

Lecture Note 7

Maximum Likelihood Estimation—Part I

Introduction:

Let y_1, \dots, y_n be an i.i.d. random sample from $N(\mu, \sigma^2)$, that is, the probability density function (pdf) of y_i is given by

$$f(y_i; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2\sigma^2}(y_i - \mu)^2\right\}.$$

Since y_1, \dots, y_n are independent the joint pdf for y_1, \dots, y_n can be written as

$$\begin{aligned} f(y_1, \dots, y_n; \mu, \sigma^2) &= \prod_{i=1}^n f(y_i; \mu, \sigma^2), \\ &= \frac{1}{\sigma^n (2\pi)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mu)^2\right\}. \end{aligned}$$

This is a pdf, conditional on the parameter vector $\theta = (\mu, \sigma^2)'$.

The joint pdf of the observations, seen as a function of the parameter vector θ , is called the *likelihood function*, denoted by $L(\theta|y)$, where $y = (y_1, \dots, y_n)'$.

The likelihood function can be used to obtain an estimator for the population parameter vector θ_0 . A natural idea is for a given set of observations to maximize the probability of those observations, i.e., maximize the joint pdf (likelihood function) w.r.t. θ . This gives the *Maximum Likelihood Estimator* (MLE), which is defined by

$$\hat{\theta}_{ML} = \arg \max_{\theta \in \Theta} L(\theta|y),$$

where Θ is the parameter space that has all possible values of the parameter vectors, and, of course $\theta_0 \in \Theta$.

In the example above

$$\begin{aligned} L(\theta|y) &= \frac{1}{\sigma^n (2\pi)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mu)^2\right\}, \quad \text{and} \\ \Theta &= \left\{(\mu, \sigma^2) : |\mu| < \infty, \text{ and } \sigma^2 > 0\right\}. \end{aligned}$$

Equivalent definition of the MLE is:

$$\hat{\theta}_{ML} = \arg \max_{\theta \in \Theta} \ln L(\theta|y),$$

because the natural logarithm is a monotonic transformation. We usually denote

$$\begin{aligned} l(\theta|y) &\equiv \ln L(\theta|y), \\ &= \sum_{i=1}^n \ln f(y_i; \mu, \sigma^2), \end{aligned}$$

and call $l(\theta|y)$ the *log likelihood function*.

It is better to work with the log likelihood function for two reasons: (a) it is easier to differentiate it to get the first-order conditions; and (b) the log likelihood is a sum of random variables to which laws of large numbers can be easily applied.

In the example above

$$l(\theta|y) = -\frac{n}{2} \ln 2\pi - \frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mu)^2. \quad (7.1)$$

Solving the first-order conditions for (7.1), we find that

$$\begin{aligned} \hat{\mu} &= \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i, \quad \text{and} \\ \hat{\sigma}^2 &= \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2. \end{aligned}$$

Note that both $\hat{\mu}$ and $\hat{\sigma}^2$ are functions of only observed variables (i.e., they are statistics), so they provide computable estimates for their population counterparts μ_0 and σ_0^2 .

Classical Linear Regression (CLR) Model with Normal Errors:

Consider the linear regression model

$$\begin{aligned} y_i &= x_i' \beta + \varepsilon_i, \quad \text{where} \\ \varepsilon_i | x_i &\sim N(0, \sigma_\varepsilon^2), \end{aligned} \quad (7.2)$$

This implies that the conditional pdf of y_i , conditional on x_i (and the parameters β and σ_ε^2) is given by

$$f(y_i; x_i, \theta) = \frac{1}{\sigma \sqrt{2\pi}} \exp \left\{ -\frac{1}{2\sigma^2} (y_i - x_i' \beta)^2 \right\}.$$

For the joint pdf of y_i and x_i we need to make some assumptions. Typically we assume that

Ass.1: $x_i, i = 1, \dots, n$, are non-stochastic $k \times 1$ vectors.

Ass.2: $x_i, i = 1, \dots, n$, are i.i.d. random variables with a pdf given by $g(x_i)$, that does not depend on the parameter vector θ .

