1 In Brief: Elements of Statistical Inference

To prepare for what will be examined in the course, we will attempt to develop a basic language concerning Statistical Estimation and Statistical Hypothesis Testing. These, along with Statistical Forecasting processes, constitute parts of the field of statistics known as Statistical Inference.

2 The Statistical Problem-Parametric Models

Statistical inference generally involves solving the following version of the "statistical problem":[1] We are given a probability space equipped with a well-defined probability distribution, the characteristics of which are unknown and of interest to us. We have at our disposal observations, i.e., values obtained from a random element following the given distribution. This random element is called a (statistical) sample, and the values observed are the values of the available sample. These values carry some information about the unknown characteristics of the underlying distribution in which we are interested. Is it possible to use the information from the sample to determine or approximate these characteristics?

The previous description of the statistical problem is quite abstract. A simple example from our social experience can help us understand it in more familiar terms. Consider political polling processes, which aim, among other things, to estimate the electoral influence of political parties

[1]This version refers to the part of the statistical science known as Statistical Inference. The other part of the statistical problem concerns the concise representation of the information contained in the sample, which is the focus of Descriptive Statistics. Despite their different purposes, these parts are not independent of each other.

within the electorate at a given point in time. The true and unknown proportions of party influence form a discrete probability distribution; the reference set is the collection of voters, equipped with the algebra formed by the collection comprising the groups of voters for each party. On this algebra, the distribution is defined, assigning to each pure group the ratio of its members to the total population size. Through a random variable that assigns each voter to their preferred political party, this distribution is transferred to a probability distribution over the set of political parties; this distribution ultimately becomes the subject of inference. The sample consists of n independent realizations of the aforementioned random variable. Its empirical distribution provides an approximation of the unknown distribution mentioned above.

Based on the above, the general framework within which we will work is composed of the following characteristics.

- The sample, which will be denoted as z_n in the following, constitutes a random element (e.g. random variable, or random vector, etc) defined on some, potentially latent, probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and takes values in some Euclidean space, say $\mathbb{R}^{k \times n}$. The k refers to the dimension of each individual sample element, while n refers to the number of these individual elements. The random variables constituting each individual sample element, as a random k-dimensional vector, may, in turn, be appropriately grouped according to the characteristics of the related distribution that are the focus of inference. For example, in the general linear model, which we will gradually develop and is widely used in econometrics, we have $z_n = (Y_n, \mathbf{X}_n)$, where Y_n is an

$n \times 1$ random vector called the dependent variable, and $\mathbf{X}_n$ is an $n \times p$ stochastic matrix whose column vectors are called independent variables or regressors. Obviously, here $k = p + 1$, while for $i = 1, \ldots, n$, the i-th individual sample element is the random row vector formed by the i-th random variable of Y_n paired with the i-th row of $\mathbf{X}_n$ in the order implied by the form of z_n. Thus, in our framework, the sample is essentially a collection of random variables grouped into random vectors and possibly stochastic matrices according to the requirements of statistical inference. Each sample element can be perceived as a random vector of suitable dimension, while we assume the researcher has n such elements at their disposal.[2] This is general enough for what we want to develop next but does not constitute the most abstract framework. Note that the sample should not be confused with the observations available to the researcher. Observations are the value that the sample, as a random element, takes when evaluated at some element of its domain Ω; this value is precisely what is available in the context of sampling and may change when the sampling is repeated.

- The object of inference: the joint distribution of z_n, which will be at least partially unknown. The goal of statistical inference is to determine (or approximate) some of the unknown characteristics of the joint distribution of the sample that are of interest and potentially feasible to estimate/approximate. These characteristics, or the entire joint distribution, or parts of it, will be symbolized, somewhat ambiguously,

[2] Somewhat abusively, we denote the i-th random k-dimensional member of the sample as z_i, for $i = 1, \ldots, n$. The abuse arises from the fact that z_n simultaneously denotes the sample and the n-th element of it. What is meant each time will be clear from the context.

as D_0. In our framework, statistical inference will be considered as the set of procedures aimed at determining (or approximating) D_0.

- (Semi-)Parametric Structure: We assume that our framework is supplemented by some exogenous structure with respect to z_n, pertaining to the unknown D_0. Specifically, we consider that the object of inference D_0 depends-possibly partially-on the unknown value of a Euclidean parameter, i.e., a vector $\theta_0 \in \Theta \subseteq \mathbb{R}^p, p \in \mathbb{N}^\star$. This is equivalent to the existence of a relationship between θ_0 and D_0, which in various cases is defined by the probabilistic properties of D_0. Knowing θ_0 precisely is equivalent to adequately knowing the characteristics of D_0 that are of interest. Therefore, the issue of determining these characteristics essentially reduces to locating θ_0.Reducing the statistical problem to the problem of locating a point in a Euclidean space can be greatly facilitating solving the statistical problem at hand; the mathematical machinery of locating Euclidean points (e.g. equations, inequalities, etc) is usable.

Example 1 (Conditional Mean - Linear Case (LLS))**.** Let us assume that in the previously mentioned framework where $z_n = (Y_n, \mathbf{X}_n)$, we are interested in the unknown conditional mean of the dependent variable given the independent variables, that is, the random vector $\mathbb{E}(Y_n/\sigma(\mathbf{X}_n))$, provided it is well-defined. If the involved random variables have marginal distributions with first-order moments, then this is indeed well-defined (why?). Note that here, and for reasons that will become clear as this example is further developed, D_0 is considered to be the conditional distribution of Y_n given the algebra $\sigma(\mathbf{X}_n)$. Thus, the aforementioned conditional mean represents

the first moment of this distribution. In the linear case of the problem, it is assumed that the unknown $\mathbb{E}(Y_n/\sigma(\mathbf{X}_n))$ is some unknown linear function of $\sigma(\mathbf{X}_n)$, so there necessarily exists some $\theta_0 \in \mathbb{R}^p$ such that $\mathbb{E}(Y_n/\sigma(\mathbf{X}_n)) := \mathbf{X}_n\theta_0$, which establishes the relationship between θ_0 and D_0. Observe that, although by construction the conditional mean depends on the sample through $\mathbf{X}_n$, the parameter value characterizing it, θ_0, is independent of the sample. In a way, θ_0 represents certain characteristics of the conditional mean that are independent of the sample. If the above specification of the conditional mean is correct, then finding the unknown θ_0 is equivalent to finding the first moment of this conditional distribution, which is the goal. Note that the above is equivalent to $Y_n = \mathbf{X}_n\theta_0 + \varepsilon_n$, where for the random vector ε_n, it necessarily holds that $\mathbb{E}(\varepsilon_n/\sigma(\mathbf{X}_n)) = \mathbf{0}_n$ (why?).

