Cambridge University Press
978-0-521-87640-7 - Econometric Modeling and Inference
Jean-Pierre Florens, Velayoudom Marimoutou and Anne Peguin-Feissolle
Excerpt
More information

Part I

Statistical Methods

1. Statistical Models

1.1 Introduction

In this chapter, we review the usual statistical terminology that introduces the fundamental notions of sample, parameters, statistical model, and likelihood function. Our presentation avoids all technical developments of probability theory, which are not strictly necessary in this book. For example, σ-fields (or σ-algebras) are not introduced, nor are measurability conditions. The mathematical rigor of the exposition is necessarily weakened by this choice, but our aim is to focus the interest of the reader on purely statistical concepts.

It is expected that the reader knows the usual concepts of probability as well as the most common probability distributions and we refer to various reference books on this theme in the bibliographic notes.

At the end of this chapter, we emphasize conditional models, whose importance is fundamental in econometrics, and we introduce important concepts such as identification and exogeneity.

1.2 Sample, Parameters, and Sampling Probability Distributions

A *statistical model* is usually defined as a triplet consisting of a sample space, a parametric space and a family of sampling probability distributions.

We denote by x the *realization* of a sample. It is always assumed that x is equal to a finite sequence $(x_i)_{i=1,\ldots,n}$ where n is the sample size and x_i is the ith observation. We limit ourselves to the case where x_i is a vector of m real numbers (possibly integers) belonging to a subset X of $\mathbb{R}^m$. Hence, the *sample space* is $X^n \subset \mathbb{R}^{mn}$. The index i of the observations may have various meanings:

- i may index a set of individuals (households, firms, areas...) observed at a given instant. These data are referred to as *cross-sectional data*.
- i may describe a set of periods. Then, the observations x_i form a *time series* (multidimensional if $m > 1$).

- i may belong to a more complex set and be, for instance, equal to a pair (ℓ, t) where ℓ represents an individual and t an observation time. Then, the observations $x_i = x_{\ell t}$ are double indexed and are called *panel data* or longitudinal data.

As the sample space X^n is always assumed to belong to $\mathbb{R}^{nm}$, it is associated with a Borel field, on which some probability will be defined.

The *parameter space* is denoted as Θ and an element of this space is usually denoted θ. The parameters are unknown elements of the statistical problem, the observations provide information about these elements. Two kinds of statistical models can be defined depending on the dimension of Θ:

- If $\Theta \subset \mathbb{R}^k$ where k is a finite integer, the statistical model is said to be *parametric* or a model with vector parameters.
- If Θ is not finite dimensional but contains a function space, the model is said to be *nonparametric* or a model with functional parameters. In some examples, although Θ is infinite dimensional, there exists a function λ of θ which is finite dimensional. Then, the model is called *semiparametric*.

In the following, a parameter will be an element of Θ, whether the dimension of this space is finite or infinite.

The third building block of a statistical model is the family of *sampling probability distributions*. They will be denoted P_n^θ and therefore, for all $\theta \in \Theta$, P_n^θ is the probability distribution on the sample space X^n. If the model is correctly specified, we assume that the available observations $(x_1, \ldots, x_n)$ are generated by a random process described by one of the sampling probability distributions.

We summarize these concepts in the following definition.

Definition 1.1 *A statistical model $\mathcal{M}_n$ is defined by the triplet*

$$\mathcal{M}_n = \left\{X^n, \Theta, P_n^\theta\right\}$$

where $X^n \subset \mathbb{R}^{nm}$ is the sample space of dimension n, Θ is a parameter space and P_n^θ is the family of sampling probability distributions. ■

We use the notation

$$x|\theta \sim P_n^\theta \tag{1.1}$$

to summarize "x is distributed according to the distribution P_n^θ if the parameter value equals θ". Equivalently, we say that x follows the distribution P_n^θ conditionally on θ. Hence, we incorporate the dependence on a parameter in a probabilistic conditioning (which would necessitate, to be rigorous, regularity assumptions not examined here).

