1 Introduction

In the past fifty years since it first came to prominence, econometrics has broadened, both in the nature of the data being worked with and in the range of issues that it addresses. Initially econometricians were primarily interested in relations between two or more series, but today they find themselves concerned with measuring volatility of financial returns, durations of events, and the conditional probabilities of decisions, *inter alia*. A textbook such as Greene's (1997) captures this expansion extremely well. The same textbook also shows that there is still an underlying unity in the way analysis proceeds, in that the method of linear regression and maximum likelihood form the tool kit of an applied econometrician.

In one area of quantitative economics, namely that concerned with the analysis of observed choices made by economic agents in the sphere of consumption and production, it has long been felt that the investigation of questions such as consistency of the data with the maximization principle, homotheticity, and separability of preferences should not be constrained by the need to make precise assumptions about the nature of preferences or production relations. Probably the earliest manifestation of this concern was Samuelson's (1938) development of the revealed preference theory, and since that time there has been a series of contributions aiming to develop a nonparametric approach to the economics of production and consumption, for example, Afriat (1967), Hanoch and Rothschild (1972), Diewert and Parkan (1978), and Varian (1984). These papers deal with nonparametric tests of the predictions of economic theory without specifying particular functional forms for underlying demand and production relations, something that would be required if one proceeded to test the predictions in the traditional framework (e.g., Greene 1997, Chapter 15). Perhaps the major limitation to these methods has been the fact that the stochastic nature of the data is ignored, although the recent work of Varian (1985) and Epstein and Yatchew (1985) attempts to remove some of these difficulties.

A different approach, more in line with traditional work, has been to locate the source of unease in particular assumptions being made during the application of

the classic techniques. In the case of linear regression, a particular concern has been with the linearity of the functional form connecting the variables appearing in it. This concern initially spawned an interest in transformations of the dependent and independent variables, leading to the use of flexible functional forms, such as those of Diewert (1971), Berndt and Khaled (1979) and Christensen et al. (1973), to approximate the unknown relation. A drawback of this literature was that it only provided local Taylor series approximations. Consequently, it is not surprising that interest arose in improving on such approximations.

To do this it was important to realize that, in many instances, one was attempting to estimate an expectation of one variable, Y, conditional upon others, X. This identification directs attention to the need to be able to estimate the density of Y conditional upon X, as the knowledge of that quantity would enable one to extract the conditional mean. Accordingly, Chapter 2 reviews the procedures that have been advocated for doing that. This is a surprisingly long chapter, partly because there are many different strategies, and partly because all of the techniques involve "tuning parameters" whose determination has been the subject of an enormous literature, which has only now settled down to reflect a consensus view.

Having determined ways of nonparametrically estimating a conditional density, the conditional mean at a point x readily follows as a weighted average $\sum_{i=1}^{n} w(x_i; x)y_i$ of the n data points $\{x_i, y_i\}$; here y_i are observations on the dependent variable and x_i on the independent variable, and $w(x_i; x)$ are a set of weights that depend upon x_i and the point x at which the conditional expectation is to be evaluated. Of course there turn out to be many weighting functions $w(x_i; x)$ that work, and some of the popular ones are detailed in Chapter 3. Broadly these correspond to whether one wishes to have a local (to the point x) or a global approximation. This chapter also shows how the procedures extend to the estimation of any higher order moment, whereas Chapter 4 considers the modification needed if interest centers upon the derivatives of the function linking Y and X, either at a point or as the average over an interval.

Perhaps the major complication in a purely nonparametric (*NP*) approach to estimation is the "curse of dimensionality." Every method has some cost associated with it and, in the instance of nonparametrics, it is the need for very large samples if an accurate measurement of the function is to be made. Moreover, the size of sample required increases rapidly with the number of variables involved in any relation. Such a feature leads to the proposition that one might well prefer to restrict some variables to have a linear impact while allowing a much smaller number to have a nonlinear one. A well-known example of this phenomenon occurs in studies of the wage paid to an individual. The wage is regarded as being influenced by the individual's personal characteristics as well as the number of years of job experience, but, whereas the impact of the personal characteristics is taken to be linear, that for experience is nonlinear. Accordingly, the first part

of Chapter 5 deals with such models, allowing the nonlinearity to be located either in the conditional mean or the conditional variance. Effectively, estimation involves a combination of parametric and nonparametric methods, leading to the estimators being described as *semiparametric* (SP).

