Chapter 2: Maximum Likelihood Estimation

Advanced Econometrics - HEC Lausanne

Christophe Hurlin

University of Orléans

December 9, 2013

Section 1

Introduction

1. Introduction

- The Maximum Likelihood Estimation (MLE) is a method of estimating the parameters of a model. This estimation method is one of the most widely used.
- The method of maximum likelihood selects the set of values of the model parameters that maximizes the likelihood function. Intuitively, this maximizes the "agreement" of the selected model with the observed data.
- The Maximum-likelihood Estimation gives an unified approach to estimation.

2. The Principle of Maximum Likelihood

- What are the main properties of the maximum likelihood estimator?
    - Is it asymptotically unbiased?
    - Is it asymptotically efficient? Under which condition(s)?
    - Is it consistent?
    - What is the asymptotic distribution?
- How to apply the maximum likelihood principle to the multiple linear regression model, to the Probit/Logit Models etc. ?

... All of these questions are answered in this lecture...

1. Introduction

The outline of this chapter is the following:

Section 2: The principle of the maximum likelihood estimation

Section 3: The likelihood function

Section 4: Maximum likelihood estimator

Section 5: Score, Hessian and Fisher information

Section 6: Properties of maximum likelihood estimators

1. Introduction

References

- Amemiya T. (1985), Advanced Econometrics. Harvard University Press.
- Greene W. (2007), Econometric Analysis, sixth edition, Pearson - Prentice Hil
- Pelgrin, F. (2010), Lecture notes Advanced Econometrics, HEC Lausanne (**a special thank**)
- Ruud P., (2000) An introduction to Classical Econometric Theory, Oxford University Press.
- Zivot, E. (2001), Maximum Likelihood Estimation, Lecture notes.

Section 2

The Principle of Maximum Likelihood

2. The Principle of Maximum Likelihood

Objectives

In this section, we present a simple example in order

1. To introduce the **notations**
2. To introduce the notion of **likelihood** and **log-likelihood**.
3. To introduce the concept of **maximum likelihood estimator**
4. To introduce the concept of **maximum likelihood estimate**

2. The Principle of Maximum Likelihood

Example

Suppose that $X_1, X_2, \cdots, X_N$ are i.i.d. discrete random variables, such that $X_i \sim Pois(\theta)$ with a **pmf** (probability mass function) defined as:

$$\Pr(X_i = x_i) = \frac{\exp(-\theta)\,\theta^{x_i}}{x_i!}$$

where θ is an unknown parameter to estimate.

2. The Principle of Maximum Likelihood

Question: What is the probability of observing the **particular sample** $\{x_1, x_2, .., x_N\}$, assuming that a Poisson distribution with as yet unknown parameter θ generated the data?

This probability is equal to

$$\Pr\left((X_1 = x_1) \cap ... \cap (X_N = x_N)\right)$$

2. The Principle of Maximum Likelihood

Since the variables X_i are *i.i.d.* this joint probability is equal to the product of the marginal probabilities

$$\Pr\left((X_1 = x_1) \cap ... \cap (X_N = x_N)\right) = \prod_{i=1}^{N} \Pr\left(X_i = x_i\right)$$

Given the pmf of the Poisson distribution, we have:

$$\begin{aligned}\Pr\left((X_1 = x_1) \cap ... \cap (X_N = x_N)\right) &= \prod_{i=1}^{N} \frac{\exp\left(-\theta\right)\theta^{x_i}}{x_i!} \\ &= \exp\left(-\theta N\right) \frac{\theta^{\sum_{i=1}^{N} x_i}}{\prod_{i=1}^{N} x_i!}\end{aligned}$$

2. The Principle of Maximum Likelihood

Definition

This joint probability is a function of θ (the unknown parameter) and corresponds to the **likelihood of the sample** $\{x_1, .., x_N\}$ denoted by

$$L_N(\theta; x_1.., x_N) = \Pr((X_1 = x_1) \cap ... \cap (X_N = x_N))$$

with

$$L_N(\theta; x_1.., x_N) = \exp(-\theta N) \times \theta^{\sum_{i=1}^{N} x_i} \times \frac{1}{\prod\limits_{i=1}^{N} x_i!}$$

2. The Principle of Maximum Likelihood

Example

Let us assume that for $N = 10$, we have a realization of the sample equal to $\{5, 0, 1, 1, 0, 3, 2, 3, 4, 1\}$, then:

$$L_N(\theta; x_1.., x_N) = \Pr((X_1 = x_1) \cap ... \cap (X_N = x_N))$$

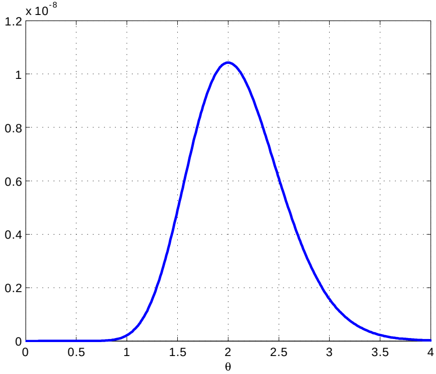
$$L_N(\theta; x_1.., x_N) = \frac{e^{-10\theta}\theta^{20}}{207,360}$$

2. The Principle of Maximum Likelihood

Question: What value of θ would make this **sample most probable**?

2. The Principle of Maximum Likelihood

This Figure plots the function $L_N(\theta; x)$ for various values of θ. It has a single mode at $\theta = 2$, which would be the maximum likelihood estimate, or MLE, of θ.



2. The Principle of Maximum Likelihood

```
%==================================================
% PURPOSE: Reproduce the Figure 1 of the Chapter 2
% Lecture: "Advanced Econometrics", HEC Lausanne
%-------------------------
% Author: Christophe Hurlin, University of Orleans
% Version: v1. October 2013
%==================================================

clear all ; clc ; close all

x=[5 0 1 1 0 3 2 3 4 1]';     % Sample

N=length(x);                 % Sample size

theta=(0:0.01:4)';           % Potential values of theta

% Likelihood Function

Ln=ones(size(theta));

for i=1:length(theta)

    Ln(i)=prod(poisspdf(x,theta(i)));

end

% Other expression in one command line

Ln2=exp(-theta*N).*(theta.^sum(x))/prod(factorial(x));
```

2. The Principle of Maximum Likelihood

Consider maximizing the likelihood function $L_N(\theta; x_1.., x_N)$ with respect to θ. Since the log function is monotonically increasing, we usually maximize $\ln L_N(\theta; x_1.., x_N)$ instead. In this case:

$$\ln L_N(\theta; x_1.., x_N) = -\theta N + \ln(\theta) \sum_{i=1}^{N} x_i - \ln\left(\prod_{i=1}^{N} x_i!\right)$$

$$\frac{\partial \ln L_N(\theta; x_1.., x_N)}{\partial \theta} = -N + \frac{1}{\theta} \sum_{i=1}^{N} x_i$$

$$\frac{\partial^2 \ln L_N(\theta; x_1.., x_N)}{\partial \theta^2} = -\frac{1}{\theta^2} \sum_{i=1}^{N} x_i < 0$$

2. The Principle of Maximum Likelihood

Under suitable regularity conditions, the maximum likelihood estimate (estimator) is defined as:

$$\widehat{\theta} = \underset{\theta \in \mathbb{R}^+}{\arg\max} \ln L_N\left(\theta; x_1 .., x_N\right)$$

$$FOC: \left.\frac{\partial \ln L_N\left(\theta; x_1 .., x_N\right)}{\partial \theta}\right|_{\widehat{\theta}} = -N + \frac{1}{\widehat{\theta}} \sum_{i=1}^{N} x_i = 0$$

$$\Longleftrightarrow \widehat{\theta} = (1/N) \sum_{i=1}^{N} x_i$$

$$SOC: \left.\frac{\partial^2 \ln L_N\left(\theta; x_1 .., x_N\right)}{\partial \theta^2}\right|_{\widehat{\theta}} = -\frac{1}{\widehat{\theta}^2} \sum_{i=1}^{N} x_i < 0$$

$\widehat{\theta}$ is a maximum.

2. The Principle of Maximum Likelihood

- The maximum likelihood estimate (**realization**) is:

$$\widehat{\theta} \equiv \widehat{\theta}\,(x) = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Given the sample $\{5, 0, 1, 1, 0, 3, 2, 3, 4, 1\}$, we have $\widehat{\theta}\,(x) = 2$.

- The maximum likelihood estimator (**random variable**) is:

$$\widehat{\theta} = \frac{1}{N}\sum_{i=1}^{N} X_i$$

2. The Principle of Maximum Likelihood

Continuous variables

- The reference to the probability of observing the given sample is not exact in a continuous distribution, since a particular sample has probability zero. Nonetheless, the principle is the same.
- The likelihood function then corresponds to the pdf associated to the **joint distribution** of $(X_1, X_2, .., X_N)$ evaluated at the point $(x_1, x_2, .., x_N)$:

$$L_N(\theta; x_1 .., x_N) = f_{X_1,..,X_N}(x_1, x_2, .., x_N; \theta)$$

2. The Principle of Maximum Likelihood

Continuous variables

- If the random variables $\{X_1, X_2, .., X_N\}$ are *i.i.d.* then we have:

$$L_N\left(\theta; x_1 .., x_N\right) = \prod_{i=1}^{N} f_X\left(x_i; \theta\right)$$

  where $f_X\left(x_i; \theta\right)$ denotes the pdf of the marginal distribution of X (or X_i since all the variables have the same distribution).

- The values of the parameters that maximize $L_N\left(\theta; x_1 .., x_N\right)$ or its log are the maximum likelihood estimates, denoted $\widehat{\theta}\left(x\right)$.

Section 3

The Likelihood function

Definitions and Notations

3. The Likelihood Function

Objectives

1. Introduce the **notations** for an estimation problem that deals with a **marginal distribution** or a **conditional distribution (model)**.
2. Define the **likelihood** and the **log-likelihood** functions.
3. Introduce the concept of **conditional log-likelihood**
4. Propose various applications

3. The Likelihood Function

Notations

- Let us consider a continuous random variable X, with a pdf denoted $f_X(x;\theta)$, for $x \in \mathbb{R}$
- $\theta = (\theta_1..\theta_K)^\intercal$ is a $K \times 1$ vector of unknown parameters. We assume that $\theta \in \Theta \subset \mathbb{R}^K$.
- Let us consider a sample $\{X_1, .., X_N\}$ of $i.i.d.$ random variables with the same arbitrary distribution as X.
- The realisation of $\{X_1, .., X_N\}$ (the data set..) is denoted $\{x_1, .., x_N\}$ or x for simplicity.

3. The Likelihood Function

Example (Normal distribution)

If $X \sim N\left(m, \sigma^2\right)$ then:

$$f_X\left(z; \theta\right) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{\left(z-m\right)^2}{2\sigma^2}\right) \qquad \forall z \in \mathbb{R}$$

with $K = 2$ and

$$\theta = \left(\begin{array}{c} m \\ \sigma^2 \end{array}\right)$$

3. The Likelihood Function

Definition (Likelihood Function)

The likelihood function is defined to be:

$$L_N : \Theta \times \mathbb{R}^N \rightarrow \mathbb{R}^+$$

$$(\theta; x_1, .., x_n) \longmapsto L_N (\theta; x_1, .., x_n) = \prod_{i=1}^{N} f_X (x_i; \theta)$$

3. The Likelihood Function

Definition (Log-Likelihood Function)

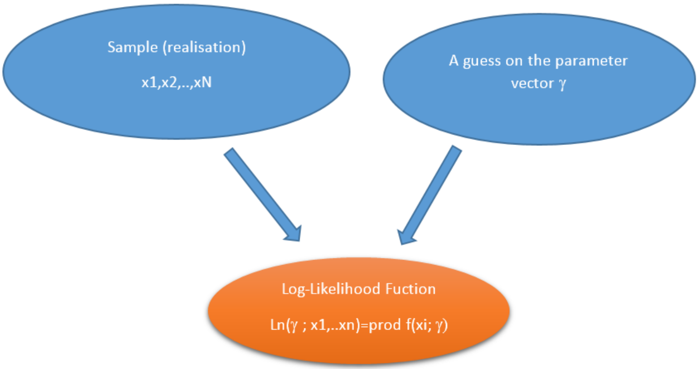
The log-likelihood function is defined to be:

$$\ell_N : \Theta \times \mathbb{R}^N \longrightarrow \mathbb{R}$$

$$(\theta; x_1, .., x_n) \longmapsto \ell_N (\theta; x_1, .., x_n) = \sum_{i=1}^{N} \ln f_X (x_i; \theta)$$

3. The Likelihood Function

Remark: the (log-)likelihood function depends on two type of arguments:



3. The Likelihood Function

Notations: In the rest of the chapter, I will use the following alternative notations:

$$L_N(\theta; x) \equiv L(\theta; x_1, .., x_N) \equiv L_N(\theta)$$

$$\ell_N(\theta; x) \equiv \ln L_N(\theta; x) \equiv \ln L(\theta; x_1, .., x_N) \equiv \ln L_N(\theta)$$

3. The Likelihood Function

Example (Sample of Normal Variables)

We consider a sample $\{Y_1, .., Y_N\}$ $\mathcal{N}.i.d.\left(m, \sigma^2\right)$ and denote the realisation by $\{y_1, .., y_N\}$ or y. Let us define $\theta = \left(m \ \sigma^2\right)^{\intercal}$, then we have:

$$\begin{aligned} L_N(\theta; y) &= \prod_{i=1}^{N} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(y_i - m)^2}{2\sigma^2}\right) \\ &= \left(\sigma^2 2\pi\right)^{-N/2} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{N} (y_i - m)^2\right) \end{aligned}$$

$$\ell_N(\theta; y) = -\frac{N}{2} \ln\left(\sigma^2\right) - \frac{N}{2} \ln(2\pi) - \frac{1}{2\sigma^2} \sum_{i=1}^{N} (y_i - m)^2$$

3. The Likelihood Function

Definition (Likelihood of one observation)

We can also define the (log-)likelihood of **one observation** x_i :

$$L_i(\theta; x) = f_X(x_i; \theta) \quad \text{with} \quad L_N(\theta; x) = \prod_{i=1}^{N} L_i(\theta; x)$$

$$\ell_i(\theta; x) = \ln f_X(x_i; \theta) \quad \text{with} \quad \ell_N(\theta; x) = \sum_{i=1}^{N} \ell_i(\theta; x)$$

3. The Likelihood Function

Example (Exponential Distribution)

Suppose that $D_1, D_2, .., D_N$ are $i.i.d.$ positive random variables (durations for instance), with $D_i \sim Exp\left(\theta\right)$ with $\theta \geq 0$ and

$$L_i\left(\theta; d_i\right) = f_D\left(d_i; \theta\right) = \frac{1}{\theta}\exp\left(-\frac{d_i}{\theta}\right)$$

$$\ell_i\left(\theta; d_i\right) = \ln\left(f_D\left(d_i; \theta\right)\right) = -\ln\left(\theta\right) - \frac{d_i}{\theta}$$

Then we have:

$$L_N\left(\theta; d\right) = \theta^{-N}\exp\left(-\frac{1}{\theta}\sum_{i=1}^{N} d_i\right)$$

$$\ell_N\left(\theta; d\right) = -N\ln\left(\theta\right) - \frac{1}{\theta}\sum_{i=1}^{N} d_i$$

3. The Likelihood Function

Remark: The (log-)likelihood and the Maximum Likelihood Estimator are always based on an assumption (bet?) about the distribution of Y.

$$Y_i \sim \text{Distribution with pdf } f_Y(y;\theta) \Longrightarrow L_N(\theta;y) \text{ and } \ell_N(\theta;y)$$

In practice, generally we have no idea about the true distribution of Y_i....

A solution: the Quasi-Maximum Likelihood Estimator

3. The Likelihood Function

Remark: We can also use the MLE to estimate the parameters of a **model (with dependent and explicative variables)** such that:

$$y = g(x; \theta) + \varepsilon$$

where β denotes the vector or parameters, X a set of explicative variables, ε and error term and $g(.)$ the link function.

In this case, we generally consider the *conditional distribution* of Y given X, which is equivalent to unconditional distribution of the error term ε :

$$Y|\, X \sim D \Longleftrightarrow \varepsilon \sim D$$

3. The Likelihood Function

Notations (model)

- Let us consider two continuous random variables Y and X
- We assume that Y has a conditional distribution given $X = x$ with a pdf denoted $f_{Y|x}(y; \theta)$, for $y \in \mathbb{R}$
- $\theta = (\theta_1 .. \theta_K)^\intercal$ is a $K \times 1$ vector of unknown parameters. We assume that $\theta \in \Theta \subset \mathbb{R}^K$.
- Let us consider a sample $\{X_1, Y_N\}_{i=1}^N$ of $i.i.d.$ random variables and a realisation $\{x_1, y_N\}_{i=1}^N$.

3. The Likelihood Function

Definition (Conditional likelihood function)

The (conditional) likelihood function is defined to be:

$$L_N\left(\theta; y \mid x\right) = \prod_{i=1}^{N} f_{Y|X}\left(y_i \mid x_i; \theta\right)$$

where $f_{Y|X}\left(y_i \mid x_i; \theta\right)$ denotes the conditional pdf of Y_i given X_i.

Remark: The conditional likelihood function is the joint conditional density of the data in which the unknown parameter is .

3. The Likelihood Function

Definition (Conditional log-likelihood function)

The (conditional) log-likelihood function is defined to be:

$$\ell_N\left(\theta; y \mid x\right) = \sum_{i=1}^{N} \ln f_{Y|X}\left(y_i \mid x_i; \theta\right)$$

where $f_{Y|X}\left(y_i \mid x_i; \theta\right)$ denotes the conditional pdf of Y_i given X_i.

3. The Likelihood Function

Remark: The conditional probability density function (pdf) can denoted by:

$$f_{Y|X}\left(y|\,x;\theta\right) \equiv f_{Y}\left(y|\,X=x;\theta\right) \equiv f_{Y}\left(y|\,X=x\right)$$

3. The Likelihood Function

Example (Linear Regression Model)

Consider the following linear regression model:

$$y_i = \mathbf{X}_i^\top \boldsymbol{\beta} + \varepsilon_i$$

where $\mathbf{X}_i$ is a $K \times 1$ vector of random variables and $\boldsymbol{\beta} = (\beta_1 .. \beta_K)^\top$ a $K \times 1$ vector of parameters. We assume that the ε_i are $i.i.d.$ with $\varepsilon_i \sim \mathcal{N}\left(0, \sigma^2\right)$. Then, the conditional distribution of Y_i given $\mathbf{X}_i = \mathbf{x}_i$ is:

$$Y_i | \, \mathbf{x}_i \sim \mathcal{N}\left(\mathbf{x}_i^\top \boldsymbol{\beta}, \sigma^2\right)$$

$$L_i\left(\boldsymbol{\theta}; y | \, \mathbf{x}\right) = f_{Y|\mathbf{x}}\left(y_i | \, \mathbf{x}_i; \boldsymbol{\theta}\right) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{\left(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}\right)^2}{2\sigma^2}\right)$$

where $\boldsymbol{\theta} = \left(\boldsymbol{\beta}^\top \; \sigma^2\right)^\top$ is $K + 1 \times 1$ vector.