Under Ass.2 we have then that

$$f(y_i, x_i; \theta) = f(y_i; x_i, \theta) g(x_i).$$

Hence,

$$L(\theta|y, X) = \prod_{i=1}^n \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2\sigma^2} (y_i - x_i'\beta)^2\right\} g(x_i),$$

or alternatively

$$= \sum_{i=1}^n \ln g(x_i) - \frac{n}{2} \ln 2\pi - \frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - x_i'\beta)^2.$$

The MLE for β_0 and σ_0^2 are, respectively,

$$\begin{aligned} \hat{\beta}_{ML} &= b = (X'X)^{-1} X'y, \quad \text{and} \\ \hat{\sigma}^2 &= \frac{1}{n} \sum_{i=1}^n e_i^2, \quad \text{where } e_i = y_i - x_i'\hat{\beta}_{ML}. \end{aligned}$$

Remarks:

1. The ML estimator for β_0 is the same as the OLS estimator for β_0 .
2. The ML estimator for σ_0^2 is

$$\begin{aligned} \hat{\sigma}_\varepsilon^2 &= \frac{n-k}{n} S^2, \quad \text{where} \\ S^2 &= \frac{1}{n-k} \sum_{i=1}^n e_i^2, \end{aligned}$$

and S^2 is an unbiased estimator for σ_ε^2 .

3. Note that as $n \rightarrow \infty$

$$\hat{\sigma}^2 \xrightarrow{p} \sigma_\varepsilon^2.$$

4. The MLE does not depend on the distribution of x_i , if that distribution does not depend on the parameter of interest θ .
5. The x_i 's need not be independent, as long as their joint distribution does not depend on θ .

Binary Logit Model:

Here we consider an example for dependent variable y that takes only two possible values that we label 0 and 1, e.g. $y = 1$ if a consumer buys a new car, and $y = 0$ otherwise.

Let the independent variables be x_i , a $k \times 1$ vector. In general we cannot predict with certainty whether $y = 0$ or $y = 1$, even if we know x . The only things that we can compute are

$$\Pr(y_i = 1|x_i) = p(x_i), \quad \text{and}$$

$$\Pr(y_i = 0|x_i) = 1 - p(x_i).$$

Suppose that we already chose a parametric distribution for $p(x)$, and it is given by the *logit* model

$$\begin{aligned} p(x_i) &= \frac{\exp\{x_i'\beta\}}{1 + \exp\{x_i'\beta\}}, \quad \text{and} \\ 1 - p(x_i) &= 1 - \frac{\exp\{x_i'\beta\}}{1 + \exp\{x_i'\beta\}} = \frac{1}{1 + \exp\{x_i'\beta\}}. \end{aligned}$$

Then,

$$f(y_i|x_i; \beta) = \left(\frac{\exp\{x_i'\beta\}}{1 + \exp\{x_i'\beta\}} \right)^{y_i} \left(\frac{1}{1 + \exp\{x_i'\beta\}} \right)^{1-y_i}.$$

Suppose also that x_i , $i = 1, \dots, n$, are i.i.d. with the common pdf $g(x)$.

Then,

$$l(\theta) = \sum_{i=1}^n \left\{ y_i \ln \left(\frac{\exp\{x_i'\beta\}}{1 + \exp\{x_i'\beta\}} \right) + (1 - y_i) \ln \left(\frac{1}{1 + \exp\{x_i'\beta\}} \right) \right\} + \sum_{i=1}^n \ln g(x_i). \quad (7.3)$$

The first-order conditions for (7.3) are:

$$\sum_{i=1}^n \left(y_i - \frac{\exp\{x_i'\hat{\beta}_{ML}\}}{1 + \exp\{x_i'\hat{\beta}_{ML}\}} \right) x_i = 0. \quad (7.4)$$

From the system in (7.4) we cannot find a close-form solution for $\hat{\beta}_{ML}$. We have a nonlinear system of equations that has to be solved numerically. Because there is no closed-form solution for $\hat{\beta}_{ML}$ the derivation of the sampling distribution cannot proceed as in the linear regression model and we have to resort to some more complicated method in order to establish the sampling distribution for $\hat{\beta}_{ML}$.