Example 1.1 *(Unidimensional normal model) Suppose that* $m = 1$ *and* $x \in \mathbb{R}^n$. *Here, the parameter vector is* $\theta = (\mu, \sigma^2) \in \Theta = \mathbb{R} \times (0, +\infty)$. *Moreover,*

$$x|\mu, \sigma^2 \sim N_n\left(\mu \mathbf{1}_n, \sigma^2 I_n\right)$$

where $\mathbf{1}_n$ *is a vector of* $\mathbb{R}^n$ *whose elements are all equal to* 1 *and* I_n *is the identity matrix of size* n. *The notation* N_n *represents the multidimensional normal distribution of dimension* n, *but often we will drop the subscript* n. □

Example 1.2 *(Binomial model) Let* $m = 1$ *and* $x \in \mathbb{R}^n$ *with* $x_i \in \{0, 1\} \subset \mathbb{R}$ *for all* $i = 1, \ldots, n$. *The parameter* θ *is now an element of* $\Theta = [0, 1]$. *The probability of a vector* x *given* θ *is then:*

$$P_n^\theta(\{x\}) = \prod_{i=1}^{n} \theta^{x_i} (1 - \theta)^{1 - x_i}.$$

It follows from this expression that, if $k = \sum_{i=1}^n x_i$,

$$P_n^\theta(k) = C_n^k \theta^k (1 - \theta)^{n-k}.$$ □

The aim of statistical inference is essentially the acquisition of knowledge on the distribution that generates the data or on the parameter θ that characterizes this distribution. In order to relate these two notions, we suppose that the statistical model is identified. This property is defined below.

Definition 1.2 *The model* $\mathcal{M}_n$ *is identified if, for any pair of (vectorial or functional) parameters* θ_1 *and* θ_2 *of* Θ, *the equality* $P_n^{\theta_1} = P_n^{\theta_2}$ *implies* $\theta_1 = \theta_2$. *In other words, the model is identified if the sampling probability distributions define an injective mapping of the elements of* Θ. ■

We will spend more time on this concept in the sequel, in particular in Chapter 16. Examples 1.1 and 1.2 define two identified models. The following model illustrates the lack of identification.

Example 1.3 *Suppose* $m = 1$ *and* $x \in \mathbb{R}^n$ *with* $\theta = (\alpha, \beta) \in \mathbb{R}^2 = \Theta$. *The sampling probability distributions satisfy*

$$x|\theta \sim N_n((\alpha + \beta)\mathbf{1}_n, I_n).$$

The model is not identified because $\theta_1 = (\alpha_1, \beta_1)$ *and* $\theta_2 = (\alpha_2, \beta_2)$ *define the same distribution as long as* $\alpha_1 + \beta_1 = \alpha_2 + \beta_2$, *which does not imply* $\theta_1 = \theta_2$. □

Given a realization x of a sample of size n, the econometrician will try to estimate θ, that is, to associate with x a value $\hat{\theta}(x)$ (or $\hat{\theta}_n(x)$) of θ, or to perform hypothesis testing, that is, to answer positively or negatively to the question whether θ belongs to a given subset of Θ. The estimation of θ can then serve to forecast new realizations of the sampling process.

In a stylized way, the statistical model empirically translates an economic theory by maintaining some assumptions through the choice of the family of sampling probability distributions. The observations do not lead to reconsidering the choice of the parameter space; instead they permit us to determine the parameter value. This vision is a bit simplistic because recent procedures have been developed to reject or validate a model as a whole, to choose between models, and to determine a model from observations.

The statistical models described here pertain to so-called *reduced forms.* The economic theory describes the complex behaviors of agents, the equilibrium relationship between these behaviors, and the link between relevant economic measures and the observable measures. It is assumed that this set of relationships is solved in order to describe the law of the data. The last part of this book, in particular Chapters 16 and 17, will detail the essential elements of this construction, whereas the first parts suppose that this task of statistical translation of the economic theory has been done.

The vector x is alternatively called *data*, *observations*, or *sample*. The two last terms refer implicitly to different learning schemes; the first one evokes a process of passive acquisition of data (macroeconomic data), whereas the second one refers to a partial or total control of the data collection procedure (poll, stratified survey, experiments). Again, these distinctions will not be exploited until the last part of this book.

Similarly, we will not discuss the choice of random formalization, which is now standard. The stochastic nature of the way observations are generated can be interpreted in various manners, either as a measurement error or an error resulting from missing variables, for instance. Moreover, the economic theory has recently provided constructions that are random per se (for instance, models describing the solution of games with imperfect information) and which we will discuss in the presentation of structural models.