3. The Likelihood Function

Example (Linear Regression Model, cont'd)

Then, if we consider an *i.i.d.* sample $\{y_i, \mathbf{x}_i\}_{i=1}^{N}$, the corresponding **conditional** (log-)likelihood is defined to be:

$$
\begin{aligned}
L_N(\boldsymbol{\theta}; y|\,\mathbf{x}) &= \prod_{i=1}^{N} f_{Y|X}(y_i|\,\mathbf{x}_i; \boldsymbol{\theta}) = \prod_{i=1}^{N} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{\left(y_i - \mathbf{x}_i^{\top}\boldsymbol{\beta}\right)^2}{2\sigma^2}\right) \\
&= \left(\sigma^2 2\pi\right)^{-N/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{N}\left(y_i - \mathbf{x}_i^{\top}\boldsymbol{\beta}\right)^2\right)
\end{aligned}
$$

$$
\ell_N(\boldsymbol{\theta}; y|\,\mathbf{x}) = -\frac{N}{2}\ln\left(\sigma^2\right) - \frac{N}{2}\ln(2\pi) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}\left(y_i - \mathbf{x}_i^{\top}\boldsymbol{\beta}\right)^2
$$

3. The Likelihood Function

Remark: Given this principle, we can derive the (conditional) likelihood and the log-likelihood functions associated to a specific sample for any type of econometric model in which the conditional distribution of the dependent variable is known.

- Dichotomic models: probit, logit models etc.
- Censored regression models: Tobit etc.
- Times series models: AR, ARMA, VAR etc.
- GARCH models
- ....

3. The Likelihood Function

Example (Probit/Logit Models)

Let us consider a dichotomic variable Y_i such that $Y_i = 1$ if the firm i is in default and 0 otherwise. $\mathbf{X}_i = (X_{i1}...X_{iK})$ denotes a a $K \times 1$ vector of individual caracteristics. We assume that the conditional probability of default is defined as:

$$\Pr\left(Y_i = 1 \middle| \mathbf{X}_i = \mathbf{x}_i \right) = F\left(\mathbf{x}_i^\top \boldsymbol{\beta}\right)$$

where $\boldsymbol{\beta} = (\beta_1..\beta_K)^\top$ is a vector of parameters and $F(.)$ is a cdf (cumlative distribution function).

$$Y_i = \begin{cases} 1 & \text{with probability } F\left(\mathbf{x}_i^\top \boldsymbol{\beta}\right) \\ 0 & \text{with probability } 1 - F\left(\mathbf{x}_i^\top \boldsymbol{\beta}\right) \end{cases}$$

3. The Likelihood Function

Remark: Given the choice of the link function $F(.)$ we get a probit or a logit model.

3. The Likelihood Function

Definition (Probit Model)

In a **probit model**, the conditional probability of the event $Y_i = 1$ is:

$$\Pr\left(Y_i = 1 \middle| \mathbf{X}_i = \mathbf{x}_i \right) = \Phi\left(x_i \beta\right) = \int_{-\infty}^{\mathbf{x}_i^\top \boldsymbol{\beta}} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right) du$$

where $\Phi\left(.\right)$ denotes the cdf of the standard normal distribution.

3. The Likelihood Function

Definition (Logit Model)

In a **logit model**, the conditional probability of the event $Y_i = 1$ is:

$$\Pr\left(Y_i = 1 \middle| \mathbf{X}_i = \mathbf{x}_i \right) = \Lambda\left(\mathbf{x}_i^\top \boldsymbol{\beta}\right) = \frac{1}{1 + \exp\left(-\mathbf{x}_i^\top \boldsymbol{\beta}\right)}$$

where $\Lambda\left(.\right)$ denotes the cdf of the logistic distribution.

3. The Likelihood Function

Example (Probit/Logit Models, cont'd)

What is the (conditional) log-likelihood of the sample $\{y_i, x_i\}_{i=1}^{N}$?
Whatever the choice of $F(.)$, the conditional distribution of Yi given $\mathbf{X}_i = \mathbf{x}_i$ is a **Bernouilli distribution** since:

$$Y_i = \begin{cases} 1 & \text{with probability } F\left(\mathbf{x}_i^\top \boldsymbol{\beta}\right) \\ 0 & \text{with probability } 1 - F\left(\mathbf{x}_i^\top \boldsymbol{\beta}\right) \end{cases}$$

Then, for $\boldsymbol{\theta} = \boldsymbol{\beta}$, we have:

$$L_i\left(\boldsymbol{\theta}; y|\, \mathbf{x}\right) = f_{Y|\mathbf{x}}\left(y_i|\, \mathbf{x}_i; \boldsymbol{\theta}\right) = \left[F\left(\mathbf{x}_i^\top \boldsymbol{\beta}\right)\right]^{y_i} \left[1 - F\left(\mathbf{x}_i^\top \boldsymbol{\beta}\right)\right]^{1-y_i}$$

where $f_{Y|x}\left(y_i|\, \mathbf{x}_i; \boldsymbol{\theta}\right)$ denotes the conditional probability mass function (pmf) of Y_i.

3. The Likelihood Function

Example (Probit/Logit Models, cont'd)

The (conditional) likelihood and log-likelihood of the sample $\{y_i, \mathbf{x}_i\}_{i=1}^N$ are defined to be:

$$L_N(\boldsymbol{\theta}; y|\mathbf{x}) = \prod_{i=1}^N f_{Y|\mathbf{x}}(y_i|\mathbf{x}_i; \boldsymbol{\theta}) = \prod_{i=1}^N \left[F\left(\mathbf{x}_i^\top \boldsymbol{\beta}\right)\right]^{y_i} \left[1 - F\left(\mathbf{x}_i^\top \boldsymbol{\beta}\right)\right]^{1-y_i}$$

$$\begin{aligned} \ell_N(\boldsymbol{\theta}; y|\mathbf{x}) &= \sum_{i=1}^N y_i \ln\left[F\left(\mathbf{x}_i^\top \boldsymbol{\beta}\right)\right] + \sum_{i=1}^N (1-y_i) \ln\left[1 - F\left(\mathbf{x}_i^\top \boldsymbol{\beta}\right)\right] \\ &= \sum_{i:y_i=1} \ln F\left(\mathbf{x}_i^\top \boldsymbol{\beta}\right) + \sum_{i:y_i=0} \ln\left[1 - F\left(\mathbf{x}_i^\top \boldsymbol{\beta}\right)\right] \end{aligned}$$

where $f_{Y|\mathbf{x}}(y_i|\mathbf{x}_i; \boldsymbol{\theta})$ denotes the **conditional probability mass function** (pmf) of Y_i.

3. The Likelihood Function

Key Concepts

1. Likelihood (of a sample) function
2. Log-likelihood (of a sample) function
3. Conditional Likelihood and log-likelihood function
4. Likelihood and log-likelihood of one observation

Section 4

Maximum Likelihood Estimator

4. Maximum Likelihood Estimator

Objectives

1. This section will be concerned with obtaining **estimates** of the parameters θ.
2. We will define the **maximum likelihood estimator (MLE).**
3. Before we begin that study, we consider the question of whether estimation of the parameters is possible at all: the question of **identification**.
4. We will introduce the **invariance principle**

4. Maximum Likelihood Estimator

Definition (Identification)

The parameter vector θ is identified (estimable) if for any other parameter vector, $\theta^* \neq \theta$, for some data y, we have

$$L_N(\theta; y) \neq L_N(\theta^*; y)$$

4. Maximum Likelihood Estimator

Example

Let us consider a **latent** (continuous and unobservable) variable Y_i^* such that:

$$Y_i^* = \mathbf{X}_i^\top \beta + \varepsilon_i$$

with $\boldsymbol{\beta} = (\beta_1 .. \beta_K)^\top$, $\mathbf{X}_i = (X_{i1} ... X_{iK})^\top$ and where the error term ε_i is $i.i.d.$ such that $\mathbb{E}(\varepsilon_i) = 0$ and $\mathbb{V}(\varepsilon_i) = \sigma^2$. The distribution of ε_i is symmetric around 0 and we denote by $G(.)$ the cdf of the standardized error term ε_i / σ. We assume that this cdf does not depend on σ or β. Example: $\varepsilon_i / \sigma \sim \mathcal{N}(0, 1)$.

4. Maximum Likelihood Estimator

Example (cont'd)

We observe a dichotomic variable Y_i such that:

$$Y_i = \begin{cases} 1 & \text{if } Y_i^* > 0 \\ 0 & \text{otherwise} \end{cases}$$

Problem: are the parameters $\boldsymbol{\theta} = \left(\boldsymbol{\beta}^\top \ \sigma^2\right)^\top$ identifiable?

4. Maximum Likelihood Estimator

Solution:

To answer to this question we have to compute the (log-)likelihood of the sample of observed data $\{y_i, \mathbf{x}_i\}_{i=1}^{N}$. We have:

$$\begin{aligned}
\Pr\left(Y_i = 1 \middle| \mathbf{X}_i = \mathbf{x}_i \right) &= \Pr\left(Y_i^* > 0 \middle| \mathbf{X}_i = \mathbf{x}_i \right) \\
&= \Pr\left(\varepsilon_i > -\mathbf{x}_i^{\top} \boldsymbol{\beta} \right) \\
&= 1 - \Pr\left(\varepsilon_i \leq -\mathbf{x}_i^{\top} \boldsymbol{\beta} \right) \\
&= 1 - \Pr\left(\frac{\varepsilon_i}{\sigma} \leq -\mathbf{x}_i^{\top} \frac{\boldsymbol{\beta}}{\sigma} \right)
\end{aligned}$$

If we denote by $G(.)$ the cdf associated to the distribution of ε_i / σ, since this distribution is symetric around 0, then we have:

$$\Pr\left(Y_i = 1 \middle| \mathbf{X}_i = \mathbf{x}_i \right) = G\left(\mathbf{x}_i^{\top} \frac{\boldsymbol{\beta}}{\sigma} \right)$$

4. Maximum Likelihood Estimator

Solution (cont'd):

For $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top \ \sigma^2)^\top$, we have

$$\ell_N(\boldsymbol{\theta}; y | \mathbf{x}) = \sum_{i=1}^{N} y_i \ln \left[G\left(\mathbf{x}_i^\top \frac{\boldsymbol{\beta}}{\sigma} \right) \right] + \sum_{i=1}^{N} (1 - y_i) \ln \left[1 - G\left(\mathbf{x}_i^\top \frac{\boldsymbol{\beta}}{\sigma} \right) \right]$$

This log-likelihood depends only on the ratio $\boldsymbol{\beta}/\sigma$. So, for $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top \ \sigma^2)^\top$ and $\boldsymbol{\theta}^* = (k \times \boldsymbol{\beta}^\top \ \ k \times \sigma)^\top$, with $k \neq 1$:

$$\ell_N(\boldsymbol{\theta}; y | \mathbf{x}) = \ell_N(\boldsymbol{\theta}^*; y | \mathbf{x})$$

The parameters β and σ^2 **cannot be identified**. We can only identify the ratio $\boldsymbol{\beta}/\sigma$.

4. Maximum Likelihood Estimator

Remark:

In this latent model, only the ratio β/σ can be identified since

$$\Pr\left(Y_i = 1 \right| \mathbf{X}_i = \mathbf{x}_i) = \Pr\left(\frac{\varepsilon_i}{\sigma} < \mathbf{x}_i^\top \frac{\beta}{\sigma} \right) = G\left(\mathbf{x}_i^\top \frac{\beta}{\sigma} \right)$$

The choice of a logit or probit model implies a **normalisation** on the variance of ε_i/σ and then on σ^2 :

$$\text{probit} : \Pr\left(Y_i = 1 \right| \mathbf{X}_i = \mathbf{x}_i) = \Phi\left(\mathbf{x}_i^\top \widetilde{\beta} \right) \quad \text{with } \widetilde{\beta} = \beta_i/\sigma, \ \mathbb{V}\left(\frac{\varepsilon_i}{\sigma} \right) = 1$$

4. Maximum Likelihood Estimator

Definition (Maximum Likelihood Estimator)

A maximum likelihood estimator $\widehat{\theta}$ of $\theta \in \Theta$ is a solution to the maximization problem:

$$\widehat{\theta} = \underset{\theta \in \Theta}{\arg\max}\ \ell_N \left(\theta; y | x \right)$$

or equivalently

$$\widehat{\theta} = \underset{\theta \in \Theta}{\arg\max}\ L_N \left(\theta; y | x \right)$$

4. Maximum Likelihood Estimator

Remarks

1. Do not confuse the maximum likelihood **estimator** $\widehat{\theta}$ (which is a random variable) and the maximum likelihood **estimate** $\widehat{\theta}(x)$ which corresponds to the realisation of $\widehat{\theta}$ on the sample x.

2. Generally, it is easier to maximise the log-likelihood than the likelihood (especially for the distributions that belong to the exponential family).

3. When we consider an unconditional likelihood, the MLE is defined by:

$$\widehat{\theta} = \underset{\theta \in \Theta}{\arg\max}\, \ell_N(\theta; x)$$

4. Maximum Likelihood Estimator

Definition (Likelihood equations)

Under suitable regularity conditions, a maximum likelihood estimator (MLE) of θ is defined to be the solution of the first-order conditions (FOC):

$$\left.\frac{\partial \ell_N\left(\theta; y \middle| x\right)}{\partial \theta}\right|_{\widehat{\theta}} = \underset{(K,1)}{0}$$

or

$$\left.\frac{\partial L_N\left(\theta; y \middle| x\right)}{\partial \theta}\right|_{\widehat{\theta}} = \underset{(K,1)}{0}$$

These conditions are generally called the **likelihood** or **log-likelihood equations**.

4. Maximum Likelihood Estimator

Notations

The first derivative (**gradient**) of the (conditional) log-likelihood evaluated at the point $\widehat{\theta}$ satisfies:

$$\left.\frac{\partial L_N\left(\theta;\, y|\, x\right)}{\partial \theta}\right|_{\widehat{\theta}} \equiv \frac{\partial L_N\left(\widehat{\theta};\, y|\, x\right)}{\partial \theta} = g\left(\widehat{\theta};\, y|\, x\right) = 0$$

Remark

The **log-likelihood equations** correspond to a linear/nonlinear system of K equations with K unknown parameters $\theta_1, .., \theta_K$:

$$\left.\frac{\partial \ell_N\left(\theta;\, Y \middle|\, x\right)}{\partial \theta}\right|_{\widehat{\theta}} = \begin{pmatrix} \left.\frac{\partial \ell_N(\theta; Y|x)}{\partial \theta_1}\right|_{\widehat{\theta}} \\ ... \\ \left.\frac{\partial \ell_N(\theta; Y|x)}{\partial \theta_K}\right|_{\widehat{\theta}} \end{pmatrix} = \begin{pmatrix} 0 \\ ... \\ 0 \end{pmatrix}$$

4. Maximum Likelihood Estimator

Definition (Second Order Conditions)

Second order condition (SOC) of the likelihood maximisation problem: the **Hessian** matrix evaluated at $\widehat{\theta}$ must be negative definite.

$$\left.\frac{\partial^2 \ell_N(\theta; y|x)}{\partial\theta\partial\theta^\top}\right|_{\widehat{\theta}} \text{ is negative definite}$$

or

$$\left.\frac{\partial^2 L_N(\theta; y|x)}{\partial\theta\partial\theta^\top}\right|_{\widehat{\theta}} \text{ is negative definite}$$

Remark:

The **Hessian matrix** (realisation) is a $K \times K$ matrix:

$$\frac{\partial^2 \ell_N \left(\theta;\, y \middle|\, x\right)}{\partial\theta\partial\theta^\top} = \begin{pmatrix} \frac{\partial^2 \ell_N(\theta; y|x)}{\partial\theta_1^2} & \frac{\partial^2 \ell_N(\theta; y|x)}{\partial\theta_1\partial\theta_2} & .. & \frac{\partial^2 \ell_N(\theta; y|x)}{\partial\theta_1\partial\theta_K} \\ \frac{\partial^2 \ell_N(\theta; y|x)}{\partial\theta_2\partial\theta_1} & \frac{\partial^2 \ell_N(\theta; y|x)}{\partial\theta_2^2} & .. & .. \\ .. & .. & .. & .. \\ \frac{\partial^2 \ell_N(\theta; y|x)}{\partial\theta_K\partial\theta_1} & .. & .. & \frac{\partial^2 \ell_N(\theta; y|x)}{\partial\theta_K^2} \end{pmatrix}$$

Reminders

- A negative definite matrix is a symetric (Hermitian if there are complex entries) matrix all of whose eigenvalues are negative.
- The $n \times n$ Hermitian matrix M is said to be negative-definite if:

$$\mathbf{x}^{\intercal}\mathbf{M}\mathbf{x} < 0$$

for all non-zero $\mathbf{x}$ in $\mathbb{R}^n$.

4. Maximum Likelihood Estimator

Example (MLE problem with one parameter)

Let us consider a real-valued random variable X with a pdf given by:

$$f_X\left(x;\sigma^2\right) = \exp\left(-\frac{x^2}{2\sigma^2}\right)\frac{x}{\sigma^2} \quad \forall x \in [0, +\infty[$$

where σ^2 is an unknown parameter. Let us consider a sample $\{X_1, .., X_N\}$ of $i.i.d.$ random variables with the same arbitrary distribution as X.

Problem: What is the maximum likelihood estimator (MLE) of σ^2?