Sampling Distribution for ML Estimator:

We develop here a general theory for the sampling distribution of the MLE $\hat{\theta}_{ML}$. This theory answers to two main questions:

1. What is the sampling distribution of $\hat{\theta}_{ML}$ if the sample size goes to ∞ ?
2. Is this asymptotic sampling distribution useful for inference (i.e., for constructing confidence intervals and tests)?

Notation:

y dependent variable

x vector of independent variables (regressors)

θ vector of parameters of interest. We will also denote the parameter space by Θ ($\theta \in \Theta$).

θ_0 the value of the parameter vector θ in the population. Note that $\theta_0 \in \Theta$.

The population density of x and y , conditional on the true parameter vector θ_0 is given by

$$f(y, x; \theta_0) = f(y|x; \theta_0) g(x),$$

where $f(y|x; \theta_0)$ represents the model and the density $g(x)$ does not depend on θ_0 .

Goal:

Our goal is to estimate the population parameter vector θ_0 and to conduct some inference. That is, to provide confidence intervals for θ_0 , and to test hypotheses regarding this parameter vector. For that purpose we collect data/observations on (y, x) that we denote by (y_i, x_i) , $i = 1, \dots, n$.

We will assume throughout that

1. The selection of an observation into the sample is random (i.e., the probability of being selected does not depend on (y, x)); and
2. The population is large.

Then, we say that (y_i, x_i) , $i = 1, \dots, n$, is a random sample from a distribution with density $f(y, x; \theta_0)$.

The *likelihood function*:

$$L(\theta) = \prod_{i=1}^n f(y_i, x_i; \theta) = \prod_{i=1}^n f(y_i|x_i; \theta) \prod_{i=1}^n g(x_i). \quad (7.5)$$

The *log likelihood function*:

$$l(\theta) = \ln L(\theta) = \sum_{i=1}^n \ln f(y_i|x_i; \theta) + \sum_{i=1}^n \ln g(x_i). \quad (7.6)$$

Note that the marginal density of x is a multiplicative factor that does not depend on θ in (7.5), and an additive factor that does not depend on θ in (7.6).

Maximum Likelihood Estimator (definition):

The maximum likelihood estimator is defined by

$$\hat{\theta} = \arg \max_{\theta \in \Theta} l(\theta).$$

Note that if the marginal distribution of x is independent of θ , as we assumed above, then the MLE is independent of the distribution of x . Hence, we need not specify this distribution at all. What we need to specify is only the conditional distribution of y , conditional on x .

In the example of the logit model we have

$$f(y|x, \theta) = \begin{cases} \left[\frac{e^{x'\theta}}{1+e^{x'\theta}} \right]^y \cdot \left[\frac{1}{1+e^{x'\theta}} \right]^{1-y} & \text{if } y = 0, 1, \\ 0 & \text{otherwise.} \end{cases}$$

The log likelihood (omitting terms that do not depend on θ) is

$$\begin{aligned} l(\theta) &= \sum_{i=1}^n \left\{ y_i \ln \left[\frac{e^{x_i'\theta}}{1+e^{x_i'\theta}} \right] + (1-y_i) \ln \left[\frac{1}{1+e^{x_i'\theta}} \right] \right\}, \\ &= \sum_{i=1}^n \left\{ y_i (x_i'\theta) - \ln (1+e^{x_i'\theta}) \right\}. \end{aligned}$$

The MLE satisfies

$$\frac{\partial l(\hat{\theta})}{\partial \theta} = \sum_{i=1}^n \left(y_i - \frac{\exp \{x_i'\hat{\beta}_{ML}\}}{1 + \exp \{x_i'\hat{\beta}_{ML}\}} \right) x_i = 0. \quad (7.7)$$

Note that from the model above we have

$$E(y_i|x_i) = \frac{\exp \{x_i'\hat{\beta}_{ML}\}}{1 + \exp \{x_i'\hat{\beta}_{ML}\}}.$$

Hence, the FOC merely state that the “residual”

$$y_i - \frac{\exp \{x_i' \hat{\beta}_{ML}\}}{1 + \exp \{x_i' \hat{\beta}_{ML}\}}$$

is orthogonal to the x_i 's.

