Econometrics I - MLE 1

Ryan T. Godwin

University of Manitoba

Maximum likelihood estimation

Least squares does not work when the dependent variable ($\boldsymbol{y}$) is:

1. The length of time it takes for something to happen (a strike, an insurance claim, an unemployment spell)
2. The number of things that happen (doctor visits, number of customers, bank failures, number of patents)
3. A yes/no dummy variable (whether person is in the labour force, whether a customer makes a purchase)

In these cases, the dependent variable is *limited* in some way. In (1) the time must be positive, $y_i \geq 0$. In (2) the counts are not continuous and non-negative, $y_i = 0, 1, 2, \ldots$. In (3) the values take on only 0 or 1. The linear model, and least-squares, has no way of recognizing or accounting for the limited nature of the dependent variable. A model estimated by LS will provide predicted values that are not allowed for y, and in most cases is misspecified so that the LS estimator is inconsistent.

In cases such as above, if we are willing to specify a *probability* distribution for y, then we can use maximum likelihood estimation (MLE). (Anytime we are willing to choose a distribution for y, we can use MLE). MLE is not the only option available: GMM, Bayesian, non-parametric, and others; but MLE has excellent properties and is a popular estimation strategy.

- MLE proposed by R. A. Fisher, 1921-1925.
- MLE is a parametric method.
- That is, we assume each sample data is generated from a known probability distribution function (pdf), $p(y_i \mid \boldsymbol{\theta})$. i.e. y_i comes from a "family".

Consider that we have random data $\boldsymbol{y} = \{y_1, \ldots, y_n\}$, and a parameter vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)'$. Our objective is to estimate $\boldsymbol{\theta}$. The probability of jointly observing the data is:

$$p(y_1, \ldots, y_n \mid \boldsymbol{\theta}) \qquad \text{"joint pdf"}$$

We can view $p\left(y_1, \ldots, y_n \mid \boldsymbol{\theta}\right)$ in two different ways:

1. As a function of $\{y_1, \ldots, y_n\}$, given $\boldsymbol{\theta}$.
2. As a function of $(\theta_1, \ldots, \theta_k)$, given $\boldsymbol{y}$. i.e., the data are given, the parameters vary.

The latter is called the **likelihood function**. Note:

$$L(\boldsymbol{\theta}) = L\left(\boldsymbol{\theta} \mid y_1, \ldots, y_n\right) = p\left(y_1, \ldots, y_n \mid \boldsymbol{\theta}\right)$$

Maximum likelihood estimator (MLE): The MLE of $\boldsymbol{\theta}$ (call it $\tilde{\boldsymbol{\theta}}$) is that value of $\boldsymbol{\theta}$ such that $L(\tilde{\boldsymbol{\theta}}) > L(\hat{\boldsymbol{\theta}})$, for all other $\hat{\boldsymbol{\theta}}$.

The idea behind MLE: "given the y_i's, what is the most likely $\boldsymbol{\theta}$ to have generated such a sample?"

Note:

- $\tilde{\boldsymbol{\theta}}$ need not be unique.
- $\tilde{\boldsymbol{\theta}}$ should locate the global max. of $L(\boldsymbol{\theta})$.
- If the sample data are independent then $L(\boldsymbol{\theta} \mid \boldsymbol{y}) = p(\boldsymbol{y} \mid \boldsymbol{\theta}) = \prod_{i=1}^{n} p\left(y_i \mid \boldsymbol{\theta}\right)$.
- Any monotonic transformation of $L(\boldsymbol{\theta})$ leaves the location of the extremum unchanged, e.g. $\log L(\boldsymbol{\theta})$

Some basic concepts and notation

1. Gradient/score vector: $\left[\frac{\partial \log L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right] \quad (k \times 1)$
2. Hessian matrix: $\left[\frac{\partial^2 \log L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\right] \quad (k \times k)$
3. Likelihood equations: $\frac{\partial \log L(\theta)}{\partial \theta} = 0 \quad (k \times 1)$

The optimization problem is:

$$\max_{\boldsymbol{\theta}} \prod_{i=1}^{n} L\left(\boldsymbol{\theta} \mid y_i\right)$$