4. Maximum Likelihood Estimator

Solution:

We have:

$$\ln f_X\left(x;\sigma^2\right) = -\frac{x^2}{2\sigma^2} + \ln\left(x\right) - \ln\left(\sigma^2\right)$$

So, the log-likelihood of the sample $\{x_1, .., x_N\}$ is:

$$\ell_N\left(\sigma^2; x\right) = \sum_{i=1}^{N} \ln f_X\left(x_i;\sigma^2\right) = -\frac{1}{2\sigma^2}\sum_{i=1}^{N} x_i^2 + \sum_{i=1}^{N} \ln\left(x_i\right) - N\ln\left(\sigma^2\right)$$

4. Maximum Likelihood Estimator

Solution (cont'd):

The maximum likelihood estimator $\widehat{\sigma}^2$ of $\sigma^2 \in \mathbb{R}^+$ is a solution to the maximization problem:

$$\widehat{\sigma}^2 = \underset{\sigma^2 \in \mathbb{R}^+}{\arg\max} \ell_N\left(\sigma^2; x\right) = \underset{\sigma^2 \in \mathbb{R}^+}{\arg\max} -\frac{1}{2\sigma^2}\sum_{i=1}^{N} x_i^2 + \sum_{i=1}^{N} \ln\left(x_i\right) - N\ln\left(\sigma^2\right)$$

$$\frac{\partial \ell_N\left(\sigma^2; x\right)}{\partial \sigma^2} = \frac{1}{2\sigma^4}\sum_{i=1}^{N} x_i^2 - \frac{N}{\sigma^2}$$

FOC **(log-likelihood equation)**:

$$\left.\frac{\partial \ell_N\left(\sigma^2; x\right)}{\partial \sigma^2}\right|_{\widehat{\sigma}^2} = \frac{1}{2\widehat{\sigma}^4}\sum_{i=1}^{N} x_i^2 - \frac{N}{\widehat{\sigma}^2} = 0 \Longleftrightarrow \widehat{\sigma}^2 = \frac{1}{2N}\sum_{i=1}^{N} x_i^2$$

4. Maximum Likelihood Estimator

Solution (cont'd):

Check that $\widehat{\sigma}^2$ is a maximum:

$$\frac{\partial \ell_N\left(\sigma^2; x\right)}{\partial \sigma^2} = \frac{1}{2\sigma^4}\sum_{i=1}^{N} x_i^2 - \frac{N}{\sigma^2} \qquad \frac{\partial^2 \ell_N\left(\sigma^2; x\right)}{\partial \sigma^4} = -\frac{1}{\sigma^6}\sum_{i=1}^{N} x_i^2 + \frac{N}{\sigma^4}$$

SOC:

$$\begin{aligned}
\left.\frac{\partial^2 \ell_N\left(\sigma^2; x\right)}{\partial \sigma^4}\right|_{\widehat{\sigma}^2} &= -\frac{1}{\widehat{\sigma}^6}\sum_{i=1}^{N} x_i^2 + \frac{N}{\widehat{\sigma}^4} \\
&= -\frac{2N\widehat{\sigma}^2}{\widehat{\sigma}^6} + \frac{N}{\widehat{\sigma}^4} \quad \text{since } \widehat{\sigma}^2 = \frac{1}{2N}\sum_{i=1}^{N} x_i^2 \\
&= -\frac{N}{\widehat{\sigma}^4} < 0
\end{aligned}$$

4. Maximum Likelihood Estimator

Conclusion:

The maximum likelihood estimator (MLE) of the parameter σ^2 is defined by:

$$\widehat{\sigma}^2 = \frac{1}{2N} \sum_{i=1}^{N} X_i^2$$

The maximum likelihood estimate of the parameter σ^2 is equal to:

$$\widehat{\sigma}^2 (x) = \frac{1}{2N} \sum_{i=1}^{N} x_i^2$$

4. Maximum Likelihood Estimator

Example (Sample of normal variables)

We consider a sample $\{Y_1, .., Y_N\}$ $N.i.d.\left(m, \sigma^2\right)$. **Problem:** what are the MLE of m and σ^2?

Solution: Let us define $\theta = \left(m \; \sigma^2\right)^{\intercal}$.

$$\widehat{\theta} = \underset{\sigma^2 \in \mathbb{R}^+, m \in \mathbb{R}}{\arg\max} \; \ell_N\left(\theta; y\right)$$

with

$$\ell_N\left(\theta; y\right) = -\frac{N}{2}\ln\left(\sigma^2\right) - \frac{N}{2}\ln\left(2\pi\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}\left(y_i - m\right)^2$$

Solution (cont'd):

$$\ell_N(\theta; y) = -\frac{N}{2}\ln\left(\sigma^2\right) - \frac{N}{2}\ln(2\pi) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - m)^2$$

The first derivative of the log-likelihood function is defined by:

$$\frac{\partial \ell_N(\theta; y)}{\partial \theta} = \begin{pmatrix} \frac{\partial \ell_N(\theta; y)}{\partial m} \\ \frac{\partial \ell_N(\theta; y)}{\partial \sigma^2} \end{pmatrix}$$

$$\frac{\partial \ell_N(\theta; y)}{\partial m} = \frac{1}{\sigma^2}\sum_{i=1}^{N}(y_i - m) \qquad \frac{\partial \ell_N(\theta; y)}{\partial \sigma^2} = -\frac{N}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{N}(y_i - m)^2$$

4. Maximum Likelihood Estimator

Solution (cont'd):

FOC (**log-likelihood equations**)

$$\left.\frac{\partial \ell_N(\theta; y)}{\partial \theta}\right|_{\widehat{\theta}} = \begin{pmatrix} \frac{1}{\widehat{\sigma}^2}\sum_{i=1}^{N}(y_i - \widehat{m}) \\ -\frac{N}{2\widehat{\sigma}^2} + \frac{1}{2\widehat{\sigma}^4}\sum_{i=1}^{N}(y_i - \widehat{m})^2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

So, the MLE correspond to the empirical mean and variance:

$$\widehat{\theta} = \begin{pmatrix} \widehat{m} \\ \widehat{\sigma}^2 \end{pmatrix}$$

with

$$\widehat{m} = \frac{1}{N}\sum_{i=1}^{N} Y_i \qquad \widehat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}\left(Y_i - \overline{Y}_N\right)^2$$

Solution (cont'd):

$$\frac{\partial \ell_N(\theta;y)}{\partial m}=\frac{1}{\sigma^2}\sum_{i=1}^{N}(y_i-m) \qquad \frac{\partial \ell_N(\theta;y)}{\partial \sigma^2}=-\frac{N}{2\sigma^2}+\frac{1}{2\sigma^4}\sum_{i=1}^{N}(y_i-m)^2$$

The Hessian matrix (realization) is:

$$\begin{aligned}\frac{\partial^2 \ell_N(\theta;y)}{\partial\theta\partial\theta^\top} &= \begin{pmatrix}\frac{\partial^2\ell_N(\theta;y)}{\partial m^2} & \frac{\partial^2\ell_N(\theta;y)}{\partial m\partial\sigma^2}\\ \frac{\partial^2\ell_N(\theta;y)}{\partial\sigma^2\partial m} & \frac{\partial^2\ell_N(\theta;y)}{\partial\sigma^4}\end{pmatrix}\\ &= \begin{pmatrix}-\frac{N}{\sigma^2} & -\frac{1}{\sigma^4}\sum_{i=1}^{N}(y_i-m)\\ -\frac{1}{\sigma^4}\sum_{i=1}^{N}(y_i-m) & \frac{N}{2\sigma^4}-\frac{1}{\sigma^6}\sum_{i=1}^{N}(y_i-m)^2\end{pmatrix}\end{aligned}$$

Solution (cont'd): SOC

$$\left.\frac{\partial^2 \ell_N(\theta; y)}{\partial\theta\partial\theta^\top}\right|_{\widehat{\theta}} = \begin{pmatrix} -\frac{N}{\widehat{\sigma}^2} & -\frac{1}{\widehat{\sigma}^4}\sum_{i=1}^N (y_i - \widehat{m}) \\ -\frac{1}{\widehat{\sigma}^4}\sum_{i=1}^N (y_i - \widehat{m}) & \frac{N}{2\widehat{\sigma}^4} - \frac{1}{\widehat{\sigma}^6}\sum_{i=1}^N (y_i - \widehat{m})^2 \end{pmatrix}$$

$$= \begin{pmatrix} -\frac{N}{\widehat{\sigma}^2} & 0 \\ 0 & \frac{N}{2\widehat{\sigma}^4} - \frac{N\widehat{\sigma}^2}{\widehat{\sigma}^6} \end{pmatrix}$$

since since $N\ \widehat{m} = \sum_{i=1}^N y_i$ and $N\ \widehat{\sigma}^2 = \sum_{i=1}^N (y_i - \widehat{m})^2$

$$\left.\frac{\partial^2 \ell_N(\theta; y)}{\partial\theta\partial\theta^\top}\right|_{\widehat{\theta}} = \begin{pmatrix} -\frac{N}{\widehat{\sigma}^2} & 0 \\ 0 & -\frac{N}{2\widehat{\sigma}^4} \end{pmatrix} \text{ is definite negative}$$

4. Maximum Likelihood Estimator

Example (Linear Regression Model)

Consider the linear regression model:

$$y_i = x_i^\top \beta + \varepsilon_i$$

where $x_i = (x_{i1} ... x_{iK})^\top$ and $\beta = (\beta_1 .. \beta_K)^\top$ are $K \times 1$ vectors. We assume that the ε_i are $\mathcal{N}.i.d.\left(0, \sigma^2\right)$. Then, the (conditional) log-likelihood of the observations (x_i, y_i) is given by

$$\ell_N(\theta; y|\, x) = -\frac{N}{2} \ln\left(\sigma^2\right) - \frac{N}{2} \ln(2\pi) - \frac{1}{2\sigma^2} \sum_{i=1}^{N} \left(y_i - x_i^\top \beta\right)^2$$

where $\theta = (\beta^\top \; \sigma^2)^\top$ is $(K+1) \times 1$ vector. **Question:** what are the MLE of β and σ^2?

4. Maximum Likelihood Estimator

Notation 1: The derivative of a scalar y by a $K \times 1$ vector $x = (x_1 ... x_K)^\top$ is $K \times 1$ vector

$$\frac{\partial y}{\partial x} = \begin{pmatrix} \frac{\partial y}{\partial x_1} \\ .. \\ \frac{\partial y}{\partial x_K} \end{pmatrix}$$

Notation 2: *If x and β are two $K \times 1$* vectors, then*:*

$$\frac{\partial \left(x^\top \beta \right)}{\partial \beta} = \underset{(K,1)}{x}$$

Solution

$$\widehat{\theta} = \underset{\beta \in \mathbb{R}^K, \sigma^2 \in \mathbb{R}^+}{\arg\max} - \frac{N}{2} \ln\left(\sigma^2\right) - \frac{N}{2} \ln\left(2\pi\right) - \frac{1}{2\sigma^2} \sum_{i=1}^{N} \left(y_i - x_i^\top \beta\right)^2$$

The first derivative of the log-likelihood function is a $(K+1) \times 1$ vector:

$$\underbrace{\frac{\partial \ell_N\left(\theta; y | x\right)}{\partial \theta}}_{(K+1)\times 1} = \begin{pmatrix} \frac{\partial \ell_N(\theta; y|x)}{\partial \beta} \\ \frac{\partial \ell_N(\theta; y|x)}{\partial \sigma^2} \end{pmatrix} = \begin{pmatrix} \frac{\partial \ell_N(\theta; y|x)}{\partial \beta_1} \\ .. \\ \frac{\partial \ell_N(\theta; y|x)}{\partial \beta_K} \\ \hline \frac{\partial \ell_N(\theta; y|x)}{\partial \sigma^2} \end{pmatrix}$$

4. Maximum Likelihood Estimator

Solution (cont'd)

$$\widehat{\theta} = \underset{\beta \in \mathbb{R}^K, \sigma^2 \in \mathbb{R}^+}{\arg\max} - \frac{N}{2} \ln\left(\sigma^2\right) - \frac{N}{2} \ln\left(2\pi\right) - \frac{1}{2\sigma^2} \sum_{i=1}^{N} \left(y_i - x_i^\top \beta\right)^2$$

The first derivative of the log-likelihood function is a $(K+1) \times 1$ vector:

$$\underbrace{\frac{\partial \ell_N\left(\theta; y \mid x\right)}{\partial \beta}}_{(K,1)} = \frac{1}{\sigma^2} \sum_{i=1}^{N} \underbrace{x_i}_{(K,1)} \underbrace{\left(y_i - x_i^\top \beta\right)}_{(1,1)}$$

$$\underbrace{\frac{\partial \ell_N\left(\theta; y \mid x\right)}{\partial \sigma^2}}_{(1,1)} = -\frac{N}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{N} \underbrace{\left(y_i - x_i^\top \beta\right)^2}_{(1,1)}$$

4. Maximum Likelihood Estimator

Solution (cont'd):

FOC (**log-likelihood equations**)

$$\left.\frac{\partial \ell_N\left(\theta; y\right| x)}{\partial \theta}\right|_{\widehat{\theta}} = \begin{pmatrix} \frac{1}{\widehat{\sigma}^2} \sum_{i=1}^{N} x_i \left(y_i - x_i^\top \widehat{\beta}\right) \\ -\frac{N}{2\widehat{\sigma}^2} + \frac{1}{2\widehat{\sigma}^4} \sum_{i=1}^{N} \left(y_i - x_i^\top \widehat{\beta}\right)^2 \end{pmatrix} = \begin{pmatrix} 0_K \\ 0 \end{pmatrix}$$

So, the MLE is defined by:

$$\widehat{\theta} = \begin{pmatrix} \widehat{\beta} \\ \widehat{\sigma}^2 \end{pmatrix}$$

$$\widehat{\beta} = \left(\sum_{i=1}^{N} X_i X_i^\top\right)^{-1} \left(\sum_{i=1}^{N} X_i Y_i\right) \qquad \widehat{\sigma}^2 = \frac{1}{N} \sum_{i=1}^{N} \left(Y_i - X_i^\top \widehat{\beta}\right)^2$$

4. Maximum Likelihood Estimator

Solution (cont'd):

The Hessian is a $(K+1) \times (K+1)$ matrix:

$$\underbrace{\frac{\partial^2 \ell_N(\theta; y|x)}{\partial\theta\partial\theta^\top}}_{(K+1)\times(K+1)} = \begin{pmatrix} \underbrace{\frac{\partial^2 \ell_N(\theta; y|x)}{\partial\beta\partial\beta^\top}}_{K\times K} & \underbrace{\frac{\partial^2 \ell_N(\theta; y|x)}{\partial\beta\partial\sigma^2}}_{K\times 1} \\ \underbrace{\frac{\partial^2 \ell_N(\theta; y|x)}{\partial\sigma^2\partial\beta^\top}}_{1\times K} & \underbrace{\frac{\partial^2 \ell_N(\theta; y|x)}{\partial\sigma^4}}_{1\times 1} \end{pmatrix}$$

4. Maximum Likelihood Estimator

Solution (cont'd):

$$\frac{\partial \ell_N\left(\theta; y \mid x\right)}{\partial \beta} = \frac{1}{\sigma^2} \sum_{i=1}^{N} x_i \left(y_i - x_i^\top \beta\right)$$

$$\frac{\partial \ell_N\left(\theta; y \mid x\right)}{\partial \sigma^2} = -\frac{N}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{N} \left(y_i - x_i^\top \beta\right)^2$$

So, the Hessian matrix (realization) is equal to:

$$\frac{\partial^2 \ell_N\left(\theta; y \mid x\right)}{\partial \theta \partial \theta^\top} = \begin{pmatrix} -\frac{1}{\sigma^2} \sum_{i=1}^{N} \underbrace{x_i}_{K \times 1} \underbrace{x_i^\top}_{1 \times K} & -\frac{1}{\sigma^4} \sum_{i=1}^{N} \underbrace{x_i}_{K \times 1} \underbrace{\left(y_i - x_i^\top \beta\right)}_{1 \times 1} \\ -\frac{1}{\sigma^4} \sum_{i=1}^{N} \underbrace{x_i^\top}_{1 \times K} \underbrace{\left(y_i - x_i^\top \beta\right)}_{1 \times 1} & \frac{N}{2\sigma^4} - \frac{1}{\sigma^6} \sum_{i=1}^{N} \underbrace{\left(y_i - x_i^\top \beta\right)^2}_{1 \times 1} \end{pmatrix}$$

4. Maximum Likelihood Estimator

Solution (cont'd):

Second Order Conditions (SOC)

$$\left.\frac{\partial^2 \ell_N(\theta)}{\partial\theta\partial\theta^\top}\right|_{\widehat{\theta}} = \begin{pmatrix} -\frac{1}{\widehat{\sigma}^2}\sum_{i=1}^N x_i x_i^\top & -\frac{1}{\widehat{\sigma}^4}\sum_{i=1}^N x_i\left(y_i - x_i^\top\widehat{\beta}\right) \\ -\frac{1}{\widehat{\sigma}^4}\sum_{i=1}^N x_i^\top\left(y_i - x_i^\top\widehat{\beta}\right) & \frac{N}{2\widehat{\sigma}^4} - \frac{1}{\widehat{\sigma}^6}\sum_{i=1}^N\left(y_i - x_i^\top\widehat{\beta}\right)^2 \end{pmatrix}$$

Since $\sum_{i=1}^N x_i^\top\left(y_i - x_i^\top\widehat{\beta}\right) = 0$ (FOC) and $N\widehat{\sigma}^2 = \sum_{i=1}^N\left(y_i - x_i^\top\widehat{\beta}\right)^2$

$$\left.\frac{\partial^2 \ell_N(\theta)}{\partial\theta\partial\theta^\top}\right|_{\widehat{\theta}} = \begin{pmatrix} -\frac{N}{\widehat{\sigma}^2}\sum_{i=1}^N x_i x_i^\top & 0 \\ 0 & \frac{N}{2\widehat{\sigma}^4} - \frac{N\widehat{\sigma}^2}{\widehat{\sigma}^6} \end{pmatrix}$$

4. Maximum Likelihood Estimator

Solution (cont'd):

Second Order Conditions (SOC).

$$\left.\frac{\partial^2 \ell_N\left(\theta;\, y\right|x)}{\partial\theta\partial\theta^\top}\right|_{\widehat{\theta}} = \begin{pmatrix} -\frac{1}{\widehat{\sigma}^2}\sum_{i=1}^N x_i x_i^\top & 0 \\ 0 & -\frac{N}{2\widehat{\sigma}^4} \end{pmatrix} \text{ is definite negative}$$

Since $\sum_{i=1}^N x_i x_i^\top$ is positive definite (assumption), the Hessian matrix is definite negative and $\widehat{\theta}$ is the MLE of the parameters θ.

4. Maximum Likelihood Estimator

Theorem (Equivariance or Invariance Principle)

Under suitable regularity conditions, the maximum likelihood estimator of a function $g(.)$ of the parameter θ is $g\left(\widehat{\theta}\right)$, where $\widehat{\theta}$ is the maximum likelihood estimator of θ.

Invariance Principle

- The MLE is invariant to one-to-one transformations of θ. Any transformation that is not one to one either renders the model inestimable if it is one to many or imposes restrictions if it is many to one.
- For the practitioner, this result is extremely useful. For example, when a parameter appears in a likelihood function in the form $1/\theta$, it is usually worthwhile to reparameterize the model in terms of $\gamma = 1/\theta$.
- Example: Olsen (1978) and the reparametrisation of the likelihood function of the Tobit Model.

