Foundations and Trends® in
Econometrics
Vol. 3, No 1 (2008) 1–88

DOI: 10.1561/0800000009

Nonparametric Econometrics: A Primer

Jeffrey S. Racine

Department of Economics, McMaster University, 1280 Main Street West, Hamilton, Ontario, Canada L8S 4M4, racinej@mcmaster.ca

Abstract

This review is a primer for those who wish to familiarize themselves with nonparametric econometrics. Though the underlying theory for many of these methods can be daunting for some practitioners, this article will demonstrate how a range of nonparametric methods can in fact be deployed in a fairly straightforward manner. Rather than aiming for encyclopedic coverage of the field, we shall restrict attention to a set of touchstone topics while making liberal use of examples for illustrative purposes. We will emphasize settings in which the user may wish to model a dataset comprised of continuous, discrete, or categorical data (nominal or ordinal), or any combination thereof. We shall also consider recent developments in which some of the variables involved may in fact be irrelevant, which alters the behavior of the estimators and optimal bandwidths in a manner that deviates substantially from conventional approaches.

1

Introduction

Nonparametric methods are statistical techniques that do not require a researcher to specify functional forms for objects being estimated. Instead, the data itself informs the resulting model in a particular manner. In a regression framework this approach is known as "nonparametric regression" or "nonparametric smoothing." The methods we survey are known as kernel[1] methods. Such methods are becoming increasingly popular for applied data analysis; they are best suited to situations involving large data sets for which the number of variables involved is manageable. These methods are often deployed after common parametric specifications are found to be unsuitable for the problem at hand, particularly when formal rejection of a parametric model based on specification tests yields no clues as to the direction in which to search for an improved parametric model. The appeal of nonparametric methods stems from the fact that they relax the parametric assumptions imposed on the data generating process and let the data determine an appropriate model.

[1] A "kernel" is simply a weighting function.

Nonparametric and semiparametric methods have attracted a great deal of attention from statisticians in the past few decades, as evidenced by the vast array of texts written by statisticians including Prakasa Rao (1983), Devroye and Györfi (1985), Silverman (1986), Scott (1992), Bickel et al. (1993), Wand and Jones (1995), Fan and Gijbels (1996), Simonoff (1996), Azzalini and Bowman (1997), Hart (1997), Efromovich (1999), Eubank (1999), Ruppert et al. (2003), Härdle et al. (2004), and Fan and Yao (2005). However, the number of texts tailored to the needs of applied econometricians is relatively scarce including, Härdle (1990), Horowitz (1998), Pagan and Ullah (1999), Yatchew (2003), and Li and Racine (2007a) being those of which we are currently aware.

The first published paper in kernel estimation appeared in 1956 (Rosenblatt (1956)), and the idea was proposed in an USAF technical report as a means of liberating discriminant analysis from rigid parametric specifications (Fix and Hodges (1951)). Since then, the field has undergone exponential growth and has even become a fixture in undergraduate textbooks (see, e.g., Johnston and DiNardo (1997, Chap. 11)), which attests to the popularity of the methods among students and researchers alike.

Though kernel methods are popular, they are but one of many approaches toward the construction of flexible models. Approaches to flexible modeling include spline, nearest neighbor, neural network, and a variety of flexible series methods, to name but a few. In this article, however, we shall restrict attention to the class of nonparametric kernel methods, and will also touch on semiparametric kernel methods as well. We shall also focus on more practical aspects of the methods and direct the interested reader to Li and Racine (2007a) and the references listed above for details on the theoretical underpinnings in order to keep this review down to a manageable size.

It bears mentioning that there are two often heard complaints regarding the state of nonparametric kernel methods, namely, (1) the lack of software, and (2) the numerical burden associated with these methods. We are of course sympathetic to both complaints. The latter may unavoidable and simply be "the nature of the beast" as they say, though see *Computational Considerations* for a discussion of the issues. However, the former is changing and recent developments hold the promise for computational breakthroughs. Many statistical software packages now contain some elementary nonparametric methods (one-dimensional density estimation, one-dimensional regression) though they often use rule-of-thumb methods for bandwidth selection which, though computationally appealing, may not be robust choices in all applications. Recently, an R (R Development Core Team (2007)) package "np" has been created that provides an easy to use and open platform for kernel estimation, and we direct the interested reader to Hayfield and Racine (2007) for details. All examples in this review were generated using the np package, and code to replicate these results is available upon request.