In the second part of Chapter 5 the nature of the conditional mean is taken to be known up to a finite number of parameters, and attention switches to the distributional properties of the error term left over after the conditional mean has been extracted. Regression analysis either explicitly or implicitly treats this error as normally distributed, except for those instances in which the dependent variable involves "count" or duration data, whereupon densities such as the negative binomial or Weibull are invoked. Consequently, it is of interest to study the estimation of the parameters of the conditional mean when the error density is unknown. This scenario also falls into the class of semiparametric problems, and a range of concepts have been introduced to categorize the properties of such SP estimators. In parametric models being estimated by maximum likelihood, questions relating to the efficiency of an estimator and its dependence on nuisance parameters are most usefully analyzed with the Cramer–Rao bound and Fisher's information matrix. There are analogous concepts in the SP literature, such as the SP efficiency bound, and the definition and construction of such quantities is laid out in Chapter 5 within the context of the simplest possible environment. Once done, it is natural to seek to design a fully efficient estimator in the face of an unknown density. It is shown that the crucial step in performing such a task is the ability to estimate the "score" of the unknown density – the ratio of the first derivative of the density to the density itself. Hence, the techniques of Chapter 2 are called upon to estimate this unknown variable.

Chapter 6 represents an excursion into the estimation of the parameters of nonlinear simultaneous equations. It has been known for many years that the optimal instruments are related to the expectation of the endogenous variables conditional upon the exogenous ones, but when the system is nonlinear there is no closed form expression for this. Nonparametric techniques therefore appeal as a way of generating the optimal instruments for later use in estimating the unknown parameters. A number of such SP estimators exist in the literature and the sections of this chapter are devoted to an enumeration of them.

The following three chapters concentrate upon some important models in econometrics where semiparametric methods are likely to be popular – binary choices (Chapter 7), censored regression (Chapter 8), and selectivity models (Chapter 9). Mostly, the unknown parameters of these models have been estimated by maximum likelihood. We therefore seek to describe estimators that do not make assumptions about the density of the observations, with particular emphasis being given to the construction of an estimator that attains the SP efficiency bound. It is not always possible to find the latter, and that leads to

the development of a range of alternatives that might be expected to produce good results without necessarily being optimal. Throughout these chapters a common strategy, first used in Chapter 5, is employed. This involves describing what a parametric estimator would look like, and then seeking to replace the unknown quantities in such an expression with nonparametric estimates. Our experience has shown this to be a valuable discipline when considering nonparametric issues. In many instances it is worth enquiring into the limitations of parametric models before engaging in the more general problems that arise with a nonparametric orientation. Finally, for the benefit of the readers, we have included an Appendix which contains basic concepts, definitions and results of Statistics and Probability which have been used in the book.

It is useful to close this introduction with a word about the scope of the book. Our primary objective is to present a survey of the NP and SP literature that practitioners might find useful. It is not our intention to provide an account of the theoretical tools that one would need to conduct research in this area. A book that did that would treat the subject with much more rigor than we have tried to do. Indeed, to specialists in the area the degree of rigor of this book may be distressingly low, but we feel that it is more important to isolate the essentials in some of the theory than to worry about it being completely rigorous. Moreover, it is our belief that, when the theory is made rigorous, it becomes almost impossible to see the "wood for the trees."

Nevertheless, the theoretical material in the book is not insignificant, being at the level of complexity of a second-year graduate econometrics course. This raises the issue of why we spend time on these matters rather than just providing a "cook book" that would describe the different approaches – as for example the excellent book on nonparametrics by Härdle (1990). Essentially, this is because we feel that something important is lost with such an orientation. There are issues raised in the NP and SP literatures that we do not come across in parametric literature and, unless one grasps these, it is hard to fully comprehend the nature of NP and SP methods. As an example, one might cite the "bias" problem of NP estimators that recurs throughout the book. In the parametric estimation context, we are used to the idea that, when suitably normalized by some function of the sample size, estimators are asymptotically normally distributed around the true value of the parameters. This is not true for NP estimators, and strategies to eliminate the bias end up accounting for many of the choices made in both the SP and NP literature. Consequently, understanding the theory can be important if one wishes to use the methods, although we feel that this can be done by capturing the flavor of the arguments rather than presenting their rigorous underpinnings.

2 Methods of Density Estimation

2.1 Introduction

This chapter describes various methods of estimating the univariate density function of a random variable, closing with extensions to the multivariate case. Some motivation needs to be given for why we should be interested in density estimation at all. An important reason is that the techniques used in, and the complications arising from, the nonparametric estimation of densities recur many times in later chapters, and it pays to study them in a simplified setting first. But, apart from this pragmatic purpose, the need to estimate densities does arise in practice sufficiently often to make a study of this literature of interest in its own right.