Example 2 (Conditional Mean - Not Necessarily Linear Case (NLLS))**.** Suppose in the framework of the previous example, we have that $\mathbb{E}(Y_n/\sigma(\mathbf{X}_n)) := g(\theta_0, \mathbf{X}_n)$, where $g : \mathbb{R}^p \times \mathbb{R}^{n\times p} \to \mathbb{R}^n$ is a known function, with θ_0 as before being unknown. Clearly, here D_0 can also be considered as the conditional distribution of Y_n given $\sigma(\mathbf{X}_n)$, and the relationship between it and θ_0 is given by the form of the conditional mean $\mathbb{E}(Y_n/\sigma(\mathbf{X}_n))$ and the specific way it depends on the parameter value. Similarly to the previous example, the above is equivalent to $Y_n = g(\theta_0, \mathbf{X}_n) + \varepsilon_n$, where again for the random vector ε_n, it necessarily holds that $\mathbb{E}(\varepsilon_n/\sigma(\mathbf{X}_n)) = \mathbf{0}_n$ (why?). Observe that this includes as a subcase the previous example when g is (multi-)linear, i.e., $g(\theta_0, \mathbf{X}_n) = \mathbf{X}_n\theta_0$. However, it also includes cases where the conditional mean is not necessarily a linear function of $\sigma(\mathbf{X}_n)$. For instance, let $g(\theta_0, X_n) = (\exp(\mathbf{X}'_{(i)}\theta_0))_{i=1,\cdots,n}$, where $\mathbf{X}_{(i)}$ denotes the i-th row of the matrix.

Finally, observe that here, just as in the previous example, $z_n = (Y_n, \mathbf{X}_n)$ and $k = 1 + p$.

Example 3 (Instrumental Variables (IV))**.** Given the linear relationship $Y_n = \mathbf{X}_n\theta_0 + \varepsilon_n$ mentioned above, we consider something generally weaker than the relationship $\mathbb{E}(\varepsilon_n/\sigma(\mathbf{X}_n)) = \mathbf{0}_n$ that appeared in Example LLS. Recall that this relationship implies, first-through the Law of Iterated Expectations- that the marginal distribution of each element of ε_n has zero mean (Exercise: prove it!). Secondly, it implies that each element of ε_n has zero covariance with any measurable transformation of any random variable composing the regressors' matrix $\mathbf{X}_n$ (Exercise: prove it!). This can be particularly strong, as, for example, in time series environments, and indicatively the AR(1) process example, where ε_n relates to X_k, $\forall k \geq n$), it may not hold. It may also fail in numerous other cases, such as when there is non-zero covariance between an element of ε_n and a nonlinear transformation of an element of $\mathbf{X}_n$. A weaker version, therefore, assumes the existence of a stochastic matrix of instrumental variables $\mathbf{W}_n$ of dimension $n \times q$, analogous to $\mathbf{X}_n$, where q essentially denotes the number of instrumental variables, such that the orthogonality condition $\mathbb{E}(\mathbf{W}'_n\varepsilon_n) = \mathbb{E}(\mathbf{W}'_n(Y_n - \mathbf{X}_n\theta_0)) = \mathbf{0}_q$ holds. In this case, the object of inference is to find the assumed linear-relative to $\mathbf{X}_n$-deviation of Y_n, for which the condition $\mathbb{E}(\mathbf{W}'_n(Y_n - \mathbf{X}_n\theta_0)) = \mathbf{0}_q$ holds. This ultimately reduces to finding the unknown θ_0. Note that the matrix of instrumental variables may include columns of the matrix $\mathbf{X}_n$. When these matrices coincide, we obtain a weaker version of Example LLS. Moreover, the above can clearly be extended to cases where $Y_n = g(\theta_0, \mathbf{X}_n)+\varepsilon_n$ and $\mathbb{E}(\mathbf{W}'_n\varepsilon_n) = \mathbb{E}(\mathbf{W}'_n(Y_n - g(\theta_0, \mathbf{X}_n))) = \mathbf{0}_q$, where g is assumed known,

while θ_0, which ultimately determines the object of inference, is unknown. Observe that in each of the aforementioned cases, the sample is (potentially) augmented by the matrix of instrumental variables, i.e., $z_n = (Y_n, \mathbf{X}_n, \mathbf{W}_n^\star)$, where $\mathbf{W}_n^\star$ is the matrix of instrumental variables remaining after excluding columns corresponding to independent variables possibly appearing within it (and which is considered empty when $\mathbf{X}_n = \mathbf{W}_n$), while $k = 1 + p + q^\star$, where $q^\star$ is the number of columns of $\mathbf{W}_n^\star$.

Further examples will be examined later. As previously mentioned, the common feature of all the examples is the parametric description of the characteristics of interest in D_0 and the reduction of finding them to locating a point in Euclidean space. This can be generalized, e.g., through the use of non-Euclidean parameters, though this is beyond the scope of the present notes.

From now on-as explicitly mentioned in Example LLS-we will assume that θ_0 is independent of n. Note that this holds in all the previously mentioned examples, while the characteristics of D_0 indicated by it may depend on n. For example, this is inherently true for the conditional means in the first and second examples. We note that the assumption of independence of θ_0 from n could be modified, though this would require various changes in the asymptotic theory developed later.