2

Density and Probability Function Estimation

The notation and the basic approaches developed in this section are intended to provide the foundation for the remaining ones, and these concepts will be reused throughout this review. More detail will therefore be presented here than elsewhere, so a solid grasp of key notions such as "generalized product kernels," kernels for categorical data, data-driven bandwidth selection and so forth ought to be helpful when digesting the material that follows.

Readers will no doubt be intimately familiar with two popular nonparametric estimators, namely the histogram and frequency estimators. The histogram is a non-smooth nonparametric method that can be used to estimate the probability density function (PDF) of a continuous variable. The frequency probability estimator is a non-smooth nonparametric method used to estimate probabilities of discrete events. Though non-smooth methods can be powerful indeed, they have their drawbacks. For an in-depth treatment of kernel density estimation we direct the interested reader to the wonderful reviews by Silverman (1986) and Scott (1992), while for mixed data density estimation we direct the reader to Li and Racine (2007a) and the references therein. We shall begin with an illustrative *parametric* example.

2.1 Parametric Density Estimators

Consider any random variable X having probability density function $f(x)$, and let $f(\cdot)$ be the object of interest. Suppose one is presented with a series of independent and identically distributed draws from the unknown distribution and asked to model the density of the data, $f(x)$. This is a common situation facing the applied researcher.

For this example we shall simulate $n = 500$ draws but immediately discard knowledge of the true data generating process (DGP) pretending that we are unaware that the data is drawn from a mixture of normals ($N(-2, 0.25)$ and $N(3, 2.25)$ with equal probability). We then (naïvely) presume the data is drawn from, say, the normal parametric family, namely

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

We then estimate this model and obtain $\hat{\mu} = 0.56$ and $\hat{\sigma} = 2.71$. Next, as is always recommended, we test for correct specification using, say, the Shapiro–Wilks test and obtain $W = 0.88$ with a p-value of $< 2.2e - 16$, rejecting this parametric model out of hand. The estimated model and true DGP are plotted in Figure 2.1.

Given that this popular parametric model is flatly rejected by this dataset, we have two choices, namely (1) search for a more appropriate parametric model or (2) use more flexible estimators.

For what follows, we shall presume that the reader has found themselves in just such a situation. That is, they have faithfully applied a parametric method and conducted a series of tests of model adequacy that indicate that the parametric model is not consistent with the underlying DGP. They then turn to more flexible methods of density estimation. Note that though we are considering density estimation at the moment, it could be virtually any parametric approach that we have been discussing, for instance, regression analysis.

2.2 Histograms and Kernel Density Estimators

Constructing a histogram is straightforward. First, one constructs a series of bins (choose an origin x_0 and bin width h). The bins are