Broadly, one can distinguish three areas in which the need to estimate densities arises. First, density estimates can be important in capturing the stylized facts that need explanation and for judging how well a potential model is likely to fit the data. For example, if it is known that the variable being examined has a density with fat tails, or strong peaks, any model of data corresponding to such a variable needs to be capable of generating a density with this characteristic. In other instances, one can efficiently learn about interrelationships between variables in large data sets from joint density estimates – a feature well illustrated in Deaton's (1989) work on rice subsidies in Thailand, in Marron and Schmitz's (1992) work on the U.K. income distribution, in Dinardo et al.'s (1996) study on the U.S. distribution of wages conditional on labor market institutions, and in Quah's (1997) cross-country analysis of the growth and convergence of economies.

Second, it is often desirable to perform a Monte Carlo analysis of a particular estimator being used in a study. Traditionally, only a few moments of this estimator are recorded or a test statistic such as Kolmogorov–Smirnov's is provided to assess departures from normality. Nonparametric density estimates, however, enable a complete picture of the distribution of the estimator and therefore seem a preferable way of summarizing the outcome of a Monte

Carlo experiment. An illustration of this point is given as an example in the concluding section of the chapter.

Finally, it is sometimes the case that parametric estimators have an asymptotic distribution that depends on a density evaluated at a specific point. For example, the median of X has variance $.25\, n^{-1} f^{-2}(0)$, where $f(0)$ is the density of X evaluated at $x = 0$. Hence, any test statistic involving the median demands an estimate of $f(0)$. Section 5.9 presents other estimators for which a density estimate at a point is required.

As before, $f = f(x)$ denotes the continuous density function of a random variable X at a point x, and $x_1, \ldots, x_n$ are the observations drawn from f. Two general methods have been advanced for the estimation of f.[1]

(i) Parametric Estimators: Parametric methods specify a form for f, say, the normal density,

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right],$$

where the mean μ and the variance σ^2 are the parameters of f. An estimator of f can be written as

$$\hat{f}(x) = \frac{1}{\hat{\sigma}\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\hat{\mu}}{\hat{\sigma}}\right)^2\right],$$

where μ and σ are estimated consistently from data as

$$\hat{\mu} = \overline{p}x = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{and} \quad \hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{p}x)^2,$$

respectively.

(ii) Nonparametric Estimators: A disadvantage of the parametric method is the need to stipulate the true parametric density of f. In the nonparametric alternative $f(x)$ is directly estimated without assuming its form. The histogram is one such estimator, and it is one of the oldest methods of density estimation (Van Ryzin (1973) and Scott (1979) among others). But, although the histogram is a useful method of density estimation, it has the drawbacks of being discontinuous and too "rough." Further, it is extremely complicated to use for two or more variables. In view of these disadvantages, in the past three decades several nonparametric estimators have been developed with the aim of producing "smooth" estimates of $f(x)$.

[1] There have been suggestions to combine the two (e.g., Olkin and Spiegelman, 1987).

Section 2.2 of the chapter sets out a variety of ways of nonparametrically computing a density estimate; Section 2.3 concentrates upon the modifications needed if it is a derivative of the density which is of interest. An example of the need for the latter arises when estimating the "score" of the density, $f^{-1}(u)\partial f/\partial u$, a quantity that appears many times in later chapters, making it important to discuss its estimation at an early stage. Sections 2.4–2.6 deal with the sampling properties of the most widely used nonparametric estimator, the *kernel* method. Many of the complications that arise in describing the distributions of nonparametric estimators occur in this simple problem, so that some time is spent studying them. As these sections demonstrate, the important elements in nonparametrics are the need to choose a smoothing function – the kernel – and a parameter – the window width – and Section 2.7 discusses the extensive literature on how to make these choices in practice. Section 2.8 outlines extensions of the ideas to multivariate density estimation, and Section 2.9 looks at the techniques that have developed for testing whether a nonparametrically estimated density has a specified parametric form or whether two estimated densities are close; that is, it focuses upon measures of the affinity of densities. Finally, Section 2.10 provides a few examples.

2.2 Nonparametric Density Estimation

There is no unique way to perform nonparametric density estimation, and some eight approaches are described in this section. Despite this variety it is possible to achieve a degree of unification by placing each estimator in a common format, namely as the sample mean of certain functions of the data.

