Binary choice – 3.3 Maximum likelihood estimation

Michel Bierlaire

Maximum likelihood estimation

We now estimate the values of the unknown parameters $\beta_1,\ldots,\beta_K$ from a sample of observations drawn at random from the population. Each observation of this sample consists of the following:

1. An indicator variable defined as
$$y_{in} = \begin{cases} 1 & \text{if individual } n \text{ chose alternative } i, \\ 0 & \text{if individual } n \text{ chose alternative } j. \end{cases}$$
2. Two vectors of explanatory variables $x_{in} = h(z_{in}, S_n)$ and $x_{jn} = h(z_{jn}, S_n)$, each containing K values.

For notational convenience, we also define $y_{jn} = 1 - y_{in}$.

As an example, consider a transportation mode choice problem (train or car), where the utility functions are specified as reported in Table 1. Consider also the sample of 3 individuals presented in Table 2.

Using the above notations, we have

$$y_{i1} = 1, y_{j1} = 0, y_{i2} = 0, y_{j2} = 1, y_{i3} = 0, y_{j3} = 1.$$

The values of the variables x are:

$$\begin{array}{lllllllll}
x_{i1} = (1 & 5 & 0 & 1.17 & 0 & 0 & 1 & 0 & 0), \\
x_{j1} = (0 & 40 & 0 & 0 & 2.5 & 0 & 0 & 0 & 0), \\
x_{i2} = (1 & 8.33 & 2 & 0 & 0 & 0 & 0 & 1 & 1), \\
x_{j2} = (0 & 7.8 & 0 & 0 & 1.75 & 1 & 0 & 0 & 0), \\
x_{i3} = (1 & 3.2 & 0 & 2.55 & 0 & 0 & 0 & 1 & 0), \\
x_{j3} = (0 & 40 & 0 & 0 & 2.67 & 0 & 0 & 0 & 0).
\end{array}$$

	Car	Train
β_1	1	0
β_2	cost of trip by car	cost of trip by train
β_3	travel time by car (hours) if trip purpose is work, 0 otherwise	0
β_4	travel time by car (hours) if trip purpose is not work, 0 otherwise	0
β_5	0	travel time by train (hours)
β_6	0	1 if first class is preferred, 0 otherwise
β_7	1 if commuter is male, 0 otherwise	0
β_8	1 if commuter is the main earner in the family, 0 otherwise	0
β_9	1 if commuter had a fixed arrival time, 0 otherwise	0

Table 1: Specification table of the binary mode choice model

The choice model is

$$P_n(i) = \frac{e^{V_{in}}}{e^{V_{in}} + e^{V_{jn}}}, \tag{1}$$

where

$$V_{in} = \sum_{k=1}^{K} \beta_k x_{ink} \tag{2}$$

$$V_{jn} = \sum_{k=1}^{K} \beta_k x_{jnk}. \tag{3}$$

Given a sample of N observations, we want to find estimates $\hat{\beta}_1, \ldots, \hat{\beta}_K$ that have some or all of the desirable properties of statistical estimators. We consider in detail the most widely used estimation procedure — maximum likelihood. The maximum likelihood estimators have the following desired properties:

	Individual 1	Individual 2	Individual 3
Train cost	40.00	7.80	40.00
Car cost	5.00	8.33	3.20
Train travel time	2.50	1.75	2.67
Car travel time	1.17	2.00	2.55
Gender	M	F	F
Trip purpose	Not work	Work	Not work
Class	Second	First	Second
Main earner	No	Yes	Yes
Arrival time	Variable	Fixed	Variable
Choice	Train	Car	Car

Table 2: A sample of three individuals

1. They are consistent in the sense of convergence to true values as sample size gets larger.

2. They are asymptotically normally distributed in the sense of the Central Limit Theorem.

3. They are asymptotically efficient, and hence their variance attains the Cramer-Rao lower bound.

The maximum likelihood estimation procedure is conceptually quite straightforward. It consists in identifying the value of the unknown parameters such that the joint probability of the observed choices as predicted by the model is the highest possible. This joint probability is called the *likelihood* of the sample. And it is a function of the parameters of the model.

In the above example, the likelihood of the sample of 3 individuals is calculated as follows:

- individual 1 has chosen the car, and this choice is predicted by the model with probability $P_1(i)$,

- individual 2 has chosen the train, and this choice is predicted by the model with probability $P_2(j)$,

- individual 3 has chosen the train, and this choice is predicted by the model with probability $P_3(j)$.

Consequently, the probability that the model predicts all three observations is

$$\mathcal{L}^*(\beta_1, \ldots, \beta_9) = P_1(i)P_2(j)P_3(j). \tag{4}$$

If this value is calculated for $\beta_k = 0$, $k = 1, \ldots, K$, we obtain

$$\mathcal{L}^* = \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = 0.125. \tag{5}$$

If this value is calculated for the values of

$$\beta = (3.04, -0.0527, -2.66, -2.22, -0.576, 0.961, -0.850, 0.383, -0.624),$$

we have

$$\mathcal{L}^* = 0.947 \cdot 0.924 \cdot 0.225 = 0.197. \tag{6}$$

This value of the likelihood is higher. But we do not know if it is the highest possible.