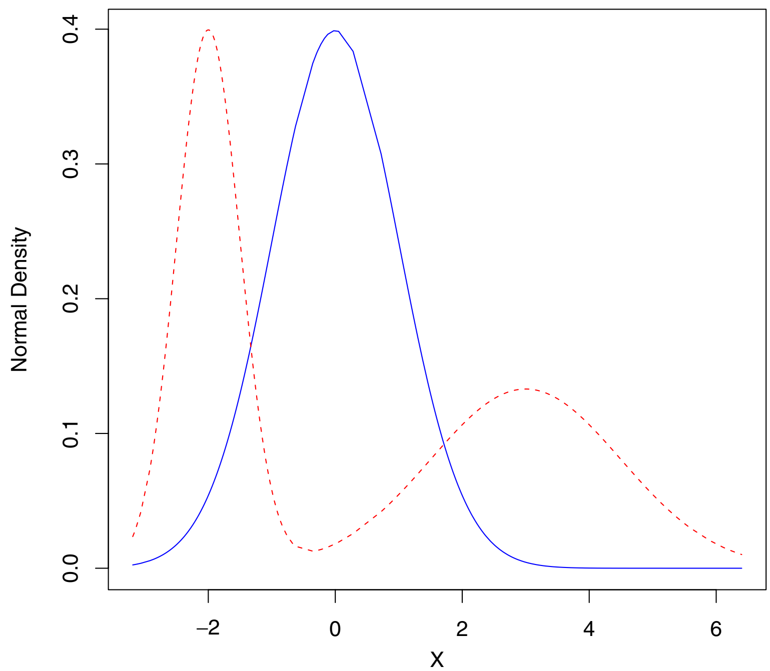

Fig. 2.1 The $N(0.56, 2.71^2)$ density estimate (unimodal, solid line) and true data generating process (bimodal, dashed line).

the intervals $[x_0 + mh, x_0 + (m+1)h)$ for positive and negative integers m. The histogram is defined as

$$\begin{aligned} \hat{f}(x) &= \frac{1}{n} \frac{(\# \text{ of } X_i \text{ in the same bin as } x)}{\text{width of bin containing } x} \\ &= \frac{1}{nh} \sum_{i=1}^{n} \mathbf{1}(X_i \text{ is in the same bin as } x), \end{aligned} \tag{2.1}$$

where $\mathbf{1}(A)$ is an indicator function taking on the value 1 if A is true, zero otherwise. The user must select the origin and bin width, and the resulting estimate is sensitive to both choices. Rules of thumb are typically used for both. Though extremely powerful, there is much room for improvement. The histogram is not particularly efficient, statistically speaking. It is discontinuous, hence any method based upon it requiring derivatives will be hampered by this property. As well, it is not centered on the point at which the density estimate is desired. Though the histogram is a wonderful tool, kernel methods provide an alternative which we shall explore.

The univariate kernel density estimator was constructed to overcome many of the limitations associated with the histogram. It involves nothing more than replacing the indicator function in (2.1) with a symmetric weight function $K(z)$, a "kernel," possessing a number of useful properties. Replacing the indicator function in (2.1) with this kernel function yields

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right). \tag{2.2}$$

This estimator is often called the Rosenblatt–Parzen estimator (Rosenblatt (1956), Parzen (1962)). Figure 2.2 presents the histogram and Rosenblatt–Parzen estimates for the simulated data used in Section 2.1, with bandwidth obtained via Sheather and Jones's (1991) plug-in method (see Section 2.3.2).

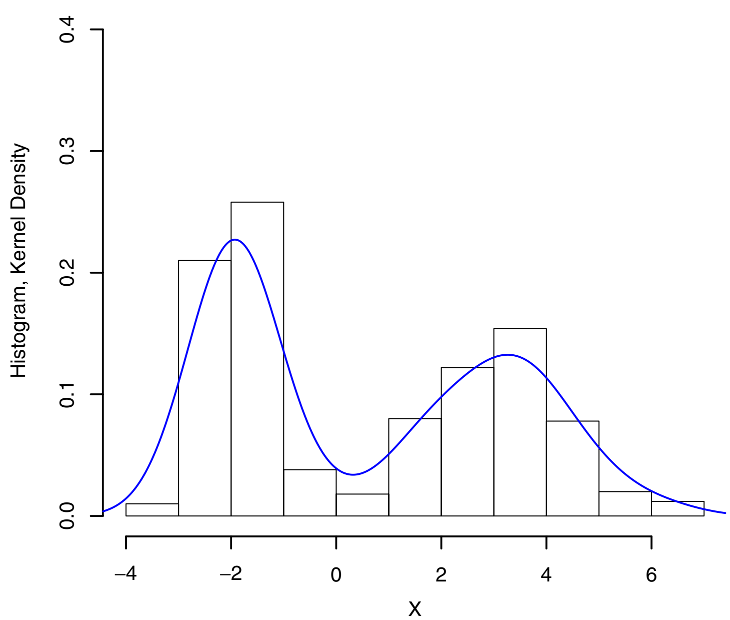

Fig. 2.2 Histogram and kernel estimates of a univariate density function.