The framework of the (semi-)parametric structure, which includes the relationship between θ_0 and D_0, may imply that there exists a corresponding relationship between the elements of Θ and a family of distributions, say $\mathcal{D}$, that (may) include as a special case the aforementioned pair (θ_0, D_0). As essentially indicated by the examples above, this relationship is constructed

based on assumptions about how the characteristics of D_0 of interest are specified with respect to the involved parameter. When the latter is allowed to take values over the entire parameter space, a mapping $\theta \to \mathcal{D}$ is formed. By construction, this mapping associates at least one distribution in $\mathcal{D}$ with each element of Θ. This ultimately gives us the definition of a statistical model within our framework:

Definition 1 (Statistical Model)**.** A statistical model in our framework is defined as the aforementioned mapping $\Theta \to \mathcal{D}$.

By construction, the statistical model is formed by (a) the choice of how the characteristics of D_0 of interest are specified with respect to the involved parameter, and (b) the choice of the possible values the parameter can take, i.e., the determination of Θ. Note that given (a), the absence of further information regarding the location of θ_0 in $\mathbb{R}^n$ may guide the choice of the parameter space to the least informative one, so that $\Theta = \mathbb{R}^p$. However, when additional information is available about where θ_0 may be located in $\mathbb{R}^n$, it may be useful to use it, in which case $\Theta \subset \mathbb{R}^p$ may be chosen.

Essentially, the statistical model represents the relationship between the possible values of the parameter and the corresponding distributions for z_n. Given the aforementioned (a) and (b), the concept includes both the parameter space Θ and the collection $\mathcal{D}$. The relationship defined by the model through the above choices describes which elements of $\mathcal{D}$ correspond to each element of Θ. In many cases-as we will see below-the set $\mathcal{D}$ may not be explicitly determined, but only implicitly.

In any case, a preliminary classification of statistical models of the above form can be made based on the properties of this mapping. If this

is functional, meaning that each element $\theta \in \Theta$ corresponds to a unique distribution in $\mathcal{D}$, then the statistical model is called fully parametric. In such cases, $\mathcal{D}$ is in a one-to-one correspondence with the parameter space. Each parameter value fully determines the corresponding element of $\mathcal{D}$. This will obviously occur when the relationship between the parameter and distributions arises from sufficiently available information incorporated into the aforementioned (a) and (b). In all other cases-i.e., when there exists $\theta \in \Theta$ such that more than one distribution in $\mathcal{D}$ corresponds to it-the model is called semi-parametric.[3]

Let us examine some examples:

Example 4 (LLS)**.** As observed in Example LLS, the conditional mean can be specified as a linear function of the matrix $\mathbf{X}_n$, whereby its dependence on the associated Euclidean parameter is immediate. This specification leads to a statistical model of the form

$$\{\mathbb{E}(Y_n/\sigma(\mathbf{X}_n)) = \mathbf{X}_n\theta, \theta \in \mathbb{R}^p\},$$

which can be viewed through the following mapping: (a) for each possible value of θ, any (conditionally corresponding) distribution on $\mathbb{R}^n$ with mean $\mathbf{X}_n\theta$ is associated, and (b) since no further information about θ_0 is available, the largest possible parameter space is chosen. Clearly, the collection $\mathcal{D}$ in this case is implicitly determined by (a) and (b), and the model is semi-parametric since each value of θ corresponds to multiple distributions (why?). If we introduce further information through (a), e.g., by assuming that the conditional covariance matrix of ε_n exists and equals $\mathbf{I}_n$, the identity

[3] In this case, the mapping is called multivalued (or correspondence).

matrix of size $n \times n$[4], the resulting model is

$$\{\mathbb{E}(Y_n/\sigma(\mathbf{X}_n)) = \mathbf{X}_n\theta, \theta \in \mathbb{R}^p, \text{Var}(Y_n/\sigma(\mathbf{X}_n)) = \mathbf{I}_n\} \text{ (Exercise: prove this!)},$$

which is clearly a subset of the previous model (why?) and is also semi-parametric (why?). If we further assume, within the framework of specification (a), that $\varepsilon_n/\sigma(\mathbf{X}_n) \sim N(\mathbf{0}_n, \mathbf{I}_n)$, we obtain a smaller subset of the above:

$$\{Y_n/\sigma(\mathbf{X}_n) \sim N(\mathbf{X}_n\theta, \mathbf{I}_n), \theta \in \mathbb{R}^p\} \text{ (Exercise: prove this!)},$$

which is fully parametric (why?). Note that the description of the Gaussian distributions composing $\mathcal{D}$ is explicit here. In this example, we specified three statistical models by choosing the relationships in (a) and selecting Θ as the largest possible. If, for instance, we also know as external information that θ_0 has integer components, this can be incorporated into the parameter space selection, resulting in smaller versions of the above by replacing $\Theta = \mathbb{R}^p$ with $\Theta = \mathbb{Z}^p$. This yields three models, where the first two are semi-parametric and the third is fully parametric (why?).

Example 5 (IV)**.** Given the specification described in Example 6.2, and similarly to the previous example, if no further information is available

[4]Note that this specification of the covariance matrix is particularly strong, as it assumes it to be diagonal (which implies further assumptions about the dependence and conditional distributions of the elements of ε_n on $\sigma(\mathbf{X}_n)$) and fully known, so no parallel statistical inference procedures are needed for it. This assumption is made without significant loss of generality for the purposes we focus on.

about the location of θ_0, we can choose a model of the form

$$\{\mathbb{E}(\mathbf{W}_n'(Y_n - \mathbf{X}_n\theta)) = \mathbf{0}_q, \theta \in \mathbb{R}^p\},$$

which can be interpreted as follows: for each value of θ, any distribution on $\mathbb{R}^{n\times k}$ that satisfies the relationship $\mathbb{E}(\mathbf{W}_n'(Y_n - \mathbf{X}_n\theta)) = \mathbf{0}_q$ is associated. Clearly, this is semi-parametric (why?). Smaller versions of this model can be obtained by allowing θ to take values in a proper (non-empty-why?) subset of $\mathbb{R}^n$, potentially utilizing relevant external information. Finally, note that when $\mathbf{W}_n = \mathbf{X}_n$, we obtain a weaker version of Example LLS (explain!).

If the specifications made in constructing a statistical model, or the choice of parameter space, reflect information related to economics, we then obtain the following classificatory definition:

Definition 2. The model $\theta \to \mathcal{D}$ is called an econometric model if and only if its specification or the choice of parameter space is based, at least partially, on exogenous information that may stem from Economic Theory or empirical characteristics of economic phenomena.

