Notes on Maximum Likelihood Estimation
(First Part)
Introduction to Econometrics
Christopher Flinn
Fall 2004

Most maximum likelihood estimation begins with the specification of an entire probability distribution for the data (i.e., the dependent variables of the analysis). We will concentrate on the case of one dependent variable, and begin with the no exogenous variables just for simplicity. When the distribution of the dependent variable is specified to depend on a finite-dimensional parameter vector, we speak of parametric maximum likelihood estimation. When not, such as the case of trying to estimate the entire distribution of the data making only some assumptions regarding the nature of its cumulative distribution function, we speak of nonparametric m.l.e. In this case the objective is to estimate an entire function, $F$, rather than a few parameters that characterize the distribution of the data under some specific functional form assumption regarding $F$.

1 Example 1. A Bernoulli random variable

Start with the simplest case of a binary dependent variable, like a coin toss. We are not sure how the coin is weighted - if it was weighted evenly then the probability of a head, $H$, on any toss is given by $.5$. If we call an $H$ a ‘1’ and a $T$ a ‘0’, then the probability distribution of the experiment is simply

<table>
<thead>
<tr>
<th>Event</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$\pi$</td>
</tr>
<tr>
<td>0</td>
<td>$1 - \pi$</td>
</tr>
</tbody>
</table>
We flip the coin \( N \) times, and the data consist of a string of \( N \) 1’s and 0’s. Let the outcome of flip \( i \) be denoted by \( d_i \). Then the joint probability of the sample is given by

\[
L(\{d_i\}_{i=1}^N, \pi) = \prod_{i=1}^{N} \pi^{d_i} (1 - \pi)^{(1-d_i)} \\
= \pi^{\sum d_i} (1 - \pi)^{\sum (1-d_i)} \\
= \pi^{N_1} (1 - \pi)^{N_0},
\]

where \( N_1 \) is the number of times ‘1’ appeared and \( N_0 \) is the number of times ‘0’ appeared, with \( N_1 + N_0 = N \).

The function \( L(\{d_i\}_{i=1}^N, \pi) \) is a joint probability function of the sample \( \{d_i\}_{i=1}^N \) given the parameter value \( \pi \). In this case, we don’t know the value of \( \pi \), and will use the one piece of information, the sample \( \{d_i\}_{i=1}^N \) to attempt to infer its value. Now we treat \( L \) as a function of the unknown parameter \( \pi \), the sample itself being implicitly an argument of the function, and choose that value of \( \pi \) that maximizes \( L \) as our estimator. Formally,

\[
\hat{\pi}_{ML} = \text{arg max}_\pi L(\pi).
\]

The value of \( \pi \) that maximizes \( L \) will also maximize any increasing monotonic function of \( L \), such as the natural logarithm. Then define

\[
\mathcal{L}(\pi) \equiv \ln L(\pi) \\
= N_1 \ln \pi + N_0 \ln (1 - \pi).
\]

Then the maximum likelihood estimator of \( \pi \) is the solution to the first order condition

\[
\frac{d\mathcal{L}(\hat{\pi}_{ML})}{d\pi} = 0 \\
\Rightarrow \frac{N_1}{\hat{\pi}_{ML}} - \frac{N_0}{1 - \hat{\pi}_{ML}} = 0 \\
\Rightarrow N_1 (1 - \hat{\pi}_{ML}) = N_0 \hat{\pi}_{ML} \\
\Rightarrow \hat{\pi}_{ML} = \frac{N_1}{N_1 + N_0} = \frac{N_1}{N}.
\]

We know that for this particular example the maximum likelihood estimator has good
“small sample” properties. For example, it is unbiased since

\[
E \hat{\pi}_{ML} = E \frac{N_1}{N} = \frac{1}{N} EN_1 = \frac{1}{N} E \sum_{i=1}^{N} d_i = \frac{1}{N} \sum_{i=1}^{N} Ed_i = \frac{1}{N} \sum_{i=1}^{N} \pi = N \frac{\pi}{N} = \pi.
\]