Example (Invariance Principle)

Suppose that the normal log-likelihood in the previous example is parameterized in terms of the precision parameter, $\gamma^2 = 1/\sigma^2$. The log-likelihood

$$\ell_N\left(m, \sigma^2; y\right) = -\frac{N}{2}\ln\left(\sigma^2\right) - \frac{N}{2}\ln\left(2\pi\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}\left(y_i - m\right)^2$$

becomes

$$\ell_N\left(m, \gamma^2; y\right) = \frac{N}{2}\ln\left(\gamma^2\right) - \frac{N}{2}\ln\left(2\pi\right) - \frac{\gamma^2}{2}\sum_{i=1}^{N}\left(y_i - m\right)^2$$

4. Maximum Likelihood Estimator

Example (Invariance Principle, cont'd)

The MLE for m is clearly still $\overline{Y}_N$. But the likelihood equation for γ^2 is now:

$$\frac{\partial \ell_N \left(m, \gamma^2; y\right)}{\partial \gamma^2} = \frac{N}{2\gamma^2} - \frac{1}{2} \sum_{i=1}^{N} \left(y_i - m\right)^2$$

and the MLE for γ^2 is now defined by:

$$\widehat{\gamma}^2 = \frac{N}{\sum_{i=1}^{N} \left(Y_i - m\right)^2} = \frac{1}{\widehat{\sigma}^2}$$

as expected.

Key Concepts

1. Identification.
2. Maximum likelihood estimator.
3. Maximum likelihood estimate.
4. Log-likelihood equations.
5. Equivariance or invariance principle.
6. Gradient Vector and Hessian Matrix (deterministic elements).

Section 5

Score, Hessian and Fisher Information

5. Score, Hessian and Fisher Information

Objectives

We aim at introducing the following concepts:

1. Score vector and gradient
2. Hessian matrix
3. Fischer information matrix of the sample
4. Fischer information matrix of one observation for marginal and conditional distributions
5. Average Fischer information matrix of one observation

5. Score, Hessian and Fisher Information

Definition (Score Vector)

The (conditional) **score vector** is a $K \times 1$ vector defined by:

$$\underset{(K,1)}{s_N\left(\theta;\, Y|\, x\right)} \equiv s\left(\theta\right) = \frac{\partial \ell_N\left(\theta;\, Y|\, x\right)}{\partial \theta}$$

5. Score, Hessian and Fisher Information

Remarks:

- The score $s_N(\theta;\, Y|\, x)$ is a vector of **random elements** since it depends on the random variables $Y_1, .., .Y_N$.
- For an unconditional log-likelihood, $\ell_N(\theta; x)$, the score is denoted by

$$s_N(\theta; X) = \partial \ell_N(\theta; X) / \partial \theta$$

- The score is a $K \times 1$ vector such that:

$$s_N(\theta;\, Y|\, x) = \begin{pmatrix} \frac{\partial \ell_N(\theta;\, Y|x)}{\partial \theta_1} \\ . \\ \frac{\partial \ell_N(\theta;\, Y|x)}{\partial \theta_K} \end{pmatrix}$$

5. Score, Hessian and Fisher Information

Corollary

By definition, the **score vector** *satisfies*

$$\mathbb{E}_{\theta}\left(s_N\left(\theta;\, Y\middle|\, x\right)\right) = 0_K$$

where $\mathbb{E}_{\theta}$ means the expectation with respect to the conditional distribution $Y|\, X = x$.

5. Score, Hessian and Fisher Information

Remark: If we consider a variable X with a pdf $f_X(x;\theta)$, $\forall x \in \mathbb{R}$, then $\mathbb{E}_\theta(.)$ means the expectation with respect to the distribution of X :

$$\mathbb{E}_\theta(s_N(\theta;X)) = \int_{-\infty}^{\infty} s_N(\theta;x) f_X(x;\theta)\, dx = 0$$

Remark: If we consider a variable Y with a conditional pdf $f_{Y|x}(y;\theta)$, $\forall y \in \mathbb{R}$, then $\mathbb{E}_\theta(.)$ means the expectation with respect to the distribution of $Y|X = x$:

$$\mathbb{E}_\theta(s_N(\theta;\, Y|\, x)) = \int_{-\infty}^{\infty} s_N(\theta;\, Y|\, x) f_{Y|x}(y;\theta)\, dy = 0$$

5. Score, Hessian and Fisher Information

Proof.

If we consider a variable X with a pdf $f_X(x;\theta)$, $\forall x \in \mathbb{R}$, then:

$$\begin{aligned}
\mathbb{E}_\theta\left(s_N(\theta;X)\right) &= \int s_N(\theta;x)\, f_X(x;\theta)\, dx \\
&= N \int \frac{\partial \ln f_X(x;\theta)}{\partial \theta} f_X(x;\theta)\, dx \\
&= N \int \frac{1}{f_X(x;\theta)} \frac{\partial f_X(x;\theta)}{\partial \theta} f_X(x;\theta)\, dx \\
&= N \frac{\partial}{\partial \theta} \int f_X(x;\theta)\, dx \\
&= N \frac{\partial 1}{\partial \theta} = 0
\end{aligned}$$

□

5. Score, Hessian and Fisher Information

Example (Exponential Distribution)

Suppose that $D_1, D_2, .., D_N$ are $i.i.d.$, positive random variable with $D_i \sim Exp(\theta)$ and $\mathbb{E}(D_i) = \theta > 0$.

$$f_D(d;\theta) = \frac{1}{\theta}\exp\left(-\frac{d}{\theta}\right), \ \forall d \in \mathbb{R}^+$$

$$\ell_N(\theta; d) = -N\ln(\theta) - \frac{1}{\theta}\sum_{i=1}^{N} d_i$$

The score (scalar) is equal to:

$$s_N(\theta; D) = -\frac{N}{\theta} + \frac{1}{\theta^2}\sum_{i=1}^{N} D_i$$

Example (Exponential Distribution, cont'd)

By definition:

$$\begin{aligned}
\mathbb{E}_\theta\left(s_N\left(\theta; D\right)\right) &= \mathbb{E}_\theta\left(-\frac{N}{\theta}+\frac{1}{\theta^2}\sum_{i=1}^{N} D_i\right) \\
&= -\frac{N}{\theta}+\frac{1}{\theta^2}\sum_{i=1}^{N}\mathbb{E}_\theta\left(D_i\right) \\
&= -\frac{N}{\theta}+\frac{N\theta}{\theta^2} \\
&= 0 \quad \square
\end{aligned}$$

5. Score, Hessian and Fisher Information

Example (Linear Regression Model)

Let us consider the previous linear regression model $y_i = x_i^\top \beta + \varepsilon_i$. The score is defined by:

$$s_N(\theta; Y|x) = \begin{pmatrix} \frac{1}{\sigma^2} \sum_{i=1}^N x_i \left(Y_i - x_i^\top \beta \right) \\ -\frac{N}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^N \left(Y_i - x_i^\top \beta \right)^2 \end{pmatrix}$$

Then, we have

$$\mathbb{E}_\theta \left(s_N(\theta; Y|x) \right) = \mathbb{E}_\theta \begin{pmatrix} \frac{1}{\sigma^2} \sum_{i=1}^N x_i \left(Y_i - x_i^\top \beta \right) \\ -\frac{N}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^N \left(Y_i - x_i^\top \beta \right)^2 \end{pmatrix}$$

5. Score, Hessian and Fisher Information

Example (Linear Regression Model, cont'd)

We know that $\mathbb{E}_\theta\left(\left. Y_i \right| x \right) = x_i^\top \beta$. So, we have:

$$\begin{aligned}
\mathbb{E}_\theta\left(\frac{1}{\sigma^2} \sum_{i=1}^N x_i \left(Y_i - x_i^\top \beta \right) \right) &= \frac{1}{\sigma^2} \sum_{i=1}^N x_i \left(\mathbb{E}_\theta\left(\left. Y_i \right| x \right) - x_i^\top \beta \right) \\
&= \frac{1}{\sigma^2} \sum_{i=1}^N x_i \left(x_i^\top \beta - x_i^\top \beta \right) \\
&= 0_K
\end{aligned}$$

5. Score, Hessian and Fisher Information

Example (Linear Regression Model, cont'd)

$$\begin{aligned}
& \mathbb{E}_\theta\left(-\frac{N}{2\sigma^2}+\frac{1}{2\sigma^4}\textstyle\sum_{i=1}^N\left(Y_i-x_i^\top\beta\right)^2\right) \\
= & -\frac{N}{2\sigma^2}+\frac{1}{2\sigma^4}\textstyle\sum_{i=1}^N\mathbb{E}_\theta\left(\left(Y_i-x_i^\top\beta\right)^2\right) \\
= & -\frac{N}{2\sigma^2}+\frac{1}{2\sigma^4}\textstyle\sum_{i=1}^N\mathbb{E}_\theta\left(\left(Y_i-\mathbb{E}_\theta\left(\left.Y_i\right|x\right)\right)^2\right) \\
= & -\frac{N}{2\sigma^2}+\frac{1}{2\sigma^4}\textstyle\sum_{i=1}^N\mathbb{V}_\theta\left(\left.Y_i\right|x\right) \\
= & -\frac{N}{2\sigma^2}+\frac{N\sigma^2}{2\sigma^4} \\
= & \; 0 \quad \square
\end{aligned}$$

5. Score, Hessian and Fisher Information

Definition (Gradient)

The **gradient vector** associated to the log-likelihood function is a $K \times 1$ vector defined by:

$$\underset{(K,1)}{g_N\left(\theta; y|\, x\right)} \equiv g\left(\theta\right) = \frac{\partial \ell_N\left(\theta; y|\, x\right)}{\partial \theta}$$

5. Score, Hessian and Fisher Information

Remarks

1. The gradient $g_N(\theta; y|x)$ is a vector of **deterministic entries** since it depends on the realisation $y_1, .., y_N$.

2. For an unconditional log-likelihood, the gradient is defined by

$$g_N(\theta; x) = \partial \ell_N(\theta; x) / \partial \theta$$

3. The gradient is a $K \times 1$ vector such that:

$$g_N(\theta; y|x) = \begin{pmatrix} \frac{\partial \ell_N(\theta; y|x)}{\partial \theta_1} \\ . \\ \frac{\partial \ell_N(\theta; y|x)}{\partial \theta_K} \end{pmatrix}$$

5. Score, Hessian and Fisher Information

Corollary

By definition of the FOC, the **gradient vector** *satisfies*

$$g_N\left(\widehat{\theta};\, y|\, x\right) = 0_K$$

where $\widehat{\theta} = \widehat{\theta}\,(x)$ *is the maximum likelihood* **estimate** *of* θ.

5. Score, Hessian and Fisher Information

Example (Linear regression model)

In the linear regression model, the gradient associated to the log-likelihood function is defined to be:

$$g_N\left(\theta; y|\, x\right) = \begin{pmatrix} \frac{1}{\sigma^2}\sum_{i=1}^{N} x_i \left(y_i - x_i^\top \beta\right) \\ -\frac{N}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{N}\left(y_i - x_i^\top \beta\right)^2 \end{pmatrix}$$

Given the FOC, we have:

$$g_N\left(\widehat{\theta}; y|\, x\right) = \begin{pmatrix} \frac{1}{\widehat{\sigma}^2}\sum_{i=1}^{N} x_i \left(y_i - x_i^\top \widehat{\beta}\right) \\ -\frac{N}{2\widehat{\sigma}^2} + \frac{1}{2\widehat{\sigma}^4}\sum_{i=1}^{N}\left(y_i - x_i^\top \widehat{\beta}\right)^2 \end{pmatrix} = \begin{pmatrix} 0_K \\ 0 \end{pmatrix}$$

5. Score, Hessian and Fisher Information

Definition (Hessian Matrix)

The Hessian matrix (deterministic) is defined as to be:

$$H_N\left(\theta; y \middle| x\right) = \frac{\partial^2 \ell_N\left(\theta; y \middle| x\right)}{\partial\theta\partial\theta^\top}$$

Remarks: The matrix $\frac{\partial^2 \ell_N(\theta; y|x)}{\partial\theta\partial\theta^\top}$ is also called the Hessian matrix, but do not confuse the two matrices $\frac{\partial^2 \ell_N(\theta; Y|x)}{\partial\theta\partial\theta^\top}$ and $\frac{\partial^2 \ell_N(\theta; y|x)}{\partial\theta\partial\theta^\top}$.

5. Score, Hessian and Fisher Information

Random Variable		Constant	
Score vector	$\frac{\partial \ell_N(\theta; Y\|x)}{\partial \theta}$	Gradient vector	$\frac{\partial \ell_N(\theta; y\|x)}{\partial \theta}$
Hessian Matrix	$\frac{\partial^2 \ell_N(\theta; Y\|x)}{\partial \theta \partial \theta^\top}$	Hessian Matrix	$\frac{\partial^2 \ell_N(\theta; y\|x)}{\partial \theta \partial \theta^\top}$

5. Score, Hessian and Fisher Information

Definition (Fisher Information Matrix)

The (conditional) Fisher information matrix associated **to the sample** $\{Y_1, .., Y_N\}$ is the variance-covariance matrix of the score vector:

$$\underbrace{I_N(\theta)}_{K \times K} = \mathbb{V}_\theta \left(s_N(\theta; Y|x) \right)$$

or equivalently:

$$I_N(\theta) = \mathbb{V}_\theta \left(\frac{\partial \ell_N(\theta; Y|x)}{\partial \theta} \right)$$

where $\mathbb{V}_\theta$ means the variance with respect to the conditional distribution $Y|X$.

5. Score, Hessian and Fisher Information

Corollary

Since by definition $\mathbb{E}_{\theta}\left(s_N\left(\theta;\, Y\middle|\, x\right)\right) = 0$, *then an alternative definition of the Fisher information matrix* **of the sample** $\{Y_1, .., Y_N\}$ *is:*

$$\underbrace{I_N\left(\theta\right)}_{K\times K} = \mathbb{E}_{\theta}\left(\underbrace{s_N\left(\theta;\, Y\middle|\, x\right)}_{K\times 1} \times \underbrace{s_N\left(\theta;\, Y\middle|\, x\right)^{\top}}_{1\times K}\right)$$

5. Score, Hessian and Fisher Information

Definition (Fisher Information Matrix)

The (conditional) Fisher information matrix **of the sample** $\{Y_1, .., Y_N\}$ is also given by:

$$I_N(\theta) = \mathbb{E}_\theta\left(-\frac{\partial^2 \ell_N(\theta;\, Y|\, x)}{\partial\theta\partial\theta^\top}\right) = \mathbb{E}_\theta\left(-H_N(\theta;\, Y|\, x)\right)$$

5. Score, Hessian and Fisher Information

Definition (Fisher Information Matrix, summary)

The (conditional) Fisher information matrix **of the sample** $\{Y_1, .., Y_N\}$ can alternatively be defined by:

$$I_N(\theta) = \mathbb{V}_\theta\left(s_N(\theta; Y|x)\right)$$

$$I_N(\theta) = \mathbb{E}_\theta\left(s_N(\theta; Y|x) \times s_N(\theta; Y|x)^\top\right)$$

$$I_N(\theta) = \mathbb{E}_\theta\left(-H_N(\theta; Y|x)\right)$$

where $\mathbb{E}_\theta$ and $\mathbb{V}_\theta$ denote the mean and the variance with respect to the conditional distribution $Y|X$, and where $s_N(\theta; Y|x)$ denotes the score vector and $H_N(\theta; Y|x)$ the Hessian matrix.

5. Score, Hessian and Fisher Information

Definition (Fisher Information Matrix, summary)

The (conditional) Fisher information matrix **of the sample** $\{Y_1, .., Y_N\}$ can alternatively be defined by:

$$I_N(\theta) = \mathbb{V}_\theta\left(\frac{\partial \ell_N(\theta; Y|x)}{\partial \theta}\right)$$

$$I_N(\theta) = \mathbb{E}_\theta\left(\frac{\partial \ell_N(\theta; Y|x)}{\partial \theta} \times \left(\frac{\partial \ell_N(\theta; Y|x)}{\partial \theta}\right)^\top\right)$$

$$I_N(\theta) = \mathbb{E}_\theta\left(-\frac{\partial^2 \ell_N(\theta; Y|x)}{\partial \theta \partial \theta^\top}\right)$$

where $\mathbb{E}_\theta$ and $\mathbb{V}_\theta$ denote the mean and the variance with respect to the conditional distribution $Y|X$.

Remarks

1. Three equivalent definitions of the Fisher information matrix, and as a consequence three different consistent estimates of the Fisher information matrix (see later).
2. The Fisher information matrix associated to the sample $\{Y_1, .., Y_N\}$ can also be defined from the Fisher information matrix for the **observation** i.

5. Score, Hessian and Fisher Information

Definition (Fisher Information Matrix)

The (conditional) Fisher information matrix associated to the i^{th} **individual** can be defined by:

$$I_i(\theta) = \mathbb{V}_\theta \left(\frac{\partial \ell_i(\theta; Y_i | x_i)}{\partial \theta} \right)$$

$$I_i(\theta) = \mathbb{E}_\theta \left(\frac{\partial \ell_i(\theta; Y_i | x_i)}{\partial \theta} \frac{\partial \ell_i(\theta; Y_i | x_i)^\top}{\partial \theta} \right)$$

$$I_i(\theta) = \mathbb{E}_\theta \left(- \frac{\partial^2 \ell_i(\theta; Y_i | x_i)}{\partial \theta \partial \theta^\top} \right)$$

where $\mathbb{E}_\theta$ and $\mathbb{V}_\theta$ denote the expectation and variance with respect to the true conditional distribution $Y_i | X_i$.

5. Score, Hessian and Fisher Information

Definition (Fisher Information Matrix)

The (conditional) Fisher information matrix associated to the i^{th} **individual** can be alternatively be defined by:

$$I_i(\theta) = \mathbb{V}_\theta \left(s_i\left(\theta; \left. Y_i \right| x_i\right)\right)$$

$$I_i(\theta) = \mathbb{E}_\theta \left(s_i\left(\theta; \left. Y_i \right| x_i\right) s_i\left(\theta; \left. Y_i \right| x_i\right)^\top \right)$$

$$I_i(\theta) = \mathbb{E}_\theta \left(-H_i\left(\theta; \left. Y_i \right| x_i\right)\right)$$

where $\mathbb{E}_\theta$ and $\mathbb{V}_\theta$ denote the expectation and variance with respect to the true conditional distribution $\left. Y_i \right| X_i$.