2.2.1 A "Local" Histogram Approach

To understand some of the density estimation techniques discussed later we begin with the situation when X is a discrete random variable. Let one of the values it can assume be x and our purpose is to estimate $f(x)$ from the data x_i, $i = 1, \ldots n$. Estimation of $f(x)$ in the discrete case is essentially the estimation of the proportion of x values in the population of X. From the data $x_1, \ldots, x_n$ an obvious and well-known consistent estimator of this is the sample proportion $\hat{f}_1(x) = n^*/n$, where n^* is the number of $x_1, \ldots, x_n$ equal to x. Alternatively, $\hat{f}_1(x) = n^{-1}\sum_{i=1}^{n} I(x_i = x)$, with $I(x_i = x)$ being an indicator function taking the value 1 if $x_i = x$ and zero otherwise.

Now, considering the case where X is a continuous random variable, the probability that x_i is equal to x is zero, and $f(x)$ will need to be estimated by averaging those x_i that are in an interval around x, say, $x \pm h/2$, where h is the

width of the interval. Thus the empirical density estimator $\hat{f}(x)$ can be written as $\hat{f}_1(x) = (nh)^{-1}\sum_{i=1}^{n} I(x - \frac{h}{2} \leq x_i \leq x + \frac{h}{2})$, where $I(\mathcal{A}) = 1$ if $\mathcal{A}$ is true and zero otherwise. Alternatively, we can write

$$\hat{f}_1(x) = \frac{1}{nh}\sum_{i=1}^{n} I\left(-1/2 \leq \frac{x_i - x}{h} \leq 1/2\right)$$

$$= \frac{1}{nh}\sum_{i=1}^{n} I\left(-1/2 \leq \psi_i \leq 1/2\right), \tag{2.1}$$

where $\psi_i = (x_i - x)/h$.

Notice that $\hat{f}_1(x)$ in (2.1) is the per unit relative frequency in the interval $(x - h/2,\ x + h/2)$ whose midpoint is x. In this sense it is exactly the ordinate of the histogram at x. Thus the estimator (2.1) can be seen to be an attempt to construct a histogram that is based on the observations "local" to x, and where every point x is the center of a sampling interval. The width of the interval h controls the amount by which the data are smoothed (averaged) to produce the estimate (2.1). The $\hat{f}_1$ is also known as the "naive" estimator, following Fix and Hodges (1951).

Clearly the indicator or weight function $I(-1/2 < \psi_i < 1/2)$ in (2.1) depends upon the distance of x_i from x. If this absolute distance is less than or equal to $1/2$ the weight is 1; otherwise it is zero. Furthermore, the weight function $I(\psi) = I(-1/2 < \psi < 1/2)$ is such that

$$\int_{-\infty}^{\infty} I(\psi)\, d\psi = \int_{-\infty}^{-1/2} I(\psi)\, d\psi + \int_{-1/2}^{1/2} I(\psi)\, d\psi + \int_{1/2}^{\infty} I(\psi)\, d\psi$$

$$= \int_{-1/2}^{1/2} I(\psi)\, d\psi = \int_{-1/2}^{1/2} d\psi = 1. \tag{2.2}$$

Thus

$$\int_{-\infty}^{\infty} \hat{f}_1(x)\, dx = \frac{1}{nh}\sum_{i=1}^{n} \int_{-\infty}^{\infty} I\left(-\frac{1}{2} < \frac{x_i - x}{h} < \frac{1}{2}\right) dx$$

$$= \frac{1}{n}\sum_{i=1}^{n} \int_{-\infty}^{\infty} I\left(-\frac{1}{2} < \psi_i < \frac{1}{2}\right) d\psi_i = 1, \tag{2.3}$$

and the density estimate is proper in that it is nonnegative and integrates to unity. A feature of (2.3) is that the integral was taken over x since it can assume values over the whole range of X.

2.2.2 *A Formal Derivation of* $\hat{f}_1(x)$

Let $F(x) = P(X \leq x)$ denote the cumulative probability distribution function of X. Then the density function $f(x)$ is defined by

$$f(x) = \frac{d}{dx}F(x) = \lim_{h\to 0}\frac{F\left(x+\frac{h}{2}\right) - F\left(x-\frac{h}{2}\right)}{h}$$
$$= \lim_{h\to 0}\frac{P\left(x-\frac{h}{2} < X < x+\frac{h}{2}\right)}{h}. \tag{2.4}$$

Our problem is to estimate $f(x)$ based on $x_1, \ldots, x_n$. For this we consider h to be a positive function of n that goes to zero as $n \to \infty$, and estimate $P(x - \frac{h}{2} < X < x + \frac{h}{2})$ by the proportion of sample observations $x_1, \ldots, x_n$ falling in $(x-\frac{h}{2}, x+\frac{h}{2})$. Then an obvious consistent estimator of $f(x)$ in (2.4) is

$$\hat{f}_2(x) = \frac{1}{nh}\left[\text{number of } x_1, \ldots, x_n \text{ in } \left(x - \frac{h}{2},\ x + \frac{h}{2}\right)\right]$$
$$= \frac{1}{nh}\left[\text{number of } \frac{x_1 - x}{h}, \ldots, \frac{x_n - x}{h} \text{ in } (-1/2,\ 1/2)\right]$$
$$= \hat{f}_1(x), \tag{2.5}$$

which is the same as (2.1). The estimator in (2.5) was first proposed by Fix and Hodges (1951).