In econometric models, the purpose of statistical inference-in our framework, approximating θ_0-is related to verifying the empirical validity of claims made by economic theory about the functioning of phenomena it addresses. More generally, it aims at probabilistic inference, i.e., verifying properties of distributions that may characterize such phenomena, to assist economic theory in understanding them. Note that constructing econometric models, studying their properties, designing statistical inference procedures for

them, analyzing the properties of these procedures and their computational complexity, and applying them to specific econometric models are processes that constitute the field of Econometrics.[5]

Example 6 (CAPM-IV)**.** Suppose that, within the framework of Example IV, Y_n consists of a time series of observed logarithmic excess returns of a stock (defined by the consecutive time differences of its logarithmic price after subtracting the logarithm of the risk-free rate at the current time), and $\mathbf{X}_n$ consists of the observed excess logarithmic returns of the market portfolio-adequately approximated, e.g., by a suitable stock index, so $p = 1$. The financial CAPM pricing model-see, for example, **?**-predicts that

$$\mathbb{E}(\mathbf{X}_n'(Y_n - \mathbf{X}_n\theta_0)) = 0$$

for some unknown θ_0 representing the relationship between the stock and systematic risk, as depicted by the market portfolio. Note that the econometric model can often-and typically does-expand, allowing the matrix $\mathbf{X}_n$ to include additional variables, e.g., a constant column of ones. This is a way to test the validity of CAPM, as if it holds true, the corresponding components of θ_0 should be zero. Clearly, the above constitutes a special case of Example IV with $\mathbf{W}_n = \mathbf{X}_n$.

Example 7 (Market Entry Game)**.** Suppose that firm $j \in \{1, 2\}$ decides whether to enter market m, with $m \in \{1, 2, \ldots, n\}$. The decision is represented by the variable $Z_{j,m}$, where $Z_{j,m} = 1$ means the firm decides to enter,

[5]We note the existence of econometric models that do not involve Euclidean parameters. These pertain to the subfield of non-parametric Econometrics, which lies outside the current scope. Interested readers may refer to **?** for more information.

and $Z_{j,m} = 0$ means it decides not to enter. The decision is based on the profit function:

$$\Pi_{j,m} = (\varepsilon_{j,m} - \theta_{0,j} Z_{-j,m}) \mathbb{I}_{\{Z_{j,m}=1\}},$$

where $\varepsilon_{j,m} \sim \text{Unif}[0,1]$ and is i.i.d. for $j = 1, 2,\ m = 1, \ldots, n$ (Exercise: is the collection $(\varepsilon_{j,m})_{j=1,2,\, m=1,\ldots,n}$ a stochastic process?), $\mathbb{I}_{\{Z_{j,m}=1\}}$ is the indicator function of the event $\{Z_{j,m} = 1\}$, and $Z_{-j,m}$ is the decision of the other firm. The random variable $\varepsilon_{j,m}$ represents the benefit of firm j from entering market m. Therefore, the parameter $\theta_{0,j}$ represents the loss of benefit due to the other firm's entry into the same market. It summarizes a form of sensitivity to the presence of a competitor, and importantly, it does not depend on the market. The parameter

$$\theta_0 = (\theta_{0,1}, \theta_{0,2}) \in \Theta = (0,1) \times (0,1)$$

is unknown. Due to the form of the profit function, we have a strategic interaction environment-a game-theoretic framework. It can be shown that this framework has the following (stochastic) Nash equilibria:

1. $(Z_{1,m}, Z_{2,m}) = (1,1)$, if $\varepsilon_{j,m} \geq \theta_{0,j}, \forall j = 1,2$,
2. $(Z_{1,m}, Z_{2,m}) = (1,0)$, if $\varepsilon_{1,m} \geq \theta_{0,1}, \varepsilon_{2,m} < \theta_{0,2}$,
3. $(Z_{1,m}, Z_{2,m}) = (0,1)$, if $\varepsilon_{1,m} < \theta_{0,1}, \varepsilon_{2,m} \geq \theta_{0,2}$, and
4. $(Z_{1,m}, Z_{2,m}) = \begin{cases} (0,1) \\ (1,0) \end{cases}$, if $\varepsilon_{j,m} < \theta_{0,j}, \forall j = 1,2$-in this case, we have multiple equilibria.

The equilibrium is stochastic and depends on the relationship between the involved random variables and parameters. Knowledge of θ_0 provides a probabilistic understanding of the firms' entry decisions in these markets.

Example 8 (GARCH(1,1))**.** The empirical characteristics of logarithmic returns for certain classes of financial securities often indicate various properties of conditional heteroskedasticity for the corresponding stochastic processes.[6] Among other things, the conditional (on an appropriate information set) variances of these processes appear to vary over time and exhibit properties that partially align with those of a GARCH(1,1) process. If a researcher has a time series of related logarithmic returns $(y_t)_{t=1,\dots,n}$ at their disposal, which are considered part of a stationary and ergodic GARCH(1,1) process with an unknown parameter vector $\theta_0 = (\omega_0, a_0, \beta_0)$, and it is known that $\omega_0 > 0$, $a_0, b_0 \geq 0$, and $a_0 + b_0 < 1$, the statistical model can be chosen as the collection of such processes related to the sample $(y_t)_{t=1,\dots,n}$, uniquely described by the following systems of stochastic recursive relationships:

$$\left\{ \begin{array}{l} \left\{ \begin{array}{ll} z_t \text{ iid}, \mathbb{E}(z_0) = 0, \mathbb{E}(z_0^2) = 1, & \\ y_t = z_t \sqrt{h_t(\theta)}, & , t = 1, \dots, n, \\ h_t(\theta) = \omega + \alpha y_{t-1}^2 + \beta h_{t-1}(\theta) & \end{array} \right. \\ \theta = (\omega, \alpha, \beta) \in \Theta, \\ \Theta = \{(\omega, \alpha, \beta) \in \mathbb{R}^3 : \omega > 0, \alpha, \beta \geq 0, \alpha + \beta < 1\}. \end{array} \right\}.$$

The above is semi-parametric since no further specification is made about

[6] There is an extensive related literature. Interested readers may refer to **?** for an example.

the distribution of z_0 beyond its first two moments. Additionally, the parameter space consists of all three-dimensional real vectors with a strictly positive first component, non-negative remaining components, and a sum less than one. The positivity constraints are natural for the strict positivity of σ_t, and the last constraint reflects external information available to the researcher about θ_0, related to the existence of $\mathbb{E}(\sigma_0^2)$. This constraint is stronger (Exercise: prove this using Jensen's inequality) than the condition ensuring the existence of a unique stationary and ergodic solution to the system.