In general, we find the MLE by solving the FOC:

$$\frac{\partial l(\hat{\theta})}{\partial \theta} = \frac{\partial l(\theta)}{\partial \theta} \Big|_{\theta=\hat{\theta}} = \frac{\partial \ln L(\theta)}{\partial \theta} \Big|_{\theta=\hat{\theta}} = 0.$$

This vector of derivatives with respect to θ is called *the score*.

We usually look at the normalized log likelihood function, that is, the likelihood function divided by n :

$$\frac{1}{n} \frac{\partial l(\hat{\theta})}{\partial \theta} = \frac{1}{n} \sum_{i=1}^n \ln f(y_i, x_i; \theta). \quad (7.8)$$

The MLE maximizes the normalized log likelihood. The normalized log likelihood is a sample average of (random) functions of θ . This sample average converges to a deterministic function, given by

$$E [\ln f(y, x; \theta)],$$

where the expectation is taken with respect to the population distribution of (y, x) with density $f(y, x; \theta_0)$.

If the limit function (of θ) has a unique maximum in θ_0 , then we expect that the MLE converges to θ_0 if the sample size becomes large. This is an application of the *analogy principle*; the most basic idea in parameter estimation. In this principle we find a characterization of the population parameter as the solution to an optimization problem (or as a solution to a set of equations) that is defined for the population. Next, we find the corresponding optimization problem for the sample, that is, we treat the sample as if it were the population.

If the sample function that is being optimized converges (in the right sense, see below) to the population function, then the sample optimand, i.e., $\hat{\theta}$, converges to the population optimand, i.e. θ_0 .

This idea works if we have the proper type of convergence of the sample function to the population function, and if the maximum for the population function is unique.

Uniform Law of Large Numbers:

Equation (7.8) is an average of the form

$$\frac{1}{n} \sum_{i=1}^n g(W_i; \theta),$$

where $W_i = (Y_i, X_i)$ simply denote the data. Here W_1, \dots, W_n is a random sample, i.e., a sequence of i.i.d. random variables.

For any value of θ (that is, consider θ as fixed), if

$$E [|g(W, \theta)|] < \infty,$$

then by the *strong law of large numbers* (SLLN) we have

$$\frac{1}{n} \sum_{i=1}^n g(W_i; \theta) \longrightarrow E [g(W, \theta)], \quad \text{almost surely.}$$

Note that almost sure (a.s.) convergence implies also convergence in probability to the same limit, that is,

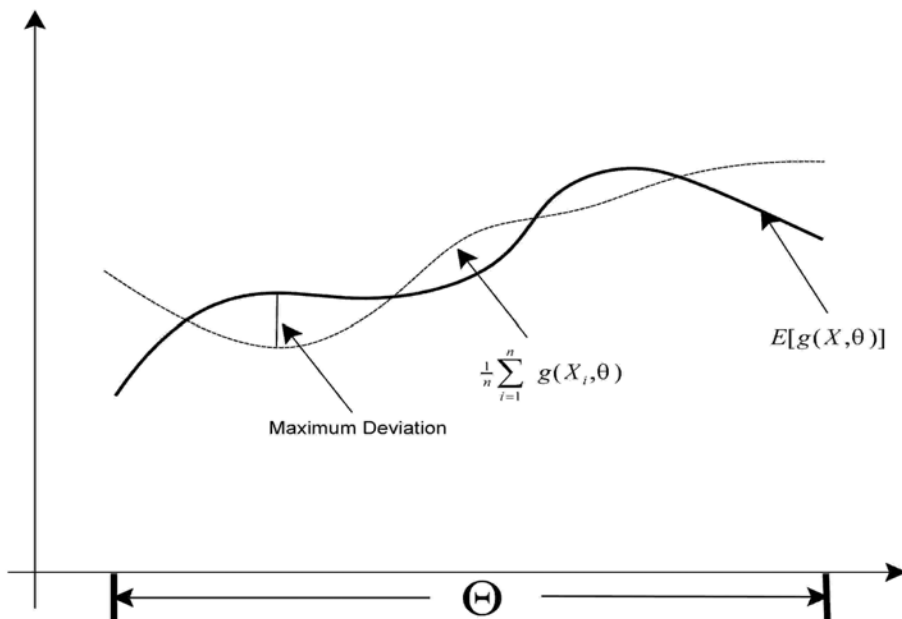
$$\frac{1}{n} \sum_{i=1}^n g(W_i; \theta) \longrightarrow E [g(W, \theta)], \quad \text{in probability.}$$