Figure 2.2 reveals that both the histogram and Rosenblatt–Parzen estimates readily reveal the bimodal nature of the underlying data, unlike the misspecified unimodal parametric model presented in Figure 2.1. The reader who compares Figures 2.1 and 2.2 will immediately notice that both the histogram and kernel estimator are *biased*, that is, they appear to underestimate the left peak in finite-samples, and indeed they will do so systematically as will be seen below when we consider the properties of the Rosenblatt–Parzen estimator. But, as n increases and h decreases in a particular manner to be outlined shortly, the kernel estimator will converge to the true DGP with probability one. The misspecified parametric model can never converge to the true DGP. Which method provides a more appropriate description of the DGP, the unimodal parametric model or the bimodal nonparametric model?[1] This issue is taken up in Section 2.7.

The kernel estimation of an unconditional cumulative distribution function (CDF) has received much less attention than that of the PDF. We direct the interested reader to the seminal paper by Bowman et al. (1998) and to Li and Racine (2007a, Chap. 1).

2.2.1 Properties of the Univariate Kernel Density Estimator

Presume the kernel function $K(z)$ is nonnegative and satisfies

$$\int K(z)\,dz = 1, \quad \int zK(z)\,dz = 0, \quad \int z^2 K(z)\,dz = \kappa_2 < \infty.$$

Unless otherwise indicated, the lower and upper limits of integration shall be $-\infty$ and ∞, respectively. This kernel is often called a "second order kernel." Parzen (1962) demonstrated that one can choose kernels that can potentially reduce the pointwise bias of $\hat{f}(x)$, however one must forgo the nonnegativity of $K(z)$ in order to do so. One drawback of using such "higher order" kernels[2] in a density

[1] G.E.P. Box's sentiment that "all models are wrong, but some are useful" is perhaps relevant here (Draper, 1987, p. 424).

[2] A general νth order kernel ($\nu \geq 2$ is an integer) must satisfy $\int K(z)\,dz = 1$, $\int z^l K(z)\,dz = 0$, $(l = 1, \dots, \nu - 1)$, and $\int z^\nu K(z) dz = \kappa_\nu \neq 0$.

context is that negative density estimates can be encountered which is clearly an undesirable side effect. Higher order kernels are sometimes encountered in multivariate settings to ensure rates of convergence necessary for establishing limit distributions. For what follows we are presuming that one is using a second-order kernel unless otherwise indicated.

The pointwise mean square error (MSE) criterion is used for assessing the properties of many kernel methods. We proceed by deriving both the bias and variance of $\hat{f}(x)$ to thereby have an expression for the MSE. Recalling that

$$\text{mse}\hat{f}(x) = E\{\hat{f}(x) - f(x)\}^2 = \text{var}\hat{f}(x) + \{\text{bias}\hat{f}(x)\}^2,$$

using a Taylor series expansion and a change of variables we can obtain the approximate bias, which is

$$\text{bias } \hat{f}(x) \approx \frac{h^2}{2} f''(x)\kappa_2, \tag{2.3}$$

and the approximate variance, which is

$$\text{var}\hat{f}(x) \approx \frac{f(x)}{nh} \int K^2(z)dz. \tag{2.4}$$

See Pagan and Ullah (1999, pp. 23–24) or Li and Racine (2007a, pp. 11–12) for a detailed derivation of these results.

Note that both the bias and variance depend on the bandwidth (bias falls as h decreases, variance rises as h decreases). The bias also increases with $f''(x)$, hence is highest in the peaks of distributions. But, as long as the conditions for consistency are met, namely $h \to 0$ as $n \to \infty$ (bias $\to 0$) and $nh \to \infty$ as $n \to \infty$ (var $\to 0$), then the bias related to $f''(x)$ will diminish as the available data increases and will vanish in the limit. Note that nh is sometimes called the "effective sample size," and the requirement that $nh \to \infty$ as $n \to \infty$ simply requires that as we get more information ($n \to \infty$) we average over a narrower region ($h \to 0$) but the amount of "local information" (nh) must increase at the same time.