5. Score, Hessian and Fisher Information

Theorem

The Fisher information matrix associated to the sample $\{Y_1, .., Y_N\}$ *is equal to the sum of individual Fisher information matrices:*

$$I_N(\theta) = \sum_{i=1}^{N} I_i(\theta)$$

Remark:

1. In the case of a **marginal** log-likelihood, the Fisher information matrix associated to the **variable** X_i is the same for the observations i :

$$I_i(\theta) = I(\theta) \quad \forall i = 1, ..N$$

2. In the case of a **conditional** log-likelihood, the Fisher information matrix associated to the **variable** Y_i given $X_i = x_i$ depends on the observation i :

$$I_i(\theta) \neq I_j(\theta) \quad \forall i \neq j$$

Example (Exponential marginal distribution)

Suppose that $D_1, D_2, .., D_N$ are $i.i.d.$, positive random variable with $D_i \sim Exp(\theta)$

$$\mathbb{E}(D_i) = \theta \qquad \mathbb{V}(D_i) = \theta^2$$

$$f_D(d;\theta) = \frac{1}{\theta}\exp\left(-\frac{d}{\theta}\right), \ \forall d \in \mathbb{R}^+$$

$$\ell_i(\theta; d_i) = -\ln(\theta) - \frac{d_i}{\theta}$$

Question: what is the Fisher information number (scalar) associated to D_i?

5. Score, Hessian and Fisher Information

Solution

$$\ell\left(\theta; d_i\right) = -\ln\left(\theta\right) - \frac{d_i}{\theta}$$

The score of the observation X_i is defined by:

$$s_i\left(\theta; D_i\right) = \frac{\partial \ell_i\left(\theta; D_i\right)}{\partial \theta} = -\frac{1}{\theta} + \frac{D_i}{\theta^2}$$

Let us use the three definitions of the information quantity $I_i\left(\theta\right)$:

$$\begin{aligned} I_i\left(\theta\right) &= \mathbb{V}_\theta\left(s_i\left(\theta; D_i\right)\right) \\ &= \mathbb{E}_\theta\left(s_i\left(\theta; D_i\right)^2\right) \\ &= \mathbb{E}_\theta\left(-H_i\left(\theta; D_i\right)\right) \end{aligned}$$

5. Score, Hessian and Fisher Information

Solution, cont'd

$$s_i\left(\theta; D_i\right) = \frac{\partial \ell_i\left(\theta; D_i\right)}{\partial \theta} = -\frac{1}{\theta} + \frac{D_i}{\theta^2}$$

First definition:

$$\begin{aligned} I_i\left(\theta\right) &= \mathbb{V}_\theta\left(s_i\left(\theta; D_i\right)\right) \\ &= \mathbb{V}_\theta\left(-\frac{1}{\theta} + \frac{D_i}{\theta^2}\right) \\ &= \frac{1}{\theta^4}\mathbb{V}_\theta\left(D_i\right) \\ &= \frac{1}{\theta^2} \end{aligned}$$

Conclusion: $I_i\left(\theta\right) = I(\theta)$ does not depend on i.

5. Score, Hessian and Fisher Information

Solution, cont'd

$$s_i(\theta; D_i) = \frac{\partial \ell_i(\theta; D_i)}{\partial \theta} = -\frac{1}{\theta} + \frac{D_i}{\theta^2}$$

Second definition:

$$\begin{aligned}
I_i(\theta) &= \mathbb{E}_\theta\left(s_i(\theta; D_i)^2\right) \\
&= \mathbb{E}_\theta\left(\left(-\frac{1}{\theta} + \frac{D_i}{\theta^2}\right)^2\right) \\
&= \mathbb{V}_\theta\left(-\frac{1}{\theta} + \frac{D_i}{\theta^2}\right) \quad \text{since } \mathbb{E}_\theta\left(-\frac{1}{\theta} + \frac{D_i}{\theta^2}\right) = 0 \\
&= \frac{1}{\theta^2}
\end{aligned}$$

Conclusion: $I_i(\theta) = I(\theta)$ does not depend on i.

5. Score, Hessian and Fisher Information

Solution, cont'd

$$s_i(\theta; D_i) = \frac{\partial \ell_i(\theta; D_i)}{\partial \theta} = -\frac{1}{\theta} + \frac{D_i}{\theta^2}$$

$$H_i(\theta; D_i) = \frac{\partial^2 \ell_i(\theta; D_i)}{\partial \theta^2} = \frac{1}{\theta^2} - \frac{2D_i}{\theta^3}$$

Third definition:

$$\begin{aligned}
I_i(\theta) &= \mathbb{E}_\theta\left(-H_i(\theta; D_i)\right) \\
&= \mathbb{E}_\theta\left(-\left(\frac{1}{\theta^2} - \frac{2D_i}{\theta^3}\right)\right) \\
&= -\frac{1}{\theta^2} + \frac{2}{\theta^3}\mathbb{E}_\theta(D_i) \\
&= -\frac{1}{\theta^2} + \frac{2}{\theta^3}\theta = \frac{1}{\theta^2}
\end{aligned}$$

Conclusion: $I_i(\theta) = I(\theta)$ does not depend on i.

5. Score, Hessian and Fisher Information

Example (Linear regression model)

We shown that:

$$\frac{\partial^2 \ell_i\left(\theta;\, Y_i \middle| x_i\right)}{\partial\theta\partial\theta^\top} = \begin{pmatrix} -\frac{1}{\sigma^2} \underbrace{x_i}_{K\times 1} \underbrace{x_i^\top}_{1\times K} & -\frac{1}{\sigma^4} \underbrace{x_i}_{K\times 1} \underbrace{\left(Y_i - x_i^\top \beta\right)}_{1\times 1} \\ -\frac{1}{\sigma^4} \underbrace{x_i^\top}_{1\times K} \underbrace{\left(Y_i - x_i^\top \beta\right)}_{1\times 1} & \frac{1}{2\sigma^4} - \frac{1}{\sigma^6} \underbrace{\left(Y_i - x_i^\top \beta\right)^2}_{1\times 1} \end{pmatrix}$$

Question: what is the Fisher information matrix associated to the observation Y_i?

5. Score, Hessian and Fisher Information

Solution

The information matrix is then defined by:

$$\underbrace{I_i(\theta)}_{K+1\times K+1} = \mathbb{E}_\theta\left(-\frac{\partial^2 \ell_i(\theta;\, Y_i|\, x_i)}{\partial\theta\partial\theta^\top}\right) = \mathbb{E}_\theta\left(-H_i(\theta;\, Y_i|\, x_i)\right)$$

where $\mathbb{E}_\theta$ means the expectation with respect to the conditional distribution $Y_i|\, X_i = x_i$

$$I_i(\theta) = \begin{pmatrix} \frac{1}{\sigma^2} x_i x_i^\top & \frac{1}{\sigma^4} x_i\left(\mathbb{E}_\theta(Y_i) - x_i^\top\beta\right) \\ \frac{1}{\sigma^4} x_i^\top\left(\mathbb{E}_\theta(Y_i) - x_i^\top\beta\right) & -\frac{1}{2\sigma^4} + \frac{1}{\sigma^6}\mathbb{E}_\theta\left(\left(Y_i - x_i^\top\beta\right)^2\right) \end{pmatrix}$$

5. Score, Hessian and Fisher Information

Solution (cont'd)

$$I_i(\theta) = \begin{pmatrix} \frac{1}{\sigma^2} x_i x_i^\top & \frac{1}{\sigma^4} x_i \left(\mathbb{E}_\theta(Y_i) - x_i^\top \beta \right) \\ \frac{1}{\sigma^4} x_i^\top \left(\mathbb{E}_\theta(Y_i) - x_i^\top \beta \right) & -\frac{1}{2\sigma^4} + \frac{1}{\sigma^6} \mathbb{E}_\theta\left(\left(Y_i - x_i^\top \beta \right)^2 \right) \end{pmatrix}$$

Given that $\mathbb{E}_\theta(Y_i) = x_i^\top \beta$ and $\mathbb{E}_\theta(\left(Y_i - x_i^\top \beta\right)^2) = \sigma^2$, then we have:

$$I_i(\theta) = \begin{pmatrix} \frac{1}{\sigma^2} x_i x_i^\top & 0 \\ 0 & \frac{1}{2\sigma^4} \end{pmatrix}$$

Conclusion: $I_i(\theta)$ depends on x_i and $I_i(\theta) \neq I_j(\theta)$ for $i \neq j$.

5. Score, Hessian and Fisher Information

Definition (Average Fisher information matrix)

For a **conditional model**, the **average** Fisher information matrix for one observation is defined by:

$$I(\theta) = \mathbb{E}_X \left(I_i(\theta) \right)$$

where $\mathbb{E}_X$ denotes the expectation with respect to X (conditioning variable).

5. Score, Hessian and Fisher Information

Summary: For a **conditional model** (and only for a conditional model), we have:

$$I(\theta) = \mathbb{E}_X \left(\mathbb{V}_\theta \left(\frac{\partial \ell_i(\theta;\, Y_i|\, X_i)}{\partial \theta} \right) \right) = \mathbb{E}_X \left(\mathbb{V}_\theta \left(s(\theta;\, Y_i|\, X_i) \right) \right)$$

$$\begin{aligned} I(\theta) &= \mathbb{E}_X \mathbb{E}_\theta \left(\frac{\partial \ell_i(\theta;\, Y_i|\, X_i)}{\partial \theta} \frac{\partial \ell_i(\theta;\, Y_i|\, X_i)^\top}{\partial \theta} \right) \\ &= \mathbb{E}_X \mathbb{E}_\theta \left(s_i(\theta;\, Y_i|\, X_i)\, s_i(\theta;\, Y_i|\, X_i)^\top \right) \end{aligned}$$

$$I(\theta) = \mathbb{E}_X \mathbb{E}_\theta \left(- \frac{\partial^2 \ell_i(\theta;\, Y_i|\, X_i)}{\partial \theta \partial \theta^\top} \right) = \mathbb{E}_X \mathbb{E}_\theta \left(-H_i(\theta;\, Y_i|\, X_i) \right)$$

5. Score, Hessian and Fisher Information

Summary: For a **marginal distribution**, we have:

$$I(\theta) = \mathbb{V}_\theta\left(\frac{\partial \ell_i(\theta; Y_i)}{\partial \theta}\right) = \mathbb{V}_\theta\left(s(\theta; Y_i)\right)$$

$$\begin{aligned} I(\theta) &= \mathbb{E}_\theta\left(\frac{\partial \ell_i(\theta; Y_i)}{\partial \theta}\frac{\partial \ell_i(\theta; Y_i)}{\partial \theta}^{\top}\right) \\ &= \mathbb{E}_\theta\left(s_i(\theta; Y_i)\, s_i(\theta; Y_i)^{\top}\right) \end{aligned}$$

$$I(\theta) = \mathbb{E}_\theta\left(-\frac{\partial^2 \ell_i(\theta; Y_i)}{\partial \theta \partial \theta^{\top}}\right) = \mathbb{E}_\theta\left(-H_i(\theta; Y_i)\right)$$

5. Score, Hessian and Fisher Information

Example (Linear Regression Model)

In the linear model, the individual Fisher information matrix is equal to:

$$I_i(\theta) = \begin{pmatrix} \frac{1}{\sigma^2} x_i x_i^\top & 0 \\ 0 & \frac{1}{2\sigma^4} \end{pmatrix}$$

and the average Fisher information Matrix for one observation is defined by:

$$I(\theta) = \mathbb{E}_X(I_i(\theta)) = \begin{pmatrix} \frac{1}{\sigma^2} \mathbb{E}_X\left(X_i X_i^\top\right) & 0 \\ 0 & \frac{1}{2\sigma^4} \end{pmatrix}$$

5. Score, Hessian and Fisher Information

Summary: in order to compute the average information matrix $I(\theta)$ for one observation:

Step 1: Compute the Hessian matrix or the score vector for **one observation**

$$H_i\left(\theta;\, Y_i \middle| x_i\right) = \frac{\partial^2 \ell_i\left(\theta;\, Y_i \middle| x_i\right)}{\partial\theta\partial\theta^\top} \qquad s_i\left(\theta;\, Y_i \middle| x_i\right) = \frac{\partial \ell_i\left(\theta;\, Y_i \middle| x_i\right)}{\partial\theta}$$

Step 2: Take the expectation (or the variance) with respect to the conditional distribution $Y_i | X_i = x_i$

$$I_i\left(\theta\right) = \mathbb{V}_\theta\left(s_i\left(\theta;\, Y_i \middle| x_i\right)\right) = \mathbb{E}_\theta\left(-H_i\left(\theta;\, Y_i \middle| x_i\right)\right)$$

Step 3: Take the expectation with respect to the conditioning variable X

$$I\left(\theta\right) = \mathbb{E}_X\left(I_i\left(\theta\right)\right)$$

5. Score, Hessian and Fisher Information

Theorem

In a sampling model (with i.i.d. observations), one has:

$$\mathtt{I}_N(\theta) = N\ \mathtt{I}(\theta)$$

5. Score, Hessian and Fisher Information

	Marginal Distribution	Cond. Distribution (model)
pdf	$f_{X_i}(\theta; x_i)$	$f_{Y_i \mid x_i}(\theta; y \mid x)$
Score Vector	$s_i(\theta; X_i)$	$s_i(\theta; Y_i \mid x_i)$
Hessian Matrix	$H_i(\theta; X_i)$	$H_i(\theta; Y_i \mid x_i)$
Information matrix	$I_i(\theta) = I(\theta)$	$I_i(\theta)$
Av. Infor. Matrix	$I(\theta) = I_i(\theta)$	$I(\theta) = \mathbb{E}_X(I_i(\theta))$

$$\text{with } I_i(\theta) = \mathbb{V}_\theta(s_i(\theta; Y_i \mid x_i)) = \mathbb{E}_\theta\left(s_i(\theta; Y_i \mid x_i)\, s_i(\theta; Y_i \mid x_i)^\top\right) = \mathbb{E}_\theta(-H_i(\theta; Y_i \mid x_i))$$

How to estimate the average Fisher Information Matrix?

- This matrix is particularly important, since we will see that its corresponds to the **asymptotic variance covariance matrix** of the MLE.
- Let us assume that we have a consistent estimator $\widehat{\theta}$ of the parameter θ, how to estimate the average Fisher information matrix?

5. Score, Hessian and Fisher Information

Definition (Estimators of the average Fisher Information Matrix)

If $\widehat{\theta}$ converges in probability to θ_0 (true value), then:

$$\widehat{I}\left(\widehat{\theta}\right) = \frac{1}{N}\sum_{i=1}^{N}\widehat{I}_i\left(\widehat{\theta}\right)$$

$$\widehat{I}\left(\widehat{\theta}\right) = \frac{1}{N}\sum_{i=1}^{N}\left(\left.\frac{\partial \ell_i\left(\theta;\, y_i|\, x_i\right)}{\partial \theta}\right|_{\widehat{\theta}} \left.\frac{\partial \ell_i\left(\theta;\, y_i|\, x_i\right)}{\partial \theta}\right|_{\widehat{\theta}}^{\top}\right)$$

$$\widehat{I}\left(\widehat{\theta}\right) = \frac{1}{N}\sum_{i=1}^{N}\left(-\left.\frac{\partial^2 \ell_i\left(\theta;\, y_i|\, x_i\right)}{\partial \theta \partial \theta^{\top}}\right|_{\widehat{\theta}}\right)$$

are three consistent estimators of the average Fisher information matrix.

5. Score, Hessian and Fisher Information

1. The first estimator corresponds to the average of the N Fisher information matrices (for $Y_1, .., Y_N$) evaluated at the estimated value $\widehat{\theta}$. This estimator will rarely be available in practice.

2. The second estimator corresponds to the average of the product of the individual score vectors evaluated at $\widehat{\theta}$. It is known as the **BHHH** (Berndt, Hall, Hall, and Hausman, 1994) estimator or **OPG** estimator (outer product of gradients).

$$\widehat{I}\left(\widehat{\theta}\right) = \frac{1}{N}\sum_{i=1}^{N}\left(g_i\left(\widehat{\theta};\, y_i \middle|\, x_i\right) g_i\left(\widehat{\theta};\, y_i \middle|\, x_i\right)^{\top}\right)$$

5. Score, Hessian and Fisher Information

3. The third estimator corresponds to the opposite of the average of the Hessian matrices evaluated at $\widehat{\theta}$.

$$\widehat{I}\left(\widehat{\theta}\right)=\frac{1}{N}\sum_{i=1}^{N}\left(-H_i\left(\widehat{\theta};\, y_i\right|x_i\right)\Big)$$

5. Score, Hessian and Fisher Information

Problem

These three estimators are asymptotically equivalent, but they could give different results in finite samples. Available evidence suggests that in small or moderate sized samples, the Hessian is preferable (Greene, 2007). However, in most cases, the BHHH estimator will be the easiest to compute.

5. Score, Hessian and Fisher Information

Dependent Variable: GRADE
Method: ML - Binary Logit
Date: 09/06/02 Time: 18:40
Sample: 1 32
Included observations: 32
Convergence achieved after 4 iterations
Covariance matrix computed using second derivatives

Variable	Coefficient	Std. Error	z-Statistic	Prob.
C	-10.65600	4.057117	-2.626497	0.0086
TUCE	0.085551	0.133185	0.642352	0.5206
GPA	2.538281	1.181851	2.147716	0.0317

Mean dependent var	0.343750	S.D. dependent var	0.482559
S.E. of regression	0.419006	Akaike info criterion	1.186968
Sum squared resid	5.091415	Schwarz criterion	1.324380
Log likelihood	-15.99148	Hannan-Quinn criter.	1.232516
Restr. log likelihood	-20.59173	Avg. log likelihood	-0.499734
LR statistic (2 df)	9.200493	McFadden R-squared	0.223403
Probability(LR stat)	0.010049		
Obs with Dep=0	21	Total obs	32
Obs with Dep=1	11		

5. Score, Hessian and Fisher Information

Example (CAPM)

The empirical analogue of the CAPM is given by:

$$\widetilde{r}_{it} = \alpha_i + \beta_i \widetilde{r}_{mt} + \varepsilon_t$$

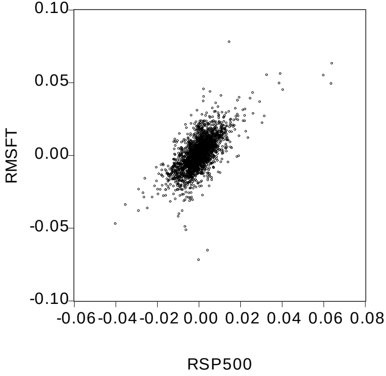
$$\widetilde{r}_{it} = \underbrace{r_{it} - r_{ft}}_{\text{excess return of security } i \text{ at time } t} \qquad \underbrace{\widetilde{r}_{mt} = (r_{mt} - r_{ft})}_{\text{market excess return at time } t}$$

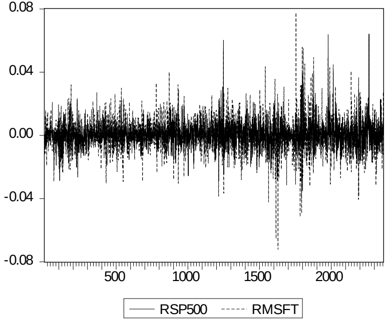
where ε_t is an $i.i.d.$ error term with:

$$\mathbb{E}(\varepsilon_t) = 0 \qquad \mathbb{V}(\varepsilon_t) = \sigma^2 \qquad \mathbb{E}(\varepsilon_t | \widetilde{r}_{mt}) = 0$$

Example (CAPM, cont'd)

Data (**data file: capm.xls**): Microsoft, SP500 and Tbill (closing prices) from 11/1/1993 to 04/03/2003





5. Score, Hessian and Fisher Information

Example (CAPM, cont'd)

We consider the CAPM model rewritten as follows

$$\widetilde{r}_{it} = \mathbf{x}_t^\top \boldsymbol{\beta} + \varepsilon_t \quad t = 1, ..T$$

where $\mathbf{x}_t = (1 \; \widetilde{r}_{mt})^\top$ is 2×1 vector of random variables, $\boldsymbol{\theta} = (\alpha_i : \beta_i : \sigma^2)^\top = \left(\boldsymbol{\beta}^\top : \sigma^2\right)^\top$ is 3×1 vector of parameters, and where the error term ε_t satisfies $\mathbb{E}(\varepsilon_t) = 0$, $\mathbb{V}(\varepsilon_t) = \sigma^2$ and $\mathbb{E}(\varepsilon_t | \widetilde{r}_{mt}) = 0$.