1.3 Independent and Identically Distributed Models

Independent and identically distributed models (*i.i.d.*) constitute the basic structure of statistical inference. Basically, they describe the arrival of a sequence of observations that are generated by the same probability distribution, independently from each other. These models do not provide a sufficient tool for the econometrician who exploits individual observations (and hence generated by different distributions dependent on the individual characteristics) or time series (and hence generally dependent from one another), but they play a fundamental role in the study of statistical procedures.

Definition 1.3 *The model $\mathcal{M}_n = \{X^n, \Theta, P_n^\theta\}$ is i.i.d. if*

a) *The observations $x_1, \ldots, x_n$ are independent in terms of the distribution P_n^θ for all θ (denoted $\perp\!\!\!\perp_{i=1}^n x_i|\theta$).*

b) *The observations $x_1, \ldots, x_n$ have the same distribution denoted Q^θ, so that $P_n^\theta = [Q^\theta]^{\otimes n}$.* ■

Example 1.4 *The model defined in Example 1.1 is i.i.d. and Q^θ is the normal distribution with mean μ and variance σ^2. This example permits us to define a new notation:*

$$\left.\begin{array}{l} \perp\!\!\!\perp_{i=1}^n x_i|\theta \\ x_i|\theta \sim N\left(\mu, \sigma^2\right) \quad \forall i \\ \theta = \left(\mu, \sigma^2\right) \end{array}\right\} \Longleftrightarrow x_i|\theta \sim i.i.N.\left(\mu, \sigma^2\right).$$

□

Example 1.5 *Example 1.2 is again an i.i.d. model satisfying:*

$$\perp\!\!\!\perp_{i=1}^n x_i|\theta \quad \text{and} \quad x_i|\theta \sim B(\theta) \quad \forall i,$$

where $B(\theta)$ denotes the Bernoulli random variable, which equals 1 and 0 with probabilities θ and $(1-\theta)$ respectively. □

Consider now some counterexamples of i.i.d. models.

Example 1.6 *Suppose that $\theta \in \mathbb{R}^n$ and $x_i \in \mathbb{R}$ with*

$$\perp\!\!\!\perp_{i=1}^n x_i|\theta \quad \text{and} \quad x_i|\theta \sim N(\theta_i, 1).$$

The random variables x_i are independent but their distributions differ. □

Example 1.7 *Suppose that $\lambda = (a, \xi, \sigma^2) \in \mathbb{R}^2 \times \mathbb{R}_*^+$ and that the sample is i.i.d. conditionally on λ such that*

$$\perp\!\!\!\perp_{i=1}^n x_i|\lambda \quad \text{and} \quad x_i|\lambda \sim N\left(a+\xi, \sigma^2\right).$$

Now, suppose ξ is an unobservable random variable generated by a normal distribution with mean 0 and variance 1. Then, the parameter of interest is $\theta = (a, \sigma^2)$. We integrate out ξ to obtain the distribution of the sample conditional on θ. It follows that

$$x|\theta \sim N(a, V) \quad \text{with} \quad V = \sigma^2 I_n + \mathbf{1}_n \mathbf{1}_n'.$$

Then, the observations x_i have the same marginal distributions but are not independent. Moreover, the distribution of x is not modified if one permutes the order of the x_i. In this case, the distribution is said to be exchangeable. □

This example, based on the presence of an unobservable variable, will also be detailed in the last part of this book.

An important example of an i.i.d. model is provided by the following nonparametric model.

Example 1.8 *The sample $x = (x_1, \ldots, x_n)$, $x_i \in \mathbb{R}^m$, is i.i.d. and each x_i is generated by an unknown distribution Q. This model is denoted as*

$$\perp\!\!\!\perp_{i=1}^{n} x_i | Q \quad \text{and} \quad x_i | Q \sim Q.$$

Here, the parameter θ is equal to Q. It is a functional parameter belonging to the family $\mathcal{P}_m$ of distributions on $\mathbb{R}^m$. We could modify this example by restricting Q (for example, Q could have zero mean or could satisfy some symmetry condition resulting in zero third moment). □

1.4 Dominated Models, Likelihood Function

The statistical model $\mathcal{M}_n = \{X^n, \Theta, P_n^\theta\}$ is *dominated* if the sampling probability distributions can be characterized by their density functions with respect to the same dominating measure. In a large number of cases, this dominating measure is Lebesgue measure on X^n (included in $\mathbb{R}^{nm}$) and the dominance property means that there exists a function $\ell(x|\theta)$ such that

$$P_n^\theta(S) = \int_S \ell(x|\theta)dx \quad S \subset X^n.$$

Example 1.9 *Return to Example 1.1. The model is dominated and we have*

$$\ell_n(x|\theta) = (2\pi)^{-\frac{n}{2}} \sigma^{-n} \exp -\frac{1}{2\sigma^2}(x - \mu \mathbf{1}_n)'(x - \mu \mathbf{1}_n). \quad \square$$