This can be generalized to a sample of N observations assumed to be independently drawn from the population. As discussed above, the likelihood of the sample is the product of the likelihoods (or probabilities) of the individual observations. It is defined as follows:

$$\mathcal{L}^*(\beta_1, \beta_2, \ldots, \beta_K) = \prod_{n=1}^{N} P_n(i)^{y_{in}} P_n(j)^{y_{jn}}, \tag{7}$$

where $P_n(i)$ and $P_n(j)$ are functions of β_1,...,β_K. Note that each factor represents the choice probability of the chosen alternative. Indeed,

$$P_n(i)^{y_{in}} P_n(j)^{y_{jn}} = \begin{cases} P_n(i) & \text{if } y_{in} = 1, y_{jn} = 0 \\ P_n(j) & \text{if } y_{in} = 0, y_{jn} = 1. \end{cases}$$

It is more convenient to analyze the logarithm of $\mathcal{L}^*$, denoted as $\mathcal{L}$ and called the *log likelihood*, because the logarithm of a product of elements is easier to manipulate, being equal to the sum of the logarithms of the elements. Moreover, the value of the likelihood is always between 0 and 1, and usually very small, especially when N is large. The range of values of the log likelihood is much larger, as it can take any negative value (from $-\infty$ to 0) and can be represented better in computers. The log likelihood is written as follows:

$$\mathcal{L}(\beta_1, \ldots, \beta_K) = \sum_{n=1}^{N} (y_{in} \ln P_n(i) + y_{jn} \ln P_n(j)). \tag{8}$$

where β is the vector with entries $\beta_1, \ldots, \beta_K$. We are looking for estimates $\hat{\beta}_1$, $\hat{\beta}_2$,...,$\hat{\beta}_K$ that solve

$$\max \mathcal{L}(\hat{\beta}) = \mathcal{L}(\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_K), \tag{9}$$

where $\hat{\beta}$ is the vector with entries $\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_K$. The optimization problem is solved using dedicated algorithms.

If a solution exists, it must satisfy the necessary first order conditions:

$$\frac{\partial \mathcal{L}}{\partial \beta_k}(\widehat{\beta}) = \sum_{n=1}^{N} \left(y_{in} \frac{\partial P_n(i)/\partial \beta_k}{P_n(i)} + y_{jn} \frac{\partial P_n(j)/\partial \beta_k}{P_n(j)} \right) = 0, \ k = 1, \ldots, K, \tag{10}$$

or in vector form

$$\frac{\partial \mathcal{L}}{\partial \beta}(\widehat{\beta}) = 0. \tag{11}$$

The term $\partial \mathcal{L}(\widehat{\beta})/\partial \beta$ is the vector of first derivatives of the log likelihood function with respect to the unknown parameters, evaluated at the estimated value of the parameters. Each entry k of the vector $\partial \mathcal{L}(\widehat{\beta})/\partial \beta$ represents the slope of the multi-dimensional log likelihood function along the corresponding kth axis. If $\widehat{\beta}$ corresponds to a maximum of the function, all these slopes must be zero, justifying (10).

Solving the optimization problem requires an iterative procedure. It starts with arbitrary values for the parameters (provided by the analyst, or all set to zero if no value can be guessed). If the first derivatives of the log likelihood function are zero, a solution has been found. If not, they provide information about the slope of the function, and a direction of "hill-climbing" can be identified. This direction is followed for a while, until a new set of values is found, corresponding to a higher log likelihood. The process is restarted from this new set of values, until convergence to the maximum is reached.

A family of algorithms commonly used in practice is called *Newton*'s method. At each iteration ℓ, a quadratic model of the log likelihood function is built around the current iterate $\beta^{(\ell)}$. This quadratic model is such that the value of the model and of its first and second derivatives are the same at $\beta^{(\ell)}$ as the log likelihood function:

$$m(\beta; \beta^{(\ell)}) = \mathcal{L}(\beta^{(\ell)}) + (\beta - \beta^{(\ell)})^T \nabla \mathcal{L}(\beta^{(\ell)}) + \frac{1}{2} (\beta - \beta^{(\ell)})^T \nabla^2 \mathcal{L}(\beta^{(\ell)}) (\beta - \beta^{(\ell)}), \tag{12}$$

where $\nabla\mathcal{L}(\beta^{(\ell)})$ is the gradient, that is the vector of the first derivatives of the log likelihood function evaluated at $\beta^{(\ell)}$, and $\nabla^2\mathcal{L}(\beta^{(\ell)})$ is the matrix of the second derivatives of the log likelihood function evaluated at $\beta^{(\ell)}$. The kth entry of $\mathcal{L}(\beta^{(\ell)})$ is $\partial\mathcal{L}(\beta^{(\ell)})/\partial\beta_k$, and the entry in the kth row and the mth column of $\nabla^2\mathcal{L}(\beta^{(\ell)})$ is

$$\frac{\partial^2\mathcal{L}(\beta^{(\ell)})}{\partial\beta_k\partial\beta_m}. \tag{13}$$