To obtain the MLE, $\widetilde{\boldsymbol{\theta}}$, we solve the likelihood equation(s) and then check the second-order condition(s) to make sure we have maximized (not minimized) $L(\boldsymbol{\theta})$. If the Hessian matrix is at least n.s.d., then $L(\boldsymbol{\theta})$ is concave, and this is sufficient for a maximum. So, MLE is accomplished by:

1. Specifying the likelihood function.
   - This involves writing down an equation which states the joint likelihood (or joint probability) of observing the sample data, conditional on the unknown parameter values of the probability distribution function.
   - Independence of the y data is usually assumed (and will be for the purposes of this course).
   - Given independence, the likelihood function is obtained by multiplying together the probability of each y_i occurring.
2. Taking the natural log of the likelihood function. This usually simplifies the next step. The location of the maximum will not change.

3. Taking the first derivative of the log-likelihood function with respect to all parameters, setting each derivative equal to zero, and solving for the parameter values. The solution of the FOCs provides the formulas for the MLEs.
4. Checking to make sure the estimator in (3) attains a maximum (not a minimum). This involves taking the second derivatives of the log-likelihood function with respect to all parameters, so as to construct the Hessian matrix. If the Hessian is n.s.d., then the MLE achieves a global max.
5. Obtaining the variance of the MLEs for use in hypothesis testing. A variance-covariance matrix can be found by inverting the negative of the expected Hessian.

Properties of MLE

- MLE has very desirable asymptotic properties.
- Namely, MLE is Best Asymptotically Normal.
- That is, under mild assumptions, ML estimators are consistent, asymptotically efficient, and asymptotically Normally distributed.
- These properties are obtained by examining the asymptotic distribution of the MLE (which we will not derive in class):

$$\sqrt{n}\left(\tilde{\theta} - \theta_0\right) \xrightarrow{d} N\left[0, IA^{-1}(\theta)\right]$$

where

$$IA^{-1}(\theta) = \lim_{n\to\infty}\left(\frac{1}{n}[-E[H(\theta)]]^{-1}\right)$$

- $IA^{-1}(\theta)$ is the asymptotic information matrix, and $H(\theta)$ is the Hessian.
- The statement of the asymptotic distribution shows that the MLEs are consistent, asymptotically normal, and asymptotically efficient.
- The efficiency result relies on the Cramer-Rao lower bound. The Cramer-Rao lower bound is a theoretical minimum variance that any estimator can obtain. The MLE attains this minimum, that is, $IA^{-1}(\theta)$ is equal to the asymptotic Cramer-Rao lower bound.

The asymptotic distribution also allows us to see the variance of the MLEs in finite samples. The variance-covariance of $\widetilde{\boldsymbol{\theta}}$ for finite samples can be solved from the asymptotic variance:

$$\text{var}[\sqrt{n}(\tilde{\theta})] = n \times \text{var}(\tilde{\theta}) = \frac{1}{n}[-E[H(\theta)]]^{-1}$$

so,

$$\text{var}(\tilde{\theta}) = [-E[H(\theta)]]^{-1}$$

The matrix $-E[H]$ is termed the "Information Matrix" and is denoted by $I(\theta)$.

A very useful property of MLEs is their "invariance." That is, the estimator for $g(\theta)$ is $g(\tilde{\theta})$. Hence, an estimator for the variance-covariance of $\tilde{\theta}$ is:

$$\widetilde{\text{var}(\widetilde{\theta})} = [-E[H(\widetilde{\theta})]]^{-1}$$

Note that if misspecification occurs (if we have selected the wrong probability density function to begin with), we are not assured of any of the asymptotic properties.

Finite sample properties of MLEs

MLEs can be biased in finite samples (and typically are). We can evaluate bias much like we have done in previous parts of the course; by taking $E(\tilde{\theta})$. This knowledge can be used to correct for any bias (as in the case of $\tilde{\sigma}^2$). However, in most cases, there is no closed-form solution for the MLE itself, and numerical methods must be used to solve for the estimate. When the estimator does not have a closed form solution, we cannot take $E(\tilde{\theta})$, and we will not be able to "see" whether or not the estimator is biased. In this case, approximations or Monte Carlo experiments may be used to evaluate bias.