We even have the small sample variance of the estimator, which is simply

\[
VAR(\hat{\pi}_{ML}) = \frac{\pi(1 - \pi)}{N},
\]

an estimate of which is given by

\[
\hat{VAR}(\hat{\pi}_{ML}) = \hat{\pi}_{ML}(1 - \hat{\pi}_{ML})/N.
\]

One of the reasons for the good small sample properties of the ML estimator in this case is that it coincides with the sample mean - which is just the proportion of successes out of the \(N\) draws. In a random sample of \(N\) draws, the sample mean is an unbiased estimator of the population mean, remember, with the variance of the estimate of the sample mean equal to the variance of the random variable divided by the sample size - exactly what we have above. Proving consistency of the maximum likelihood estimator in this case is straightforward, since the estimator is unbiased and the limiting value of the variance is 0.

### 2 Example 2. The Negative Exponential distribution

In many cases the dependent variable is continuous, unlike in the previous example, but is restricted to a subset of the real line - most often the positive real line \(\mathbb{R}_+\). The simplest distribution defined on this interval is the negative exponential, with cumulative distribution function (c.d.f.) and probability distribution function (p.d.f.) given by

\[
F(t) = 1 - \exp(-\alpha t), \quad \alpha > 0, \ t > 0,
\]

\[
f(t) = \alpha \exp(-\alpha t).
\]
Say that we wanted to estimate the parameter $\alpha$, which given our functional form assumption, would serve to completely characterize the distribution of the random variable $T$. Let’s think of $T$ as the duration of time an individual spend unemployed while looking for a job.

We have a random sample of $N$ individuals, and we consider each individual’s unemployment duration as an independent draw from the same distribution $F$. In this case, the joint p.d.f. of all of the sample draws, the likelihood function, is given by

$$L(t_1, ..., t_N, \alpha) = \prod_{i=1}^{N} \alpha \exp(-\alpha t_i)$$

$$= \alpha^N \exp(-\alpha \sum_{i=1}^{N} t_i).$$

Then the log likelihood function is given by

$$\mathcal{L}(\alpha) = N \ln \alpha - \alpha \sum_{i=1}^{N} t_i.$$  

We want to maximize this function with respect to $\alpha$. The solution, if there is exactly one, is the maximum likelihood estimator for the problem. Does this function have a unique solution? The first derivative is

$$\frac{d\mathcal{L}(\alpha)}{d\alpha} = \frac{N}{\alpha} - \sum_{i=1}^{N} t_i,$$

and the second derivative is

$$\frac{d^2\mathcal{L}(\alpha)}{d\alpha^2} = -\frac{N}{\alpha^2} < 0.$$  

Since the second derivative is negative, the solution to the first order condition determines the unique maximizer of the function, and is the maximum likelihood estimator.

$$\frac{d\mathcal{L}(\hat{\alpha}_{ML})}{d\alpha} = 0$$

$$\Rightarrow \frac{N}{\hat{\alpha}_{ML}} - \sum_{i=1}^{N} t_i = 0$$

$$\Rightarrow \hat{\alpha}_{ML} = \frac{N}{\sum_{i=1}^{N} t_i} = \frac{1}{\bar{t}_N},$$

where $\bar{t}_N$ is the sample mean.
3 Example 2 continued with Observable Heterogeneity

Often times we have access to information on observable characteristics of individuals in addition to the value of the dependent variable for each one. For example, in the unemployment case discussed above, we may know each individual’s gender, years of schooling completed, region of the country in which they live, etc. Let the row vector of these characteristics for individual $i$ be given by $X_i$. We assume that there are $K$ columns in $X_i$, one of which may be a ‘1’ for each individual (corresponding to a sort of intercept term).