The two previous examples established econometric models through which we gained an initial sense of how-firstly-claims of economic theory, and-secondly-empirical characteristics of economic variables, can guide the construction of statistical models. Moreover, all the examples we have discussed so far indicate the following: statistical models can be represented in various ways, e.g., using the distributions that compose them directly, or describing these distributions through their moments, recursive equations whose unique solutions are random elements following these distributions, and so on. This is clearly related to the numerous ways in which a probability distribution can be represented, which we briefly explored in the previous chapters of the first part.

A final concept that concerns us relates to the relationship between D_0 and the involved statistical model. In describing the models in the previous examples, the phrase "[...] the statistical model can be chosen [...]" was sometimes used verbatim, implying that the statistical/econometric model we work with in each case is essentially a matter of choice. Re-

call that its construction was based on (a) relationships (in what we do parametrically) specifying the characteristics of the unknown distribution of interest, and (b) the selection of permissible parameter values. Both aspects may be influenced by (exogenous to the sample) information available to the researcher, natural constraints related to the properties of the involved random elements-e.g., the positivity constraints in the last example-information derived from theoretical claims related to the sample, and so on. In any case, deciding which of the available information to use in specifying the model is a matter of decision and may relate to the researcher's preferences regarding the risk that $D_0 \notin \mathcal{D}$,i.e., the unknown distribution having characteristics not accounted for by the chosen model. This is called specification risk of the model and is typically present. Thus, the specification and choice of a statistical model can itself constitute a decision problem, for which statistical inference procedures can be employed to resolve. In this part of the book, we will not delve deeply into this, but in a later chapter, we will briefly explore a case where this risk materializes and a statistical test for the specification of certain semi-parametric models. For now, we complete our terminology with the following definition:

Definition 3. A statistical model is called well-specified if and only if $D_0 \in \mathcal{D}$. Otherwise, it is called misspecified.

Note that, based on how the parametricity of our framework was constructed, misspecification can arise for two reasons: (a) the specification relationships leading to the parametrization of the characteristics of D_0 are incorrect. For example, within the framework of the linear model, the conditional mean is not actually a linear function of $\mathbf{X}_n$, or $\mathbf{X}_n$ does not

include all relevant independent variables. For instance, within the framework of the last example, the process constituting the sample is not of the GARCH(1,1) type, and so on. In such cases, θ_0 has no meaning, although a parameter value may exist that relates to D_0 in some "optimal way." (b) Although the parametric specification is correct, the parameter space-due to, e.g., erroneous external information-was chosen such that $\theta_0 \notin \Theta$. For instance, in the last example, it may not hold that $a_0 + b_0 < 1$, but rather the more general $\mathbb{E}(\ln(a_0 z_0^2 + b_0)) < 0$ holds.

In any case, in what follows, we will deal minimally with specification issues. The statistical model will be considered given and (mostly) well-specified.

Assumption 1. *For $\Theta \to \mathcal{D}$, there exists $\theta_0 \in \Theta$ such that $\theta_0 \to D_0$.*

From the preceding discussion, it should be apparent that in (semi-)parametric models, the risk of misspecification cannot exceed that of their parametric counterparts; the former are supersets of the latter. Consequently, in various cases in Econometrics, semi-parametric models-when available-are considered a golden mean between models without Euclidean parameters (with their accompanying difficulties, e.g., in the properties of statistical inference procedures) and parametric models with their higher risk of misspecification.

3 Summary: Issues in Statistical Inference

In this section, we will briefly address some fundamental characteristics of point estimation and hypothesis testing procedures. Within the parametric

framework, we will examine general definitions and attempt to describe-by no means exhaustively-the minimum properties that these procedures should ideally possess. Note that we will not delve into set estimation issues (e.g., confidence intervals), although such estimations can be derived from the hypothesis testing procedures discussed. The properties we describe will be asymptotic properties, focusing on the behavior of these procedures as the sample size grows to infinity ($n \to \infty$), leading to unbounded growth in the information available to the inferential procedures in order to locate θ_0. These properties necessarily involve concepts of stochastic convergence. These ideas will be extensively used later on, where we will examine a rather general class of statistical inference procedures of this kind.

3.1 Elements of Point Estimation

Within our framework, point estimators are procedures that use the sample to leverage relevant information to approximate θ_0 within Θ. Mathematically, they are functions defined on the sample space, $\mathbb{R}^{k\times n}$ in our framework, with values in Θ. Their properties-pertaining to the quality of this approximation-derive from the properties of their distributions, provided the latter are well-defined. To ensure this, we require that they are also Borel measurable functions, i.e. effectively random vectors:

Definition 4. An estimator of θ_0 is any Borel measurable function $\theta_n : \mathbb{R}^{k\times n} \to \Theta$.[7]

[7]The term "estimate" refers to the value assigned by the estimator when computed for a specific sample.