The definition of dominance by Lebesgue measure is insufficient because it does not cover in particular the models with discrete sampling space. In such cases, we usually refer to the dominance by the counting measure. If X is discrete (for example $X = \{0, 1\}$), the counting measure associates all sets of X with the number of their elements. A probability distribution on X is characterized by the probability of the points x; these probabilities can be considered as the density function with respect to the counting measure.

Example 1.10 *In Example 1.2, we have*

$$P_n^\theta(\{x\}) = \ell_n(x|\theta) = \prod_{i=1}^n \theta^{x_i}(1-\theta)^{1-x_i}. \qquad \square$$

Definition 1.4 *A model $\mathcal{M}_n = \{X^n, \Theta, P_n^\theta\}$ is said to be dominated if there exists a measure ν on X (independent of θ) such that there exists $\ell(x|\theta)$ satisfying*

$$\forall \theta \in \Theta \quad P_n^\theta(S) = \int_S \ell_n\,(x|\theta)\nu(dx). \tag{1.2}$$

The function ℓ_n of $X \times \Theta$ in $\mathbb{R}^+$ is called density (function) of the observations or likelihood function depending on whether it is considered as a function of x for a fixed θ or as a function of θ for a fixed x. ■

The dominance property is actually related to the dimension of the statistical model. If the family P_n^θ is finite, that is if Θ is finite in the identified case, the model is always dominated by the probability $\frac{1}{n}\sum_{\theta\in\Theta} P_n^\theta$. This property is not true if Θ is infinite dimensional: the nonparametric model of Example 1.8 is not dominated. A parametric model (in the sense of a finite dimensional Θ) is not always dominated as shown by the following example.

Example 1.11 *Let $n = 1$, $X = [0, 1]$ and $\Theta = [0, 1]$. Let*

$$P_1^\theta = \delta_\theta$$

where δ_θ is the Dirac measure at θ defined by the property

$$\delta_\theta(S) = \left| \begin{array}{ll} 1 & \text{if } \theta \in S \\ 0 & \text{if } \theta \notin S. \end{array} \right.$$

We also use the notation

$$\delta_\theta(S) = \mathbb{I}(\theta \in S),$$

where the function $\mathbb{I}(.)$ equals 1 if the condition in parentheses is true and 0 otherwise. This model is not dominated but the proof of this result requires more advanced measure theory than we wish to use here. □

The dominance property is particularly useful in i.i.d. models. Suppose that $\mathcal{M}_n = \{X^n, \Theta, P_n^\theta\}$ is i.i.d. and that each observation is generated by the probability distribution, Q^θ. If Q^θ is dominated and admits a density $f(x_i|\theta)$, the independence and the identity of distributions imply that $\mathcal{M}_n$ is dominated and that the density of the observations can be written as

$$\ell_n(x|\theta) = \prod_{i=1}^{n} f(x_i|\theta). \tag{1.3}$$

The logarithm of the likelihood function plays an important role, it is also called *log-likelihood* and is defined as

$$L_n(x, \theta) = \ln \ell_n(x|\theta). \tag{1.4}$$

In the i.i.d. case, it satisfies the property:

$$L_n(x, \theta) = \sum_{i=1}^{n} \ln f(x_i|\theta). \tag{1.5}$$

Example 1.12 *(Multidimensional normal model) Let $\theta = (\mu, \Sigma)$ where $\mu \in \mathbb{R}^m$ and Σ is a symmetric positive definite matrix of dimension $m \times m$. Hence, $\Theta = \mathbb{R}^m \times \mathcal{C}_m$ where $\mathcal{C}_m$ is the cone of symmetric positive definite matrices of size $m \times m$. Moreover, $X = \mathbb{R}^{nm}$ and the model is i.i.d. with*

$$x_i|\theta \sim N_n(\mu, \Sigma) \quad x_i \in \mathbb{R}^n.$$

Therefore, the model is dominated. We have

$$\begin{aligned}\ell_n(x|\theta) &= \prod_{i=1}^{n} (2\pi)^{-\frac{m}{2}} |\Sigma|^{-\frac{1}{2}} \exp -\frac{1}{2}(x_i - \mu)'\Sigma^{-1}(x_i - \mu) \\ &= (2\pi)^{-\frac{nm}{2}} |\Sigma|^{-\frac{n}{2}} \exp -\frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)'\Sigma^{-1}(x_i - \mu).\end{aligned}$$ □