Say we continue to assume that the distribution of times in unemployment is negative exponential, and the one individual’s unemployment duration is unrelated (i.e., independent) of another’s. We just want to relax the “identical” part of the i.i.d. assumption. The easiest way to accomplish this, but by no means the only way, is to write

$$\alpha_i = \exp(X_i\beta),$$

where $\beta$ is an unknown $K \times 1$ parameter vector to be estimated. The likelihood function is now defined in terms of the parameter vector $\beta$, and we have

$$L(t_1, ..., t_N, \beta|X_1, ..., X_N) = \prod_{i=1}^{N} \exp(X_i\beta) \exp(-\exp(X_i\beta)t_i),$$

and the log likelihood function is written as

$$\mathcal{L}(\beta) = \sum_{i=1}^{N} \{X_i\beta - \exp(X_i\beta)t_i\}.$$ 

The first partial derivative of this function with respect to the parameter vector $\beta$ is

$$\frac{\partial \mathcal{L}(\beta)}{\partial \beta} = \sum_{i=1}^{N} \{X'_i - \exp(X_i\hat{\beta}_{ML})X'_i t_i\},$$

and the second partial of the log likelihood function with respect to $\beta$ is

$$\frac{\partial^2 \mathcal{L}(\beta)}{\partial \beta \partial \beta'} = \sum_{i=1}^{N} \{-\exp(X_i\beta)X'_i X_i t_i\} < 0,$$

so that there is a unique maximum likelihood estimator of $\beta$. The estimator solves the first order conditions, or

$$\frac{\partial \mathcal{L}(\hat{\beta}_{ML})}{\partial \beta} = \sum_{i=1}^{N} \{X'_i - \exp(X_i\hat{\beta}_{ML})X'_i t_i\} = 0.$$

There is no “closed form” solution to the first order conditions, and the solution must be obtained iteratively using numerical methods. We will say a few words on this topic below.
4 Example 3 The Latent Variable Discrete Choice Framework

We saw that there were some interpretive problems with the linear probability model - namely, the predicted probability of a “success” was often outside of the interval (0,1) - which is not a good thing for a probability. Instead we can set the problem up in a latent variable framework. We illustrate the technique in the binary choice case.

Think of a case in which choice is to buy one unit of some good or not, for example, a car. There is a utility gain from buying the car, but of course the cost of the car reduces resources available to spend on other goods. Let \( U^*(X) \) be the net utility gain from buying the car, which depends on characteristics \( X \), such as the price of the car, the individual’s residence and workplace (as indicators of how much she will use it), characteristics of the car such as color, type, etc., and the individual’s total income. We will assume a functional form for this indirect utility function, typically

\[
U^*_i = X_i \beta + \varepsilon_i,
\]

where \( X_i \) are the characteristics of the individual \( i \) and the good in question and \( \varepsilon_i \) is a random “preference shock.” We will make a parametric assumption regarding \( \varepsilon_i \), and assume that is independently and identically distributed across consumers, with c.d.f. \( F(\varepsilon) \). We will assume that \( E(\varepsilon) = 0 \).

The individual purchases the good if and only if the net utility from doing so is positive. If \( d_i = 1 \) when individual \( i \) purchases the good, then

\[
d_i = \begin{cases} 
1 & \text{iff } U^*_i > 0 \\
0 & \text{iff } U^*_i \leq 0 
\end{cases}.
\]

Then the probability that the individual purchases the good is given by

\[
p(d_i = 1|X_i) = p(X_i \beta + \varepsilon > 0) = p(\varepsilon > -X_i \beta) = 1 - F(-X_i \beta),
\]

and of course

\[
p(d_i = 0|X_i) = F(-X_i \beta).
\]

If we have a random sample of observations on \( \{d_i, X_i\}_{i=1}^N \), we define the likelihood function as

\[
L(d_1, ..., d_N, \beta | X_1, ..., X_N) = \prod_{i=1}^N p(d_i = 1|X_i)^{d_i} p(d_i = 0|X_i)^{1-d_i}
= \prod_{i=1}^N (1 - F(-X_i \beta))^{d_i} F(-X_i \beta)^{1-d_i},
\]
with the log likelihood function given by

\[ L(\beta) = \sum_{i=1}^{N} \{d_i \ln(1 - F(-X_i \beta)) + (1 - d_i) \ln(F(-X_i \beta))\}. \]