2.2.3 *Rosenblatt–Parzen Kernel Estimator*

The density estimator produced by the indicator function in (2.1) has the property that it integrates to unity, but has the disadvantage of being "rough." Also, $\hat{f}_1(x)$ is not a continuous function but has jumps at the points $x_i \pm h/2$ with zero derivative elsewhere. This gives estimates a stepwise nature, and one might prefer a smoother set of weights. Rosenblatt (1956b) addressed this issue by replacing the indicator function in (2.1) with a real positive kernel function K satisfying

$$\int_{-\infty}^{\infty} K(\psi)\, d\psi = 1. \tag{2.6}$$

His general "kernel" estimator is

$$\hat{f}(x) = \hat{f}_3(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x_i - x}{h}\right) = \frac{1}{nh}\sum_{i=1}^{n} K(\psi_i), \tag{2.7}$$

where $\psi_i = h^{-1}(x_i - x)$ as in (2.1) and h, the window-width (also called the smoothing parameter or band width), is a function of the sample size n and goes to zero as $n \to \infty$.

A number of features that a kernel should possess can be inferred from the nature of the indicator function. First, for large values of $|\psi_i|$ (i.e., x_i lies far from x) $K(\psi_i)$ should be small, as very small weights need to be assigned to such data points in constructing the density estimate. In particular, because $h \to 0$ when $n \to \infty$, it follows that $|\psi_i| \to \infty$ for any $x_i \neq x$, and therefore $K(-\infty) = K(\infty) = 0$, which is implied by the requirement (2.6). This feature reproduces the "zero" property of the indicator function, whereas the "unity" part is exhibited by having $\int_{-\infty}^{\infty} K(\psi)\, d\psi = 1$. This amounts to replacing a square centered on x with length of unity by a smooth curve, also centered on x with the same area but no longer necessarily having bounded support. Moreover, because these features are those of a density function, kernels are frequently chosen to be well-known density functions, for example, the standard normal $K(\psi) = (2\pi)^{-1/2} \exp(-.5\psi^2)$. In this vein the indicator function could be thought of as a kernel estimator with $K(\cdot)$ being the uniform density over $[-1/2,\ 1/2]$. Parzen (1962) pointed out that allowing $K(\cdot)$ to be negative could reduce the bias of the estimator $\hat{f}$, a theme taken up in Section 2.4.3. A disadvantage with allowing the kernel to be negative is that $\hat{f}_3(x)$ may now be negative, and this may be unsatisfactory for some purposes. There is a vast literature on kernels. Silverman (1986) and Härdle (1990) are very good guides to this. Mostly the nature of K is not critical to analysis, and the "optimal" kernel, discussed in Section 2.4.2, will be found to yield only modest improvements in the performance of $\hat{f}(x)$ over selections such as the standard normal.

Suppose that $K(\cdot)$ is restricted to be the standard normal. Then it is well known that $K(\psi) \approx 0$ for $|\psi| \geq 3$, and it is apparent that the weights used in (2.7) depend vitally upon the window width h. For an $x_i \neq x$, whether $\psi_i = (x_i - x)/h$ is less than or greater than 3 will depend solely upon h. Although it is true that $h \to 0$ as $n \to \infty$, in practice this still leaves the problem of determining exactly how h should vary with n. Section 2.4 contains an extensive analysis of this issue.

Usually, but not always, K will be a symmetric density function (e.g., the standard normal density). Moreover, as long as it is everywhere nonnegative and satisfies $\int K(\psi)\, d\psi = 1$, $\hat{f}_3$, like $\hat{f}_1$, will be a probability density in that $\int \hat{f}_3(x)\, dx = 1$. The kernel estimator $\hat{f}_3$ also possesses all the continuity and differentiability properties of the kernel K. This is unlike $\hat{f}_1$, which has jumps at $x_i \pm h/2$ and zero derivatives everywhere else. It is the fact that it produces a smooth function out of a discontinuous one that makes the kernel attractive, and a number of times in later chapters it will prove to be advantageous to replace indicator functions by appropriate kernels, even when the context is not specifically density estimation.