For example, within any version of the linear model in Example 6.5, when $\Theta = \mathbb{R}^p$, both $\theta_n = \mathbf{0}_p$ and what we later call the Least Squares Estimator, $\theta_n = (\mathbf{X}_n'\mathbf{X}_n)^{-1}\mathbf{X}_n Y_n$, fall within this definition (Exercise: prove this!). The former is a constant function of the sample, while the latter is linear in Y_n. Clearly, there is an abundance of estimators for θ_0, posing a selection problem. This can be addressed by crystallizing desirable properties for the estimators we work with. These properties are essentially properties of their distributions. For instance, θ_n is called unbiased if and only if its distribution has a first moment equal to θ_0, i.e., $\mathbb{E}(\theta_n) = \theta_0$. This property is often loosely interpreted as meaning that the estimator "on average over possible sample values" identifies θ_0." Similarly, it might be desirable for the estimator's distribution to exhibit "maximum possible concentration" around θ_0, characterized by some efficiency property. Note that such properties, in all but the simplest econometric models, are either particularly difficult to derive due to the unknown analytical form of the estimators as functions of the sample or generally do not hold due to the complexity of this (even unknown) form. In the following, we will describe some properties that are usually easier to verify, as they concern the asymptotic behavior of estimators as $n \to \infty$, enabling], facilitating forms of asymptotic approximation to the structure of estimators.

> It is often convenient-particularly in proving asymptotic properties- to use concise expressions to denote asymptotic properties, such as convergence in probability or asymptotic tightness of a sequence of random variables. The expression $O_p(1)$ denotes that the sequence is asymptotically tight, while $o_p(1)$ denotes convergence in probability to

zero. More generally, if x_n, y_n, and z_n represent the general terms of related sequences, then $x_n = O_p(z_n)$ means $x_n = y_n z_n$ and $y_n = O_p(1)$. Similarly, $x_n = o_p(z_n)$ means $x_n = y_n z_n$ and $y_n = o_p(1)$. It is clear from the above definitions that $O_p(z_n) = z_n O_p(1)$, $o_p(z_n) = z_n o_p(1)$, and it is easy to prove properties like $o_p(1) + o_p(1) = o_p(1)$, $O_p(1) + O_p(1) = O_p(1)$, $o_p(1) + O_p(1) = O_p(1)$, $o_p(1)o_p(1) = o_p(1)$, $O_p(1)O_p(1) = O_p(1)$, $o_p(1)O_p(1) = o_p(1)$, and so on.

Next, we will work with appropriate limits as the sample size grows unbounded, with $\|\cdot\|$ denoting the Euclidean norm:

Definition 5. An estimator θ_n is called weakly consistent if and only if $\theta_n \xrightarrow{\text{p}} \theta_0$, i.e., if and only if $\forall \varepsilon > 0, \lim_{n\to\infty} \mathbb{P}(\|\theta_n - \theta_0\| > \varepsilon) = 0$.

Thus, an estimator is weakly consistent if and only if it converges in probability to θ_0. This is a relatively weak property that is desirable to satisfy, and its validity can (among many other methods!) be facilitated by relevant Laws of Large Numbers. A stronger version of the above applies when the convergence is almost sure instead of in probability, leading to the property of strong consistency.

Given consistency, the rate at which the estimator converges to the unknown parameter value may be of interest. This rate can be represented by some unbounded real sequence, say $r_n \to \infty$, such that if the distance of the estimator from θ_0 is pointwise multiplied by r_n, we obtain a random variable that exhibits asymptotic tightness:

Definition 6. If θ_n is weakly consistent and there exists $r_n \to \infty$ such that

$\forall \varepsilon > 0, \exists M_\varepsilon > 0$ such that

$$\lim_{n\to\infty} \mathbb{P}(r_n \|\theta_n - \theta_0\| > M_\varepsilon) \leq \varepsilon,$$

i.e., equivalently $r_n \|\theta_n - \theta_0\| = O_p(1)$, then the sequence (r_n) is called the rate of convergence of the estimator.

The property of asymptotic tightness may arise for more than one rate, so the rate may need to be optimally chosen Among consistent estimators, it seems reasonable to prefer the one (if it exists) with the fastest optimal rate of convergence. In many cases (but not always!), the optimal rate of convergence (when it exists) in our (semi-)parametric framework for many cases of estimators is $\sqrt{n}$. Given the optimal rate of convergence, the issue of the weak convergence of $r_n(\theta_n - \theta_0)$ becomes interesting:

Definition 7. If θ_n is weakly consistent and has an optimal rate $r_n \to \infty$, then it is called asymptotically Gaussian if and only if there exists a random vector Z and a positive definite matrix V of dimensions $p \times p$ such that $Z \sim N(\mathbf{0}_p, V_{p\times p})$ and

$$r_n(\theta_n - \theta_0) \rightsquigarrow Z.$$

The covariance matrix of the distribution of Z is called the asymptotic variance of the estimator.

The convergence in distribution of $r_n(\theta_n - \theta_0)$ may hold even if the asymptotic distribution is not Gaussian. It is shown that in specific frameworks, which largely encompass what will be discussed later, having θ_n asymptotically Gaussian makes it preferable, in terms of certain categories of

preferences regarding the risk of "deviation of the estimator from θ_0," to estimators that are not asymptotically Gaussian but may share the properties of consistency and the same rate of convergence. Among estimators that have the aforementioned properties and are also asymptotically Gaussian, the preferred one (if it exists) will be the one with the smaller asymptotic variance.[8] It is noted that deriving the optimal rate and the property of asymptotic normality can be facilitated (among many other methods!) by the action of some Central Limit Theorem.

3.2 Elements of Hypothesis Testing

In this section, for $\Theta^\star$ being a non-empty proper subset of Θ, we consider a researcher interested in determining whether $\theta_0 \in \Theta^\star$. This may stem from (possibly additional) claims of relevant theory or from some form of external information available about the location of θ_0. For example, in Example 6.8 above, the researcher might be interested in testing whether there is no temporal variation in the evolution of the conditional variance σ_t^2, but that it remains constant over time and equal to ω_0. In this case, we would have $\Theta^\star = \{\theta \in \Theta : a = b = 0\}$. If the researcher tends to trust the hypothesis $\theta_0 \in \Theta^\star$, they can use empirical information from the sample to test the following structure of hypotheses:

$$\begin{aligned} &\mathbb{H}_0 : \theta_0 \in \Theta^\star, \\ &\mathbb{H}_1 : \theta_0 \in \Theta^\circ. \end{aligned} \tag{1}$$

[8] To be elaborated!