5. Score, Hessian and Fisher Information

Example (CAPM, cont'd)

Question: Compute three alternative estimators of the asymptotic variance covariance matrix of the MLE estimator $\widehat{\boldsymbol{\theta}} = \left(\widehat{\alpha}_i \; \widehat{\beta}_i \; \widehat{\sigma}^2 \right)^\top$

$$\widehat{\boldsymbol{\beta}} = \begin{pmatrix} \widehat{\alpha}_i \\ \widehat{\beta}_i \end{pmatrix} = \left(\sum_{t=1}^{T} \mathbf{x}_t \mathbf{x}_t^\top \right)^{-1} \left(\sum_{t=1}^{T} \mathbf{x}_t \widetilde{r}_{it} \right)$$

$$\widehat{\sigma}^2 = \frac{1}{T} \sum_{t=1}^{T} \left(\widetilde{r}_{it} - \mathbf{x}_t^\top \widehat{\boldsymbol{\beta}} \right)^2$$

5. Score, Hessian and Fisher Information

Solution The ML estimator is defined by:

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\beta} \in \mathbb{R}^2, \sigma^2 \in \mathbb{R}^+}{\arg\max} - \frac{T}{2} \ln\left(\sigma^2\right) - \frac{T}{2} \ln\left(2\pi\right) - \frac{1}{2\sigma^2} \sum_{t=1}^{T} \left(\widetilde{r}_{it} - \mathbf{x}_t^{\top} \widehat{\boldsymbol{\beta}}\right)^2$$

The problem is regular, so we have:

$$\sqrt{T}\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, I^{-1}\left(\boldsymbol{\theta}_0\right)\right)$$

or equivalently

$$\widehat{\boldsymbol{\theta}} \overset{asy}{\approx} \mathcal{N}\left(\boldsymbol{\theta}_0, \frac{1}{T} I^{-1}\left(\boldsymbol{\theta}_0\right)\right)$$

The asymptotic variance covariance matrix of $\widehat{\boldsymbol{\theta}}$ is

$$\mathbb{V}\left(\widehat{\boldsymbol{\theta}}\right) = \frac{1}{T} I^{-1}\left(\boldsymbol{\theta}_0\right)$$

5. Score, Hessian and Fisher Information

Solution (cont'd)

First estimator: The information matrix at time t is defined by (third definition):

$$I_t(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}}\left(-\frac{\partial^2 \ell_t\left(\boldsymbol{\theta};\, \widetilde{R}_{it}\middle|\, x_t\right)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}\right) = \mathbb{E}_{\boldsymbol{\theta}}\left(-H_t\left(\boldsymbol{\theta};\, \widetilde{R}_{it}\middle|\, x_t\right)\right)$$

where $\mathbb{E}_{\boldsymbol{\theta}}$ means the expectation with respect to the conditional distribution $\widetilde{R}_{it}\Big| X_t = x_t$

$$I_t(\boldsymbol{\theta}) = \begin{pmatrix} \frac{1}{\sigma^2}\mathbf{x}_t\mathbf{x}_t^\top & \frac{1}{\sigma^4}\mathbf{x}_t\left(\mathbb{E}_{\boldsymbol{\theta}}\left(\widetilde{R}_{it}\right) - \mathbf{x}_t^\top\boldsymbol{\beta}\right) \\ \frac{1}{\sigma^4}\mathbf{x}_t^\top\left(\mathbb{E}_{\boldsymbol{\theta}}\left(\widetilde{R}_{it}\right) - \mathbf{x}_t^\top\boldsymbol{\beta}\right) & -\frac{1}{2\sigma^4} + \frac{1}{\sigma^6}\mathbb{E}_{\boldsymbol{\theta}}\left(\left(\widetilde{R}_{it} - \mathbf{x}_t^\top\boldsymbol{\beta}\right)^2\right) \end{pmatrix}$$

Solution (cont'd)

First estimator:

$$I_t(\boldsymbol{\theta}) = \begin{pmatrix} \frac{1}{\sigma^2}\mathbf{x}_t\mathbf{x}_t^\top & \frac{1}{\sigma^4}\mathbf{x}_t\left(\mathbb{E}_{\boldsymbol{\theta}}\left(\widetilde{R}_{it}\right) - \mathbf{x}_t^\top\boldsymbol{\beta}\right) \\ \frac{1}{\sigma^4}\mathbf{x}_t^\top\left(\mathbb{E}_{\boldsymbol{\theta}}\left(\widetilde{R}_{it}\right) - \mathbf{x}_t^\top\boldsymbol{\beta}\right) & -\frac{1}{2\sigma^4} + \frac{1}{\sigma^6}\mathbb{E}_{\boldsymbol{\theta}}\left(\left(\widetilde{R}_{it} - \mathbf{x}_t^\top\boldsymbol{\beta}\right)^2\right) \end{pmatrix}$$

Given that $\mathbb{E}_{\boldsymbol{\theta}}\left(\widetilde{R}_{it}\right) = \mathbf{x}_t^\top\boldsymbol{\beta}$ and $\mathbb{E}_{\boldsymbol{\theta}}\left(\left(\widetilde{R}_{it} - \mathbf{x}_t^\top\boldsymbol{\beta}\right)^2\right) = \sigma^2$, then we have:

$$I_t(\boldsymbol{\theta}) = \begin{pmatrix} \frac{1}{\sigma^2}\mathbf{x}_t\mathbf{x}_t^\top & \mathbf{0}_{2\times 1} \\ \mathbf{0}_{1\times 2} & \frac{1}{2\sigma^4} \end{pmatrix}$$

5. Score, Hessian and Fisher Information

Solution (cont'd)

First estimator:

$$I_t(\boldsymbol{\theta}) = \begin{pmatrix} \frac{1}{\sigma^2}\mathbf{x}_t\mathbf{x}_t^\top & \mathbf{0}_{2\times 1} \\ \mathbf{0}_{1\times 2} & \frac{1}{2\sigma^4} \end{pmatrix}$$

An estimator of the asymptotic variance covariance matrix of $\widehat{\boldsymbol{\theta}}$ is given by:

$$\widehat{\mathbb{V}}_{asy}\left(\widehat{\boldsymbol{\theta}}\right) = \frac{1}{T}\widehat{I}^{-1}\left(\widehat{\boldsymbol{\theta}}\right)$$

$$\widehat{I}\left(\widehat{\boldsymbol{\theta}}\right) = \frac{1}{T}\sum_{t=1}^{T} I_t\left(\widehat{\boldsymbol{\theta}}\right) = \begin{pmatrix} \frac{1}{T\widehat{\sigma}^2}\sum_{t=1}^{T}\mathbf{x}_t\mathbf{x}_t^\top & \mathbf{0}_{2\times 1} \\ \mathbf{0}_{1\times 2} & \frac{1}{2\widehat{\sigma}^4} \end{pmatrix} \quad \square$$

Solution (cont'd)

Second definition (BHHH):

$$\widehat{\mathbb{V}}_{asy}\left(\widehat{\boldsymbol{\theta}}\right) = \frac{1}{T}\widehat{I}^{-1}\left(\widehat{\boldsymbol{\theta}}\right)$$

$$\widehat{I}\left(\widehat{\boldsymbol{\theta}}\right) = \frac{1}{T}\sum_{t=1}^{T}\left(\left.\frac{\partial \ell_t\left(\boldsymbol{\theta}; \widetilde{r}_{it} \middle| x_t\right)}{\partial \boldsymbol{\theta}}\right|_{\widehat{\theta}} \times \left.\frac{\partial \ell_t\left(\boldsymbol{\theta}; \widetilde{r}_{it} \middle| x_t\right)}{\partial \boldsymbol{\theta}}\right|_{\widehat{\theta}}^{\top}\right)$$

with

$$\left.\frac{\partial \ell_t\left(\boldsymbol{\theta}; \widetilde{r}_{it} \middle| x_t\right)}{\partial \boldsymbol{\theta}}\right|_{\widehat{\theta}} = \begin{pmatrix} \frac{1}{\widehat{\sigma}^2}\mathbf{x}_t\left(\widetilde{r}_{it} - \mathbf{x}_t^{\top}\widehat{\boldsymbol{\beta}}\right) \\ -\frac{1}{2\widehat{\sigma}^2} + \frac{1}{2\widehat{\sigma}^4}\left(\widetilde{r}_{it} - \mathbf{x}_t^{\top}\widehat{\boldsymbol{\beta}}\right)^2 \end{pmatrix} = \begin{pmatrix} \frac{1}{\widehat{\sigma}^2}\mathbf{x}_t\widehat{\varepsilon}_t \\ -\frac{1}{2\widehat{\sigma}^2} + \frac{1}{2\widehat{\sigma}^4}\widehat{\varepsilon}_t^2 \end{pmatrix}$$

Solution (cont'd)

Second definition (BHHH):

$$\left.\frac{\partial \ell_t\left(\boldsymbol{\theta};\left.\widetilde{r}_{it}\right|x_t\right)}{\partial \boldsymbol{\theta}}\right|_{\widehat{\theta}} \times \left.\frac{\partial \ell_t\left(\boldsymbol{\theta};\left.\widetilde{r}_{it}\right|x_t\right)}{\partial \boldsymbol{\theta}}\right|_{\widehat{\theta}}^{\top}$$

$$= \begin{pmatrix} \frac{1}{\widehat{\sigma}^2}\mathbf{x}_t\widehat{\varepsilon}_t \\ -\frac{1}{2\widehat{\sigma}^2}+\frac{1}{2\widehat{\sigma}^4}\widehat{\varepsilon}_t^2 \end{pmatrix} \times \begin{pmatrix} \frac{1}{\widehat{\sigma}^2}\mathbf{x}_t^{\top}\widehat{\varepsilon}_t & -\frac{1}{2\widehat{\sigma}^2}+\frac{1}{2\widehat{\sigma}^4}\widehat{\varepsilon}_t^2 \end{pmatrix}$$

$$= \begin{pmatrix} \frac{1}{\widehat{\sigma}^4}\mathbf{x}_t\mathbf{x}_t^{\top}\widehat{\varepsilon}_t^2 & \frac{1}{\widehat{\sigma}^2}\mathbf{x}_t\widehat{\varepsilon}_t\left(-\frac{1}{2\widehat{\sigma}^2}+\frac{1}{2\widehat{\sigma}^4}\widehat{\varepsilon}_t^2\right) \\ \frac{1}{\widehat{\sigma}^2}\mathbf{x}_t^{\top}\widehat{\varepsilon}_t\left(-\frac{1}{2\widehat{\sigma}^2}+\frac{1}{2\widehat{\sigma}^4}\widehat{\varepsilon}_t^2\right) & \left(-\frac{1}{2\widehat{\sigma}^2}+\frac{1}{2\widehat{\sigma}^4}\widehat{\varepsilon}_t^2\right)^2 \end{pmatrix}$$

Solution (cont'd)

Second definition (BHHH): so we have

$$\widehat{\mathbb{V}}_{asy}\left(\widehat{\theta}\right)=\frac{1}{T}\widehat{I}^{-1}\left(\widehat{\theta}\right)$$

with

$$\widehat{I}\left(\widehat{\theta}\right)=\frac{1}{T}\sum_{t=1}^{T}\begin{pmatrix}\frac{1}{\widehat{\sigma}^4}\mathbf{x}_t\mathbf{x}_t^{\top}\widehat{\varepsilon}_t^2 & \frac{1}{\widehat{\sigma}^2}\mathbf{x}_t\widehat{\varepsilon}_t\left(-\frac{1}{2\widehat{\sigma}^2}+\frac{1}{2\widehat{\sigma}^4}\widehat{\varepsilon}_t^2\right)\\ \frac{1}{\widehat{\sigma}^2}\mathbf{x}_t^{\top}\widehat{\varepsilon}_t\left(-\frac{1}{2\widehat{\sigma}^2}+\frac{1}{2\widehat{\sigma}^4}\widehat{\varepsilon}_t^2\right) & \left(-\frac{1}{2\widehat{\sigma}^2}+\frac{1}{2\widehat{\sigma}^4}\widehat{\varepsilon}_t^2\right)^2\end{pmatrix}$$

Solution (cont'd)

Third definition (inverse of the Hessian): we know that

$$\widehat{\mathbb{V}}_{asy}\left(\widehat{\boldsymbol{\theta}}\right) = \frac{1}{T}\widehat{I}^{-1}\left(\widehat{\boldsymbol{\theta}}\right)$$

$$\widehat{I}\left(\widehat{\theta}\right) = \frac{1}{T}\sum_{t=1}^{T}\left(-H_t\left(\widehat{\boldsymbol{\theta}}; \widetilde{r}_{it} \middle| x_t\right)\right)$$

$$H_t\left(\widehat{\boldsymbol{\theta}}; \widetilde{r}_{it} \middle| x_t\right) = \begin{pmatrix} -\frac{1}{\widehat{\sigma}^2}\mathbf{x}_t\mathbf{x}_t^\top & -\frac{1}{\widehat{\sigma}^4}\mathbf{x}_t\left(\widetilde{r}_{it} - \mathbf{x}_t^\top\widehat{\boldsymbol{\beta}}\right) \\ -\frac{1}{\widehat{\sigma}^4}\mathbf{x}_t^\top\left(\widetilde{r}_{it} - \mathbf{x}_t^\top\widehat{\boldsymbol{\beta}}\right) & \frac{1}{2\widehat{\sigma}^4} - \frac{1}{\widehat{\sigma}^6}\left(\widetilde{r}_{it} - \mathbf{x}_t^\top\widehat{\boldsymbol{\beta}}\right)^2 \end{pmatrix}$$

Solution (cont'd)

Third definition (inverse of the Hessian):

$$H_t\left(\widehat{\boldsymbol{\theta}}; \widetilde{r}_{it} \middle| x_t\right) = \begin{pmatrix} -\frac{1}{\widehat{\sigma}^2}\mathbf{x}_t\mathbf{x}_t^\top & -\frac{1}{\widehat{\sigma}^4}\mathbf{x}_t\left(\widetilde{r}_{it} - \mathbf{x}_t^\top\widehat{\boldsymbol{\beta}}\right) \\ -\frac{1}{\widehat{\sigma}^4}\mathbf{x}_t^\top\left(\widetilde{r}_{it} - \mathbf{x}_t^\top\widehat{\boldsymbol{\beta}}\right) & \frac{1}{2\widehat{\sigma}^4} - \frac{1}{\widehat{\sigma}^6}\left(\widetilde{r}_{it} - \mathbf{x}_t^\top\widehat{\boldsymbol{\beta}}\right)^2 \end{pmatrix}$$

Given the FOC (log-likelihood equations), $\sum_{t=1}^{T}\mathbf{x}_t\left(\widetilde{r}_{it} - \mathbf{x}_t^\top\widehat{\boldsymbol{\beta}}\right) = \mathbf{0}$ and $\left(\widetilde{r}_{it} - \mathbf{x}_t^\top\widehat{\boldsymbol{\beta}}\right)^2 = T\widehat{\sigma}^2$.

$$\sum_{t=1}^{T} H_t\left(\widehat{\boldsymbol{\theta}}; \widetilde{r}_{it} \middle| x_t\right) = \begin{pmatrix} -\frac{1}{\widehat{\sigma}^2}\sum_{t=1}^{T}\mathbf{x}_t\mathbf{x}_t^\top & \mathbf{0}_{2\times 1} \\ \mathbf{0}_{1\times 2} & -\frac{T}{2\widehat{\sigma}^4} \end{pmatrix}$$

Solution (cont'd)

Third definition (inverse of the Hessian):

So, in this case, the third estimator of $\widehat{I}\left(\widehat{\theta}\right)$ coïncides with the first one:

$$\widehat{\mathbb{V}}_{asy}\left(\widehat{\boldsymbol{\theta}}\right)=\frac{1}{T}\widehat{I}^{-1}\left(\widehat{\boldsymbol{\theta}}\right)$$

$$\widehat{I}\left(\widehat{\theta}\right)=\frac{1}{T}\sum_{t=1}^{T}\left(-H_t\left(\widehat{\boldsymbol{\theta}};\widetilde{r}_{it}\middle|\,x_t\right)\right)=\begin{pmatrix}-\frac{1}{T\widehat{\sigma}^2}\sum_{t=1}^{T}\mathbf{x}_t\mathbf{x}_t^\top & \mathbf{0}_{2\times 1}\\ \mathbf{0}_{1\times 2} & -\frac{1}{2\widehat{\sigma}^4}\end{pmatrix}$$

Solution (cont'd)

These three estimates of the asymptotic variance covariance matrix are asymptotically equivalent, but can be largely different in finite sample...

$$\widehat{\mathbb{V}}_{asy}\left(\widehat{\boldsymbol{\theta}}\right) = \frac{1}{T}\widehat{I}^{-1}\left(\widehat{\boldsymbol{\theta}}\right)$$

with

$$\widehat{I}\left(\widehat{\boldsymbol{\theta}}\right) = \frac{1}{T}\sum_{t=1}^{T} I_t\left(\widehat{\boldsymbol{\theta}}\right)$$

$$\widehat{I}\left(\widehat{\boldsymbol{\theta}}\right) = \frac{1}{T}\sum_{t=1}^{T}\left(\left.\frac{\partial \ell_t\left(\boldsymbol{\theta}; \widetilde{r}_{it} \mid x_t\right)}{\partial \boldsymbol{\theta}}\right|_{\widehat{\theta}} \times \left.\frac{\partial \ell_t\left(\boldsymbol{\theta}; \widetilde{r}_{it} \mid x_t\right)}{\partial \boldsymbol{\theta}}\right|_{\widehat{\theta}}^{\top}\right)$$

$$\widehat{I}\left(\widehat{\theta}\right) = \frac{1}{T}\sum_{t=1}^{T}\left(-H_t\left(\boldsymbol{\theta}; \widetilde{r}_{it} \mid x_t\right)\right)$$