The approximation of the log likelihood function by the quadratic model is illustrated in Figure 1 for a log likelihood function with only one parameter, where both the log likelihood function and the quadratic model at $\beta^{(\ell)}$ are displayed. Note that both functions coincide at $\beta^{(\ell)}$, and have the same slope (first derivative) and curvature (second derivative) at that point. The next iterate is selected as the value of the parameters maximizing the quadratic model, that is

$$\beta^{(\ell+1)} = \beta^{(k)} - \nabla^2\mathcal{L}(\beta^{(\ell)})^{-1}\nabla(\beta^{(\ell)}), \tag{14}$$

as illustrated in Figures 1 and 2 for two successive iterations.

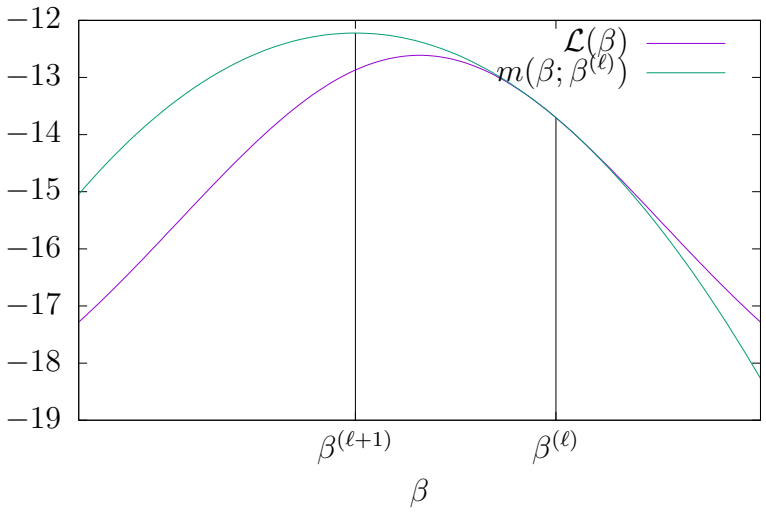

Figure 1: Illustration of Newton's method for optimization

It is numerically obtained by solving the system of linear equations

$$\nabla^2\mathcal{L}(\beta^{(\ell)})d = -\nabla(\beta^{(\ell)}), \tag{15}$$

to obtain the direction d, and then calculating

$$\beta^{(\ell+1)} = \beta^{(\ell)} + d. \tag{16}$$

The procedure continues until the gradient is sufficiently close to zero, depending on the level of precision that is required. In practice, it happens when the norm of the gradient is below a user-specified threshold Γ, that is

$$\left\| \frac{\partial \mathcal{L}(\beta)}{\partial \beta} \right\| = \sqrt{\sum_k \left(\frac{\partial \mathcal{L}(\beta)}{\partial \beta_k} \right)^2} \leq \Gamma.$$

A typical value for Γ is 10^{-6}.

Actually, the method described above is not guaranteed to converge, and variants involving a scaled version of d have to be used, that is

$$\beta^{(\ell+1)} = \beta^{(\ell)} + \alpha d, \ \alpha > 0. \tag{17}$$

We refer the reader to Bierlaire (2015) for more details on optimization algorithms.

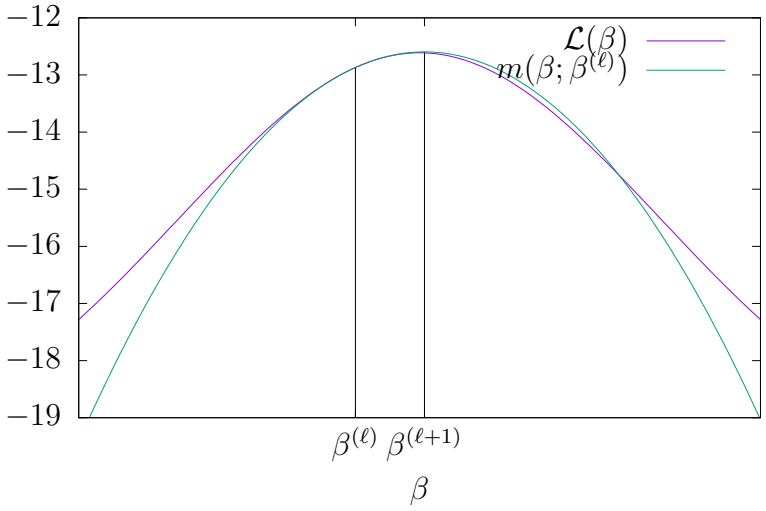

Figure 2: Illustration of Newton's method for optimization: second iteration

References

Bierlaire, M. (2015). *Optimization: Principles and Algorithms*, EPFL Press.