1.5 Marginal and Conditional Models

From a statistical model, one can build other models through the usual operations of probability calculus which are marginalization and conditioning. The concept of a conditional model is particularly fundamental in econometrics and allows us to build a first extension of the i.i.d. model which is too restrictive to model economic phenomena. First, we will derive the conditional model as a byproduct of the joint model, but in practice the conditional model is often directly specified and the underlying joint model is not explicitly defined.

Let $x = (x_i)_{i=1,\ldots,n}$ be the sample. It is assumed that, for each observation i, x_i can be partitioned into (y_i, z_i) with respective dimensions p and q (with $p + q = m$). Let us denote $y = (y_i)_{i=1,\ldots,n}$ and $z = (z_i)_{i=1,\ldots,n}$. Moreover, the space X is factorized into $Y \times Z$ with $y_i \in Y$ and $z_i \in Z$. This splitting of x facilitates the presentation, but in some examples, y_i and z_i are two functions of x_i defining a bijective (one-to-one and onto) mapping between x_i and the pair (y_i, z_i). By a relabelling of x_i, one can get back to the current presentation.

Definition 1.5 *From the model $\mathcal{M}_n = \{X^n, \Theta, P_n^\theta\}$, one obtains:*

- *the marginal model on Z^n, denoted $\mathcal{M}_{nz} = \{Z^n, \Theta, P_{nz}^\theta\}$, with sample space Z^n, parameter space Θ, and sampling probability distribution P_{nz}^θ which is the marginal probability of P_n^θ on Z.*
- *the conditional model given Z, denoted $\mathcal{M}_{ny}^z = \{Y^n \times Z^n, \Theta, P_{ny}^{\theta z}\}$, with sample space $Y^n \times Z^n$, parameter space Θ, but which sampling probability distribution is the conditional distribution of Y^n given $z \in Z^n$.*
  In a dominated model (by Lebesgue measure to simplify) with the density of observations denoted $\ell_n(x|\theta)$, the marginal and conditional models are dominated and their densities satisfy:

$$\begin{cases} \ell_{n\ marg}(z|\theta) & = \int \ell_n(y, z|\theta)dy \\ \ell_{n\ cond}(y|z, \theta) = \frac{\ell_n(y,z|\theta)}{\ell_{n\ marg}(z|\theta)}. \end{cases} \tag{1.6}$$

■

Example 1.13 *Consider an i.i.d. model with sample $x_i \in \mathbb{R}^2$ that satisfies*

$$x_i|\theta \sim i.i.N.\left(\begin{pmatrix}\eta \\ \zeta\end{pmatrix}, \Sigma\right)$$

with

$$\theta = (\eta, \zeta, \Sigma) \quad and \quad \Sigma = \begin{pmatrix}\sigma_{yy} & \sigma_{yz} \\ \sigma_{yz} & \sigma_{zz}\end{pmatrix}.$$

Then, $\theta \in \Theta = \mathbb{R}^2 \times \mathcal{C}_2$. We can decompose this model into a marginal model of Z which remains i.i.d. and satisfies

$$\perp\!\!\!\perp_{i=1}^n z_i|\theta \quad and \quad z_i|\theta \sim N(\zeta, \sigma_{zz})$$

and a conditional model characterized by

$$y_i|z_i, \theta \sim N(\alpha + \beta z_i, \sigma^2)$$

with

$$\beta = \frac{\sigma_{yz}}{\sigma_{zz}}, \quad \alpha = \eta - \frac{\sigma_{yz}}{\sigma_{zz}}\zeta \quad and \quad \sigma^2 = \sigma_{yy} - \frac{\sigma_{yz}^2}{\sigma_{zz}}.$$

□

This example has the property that the parameter θ of the original model can be decomposed into two functions of θ,

$$\theta_{marg} = (\zeta, \sigma_{zz}) \quad \text{and} \quad \theta_{cond} = (\alpha, \beta, \sigma^2),$$