Now let’s look at the first partials of the log likelihood function, which are

\[ \frac{\partial L(\beta)}{\partial \beta} = \sum_{i=1}^{N} \left\{ \frac{d_i}{1 - F(-X_i \beta)} - \frac{(1 - d_i)}{F(-X_i \beta)} \right\} f(-X_i \beta)X_i'. \]

Under many distributional assumptions regarding \( F \) it is possible to prove that the matrix of second partial derivatives is negative definite, and so the solutions to the first order conditions give the unique maximum likelihood estimates of the parameters \( \beta \), if these parameters are “identified.” We will see what this means in the context of the probit model, where we assume that \( F \) is a normal distribution.

Specifically, we will begin by assuming that \( \varepsilon \) follows the \( N(0, \sigma^2_\varepsilon) \) distribution. Then we have that the probability of \( d = 1 \) is given by

\[ p(d_i = 1 | X_i) = 1 - F(-X_i \beta) = 1 - \Phi\left(\frac{-X_i \beta}{\sigma_\varepsilon}\right), \]

where \( \Phi \) is the standard normal cumulative distribution function. Because the normal is a symmetric distribution around the mean, and the mean is 0 in this case, we know that

\[ 1 - \Phi(-z) = \Phi(z), \]

so that we can write

\[ p(d_i = 1 | X_i) = \Phi\left(\frac{X_i \beta}{\sigma_\varepsilon}\right). \]

It follows that the log likelihood is written as

\[ L(\beta, \sigma^2_\varepsilon) = \sum_{i=1}^{N} \{d_i \ln(\Phi(X_i \beta/\sigma_\varepsilon)) + (1 - d_i) \ln(1 - \Phi(X_i \beta/\sigma_\varepsilon))\}. \]

We note something immediately from the log likelihood function concerning its dependence on the unknown parameters of the model. By assuming that \( \varepsilon \) was normally distributed with a variance equal to \( \sigma^2_\varepsilon \), we added an additional parameter to estimate \( \sigma^2_\varepsilon \). However, we can see from inspection of \( L(\beta, \sigma^2_\varepsilon) \) that \( \beta \) and \( \sigma^2_\varepsilon \) always appear in the same combination \( \beta/\sigma_\varepsilon \) everywhere in \( L \). This means that these parameters cannot be separately identified. Whatever values are taken by \( \beta \) and \( \sigma_\varepsilon \), if we multiply each parameter by some scalar \( k \) we will end up with the same ratio - \( \beta/\sigma_\varepsilon \). Thus only the ratio of \( \beta \) to
\(\sigma_\varepsilon\) can be estimated, and not each term individually. Sometimes we simply “normalize” the standard deviation of \(\varepsilon\) to be equal to 1, and in that way say that we have estimated \(\beta\) uniquely. But the important thing to remember is that the estimate is not really unique unless we have fixed the value of \(\sigma_\varepsilon\) by assumption.

In light of this discussion, let’s simply assume that \(\sigma_\varepsilon = 1\) so that we only have the \(K\) parameters in \(\beta\) to estimate. The first order conditions in this case are given by

\[
\frac{\partial L(\hat{\beta}_{ML})}{\partial \beta} = 0 = \sum_{i=1}^{N} \left\{ \frac{d_i}{\Phi(X_i\hat{\beta}_{ML})} - (1 - d_i) \frac{1}{(1 - \Phi(X_i\hat{\beta}_{ML}))} \right\} \phi(X_i\hat{\beta}_{ML}) X_i',
\]

where \(\phi\) denotes the standard normal probability density function. This can be rewritten as

\[
0 = \sum_{i=1}^{N} \left\{ \frac{d_i - \Phi(X_i\hat{\beta}_{ML})}{\Phi(X_i\hat{\beta}_{ML})(1 - \Phi(X_i\hat{\beta}_{ML}))} \right\} \phi(X_i\hat{\beta}_{ML}) X_i'.
\]

Note that the maximum likelihood estimator is chosen to set a certain average of the residuals equal to 0 in the sample, where the residual is

\(d_i - \Phi(X_i\hat{\beta}_{ML})\),

since the expected value of \(d_i\) is given by \(\Phi(X_i\hat{\beta}_{ML})\).