We can make this statement for any individual θ in the parameter space, that is,

$$\frac{1}{n} \sum_{i=1}^n g(W_i; \theta) \longrightarrow E [g(W, \theta)], \quad \text{almost surely, } \textit{pointwise}.$$

However, we want something stronger (see Figure 7.1) below.

Figure 7.1: Uniform convergence



Instead of considering deviations for alternative values of θ separately, we consider the maximal deviation across *all* θ 's. This makes a difference, if θ varies continuously in Θ , i.e., if θ takes on an (uncountably) infinite number of values. The maximal deviation is

$$D_n = \sup_{\theta \in \Theta} \left| \frac{1}{n} \sum_{i=1}^n g(W_i; \theta) - E[g(W, \theta)] \right|,$$

where D_n , $n > 0$, is a sequence of random variables.

If $D_n \rightarrow 0$ a.s., then we say that

$$\frac{1}{n} \sum_{i=1}^n g(W_i; \theta) \longrightarrow E[g(W, \theta)], \quad \text{almost surely, uniformly on } \Theta. \quad (7.9)$$

Again, the a.s. convergence in (7.9) implies uniform convergence in probability.

Theorem (Uniform Strong Law of Large Numbers):

Let W_1, \dots, W_n, \dots be a random sample, and let $g(W, \theta)$ be a continuous function of θ , with $\theta \in \Theta$, a compact (closed and bounded) set. If

$$E \left[\sup_{\theta \in \Theta} |g(W, \theta)| \right] < \infty,$$

then

$$\frac{1}{n} \sum_{i=1}^n g(W_i; \theta) \longrightarrow E[g(W, \theta)], \quad \text{almost surely, uniformly on } \Theta,$$

and $E[\sup_{\theta \in \Theta} |g(W, \theta)|]$ is continuous in θ .

We already assumed that $E[|g(W, \theta)|]$ exists, that is, $g(W, \theta)$ is a measurable function of x .

Now, we need to ensure that $E[\sup_{\theta \in \Theta} |g(W, \theta)|]$ exists, i.e., that $\sup_{\theta \in \Theta} |g(W, \theta)|$ is a measurable function of x . A sufficient condition for the latter is that $g(W, \theta)$ is a continuous function of x for all $\theta \in \Theta$.

An alternative is to assume that $\sup_{\theta \in \Theta} |g(W, \theta)|$ is bounded by some function $M(W)$, with $E[M(W)] < \infty$, using the *dominating convergence theorem*.

We need to apply the uniform SLLN to the normalized log likelihood function. Hence, we must consider

$$\begin{aligned} \sup_{\theta \in \Theta} |\ln f(y, x; \theta)| &= \sup_{\theta \in \Theta} |\ln f(y|x; \theta) + \ln g(x)| \\ &\leq \sup_{\theta \in \Theta} |\ln f(y|x; \theta)| + |\ln g(x)|, \end{aligned}$$

where the inequality is due to the triangular inequality property.

Now we use some additional assumptions. We discuss these assumptions in turn below.

Assumption X: *The marginal distribution of x is such that*

$$E [|\ln g(x)|] < \infty.$$

So, we only need to check that

$$E \left[\sup_{\theta \in \Theta} |\ln f(y|x; \theta)| \right] < \infty. \quad (7.10)$$

In the example for the binary logit model we need to verify that the SLLN holds. This will hold if (7.10) holds.

