, and Jensen's inequality states that $E\log(Z) < \log(EZ)$ for *any* random variable. Using Jensen for $Z = \frac{\pounds(\theta;X)}{\pounds(\theta_0;X)}$ implies that

$$E_0 \log[\frac{\pounds(\theta;X)}{\pounds(\theta_0;X)}] < \log[E_0 \frac{\pounds(\theta;X)}{\pounds(\theta_0;X)}]\,,$$

with equality only for $\theta = \theta_0$ where the ratio is the constant 1. Now use that the data are generated by the density $\pounds(\theta_0;x)$. The expected value with of any function, say $EG(X)$ of any random variable from the distribution we consider, is $\int G(x)f(x;\theta_0)dx$, so we have

$$E_0 \frac{\pounds(\theta;X)}{\pounds(\theta_0;X)} = \int \frac{\pounds(\theta;x)}{\pounds(\theta_0;x)} f(x;\theta_0)dx = \int \frac{\pounds(\theta;x)}{\pounds(\theta_0;x)} \pounds(\theta_0;x)\,dx = \int \pounds(\theta;x)dx = 1\,,$$

because $\pounds(\theta;x)$ is a density for any θ, and densities integrate to unity. We then get

$$E_0 \log[\frac{\pounds(\theta;X)}{\pounds(\theta_0;X)}] = E_0 \log f(X;\theta) - E_0 \log f(X;\theta_0) \leq 0\,.$$

That means that the true parameter maximizes the expected value of the log of the density function. As

$$plim \frac{1}{N} l(\theta) = E_0 \log f(X, \theta) \,,$$

maximizing the empirical likelihood function will be similar to maximizing the expected value of the likelihood function and the ML estimate will be close to θ_0. This is the intuition and how the proof goes. To make it fully rigorous would take a lot more effort with little further intuition (convergence to a function is different from convergence to a constant, so there are traps, but for the functions we usually use, the fact that the plim holds point for point by the LLN, implies that it holds for the likelihood function—often times rigorous proof are complicated because authors want them to cover as many cases as possible).

Define the gradient of the log-likelihood function, $g(\theta)$, that most econometricians, including me, prefers to call the "Score" $S(\theta)$, as the derivative (a column vector of k partial derivatives),

$$S_N(\theta) = \frac{d\, l(\theta)}{d\, \theta} = \Sigma_{i=1}^{N} \frac{d}{d\, \theta} l(\theta; x_i) \,.$$

(I don't know why it is called the score, maybe to avoid saying "gradients of the likelihood function".) Often we omit the subscript N. Note: The derivative of an r-dimensional function with respect to k arguments is normally an $r \times k$ matrix. For the derivatives of log density (including the score), I write them as column vectors in the note for convenience. Usually, we work with likelihood function that has an internal maximum (this is the assumption when nothing else is said), so that

$$S_N(\hat{\theta}) = 0 \,.$$

The term

$$G_i(\theta) = \frac{d}{d\, \theta} l(\theta; x_i) \,,$$

is called the "contribution to the gradient." (I also define this a column vector.)

The (finite sample) information matrix is the variance ($k \times k$ matrix) of the score (vector) evaluated at the true value:

$$I_N(\theta_0) = ES_N(\theta_0)\, S_N(\theta_0)' = \Sigma_{i=1}^N EG_i(\theta_0)G_i(\theta_0)' \ .$$

Notice that this is a variance, which for a vector X is $E(X-\mu)(X-\mu)'$, but here the score has mean 0, so there is no "μ" term. The variance of the sum is the sum of the variances because the contributions to the score are all independent (because the X's are independent—recall that if X and Y are independent, then any function of X is independent of any function of Y). In the latter sum, all the expectations are the same. The asymptotic information matrix, evaluated at θ_0 is

$$\mathcal{I} = \lim \frac{1}{N} I_N = \mathrm{Var}(S(\theta_0))\,.$$

We show below that the asymptotic information matrix is the inverse of the asymptotic variance of $\hat{\theta}$. But we first need to find some expressions for the asymptotic variance. Define the (finite sample) Hessian as the $k \times k$ matrix as the first derivative of the score vector or, what is the same, as the second derivative of the log-likelihood function

$$H(\theta) = \Sigma_i \frac{d}{d\,\theta} S(\theta; x_i) = \Sigma_i \frac{d^2}{d\,\theta^2} l(\theta; x_i)\,.$$

The asymptotic Hessian is

$$\mathcal{H} = plim \frac{1}{N} H(\theta; x_1, ..., x_N) = E_0 H(\theta, X)\,.$$

Notice that the limit is taken over H as a function of the observed data, while the last expression is the expectation wrt. a random variable (in the last term, you will use the Hessian for just one term—the distribution is the same for each i).