The $\mathbb{H}_0$ is called the null hypothesis, and its complement $\mathbb{H}_1$ is the alternative hypothesis.[9]

A testing procedure usually involves selecting a test statistic-a measurable function of the sample taking values in the reals-choosing a rejection region for the null hypothesis, which is a measurable subset of the reals, constructed based on the properties of the statistic under the null hypothesis,[10] and determining the acceptable probability of rejecting the null hypothesis when H_0 is true. This probability is called the level of statistical significance.

Definition 8. Given the significance level $\alpha \in (0, 1)$, a test for the hypothesis structure 1 is any measurable function $\tau_n(\alpha) : \mathbb{R}^{k \times n} \to \{\text{fail to reject } \mathbb{H}_0, \text{reject } \mathbb{H}_0\}$.

The definition describes such a testing procedure as a decision-making process, which is a function of the sample that takes values in the set of possible decisions:

$$\{\text{fail to reject } \mathbb{H}_0, \text{reject } \mathbb{H}_0\}.$$

The quality of this procedure depends on the properties of this function and the way in which the probabilities of possible errors are formed in the decision-making process. Specifically, these are $\mathbb{P}(\text{reject } \mathbb{H}_0 \mid \mathbb{H}_0 \text{ true})$[11] and $\mathbb{P}(\text{fail to reject } \mathbb{H}_0 \mid \mathbb{H}_0 \text{ not true})$[12], as well as on the form of these probabilities change on the boundary between the hypotheses.

Definition 9 (First Order Properties)**.** The $\tau_n(\alpha)$ is called asymptotically

[9] It must be the case that $\Theta^\star \cap \Theta^\circ = \emptyset$. In several circumstances $\Theta^\circ = \Theta - \Theta^\star$.
[10] This represents the confidence in the null hypothesis.
[11] This is the probability of a Type I error.
[12] This is the probability of a Type II error.

conservative if and only if

$$\limsup_{n\to\infty} \mathbb{P}(\textbf{reject } \mathbb{H}_0 \mid \mathbb{H}_0 \textbf{ true}) < \alpha.^{13}$$

It is called asymptotically exact if and only if

$$\lim_{n\to\infty} \mathbb{P}(\textbf{reject } \mathbb{H}_0 \mid \mathbb{H}_0 \textbf{ true}) = \alpha.$$

Finally, it is called consistent if and only if

$$\liminf_{n\to\infty} \mathbb{P}(\textbf{reject } \mathbb{H}_0 \mid \mathbb{H}_0 \textbf{ not true}) = 1.^{14}$$

Asymptotic conservativeness refers to the case where the probability of making a Type I error in the decision-making process is asymptotically strictly less than the nominal level chosen for this. This is interpreted as greater asymptotic confidence in the null hypothesis than the nominal level. In the case of asymptotic exactness, this is not true. The test asymptotically recalls the nominal significance level. Finally, when the test is consistent, the probability of making a Type II error asymptotically vanishing.

[13] The term $\limsup$ refers to the limit of the decreasing sequence of upper bounds of the successive truncations of this sequence of probabilities. When the sequence has a limit, it coincides with this.

[14] Dual to $\limsup$, the term $\liminf$ refers to the limit of the increasing sequence of lower bounds of the successive truncations of this sequence of probabilities. When the sequence has a limit, it coincides with this.

3.3 Epilogue

Statistical models are collections of objects concerning probability distributions. Statistical inference procedures are random elements taking values in, for example, parameter spaces, decision spaces, etc. The probability distributions of these are (co-)shaped by the mapping of the unknown distribution of the sample onto the corresponding decision space. Their properties are essentially determined by the properties of this mapping. The asymptotic behavior of the corresponding sequences of estimators, testing procedures, etc, as the sample size grows to infinity, provide with-in many cases convenient-large sample approximations of those properties. Those approximations can in turn be used to further shape the inferential procedures at hand.

The process of selecting a statistical model, estimator, test, etc., is essentially a decision problem under risk (e.g., of making incorrect decisions). For example, some preferences towards risk, could allow for the selection of an estimator that is "slightly" inconsistent but more "precise." Note that solving optimal selection problems may be performed via the use of auxiliary statistical inference procedures, whose selection itself may also constitute such problems. The asymptotic properties mentioned above are rudimentary. Finer aspects of asymptotic theory, such as the study of how an asymptotically normal estimator approaches the corresponding distribution, form additional criteria for selection.

Appendix: Further Examples

Brief Introduction to Binary Response Models

Binary response models are a class of statistical models used to analyze outcomes that take one of two possible values, such as success/failure, yes/no, or 1/0. These models are designed to estimate the probability of one of the outcomes occurring, given a set of explanatory variables.

Formally, using the notation employed in the LLS and NLLS examples, Y_i represents the binary response variable, i.e. $Y_i \in \{0, 1\}$. The probability of $Y_i = 1$, given the regressors' matrix $\mathbf{X}_n$, is modeled as:

$$\mathbb{P}(Y_i = 1 \mid \sigma(\mathbf{X}_n)) = F(\mathbf{X}_i'\theta_0),\ i = 1, \ldots, n,$$

where $F(\cdot) : \mathbb{R} \to [0, 1]$ is a link function that ensures the probability remains in the interval $[0, 1]$, and θ_0 is the latent true value of the parameter to be estimated.

This is interpretable via the existence of a latent regression model, of the form $Y_n^\star = \mathbf{X}_n\theta + \varepsilon_n$, where $Y_i^\star$ is unobservable (at least for some i), and what is instead observed, is the event $Y_i := \mathbb{I}(Y_i^\star > 0) := \begin{cases} 1, Y_i^\star > 0 \\ 0, Y_i^\star \leq 0 \end{cases}$. Then $\mathbb{P}(Y_i = 1 \mid \sigma(\mathbf{X}_n)) = \mathbb{E}(Y_i \mid \sigma(\mathbf{X}_n) = \mathbb{E}(\mathbb{I}(Y_i^\star > 0) \mid \sigma(\mathbf{X}_n) = \mathbb{P}(Y_i^\star > 0 \mid \sigma(\mathbf{X}_n)) = \mathbb{P}(\mathbf{X}_i'\theta_0) + \varepsilon_i > 0 \mid \sigma(\mathbf{X}_n)) = \mathbb{P}(\varepsilon_i > -\mathbf{X}_i'\theta_0) \mid \sigma(\mathbf{X}_n)) = 1 - \mathbb{P}(\varepsilon_i \leq -\mathbf{X}_i'\theta_0) \mid \sigma(\mathbf{X}_n))$. If the distribution of ε_i conditionally on $\sigma(\mathbf{X}_n)$ is independent of i (conditionally on $\sigma(\mathbf{X}_n)$) the elements of ε are more generally homogeneous), and G is the cdf of this common distribution, then it is obtained that the latter probability is also expressible as $1 - G(-\mathbf{X}_i'\theta_0)$. Thus $F(z) := 1 - G(-z)$. If the distribution