The logit model is developed in exactly the same way, except that we assume that the distribution of \(\varepsilon\) is logistic instead of normal. The inferences one derives from a binary choice model are typically not much different no matter what distribution of \(\varepsilon\) is assumed.

5 Computation of Maximum Likelihood Estimators

In most cases the computation of maximum likelihood estimators since the first order conditions do not have closed form solutions. We typically solve for the solution to the likelihood equations - that is, the first order conditions - iteratively using methods based on low order Taylor series expansions.

Let’s work with a scalar parameter example for simplicity, which we will denote by \(\theta\). The maximum likelihood estimate of \(\theta\), being the solution to the first order condition, sets

\[
\frac{\partial L(\hat{\theta}_{ML})}{\partial \theta} = 0.
\]

Assuming that \(L\) is continuously differentiable, let us expand this function around some
initial guess of the parameter value, which we will denote by \( \theta_0 \). Then write
\[
0 = \frac{\partial L(\hat{\theta}_{ML})}{\partial \theta} \\
\simeq \frac{\partial L(\theta_0)}{\partial \theta} + \left( \hat{\theta}_{ML} - \theta_0 \right) \frac{\partial^2 L(\theta_0)}{\partial \theta^2}.
\]
We can rearrange this last expression and write
\[
\hat{\theta}_{ML} = \theta_0 - \left[ \frac{\partial^2 L(\theta_0)}{\partial \theta^2} \right]^{-1} \frac{\partial L(\theta_0)}{\partial \theta}.
\tag{1}
\]
Now because this first-order Taylor series does not hold exactly, unless \( L \) happens to be quadratic in \( \theta \), the value on the left hand side of [1] does not actually satisfy the first order conditions. Instead, we would take the value value labelled \( \hat{\theta}_{ML} \) and call it \( \theta_1 \). If the log likelihood function is well-behaved, i.e., globally concave in \( \theta \), then \( \theta_1 \) should be closer to the true solution to the likelihood equation than \( \theta_0 \). We would proceed with our iteration of [1] until the values \( \theta_0, \theta_1, ..., \theta_m \), converged according to some criterion we would have to define. This criterion usually involves the first derivative \( \frac{\partial L(\theta_m)}{\partial \theta} \) being sufficiently close to 0, small changes in the value of the log likelihood between successive iterations, or small relative changes in the parameter guesses \( \theta_m \) and \( \theta_{m+1} \) between iterations.

5.1 Example: MLE for the Negative Exponential Distribution Parameter \( \alpha \).

We know from our earlier example that the MLE for \( \alpha \) is simply the inverse of the sample mean. Let’s look at the computation of \( \alpha \). Say that we have a sample of \( N = 100 \), and that the total time in unemployment of the sample members was \( T = \sum_{i=1}^{N} t_i = 148 \) months. Then since
\[
\frac{dL(\alpha)}{d\alpha} = \frac{N}{\alpha} - \sum_{i=1}^{N} t_i \\
= \frac{100}{\alpha} - 148,
\]
and
\[
\frac{d^2 L(\alpha)}{d\alpha^2} = -\frac{100}{\alpha^2},
\]
the updating equation is written as
\[
\alpha_{m+1} = \alpha_m - \left( \frac{\alpha_m^2}{100} \right) \left( \frac{100}{\alpha} - 148 \right).
\]
For illustration we chose an initial guess of $\alpha_0$ far away from the true value, which we know is $100/148$. We set $\alpha_0 = 5$, and found that:

\[
\begin{align*}
\alpha_1 & = 0.630000 \\
\alpha_2 & = 0.672588 \\
\alpha_3 & = 0.675662 \\
\alpha_4 & = 0.675676 \\
\alpha_5 & = 0.675676
\end{align*}
\]

Thus convergence was very rapid using this algorithm for this simple model. In practice, to solve the likelihood equations in more complicated cases, certain modifications are necessary to obtain good convergence results.