5. Score, Hessian and Fisher Information

```
%=====================================================
% PURPOSE: Chapter 2 - Asymptotic Variance Covariance Matrices
% Lecture: "Advanced Econometrics", HEC Lausanne
%-------------------------
% Author: Christophe Hurlin, University of Orleans
% Version: v1. November 2013
%=====================================================

clear all ; clc ; close all

data=xlsread('capm.xls');
r_tbill=data(2:end,9);                  % Return on the Tbill
r_msft=data(2:end,10);                  % Return on MSFT
r_sp500=data(2:end,11);                 % Return on the SP500
y=r_msft-r_tbill;                       % Excess return on MSFT
x=r_sp500-r_tbill;                      % Excess return on MSFT
T=length(y);                            % Sample size
X=[ones(T,1) x];                        % Matrix X (explicative variables)

beta=X\y;                               % Beta MLE-OLS estimates
s2=sum((y-X*beta).^2)/T;                % MLE Estimate of sigma^2
eps=y-X*beta;                           % Residuals
disp('Estimated Beta')
disp(beta)
```

5. Score, Hessian and Fisher Information

```
% First Estimator of the asymptotic variance covariance matrix
I1=[ (X'*X)/(T*s2)  zeros(2,1) ; zeros(1,2) 1/(2*s2^2)];
Vasy1=(1/T)*inv(I1);
std1=sqrt(diag(Vasy1));     % Standard errors

% Second Estimator of the asymptotic variance covariance matrix
grad=[X.*repmat(eps,1,2)./s2 (-1/(2*s2)+1/(2*s2^2)*eps.^2)];
I2=grad'*grad/T;
Vasy2=(1/T)*inv(I2);
std2=sqrt(diag(Vasy2));    % Standard errors

% Third Estimator of the asymptotic variance covariance matrix
Hessian=(1/T)*[-(X'*X)/s2 -sum((1/(s2^2))*X'*repmat(eps,1,2))'...
    ; -sum((1/(s2^2))*X'*repmat(eps,1,2)) T/(2*(s2^2))-sum(eps.^2)/(s2^3)];
I3=-Hessian;
Vasy3=(1/T)*inv(I3);
std3=sqrt(diag(Vasy3));    % Standard errors
```

5. Score, Hessian and Fisher Information

```
Estimated Beta
   0.000274089254513
   1.125056007502154


std1 =

   0.000178737001840
   0.025360248763542
   0.000002191650383


std2 =

   0.000180848106625
   0.022131118141298
   0.000001094385739


std3 =

   0.000178737001840
   0.025360248763542
   0.000002191650383
```

Key Concepts

1. Gradient and Hessian Matrix (deterministic elements).
2. Score Vector (random elements).
3. Hessian Matrix (random elements).
4. Fisher information matrix associated to the sample.
5. (Average) Fisher information matrix for one observation.

Section 6

Properties of Maximum Likelihood Estimators

6. Properties of Maximum Likelihood Estimators

Objectives

- MLE is a good estimator? Under which conditions the MLE is unbiased, consistent and corresponds to the BUE (Best Unbiased Estimator)? => **regularity conditions**
- Is the MLE **consistent**?
- Is the MLE **optimal** or **efficient**?
- What is the asymptotic distribution of the MLE? The **magic** of the MLE...

6. Properties of Maximum Likelihood Estimators

Definition (Regularity conditions)

Greene (2007) identify three regularity conditions

R1 The first three derivatives of $\ln f_X(\theta; x_i)$ with respect to θ are **continuous** and **finite** for almost all x_i and for all θ. This condition ensures the existence of a certain Taylor series approximation and the finite variance of the derivatives of $\ell_i(\theta; x_i)$.

R2 The conditions necessary to obtain the **expectations** of the first and second derivatives of $\ln f_X(\theta; X_i)$ are met.

R3 For all values of θ, $\left|\partial^3 \ln f_X(\theta; x_i) / \partial\theta_i \partial\theta_j \partial\theta_k\right|$ is less than a function that has a finite expectation. This condition will allow us to truncate the Taylor series.

6. Properties of Maximum Likelihood Estimators

Definition (Regularity conditions, Zivot 2001)

A pdf $f_X(\theta; x)$ is regular if and only of:

R1 The support of the random variables X, $SX = \{x : f_X(\theta; x) > 0\}$, does not depend on θ.

R2 $f_X(\theta; x)$ is at least three times differentiable with respect to θ, and these derivatives are continuous.

R3 The true value of θ lies in a compact set Θ.

6. Properties of Maximum Likelihood Estimators

Under these regularity conditions, the maximum likelihood estimator $\widehat{\theta}$ possesses many appealing properties:

1. The maximum likelihood estimator is **consistent**.
2. The maximum likelihood estimator is **asymptotically normal** (the magic of the MLE..).
3. The maximum likelihood estimator is **asymptotically optimal** or efficient.
4. The maximum likelihood estimator is **equivariant**: if $\widehat{\theta}$ is an estimator of θ then $g(\widehat{\theta})$ is an estimator of $g(\theta)$.

6. Properties of Maximum Likelihood Estimators

Theorem (Consistency)

Under regularity conditions, the maximum likelihood estimator is **consistent**

$$\widehat{\theta} \underset{N \to \infty}{\overset{p}{\longrightarrow}} \theta_0$$

or equivalently:

$$\underset{N \to \infty}{p \lim} \widehat{\theta} = \theta_0$$

where θ_0 denotes the true value of the parameter θ.

6. Properties of Maximum Likelihood Estimators

Sketch of the proof (Greene, 2007)

Because $\widehat{\theta}$ is the MLE, in any finite sample, for any $\theta \neq \widehat{\theta}$ (including the true θ_0) it must be true that

$$\ln L_N\left(\widehat{\theta};\, y \middle|\, x\right) \geq \ln L_N\left(\theta;\, y \middle|\, x\right)$$

Consider, then, the random variable $L_N\left(\theta;\, Y \middle|\, x\right) / L_N\left(\theta_0;\, Y \middle|\, x\right)$. Because the log function is strictly concave, from Jensen's Inequality, we have

$$\mathbb{E}_\theta\left(\ln\left(\frac{L_N\left(\theta;\, Y \middle|\, x\right)}{L_N\left(\theta_0;\, Y \middle|\, x\right)}\right)\right) \leq \ln\left(\mathbb{E}_\theta\left(\frac{L_N\left(\theta;\, Y \middle|\, x\right)}{L_N\left(\theta_0;\, Y \middle|\, x\right)}\right)\right)$$

Sketch of the proof, cont'd

The expectation on the right-hand side is exactly equal to one, as

$$\begin{aligned}\mathbb{E}_{\theta}\left(\frac{L_N(\theta;\, Y|\, x)}{L_N(\theta_0;\, Y|\, x)}\right) &= \int \left(\frac{L_N(\theta;\, y|\, x)}{L_N(\theta_0;\, y|\, x)}\right) L_N(\theta_0;\, y|\, x)\, dy \\ &= \int L_N(\theta;\, y|\, x)\, dy \\ &= 1\end{aligned}$$

is simply the integral of a joint density.

Sketch of the proof, cont'd

So we have

$$\mathbb{E}_{\theta}\left(\ln\left(\frac{L_N\left(\theta;\, Y\middle|\, x\right)}{L_N\left(\theta_0;\, Y\middle|\, x\right)}\right)\right) \leq \ln\left(\mathbb{E}_{\theta}\left(\frac{L_N\left(\theta;\, Y\middle|\, x\right)}{L_N\left(\theta_0;\, Y\middle|\, x\right)}\right)\right) = \ln\left(1\right) = 0$$

Divide the left hand side of this equation by N to produce

$$\mathbb{E}_{\theta}\left(\frac{1}{N}\ln L_N\left(\theta;\, Y\middle|\, x\right)\right) \leq \mathbb{E}_{\theta}\left(\frac{1}{N}\ln L_N\left(\theta_0;\, Y\middle|\, x\right)\right)$$

This produces a central result:

6. Properties of Maximum Likelihood Estimators

Theorem (Likelihood Inequality)

The expected value of the log-likelihood is maximized at the true value of the parameters. For any θ, including $\widehat{\theta}$:

$$\mathbb{E}_{\theta}\left(\frac{1}{N}\ell_N\left(\theta_0;\ Y_i\right|x_i)\right) \geq \mathbb{E}_{\theta}\left(\frac{1}{N}\ell_N\left(\theta;\ Y_i\right|x_i)\right)$$

6. Properties of Maximum Likelihood Estimators

Sketch of the proof, cont'd

Notice that

$$\frac{1}{N}\ell_N\left(\theta;\, Y_i\middle|\, x_i\right) = \frac{1}{N}\sum_{i=1}^{N}\ell_i\left(\theta;\, Y_i\middle|\, x_i\right)$$

where the elements $\ell_i\left(\theta;\, Y_i\middle|\, x_i\right)$ for $i = 1, ..N$ are $i.i.d.$. So, using a law of large numbers, we get:

$$\frac{1}{N}\ell_N\left(\theta;\, Y_i\middle|\, x_i\right) \underset{N\rightarrow\infty}{\overset{p}{\longrightarrow}} \mathbb{E}_\theta\left(\frac{1}{N}\ell_N\left(\theta;\, Y_i\middle|\, x_i\right)\right)$$

Sketch of the proof, cont'd

The Likelihood inequality for $\theta = \widehat{\theta}$ implies

$$\mathbb{E}_{\theta}\left(\frac{1}{N}\ell_N\left(\theta_0;\, Y_i \mid x_i\right)\right) \geq \mathbb{E}_{\theta}\left(\frac{1}{N}\ell_N\left(\widehat{\theta};\, Y_i \mid x_i\right)\right)$$

with

$$\frac{1}{N}\ell_N\left(\theta_0;\, Y_i \mid x_i\right) \underset{N\rightarrow\infty}{\overset{p}{\longrightarrow}} \mathbb{E}_{\theta}\left(\frac{1}{N}\ell_N\left(\theta_0;\, Y_i \mid x_i\right)\right)$$

$$\frac{1}{N}\ell_N\left(\widehat{\theta};\, Y_i \mid x_i\right) \underset{N\rightarrow\infty}{\overset{p}{\longrightarrow}} \mathbb{E}_{\theta}\left(\frac{1}{N}\ell_N\left(\widehat{\theta};\, Y_i \mid x_i\right)\right)$$

and thus

$$\lim_{N\rightarrow\infty} \Pr\left(\frac{1}{N}\ell_N\left(\theta_0;\, Y_i \mid x_i\right) \geq \frac{1}{N}\ell_N\left(\widehat{\theta};\, Y_i \mid x_i\right)\right) = 1$$

6. Properties of Maximum Likelihood Estimators

Sketch of the proof, cont'd So we have two results:

$$\lim_{N\rightarrow\infty} \Pr\left(\frac{1}{N}\ell_N\left(\theta_0;\, Y_i|\, x_i\right) \geq \frac{1}{N}\ell_N\left(\widehat{\theta};\, Y_i|\, x_i\right)\right) = 1$$

$$\frac{1}{N}\ell_N\left(\widehat{\theta};\, Y_i|\, x_i\right) \geq \frac{1}{N}\ell_N\left(\theta_0;\, Y_i|\, x_i\right) \qquad \forall N$$

It necessarily implies that

$$\frac{1}{N}\ell_N\left(\widehat{\theta};\, Y_i|\, x_i\right) \underset{N\rightarrow\infty}{\overset{p}{\longrightarrow}} \frac{1}{N}\ell_N\left(\theta_0;\, Y_i|\, x_i\right)$$

If θ is a scalar, we have immediatly:

$$\widehat{\theta} \underset{N\rightarrow\infty}{\overset{p}{\longrightarrow}} \theta_0$$

For a more general case with $\dim\left(\theta\right) = K$, see a formal proof in Amemiya (1985).

Amemiya T., (1985) Advanced Econometrics. Harvard University Press

Remark

The proof of the consistency of the MLE is largely easiest when we have a formal expression for the maximum likelihood estimator $\widehat{\theta}$

$$\widehat{\theta} = \widehat{\theta}\left(X_1, .., X_N\right)$$

6. Properties of Maximum Likelihood Estimators

Example

Suppose that $D_1, D_2, .., D_N$ are $i.i.d.$, positive random variable with $D_i \sim Exp\left(\theta_0\right)$, with

$$f_D\left(d;\theta\right) = \frac{1}{\theta}\exp\left(-\frac{d}{\theta}\right),\ \forall d \in \mathbb{R}^+$$

$$\mathbb{E}_\theta\left(D_i\right) = \theta_0 \qquad \mathbb{V}_\theta\left(D_i\right) = \theta_0^2$$

where θ_0 is the true value of θ. **Question:** show that the MLE is consistent.

Solution

The log-likelihood function associated to the sample $\{d_1, .., d_N\}$ is defined by:

$$\ell_N(\theta; d) = -N \ln(\theta) - \frac{1}{\theta} \sum_{i=1}^{N} d_i$$

We admit that maximum likelihood estimator corresponds to the sample mean:

$$\widehat{\theta} = \frac{1}{N} \sum_{i=1}^{N} D_i$$

6. Properties of Maximum Likelihood Estimators

Solution, cont'd

Then, we have:

$$\mathbb{E}_\theta\left(\widehat{\theta}\right) = \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_\theta\left(D_i\right) = \theta \qquad \widehat{\theta} \text{ is unbiased}$$

$$\mathbb{V}_\theta\left(\widehat{\theta}\right) = \frac{1}{N^2}\sum_{i=1}^{N}\mathbb{V}_\theta\left(D_i\right) = \frac{\theta^2}{N}$$

As a consequence

$$\mathbb{E}_\theta\left(\widehat{\theta}\right) = \theta \qquad \lim_{N\to\infty}\mathbb{V}_\theta\left(\widehat{\theta}\right) = 0$$

and

$$\widehat{\theta} \underset{N\to\infty}{\overset{p}{\longrightarrow}} \theta$$

6. Properties of Maximum Likelihood Estimators

Lemma

Under stronger conditions, the maximum likelihood estimator converges almost surely to θ_0

$$\widehat{\theta} \underset{N \to \infty}{\overset{a.s.}{\longrightarrow}} \theta_0 \implies \widehat{\theta} \underset{N \to \infty}{\overset{p}{\longrightarrow}} \theta_0$$

6. Properties of Maximum Likelihood Estimators

1. If we restrict ourselves to the class of unbiased estimators (linear and nonlinear) then we define the **best estimator** as the one with the **smallest variance**.

2. With linear estimators (next chapter), the Gauss-Markov theorem tells us that the ordinary least squares (OLS) estimator is best (BLUE).

3. When we expand the class of estimators to include linear and nonlinear estimators it turns out that we can establish an absolute lower bound on the variance of any unbiased estimator $\widehat{\theta}$ of θ under certain conditions.

4. Then if an **unbiased estimator** $\widehat{\theta}$ has a variance that is equal to the lower bound then we have found the **best unbiased estimator (BUE)**.

6. Properties of Maximum Likelihood Estimators

Definition (Cramer-Rao or FDCR bound)

Let $X_1, .., X_N$ be an $i.i.d.$ sample with pdf $f_X(\theta; x)$. Let $\widehat{\theta}$ be an unbiased estimator of θ; i.e., $\mathbb{E}_\theta(\widehat{\theta}) = \theta$. If $f_X(\theta; x)$ is regular then

$$\mathbb{V}_\theta\left(\widehat{\theta}\right) \geq I_N^{-1}(\theta_0) \qquad \text{FDCR or Cramer-Rao bound}$$

where $I_N(\theta_0)$ denotes the Fisher information number for the **sample** evaluated at the true value θ_0.

6. Properties of Maximum Likelihood Estimators

Remarks

1. Hence, the Cramer-Rao Bound is the inverse of the information matrix associated to the sample. Reminder: three definitions for $I_N(\theta_0)$.

$$I_N(\theta_0) = \mathbb{V}_\theta \left(\left. \frac{\partial \ell_N(\theta; Y|x)}{\partial \theta} \right|_{\theta_0} \right)$$

$$I_N(\theta_0) = \mathbb{E}_\theta \left(\left. \frac{\partial \ell_N(\theta; Y|x)}{\partial \theta} \right|_{\theta_0} \left. \frac{\partial \ell_N(\theta; Y|x)^\top}{\partial \theta} \right|_{\theta_0} \right)$$

$$I_N(\theta_0) = \mathbb{E}_\theta \left(- \left. \frac{\partial^2 \ell_N(\theta; Y|x)}{\partial \theta \partial \theta^\top} \right|_{\theta_0} \right)$$

2. If θ is a vector then $\mathbb{V}_\theta\left(\widehat{\theta}\right) \geq I_N^{-1}(\theta_0)$ means that $\mathbb{V}_\theta\left(\widehat{\theta}\right) - I_N^{-1}(\theta_0)$ is positive semi-definite

6. Properties of Maximum Likelihood Estimators

Theorem (Efficiency)

Under regularity conditions, the maximum likelihood estimator is **asymptotically efficient** *and attains the* **FDCR** *(Frechet - Darnois - Cramer - Rao) or* **Cramer-Rao bound***:*

$$\mathbb{V}_{\theta}\left(\widehat{\theta}\right) = I_N^{-1}\left(\theta_0\right)$$

where $I_N\left(\theta_0\right)$ denotes the Fisher information matrix associated to the **sample** *evaluated at the true value θ_0.*

Example (Exponential Distribution)

Suppose that $D_1, D_2, .., D_N$ are $i.i.d.$, positive random variable with $D_i \sim Exp(\theta_0)$, with

$$f_D(d;\theta) = \frac{1}{\theta}\exp\left(-\frac{d}{\theta}\right), \ \forall d \in \mathbb{R}^+$$

$$\mathbb{E}_\theta(D_i) = \theta_0 \qquad \mathbb{V}_\theta(D_i) = \theta_0^2$$

where θ_0 is the true value of θ. **Question:** show that the MLE is efficient.