For simplicity of illustration consider the case in which we have only one parameter, and therefore only one explanatory variable x . Let also $\Theta = [\underline{\theta}, \bar{\theta}]$, with

$$-\infty < \underline{\theta} < \bar{\theta} < \infty.$$

That is, Θ is a closed and bounded interval in \Re . Then,

$$\begin{aligned} & \sup_{\theta \in [\underline{\theta}, \bar{\theta}]} \left| y \ln \left[\frac{e^{\theta x}}{1 + e^{\theta x}} \right] + (1 - y) \ln \left[\frac{1}{1 + e^{\theta x}} \right] \right| = \\ & - \inf_{\theta \in [\underline{\theta}, \bar{\theta}]} \left\{ y \ln \left[\frac{e^{\theta x}}{1 + e^{\theta x}} \right] + (1 - y) \ln \left[\frac{1}{1 + e^{\theta x}} \right] \right\} \end{aligned}$$

Define $I_0(x) = I(x > 0)$. Then, the last term equals to

$$\begin{aligned} & - \left\{ y \ln \left[\frac{\exp \left\{ \left(\underline{\theta} I_0 + \bar{\theta} (1 - I_0) \right) x \right\}}{1 + \exp \left\{ \left(\underline{\theta} I_0 + \bar{\theta} (1 - I_0) \right) x \right\}} \right] \right. \\ & \left. + (1 - y) \ln \left[\frac{1}{1 + \exp \left\{ \left(\underline{\theta} I_0 + \bar{\theta} (1 - I_0) \right) x \right\}} \right] \right\} \\ & = - \{ y \ln p(x) + (1 - y) \ln(1 - p(x)) \} \\ & = - \ln(1 - p(x)) - y \ln \left[\frac{p(x)}{1 - p(x)} \right] \\ & \leq - \ln(1 - p(x)) - I_0 \left(p(x) < \frac{1}{2} \right) \ln \left[\frac{p(x)}{1 - p(x)} \right]. \end{aligned} \quad (7.11)$$

So, we can see that if

$$0 < \underline{p} \leq p(x) \leq \bar{p} < 1, \quad \text{for all } x,$$

then the function in (7.11) is a bounded function of x . Hence, its expectation w.r.t. the distribution of x is finite.

Usually we make assumptions directly about $f(y, x; \theta)$ instead of making an assumption about the distribution of x and then verifying the condition for $f(y|x; \theta)$.

Assumption SUP:

$$E_0 \left[\sup_{\theta \in \Theta} |\ln f(y, x; \theta)| \right] < \infty, \quad (7.12)$$

where the expectation is taken with respect to the true population pdf of (y, x) , that is, $f(y, x; \theta_0)$.

Alternatively we can assume that $\sup_{\theta \in \Theta} |\ln f(y, x; \theta)|$ is bounded by some function $M(Y, X)$, with $E[M(Y, X)] < \infty$, and then use the dominating convergence theorem to show (7.12).

Hence, since we actually have a.s. convergence, we have that

$$\frac{1}{n} l(\theta) = \frac{1}{n} \sum_{i=1}^n \ln f(y_i, x_i; \theta) \xrightarrow{p} E_0 [\ln f(y_i, x_i; \theta)], \quad \text{uniformly on } \Theta.$$

Note now that

$$\begin{aligned} E_0 \left[\ln \frac{f(y_i, x_i; \theta)}{f(y_i, x_i; \theta_0)} \right] &\leq \ln E_0 \left[\frac{f(y_i, x_i; \theta)}{f(y_i, x_i; \theta_0)} \right] \\ &= \ln \left[\int_y \int_x f(y_i, x_i; \theta) dx dy \right] \\ &= 0. \end{aligned} \quad (7.13)$$

Note that the inequality is due to the Jensen inequality since the logarithm is a concave function.

The upper bound of (7.13) is achieved at the point $\theta = \theta_0$. Hence,

$$\theta_0 = \arg \max_{\theta \in \Theta} E_0 [\ln f(y, x; \theta)]. \quad (7.14)$$

If this maximum is unique, then θ_0 in (7.14) provides the definition for the population parameter. Note also that the expectation on the right-hand side of (7.14) is the limit of the normalized -log-likelihood. Hence, maximization of the log-likelihood should give, at least in a large enough sample, an estimate which is close to θ_0 .