To find the asymptotic distribution of the ML estimator, we do a first-order Taylor series expansion of $S(\theta)$ around the true value θ_0:

$$S(\hat{\theta}) = S(\theta_0) + H(\overline{\theta})\,(\hat{\theta} - \theta_0),$$

where $\overline{\theta}$ is a vector between $\hat{\theta}$ and θ_0 (which converges to θ_0 because $\hat{\theta}$ does). When the model is well specified, the Hessian has full rank and we get

$$\hat{\theta} = \theta_0 - H^{-1}S(\theta_0) \ .$$

(Remember that $S(\hat{\theta})$ is zero.) Now we will use the LLN on H and the CLT on S. We get

$$\sqrt{N}(\hat{\theta} - \theta_0) = -(\frac{1}{N}H)^{-1}N^{-1/2}S(\theta_0) \ .$$

where $N^{-1/2}S(\theta_0)$ by the CLT is asymptotically normally distributed:

$$N^{-1/2}S(\theta_0) \approx N(0, \mathcal{I}(\theta_0)) \, .$$

Then we have

$$\sqrt{N}(\hat{\theta} - \theta_0) \approx N(0, \mathcal{H}^{-1}\mathcal{I}\,\mathcal{H}^{-1}) \ .$$

It turns out that the negative of the asymptotic Hessian is identical to the asymptotic information matrix. The proof follows from the fact that the density integrates to unity for all θ:

$$\frac{d}{d\,\theta}\int \pounds(x;\theta)dx = 0$$

or, because we want things in terms of $l(\theta)$

$$\frac{d}{d\,\theta}\int \exp(l(\theta))dx = 0$$

Because $\frac{d}{d\,\theta}\pounds(x;\theta) = \frac{d}{d\,\theta}\exp(l(\theta)) = \exp(l(\theta))\frac{dl(\theta)}{d\,\theta}$, we have

$$\int \exp(l(\theta))\frac{dl(\theta)}{d\,\theta}dx = 0 \ .$$

This is constantly equal to zero so the derivative of this is also zero:

$$\int \exp(l(\theta))\frac{dl(\theta)}{d\,\theta}\frac{dl(\theta)'}{d\,\theta} + \exp(l(\theta))\frac{d^2l(\theta)}{d\,\theta^2}dx = 0 \ .$$

This holds for all θ; but in the case of the true parameter, integrating with respect to the density is the expectation and the expectation of $\frac{dl(\theta)}{d\,\theta}\frac{dl(\theta)'}{d\,\theta}$ is the asymptotic information matrix, so we have

$$\mathcal{I} + \mathcal{H} = 0 \, ;$$

which is the information matrix equality:

$$\mathcal{I} = -\mathcal{H}\,.$$

So

$$\sqrt{N}(\hat{\theta} - \theta_0) \approx N(0, -\mathcal{H}^{-1})\ .$$

(H has to negative, think of the scalar case, for a concave function to be a a strict maximum, the first derivative will be 0 and the second derivative will be negative), or, equivalently

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

Note that by the definition of the information matrix, we now can get an approximation to the variance of the ML estimator by using the so called "outer product of the gradients:"

$$\mathcal{I} = \lim \frac{1}{N} I_N(\hat{\theta}) = \lim \frac{1}{N} \Sigma_{i=1}^{N} G_i(\hat{\theta}) G_i(\hat{\theta})'\ ,$$

which converges to the asymptotic information matrix because $\hat{\theta}$ is consistent and by the law of large numbers.

Trinity of Tests The trinity of tests for testing $\theta = \theta_0$ where θ_0 satisfies $r(\theta_0) = 0$ (where we assume that the dimension of θ is k and the number of restrictions (dimension of r) is m where $m \leq k$

- The Likelihood Ratio (LR) test: $2{*}(l(\hat{\theta^u}) - l(\hat{\theta^r}))$ (where θ^u maximizes the likelihood, while θ^r maximizes the likelihood under the constraint).
- The Wald test: $N\, r(\hat{\theta^u})'\, (R(\hat{\theta^u})\, I_N^{-1} R(\hat{\theta^u})')^{-1}\, r(\hat{\theta^u})$
- The LM test: $S'(\hat{\theta^r}) I_N^{-1} S(\hat{\theta^r})$. Here, the score has to include the derivative of all the parameters as if they were unrestricted. In the restricted model, the score will be zero. (I found this a bit confusing until I worked through an example.)

These tests are all asymptotically $\chi^2(m)$ distributed under the null hypothesis. Note: R is the $k \times m$ matrix of derivatives of r (from the Δ-rule). Also note: I have not quite used the notation of the Davidson-MacKinnon book—strictly speaking I_N was defined for the true parameters, but when we use the outer product of the gradients we can only use what we have; namely, the estimated parameters. Under the null, this converges to the true parameter, and our tests are asymptotic, so it doesn't matter. And note: you can actually have as many restrictions as parameters—the maximized likelihood under the constraints is just the value of likelihood you get when you plug in the constrained values in this case.