Solution

We shown that the maximum likelihood estimator corresponds to the sample mean,

$$\widehat{\theta} = \frac{1}{N} \sum_{i=1}^{N} D_i$$

$$\mathbb{V}_{\theta}\left(\widehat{\theta}\right) = \frac{\theta_0^2}{N}$$

$$\mathbb{E}_{\theta}\left(\widehat{\theta}\right) = \theta_0$$

Solution, cont'd

The log-likelihood function is

$$\ell_N(\theta; d) = -N \ln(\theta) - \frac{1}{\theta} \sum_{i=1}^{N} d_i$$

The score vector is defined by:

$$s_N(\theta; D) = \frac{\partial \ell_N(\theta; D)}{\partial \theta} = -\frac{N}{\theta} + \frac{1}{\theta^2} \sum_{i=1}^{N} D_i$$

6. Properties of Maximum Likelihood Estimators

Solution, cont'd

Let us use one of the three definitions of the information quantity $I_N(\theta)$:

$$\begin{aligned} I_N(\theta) &= \mathbb{V}_\theta\left(\frac{\partial \ell_N(\theta; D)}{\partial \theta}\right) \\ &= \mathbb{V}_\theta\left(-\frac{N}{\theta} + \frac{1}{\theta^2}\sum_{i=1}^N D_i\right) \\ &= \frac{1}{\theta^4}\textstyle\sum_{i=1}^N \mathbb{V}_\theta(D_i) \\ &= \frac{N\theta^2}{\theta^4} = \frac{N}{\theta^2} \end{aligned}$$

Then, $\widehat{\theta}$ is efficient and attains the Cramer-Rao bound.

$$\mathbb{V}_\theta\left(\widehat{\theta}\right) = I_N^{-1}(\theta_0) = \frac{\theta^2}{N} \square$$

6. Properties of Maximum Likelihood Estimators

Theorem (Convergence of the MLE)

Under suitable regularity conditions, the MLE is **asymptotically normally distributed** *with*

$$\sqrt{N}\left(\widehat{\theta}-\theta_0\right)\xrightarrow{d}\mathcal{N}\left(0,I^{-1}\left(\theta_0\right)\right)$$

where θ_0 denotes the true value of the parameter and $I(\theta_0)$ the (average) Fisher information matrix for one observation.

6. Properties of Maximum Likelihood Estimators

Corollary

Another way, to write this result, is to say that for large sample size N, the MLE $\widehat{\theta}$ is approximatively distributed according a normal distribution

$$\widehat{\theta} \overset{asy}{\approx} \mathcal{N}\left(\theta_0, N^{-1}\ I^{-1}\left(\theta_0\right)\right)$$

or equivalently

$$\widehat{\theta} \overset{asy}{\approx} \mathcal{N}\left(\theta_0,\ I_N^{-1}\left(\theta_0\right)\right)$$

where $I_N\left(\theta_0\right) = N \times I(\theta_0)$ denotes the Fisher information matrix associated to the sample.

6. Properties of Maximum Likelihood Estimators

Definition (Asymptotic Variance)

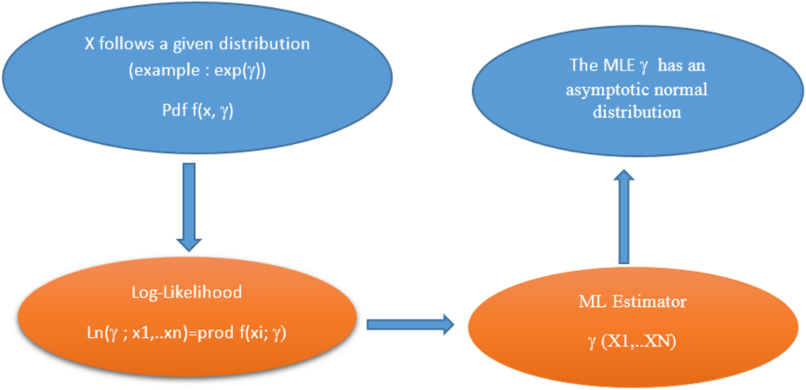
The asymptotic variance of the MLE is defined by:

$$\mathbb{V}_{asy}\left(\widehat{\theta}\right) = I_N^{-1}\left(\theta_0\right)$$

where $I_N\left(\theta_0\right)$ denotes the Fisher information matrix associated to the sample. This asymptotic variance of the MLE corresponds to the Cramer-Rao or FDCR bound.

6. Properties of Maximum Likelihood Estimators

The magic of the MLE



Proof (MLE convergence)

At the maximum likelihood estimator, the gradient of the log-likelihood equals zero (FOC):

$$\underset{(K,1)}{g_N\left(\widehat{\theta}\right)} \equiv g_N\left(\widehat{\theta}; y|\, x\right) = \left.\frac{\partial \ell_N\left(\theta; y|\, x\right)}{\partial \theta}\right|_{\widehat{\theta}} = 0_K$$

where $\widehat{\theta} = \widehat{\theta}\left(x\right)$ denotes here the ML **estimate**. Expand this set of equations in a Taylor series around the true parameters θ_0. We will use the **mean value theorem** to truncate the Taylor series at the second term:

$$g_N\left(\widehat{\theta}\right) = g_N\left(\theta_0\right) + H_N\left(\overline{\theta}\right)\left(\widehat{\theta} - \theta_0\right) = 0$$

The Hessian is evaluated at a point $\overline{\theta}$ that is between $\widehat{\theta}$ and θ_0, for instance $\overline{\theta} = \omega\widehat{\theta} + \left(1-\omega\right)\theta_0$ for some $0 < \omega < 1$.

Proof (MLE convergence, cont'd)

We then rearrange this equation and multiply the result by $\sqrt{N}$ to obtain:

$$\sqrt{N}\left(\widehat{\theta}-\theta_0\right)=\left(-H_N\left(\overline{\theta}\right)\right)^{-1}\left(\sqrt{N}g_N\left(\theta_0\right)\right)$$

By dividing $H_N\left(\overline{\theta}\right)$ and $g_N\left(\theta_0\right)$ by N, we obtain:

$$\begin{aligned}\sqrt{N}\left(\widehat{\theta}-\theta_0\right) &= \left(-\frac{1}{N}H_N\left(\overline{\theta}\right)\right)^{-1}\left(\sqrt{N}\frac{1}{N}g_N\left(\theta_0\right)\right)\\ &= \left(-\frac{1}{N}H_N\left(\overline{\theta}\right)\right)^{-1}\left(\sqrt{N}\overline{g}\left(\theta_0\right)\right)\end{aligned}$$

where $\overline{g}\left(\theta_0\right)$ denotes the sample mean of the individual gradient vectors

$$\overline{g}\left(\theta_0\right)=\frac{1}{N}g_N\left(\theta_0\right)=\frac{1}{N}\sum_{i=1}^{N}g_i\left(\theta_0;\, y_i\middle|\, x_i\right)$$

Proof (MLE convergence, cont'd)

Let us now consider the same expression in terms of random variables: $\widehat{\theta}$ now denotes the ML estimator, $H_N\left(\overline{\theta}\right) = H_N\left(\overline{\theta};\, Y\middle|\, x\right)$ and $s_N\left(\theta_0;\, Y\middle|\, x\right)$ the score vector. We have:

$$\sqrt{N}\left(\widehat{\theta}-\theta_0\right) = \left(-\frac{1}{N}H_N\left(\overline{\theta};\, Y\middle|\, x\right)\right)^{-1}\left(\sqrt{N}\overline{s}\left(\theta_0;\, Y\middle|\, x\right)\right)$$

where the score vectors associated to the variables Y_i are $i.i.d$.

$$\overline{s}\left(\theta_0;\, Y\middle|\, x\right) = \frac{1}{N}\sum_{i=1}^{N} s_i\left(\theta_0;\, Y_i\middle|\, x_i\right)$$

6. Properties of Maximum Likelihood Estimators

Proof (MLE convergence, cont'd)

Let us consider the first element:

$$\overline{s}(\theta_0) = \frac{1}{N} \sum_{i=1}^{N} s_i(\theta_0; Y_i | x_i)$$

The individual scores $s_i(\theta_0; Y_i | x_i)$ are $i.i.d.$ with

$$\mathbb{E}_\theta(s_i(\theta_0; Y_i | x_i)) = 0$$

$$\mathbb{E}_x \mathbb{V}_\theta(s_i(\theta_0; Y_i | x_i)) = \mathbb{E}_x(I_i(\theta_0)) = I(\theta_0)$$

By using the **Lindberg-Levy Central Limit Theorem**, we have:

$$\sqrt{N}\overline{s}(\theta_0) \xrightarrow{d} \mathcal{N}(0, I(\theta_0))$$

6. Properties of Maximum Likelihood Estimators

Proof (MLE convergence, cont'd)

We known that:

$$-\frac{1}{N} H_N\left(\overline{\theta};\, Y \middle|\, x\right) = -\frac{1}{N} \sum_{i=1}^{N} H_i\left(\overline{\theta};\, Y_i \middle|\, x_i\right)$$

where the hessian matrices $H_i\left(\overline{\theta};\, Y_i \middle|\, x_i\right)$ are $i.i.d$. Besides, because $\text{plim}\left(\widehat{\theta} - \theta_0\right) = 0$, $\text{plim}\left(\overline{\theta} - \theta_0\right) = 0$ as well. By applying a law of large numbers, we get:

$$-\frac{1}{N} H_N\left(\overline{\theta};\, Y \middle|\, x\right) \xrightarrow{p} \mathbb{E}_X \mathbb{E}_\theta\left(-H_i\left(\theta_0;\, Y_i \middle|\, x_i\right)\right)$$

with

$$\mathbb{E}_X \mathbb{E}_\theta\left(-H_i\left(\theta_0;\, Y_i \middle|\, x_i\right)\right) = \mathbb{E}_X \mathbb{E}_\theta\left(-\frac{\partial^2 \ell_i\left(\theta;\, Y_i \middle|\, x_i\right)}{\partial\theta\partial\theta^\top}\right) = I\left(\theta_0\right)$$

6. Properties of Maximum Likelihood Estimators

Reminder:

If X_N and Y_N verify

$$\underset{(K,K)}{X_N} \xrightarrow{p} \underset{(K,K)}{X}$$

$$\underset{(K,1)}{Y_N} \xrightarrow{d} \mathcal{N}\left(\underset{(K,1)}{0}, \underset{(K,K)}{\Sigma}\right)$$

then

$$\underset{(K,K)}{X_N}\underset{(K,1)}{Y_N} \xrightarrow{d} \mathcal{N}\left(\underset{(K,1)}{0}, \underset{(K,K)}{X}\underset{(K,K)}{\Sigma}\underset{(K,K)}{X^{\top}}\right)$$

6. Properties of Maximum Likelihood Estimators

Proof (MLE convergence, cont'd)

Here we have

$$\sqrt{N}\left(\widehat{\theta}-\theta_0\right)=\left(-\frac{1}{N}H_N\left(\overline{\theta};\,Y|\,x\right)\right)^{-1}\left(\sqrt{N}\overline{s}\left(\theta_0;\,Y|\,x\right)\right)$$

$$\left(-\frac{1}{N}H_N\left(\overline{\theta};\,Y|\,x\right)\right)^{-1}\xrightarrow{p} I^{-1}\left(\theta_0\right)\quad\text{symmetric matrix}$$

$$\sqrt{N}\overline{s}\left(\theta_0\right)\xrightarrow{d}\mathcal{N}\left(0,I\left(\theta_0\right)\right)$$

Then, we get:

$$\sqrt{N}\left(\widehat{\theta}-\theta_0\right)\xrightarrow{d}\mathcal{N}\left(0,I^{-1}\left(\theta_0\right)I\left(\theta_0\right)I^{-1}\left(\theta_0\right)\right)$$

Proof (MLE convergence, cont'd)

And finally....

$$\sqrt{N}\left(\widehat{\theta}-\theta_0\right) \xrightarrow{d} \mathcal{N}\left(0, I^{-1}\left(\theta_0\right)\right)$$

The magic of the MLE.....

6. Properties of Maximum Likelihood Estimators

Example (Exponential Distribution)

Suppose that $D_1, D_2, .., D_N$ are *i.i.d.*, positive random variable with $D_i \sim Exp(\theta_0)$, with

$$f_D(d;\theta) = \frac{1}{\theta}\exp\left(-\frac{d}{\theta}\right), \ \forall d \in \mathbb{R}^+$$

$$\mathbb{E}_\theta(D_i) = \theta_0 \qquad \mathbb{V}_\theta(D_i) = \theta_0^2$$

where θ_0 is the true value of θ. **Question:** what is the asymptotic distribution of the MLE? Propose a consistent estimator of the asymptotic variance of $\widehat{\theta}$.

Solution

We shown that $\widehat{\theta} = (1/N) \sum_{i=1}^{N} D_i$ and:

$$s_i(\theta; D_i) = \frac{\partial \ell_i(\theta; D_i)}{\partial \theta} = -\frac{1}{\theta} + \frac{D_i}{\theta^2}$$

The (average) Fisher information matrix associated to D_i is:

$$I(\theta) = \mathbb{V}_\theta \left(-\frac{1}{\theta} + \frac{D_i}{\theta^2} \right) = \frac{1}{\theta^4} \mathbb{V}_\theta (D_i) = \frac{1}{\theta^2}$$

Then, the asymptotic distribution of $\widehat{\theta}$ is:

$$\sqrt{N} \left(\widehat{\theta} - \theta_0 \right) \xrightarrow{d} \mathcal{N}\left(0, \theta^2\right)$$

or equivalently

$$\widehat{\theta} \overset{asy}{\approx} \mathcal{N}\left(\theta_0, \frac{\theta^2}{N}\right)$$

Solution, cont'd

The asymptotic variance of $\widehat{\theta}$ is:

$$\mathbb{V}_{asy}\left(\widehat{\theta}\right) = \frac{\theta^2}{N}$$

A consistent estimator of $\mathbb{V}_{as}\left(\widehat{\theta}\right)$ is simply defined by:

$$\widehat{\mathbb{V}}_{asy}\left(\widehat{\theta}\right) = \frac{\widehat{\theta}^2}{N} \square$$

6. Properties of Maximum Likelihood Estimators

Example (Linear Regression Model)

Let us consider the previous linear regression model $y_i = x_i^\top \beta + \varepsilon_i$, with ε_i $\mathcal{N}.i.d.\left(0, \sigma^2\right)$. Let us denote θ the $K + 1 \times 1$ vector defined by $\theta = \left(\beta^\top \; \sigma^2\right)^\top$. The MLE estimator of θ is defined by:

$$\widehat{\theta} = \left(\begin{array}{c} \widehat{\beta} \\ \widehat{\sigma}^2 \end{array} \right)$$

$$\widehat{\beta} = \left(\sum_{i=1}^{N} X_i X_i^\top \right)^{-1} \left(\sum_{i=1}^{N} X_i^\top Y_i \right) \qquad \widehat{\sigma}^2 = \frac{1}{N} \sum_{i=1}^{N} \left(Y_i - X_i^\top \widehat{\beta} \right)^2$$

Question: what is the asymptotic distribution of $\widehat{\theta}$? Propose an estimator of the asymptotic variance.

Solution

This model satisfy the regularity conditions. We shown that the average Fisher information matrix is equal to:

$$I(\theta) = \begin{pmatrix} \frac{1}{\sigma^2}\mathbb{E}_X\left(X_i X_i^\top\right) & 0 \\ 0 & \frac{1}{2\sigma^4} \end{pmatrix}$$

From the MLE convergence theorem, we get immediately:

$$\sqrt{N}\left(\widehat{\theta} - \theta_0\right) \xrightarrow{d} \mathcal{N}\left(0, I^{-1}(\theta_0)\right)$$

where θ_0 is the true value of θ.

Solution, cont'd

The asymptotic variance covariance matrix of $\widehat{\theta}$ is equal to:

$$\mathbb{V}_{asy}\left(\widehat{\theta}\right) = N^{-1}\ I^{-1}\left(\theta_0\right) = I_N^{-1}\left(\theta_0\right)$$

with

$$I_N\left(\theta\right) = \begin{pmatrix} \frac{N}{\sigma^2}\mathbb{E}_X\left(X_i X_i^{\top}\right) & 0 \\ 0 & \frac{N}{2\sigma^4} \end{pmatrix}$$

Solution, cont'd

A consistent estimate of $I_N(\theta)$ is:

$$\widehat{I}_N(\theta) = \widehat{\mathbb{V}}_{asy}^{-1}\left(\widehat{\theta}\right) = \begin{pmatrix} \frac{N}{\widehat{\sigma}^2}\widehat{Q}_X & 0 \\ 0 & \frac{N}{2\widehat{\sigma}^4} \end{pmatrix}$$

with

$$\widehat{Q}_X = \frac{1}{N}\sum_{i=1}^{N} x_i x_i^\top$$

Solution, cont'd

Thus we get:

$$\widehat{\beta} \overset{asy}{\approx} \mathcal{N}\left(\beta_0,\ \widehat{\sigma}^2\left(\sum_{i=1}^{N} x_i x_i^\top\right)^{-1}\right)$$

$$\widehat{\sigma}^2 \overset{asy}{\approx} \mathcal{N}\left(\sigma_0^2,\ \frac{2\widehat{\sigma}^4}{N}\right)$$

Summary

Under regular conditions

1. The MLE is **consistent.**
2. The MLE is **asymptotically efficient** and its variance attains the FDCR or Cramer-Rao bound.
3. The MLE is **asymptotically normally distributed.**

6. Properties of Maximum Likelihood Estimators

But, **finite sample properties** can be very different from large sample properties:

1. The maximum likelihood estimator is consistent but can be severely biased in finite samples
2. The estimation of the variance-covariance matrix can be seriously doubtful in finite samples.

6. Properties of Maximum Likelihood Estimators

Theorem (Equivariance)

Under regular conditions and if $g(.)$ is a continuously differentiable function of θ and is defined from $\mathbb{R}^K$ to $\mathbb{R}^P$, then:

$$g\left(\widehat{\theta}\right) \xrightarrow{p} g(\theta_0)$$

$$\sqrt{N}\left(g\left(\widehat{\theta}\right) - g(\theta_0)\right) \xrightarrow{d} \mathcal{N}\left(0, G(\theta_0)\ I^{-1}(\theta_0)\ G(\theta_0)^\top\right)$$

where θ_0 is the true value of the parameters and the matrix $G(\theta_0)$ is defined by

$$\underset{(P,K)}{G(\theta)} = \frac{\partial g(\theta)}{\partial \theta^\top}$$

Key Concepts of the Chapter 2

1. Likelihood and log-likelihood function
2. Maximum likelihood estimator (MLE) and Maximum likelihood estimate
3. Gradient and Hessian Matrix (deterministic elements)
4. Score Vector and Hessian Matrix (random elements)
5. Fisher information matrix associated to the sample
6. (Average) Fisher information matrix for one observation
7. FDCR or Cramer Rao Bound: the notion of efficiency
8. Asymptotic distribution of the MLE
9. Asymptotic variance of the MLE
10. Estimator of the asymptotic variance of the MLE

End of Chapter 2

Christophe Hurlin (University of Orléans)