In the binary logit model example we have

$$\begin{aligned} E_0 [\ln f(y, x; \theta)] &= E_x [E [\ln f(y, x; \theta_0) | x]] \\ &= E_x \left[\frac{\exp \{x' \theta_0\}}{1 + \exp \{x' \theta_0\}} \cdot \ln \left(\frac{\exp \{x' \theta_0\}}{1 + \exp \{x' \theta_0\}} \right) + \frac{1}{1 + \exp \{x' \theta_0\}} \cdot \ln \left(\frac{1}{1 + \exp \{x' \theta_0\}} \right) \right] \\ &\quad + E_x [\ln g(x)]. \end{aligned} \quad (7.15)$$

Now, consider again the case with scalar θ_0 , i.e., with only one regressor.

Note that if

$$F(y) = \frac{e^y}{1 + e^y},$$

then

$$f(y) = F'(y) = F(y) (1 - F(y)). \quad (7.16)$$

Also note that

$$F(\theta x) = \frac{e^{\theta x}}{1 + e^{\theta x}}. \quad (7.17)$$

From (7.16) and (7.17) the second order derivative of the first term in (7.15) w.r.t. θ is given by

$$\frac{\partial [F(\theta x) \ln (F(\theta x)) + (1 - F(\theta x)) \ln (1 - F(\theta x))]}{\partial \theta} = -F(\theta x) (1 - F(\theta x)) x^2. \quad (7.18)$$

Now, if

$$0 < \underline{F} \leq F(\theta x) \leq \overline{F} < 1,$$

which is the case if the condition for uniform SLLN for the normalized log-likelihood is satisfied, then the second-order derivative in (7.18) is strictly negative. Hence, it follows that $E_0 [\ln f(y, x; \theta)]$ is strictly concave function in θ . Hence, we establish that in this particular example the maximum is unique, if Θ is compact.

From this point we will simply assume that θ_0 is a unique maximizer of $E_0 [\ln f(y, x; \theta)]$ over all $\theta \in \Theta$.

Assumption Uniqueness:

We assume that for all $\varepsilon > 0$,

$$\sup_{|\theta - \theta_0| \geq \varepsilon} E_0 [\ln f(y, x; \theta)] \leq E_0 [\ln f(y, x; \theta_0)]. \quad (7.19)$$

If: (i) the parameter space Θ is compact; and (ii) $f(y, x; \theta)$ is continuous in θ for almost all (y, x) , then a sufficient condition for (7.19) is that there is a set A such that $f(y, x; \theta) \neq f(y, x; \theta_0)$ for all $(y, x) \in A$ and

$$\int_A f(y, x; \theta_0) dx dy > 0.$$

If there does not exist such a set A , then there are no data that would allow us to distinguish between θ and θ_0 . That is, such θ is observationally equivalent to θ_0 . If this is *not* the case, then we say that θ_0 is *identified*. If the model is not identified, then we cannot estimate the parameter, because there is no one parameter to estimate, and more than one parameter can generate identically the same data.

Suppose now that θ is not observationally equivalent. Note that since for all $x > 0$, $\ln x \leq 2(\sqrt{x} - 1)$, we have

$$\begin{aligned}
E_0 \left[\ln \frac{f(y_i, x_i; \theta)}{f(y_i, x_i; \theta_0)} \right] &\leq 2E_0 \left[\sqrt{\frac{f(y_i, x_i; \theta)}{f(y_i, x_i; \theta_0)}} - 1 \right] & (7.20) \\
&= 2 \left(\int \sqrt{f(y_i, x_i; \theta)f(y_i, x_i; \theta_0)} dx dy - 1 \right) \\
&\leq - \int_A \left(\sqrt{f(y_i, x_i; \theta)} - \sqrt{f(y_i, x_i; \theta_0)} \right)^2 dx dy \\
&< 0.
\end{aligned}$$

Because for all $\varepsilon > 0$ the set $\{\theta \in \Theta: |\theta - \theta_0| \geq \varepsilon\}$ is compact, equation (7.20) holds for all θ in this set. Hence, the maximum is unique.

From hereon we will only consider models for which the maximum is unique.