is symmetric, i.e. it holds that for all $z \in \mathbb{R}$, $1 - G(-z) = G(z)$, then the link function is actually the cdf of the conditional on the regressors, marginal distribution of the errors, evaluated at $\mathbf{X}_i'\theta_0$. In both the considered cases below symmetry holds.

Common Binary Response Models and Extensions

Two widely used binary response models are:

- **Logistic Regression**: In this model, the link function $F(\cdot)$ is the cdf of the standard logistic distribution:

$$F(z) = \frac{1}{1 + e^{-z}}.$$

  Logistic regression is popular due to its simplicity and interpretability.

- **Probit Model**: Here, the link function $F(\cdot)$ is the cumulative distribution function (CDF) of the standard normal distribution:

$$F(z) = \Phi(z) := \int_{-\infty}^{z} (2\pi)^{-1} \exp(-\frac{x^2}{2}) dx.$$

  Probit models are often used when the error term in the latent regression is assumed to follow a normal distribution.

Both of these models are parametric (why?). Semi-parametric flavors of binary response models exist; those necessarily leave the link function F, partially unspecified. E.g. some of those models may only assume that G

is symmetric, and/or that it has several analytical properties, e.g. satisfy some form of continuity, it is differentiable/smooth, etc.

Part of the specification of these models is found in the LLS structure for the latent variable $Y_n^\star$. Could this be generalized so as to conform to the NLLS structure? (exercise!)

Applications

Binary response models are widely applied in fields such as:

- *Economics*: Modeling consumer choice, such as whether a product is purchased, or binary decisions of whether to enter or not a labour market, etc.
- *Medicine*: Predicting disease presence or absence based on patient characteristics.
- *Finance*: Assessing credit default risk.

These models provide a flexible framework for analyzing dichotomous outcomes and are a cornerstone of predictive modeling in many domains.

Brief Introduction to Linear Causal Processes and ARMA Models

Linear causal processes and ARMA (AutoRegressive Moving Average) models are fundamental tools in time series analysis, allowing researchers to model, analyze, and predict stochastic processes evolving in time. These models

provide a structured approach to capturing the dependence and dynamics within time-dependent data.

Linear Causal Processes

A linear causal process models a time series as a function of past values and random disturbances. For a stationary process $\{X_t\}$, it can often be expressed as:

$$X_t = \sum_{j=0}^{\infty} \psi_j \epsilon_{t-j},$$

where:

- ϵ_t are independent, identically distributed random shocks, or more generally a white noise with zero mean and finite variance σ^2).
- ψ_j are coefficients determining the impact of past shocks.
- Convergence of the series $\sum_{j=0}^{\infty} \psi_j^2 < \infty$, that the process is well defined with zero mean and variance equal to $\sigma^2 \sum_{j=0}^{\infty} \psi_j^2$. If the white noise process is additionally stationary and ergodic (e.g. iid), then the resulting process is also stationary and ergodic.

This formulation captures how past shocks influence the current state, emphasizing causality in time series evolution.

ARMA Models

ARMA models are a specific class of linear models combining autoregressive (AR) and moving average (MA) components. An ARMA(p, q) process is defined

as:

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + \epsilon_t + \theta_1 \epsilon_{t-1} + \cdots + \theta_q \epsilon_{t-q},$$

where:

- ϕ_i are the autoregressive parameters, capturing dependence on past values of X_t.
- θ_j are the moving average parameters, capturing dependence on past errors ϵ_t.
- ϵ_t represents white noise.

ARMA models provide a flexible yet parsimonious framework for representing stationary time series with short-term dependence. Under stationarity and ergodicity for the white noise process, appropriate restrictions on the autoregressive parameters, ensure that the recurrence equation defining the process, has a unique stationary and ergodic solution in the form of a linear causal process with coefficients depending polynomially of the ARMA coefficients.

Applications

Linear causal processes and ARMA models are widely applied in:

- *Economics*: Modeling and forecasting macroeconomic indicators such as GDP or inflation.
- *Finance*: Analyzing stock prices or interest rate movements.
- *Engineering*: Signal processing and control systems.

- *Environmental Science*: Modeling climate or weather patterns.

A statistical model involving ARMA(p,q) processes, is a collection of suchlike processes, for different values of the associated parameters. For example a statistical model involving stationary and ergodic AR(1) processes, is a collection of the form

$$\{(X_t)_{t\in}, \text{where } X_t = \phi X_{t-1} + \varepsilon_t, \phi \in (-1,1), (\varepsilon_t)_{t\in} \text{ is stationary ergodic White Noise}\}.$$

This is semi-parametric (why?), and adheres to the structure of the LLS example for $Y_n := (X_t)_{t=1,\dots,n}, \mathbf{X}_n := (X_{t-1})_{t=1,\dots,n}, \varepsilon_n = (\varepsilon_t)_{t=1,\dots,n}$, for which however $\mathbb{E}(\varepsilon_t / X_{t-j}, j > 0) = 0$ holds, instead of the stricter $\mathbb{E}(\varepsilon_n / \sigma(\mathbf{X}_n)) = \mathbf{0}_n$. A parametric subset of this statistical model could be obtained by further restricting $\varepsilon_0 \sim N(0, \sigma^2)$, which would directly imply that $X_t / X_{t-j}, j > 0 \sim N(\phi X_{t-1}, \sigma^2)$. Analogous reductions to versions of the LLS examples hold for any model comrpised of AR(p) parameters, for any $p > 0$, but not for models involving non-trivial MA parts, due to the latency of the white noise process.