The above formulas for the bias, variance, and mean square error are *pointwise* properties, i.e., they hold at any point x. The integrated mean square error (IMSE), on the other hand aggregates the MSE over the entire domain of the density yielding a global error measure, and using the approximate bias and variance expressions given above can be defined as

$$
\begin{aligned}
\operatorname{imse}\hat{f}(x) &= \int \operatorname{mse}\hat{f}(x)dx \\
&= \int \operatorname{var}\hat{f}(x)\,dx + \int \left\{\operatorname{bias}\hat{f}(x)\right\}^2 dx \\
&\approx \int \left[\frac{f(x)}{nh} \int K^2(z)dz + \left\{ \frac{h^2}{2} f''(x)\kappa_2 \right\}^2 \right] dx \\
&= \frac{1}{nh} \int K^2(z)dz \int f(x)\,dx + \left\{ \frac{h^2}{2}\kappa_2 \right\}^2 \int \left\{ f''(x) \right\}^2 dx \\
&= \frac{\Phi_0}{nh} + \frac{h^4}{4}\kappa_2^2\Phi_1,
\end{aligned}
\tag{2.5}
$$

where $\Phi_0 = \int K^2(z)dz$ and $\Phi_1 = \int\{f''(x)\}^2 dx$. See Pagan and Ullah (1999, p. 24) or Li and Racine (2007a, p. 13) for a detailed derivation of this result.

We can now minimize this with respect to the bandwidth and kernel function to obtain “optimal bandwidths” and “optimal kernels.” This expression also provides a basis for data-driven bandwidth selection. Note that by using IMSE rather than MSE we are selecting the bandwidth to provide a good “overall” estimate rather than one that is good for just one point.

We obtain a bandwidth which globally balances bias and variance by minimizing IMSE with respect to h, i.e.,

$$
\begin{aligned}
h_{\text{opt}} &= \Phi_0^{1/5}\kappa_2^{-2/5}\Phi_1^{-1/5}n^{-1/5} \\
&= \left\{ \frac{\int K^2(z)dz}{\left(\int z^2K(z)dz\right)^2 \int \left\{f''(x)\right\}^2 dx} \right\}^{1/5} n^{-1/5} = cn^{-1/5}.
\end{aligned}
\tag{2.6}
$$

Note that the constant c depends on $f''(x)$ and $K(\cdot)$, and that if $h \propto n^{-1/5}$ then

$$
o\left(\frac{1}{nh}\right) = o\left(\frac{1}{n^{4/5}}\right)
$$

that is, using the optimal window width yields an estimator $\hat{f}(x)$ which has IMSE of order $n^{-4/5}$, i.e.,

$$\hat{f}(x) - f(x) = O_p(n^{-2/5}),$$

where $O_p(\cdot)$ is defined in *Background Material*. Note that for a *correctly specified* parametric estimator, say $\hat{f}(x, \theta)$, we would have

$$\hat{f}(x, \theta) - f(x) = O_p(n^{-1/2}),$$

which is a faster rate of convergence than the nonparametric rate which is why such models are called $\sqrt{n}$-consistent. Of course, if the parametric model is misspecified, the parametric model is no longer consistent, which is why (Robinson, 1988, p. 933) refers to such models as "$\sqrt{n}$-inconsistent."

Having obtained the optimal bandwidth, we next consider obtaining an optimal kernel function. The primary role of the kernel is to impart smoothness and differentiability on the resulting estimator. In a different setting, Hodges and Lehmann (1956) first demonstrated that a weighting function that is IMSE-optimal is given by

$$K_e(z) = \begin{cases} \frac{3}{4\sqrt{5}}\left(1 - \frac{1}{5}z^2\right) & -\sqrt{5} \le z \le \sqrt{5} \\ 0 & \text{otherwise.} \end{cases}$$

This result is obtained using calculus of variations, and a derivation can be found in Pagan and Ullah (1999, pp. 27–28). This was first suggested in the density estimation context by Epanechnikov (1969) and is often called the "Epanechnikov kernel." It turns out that a range of kernel functions result in estimators having similar relative efficiencies,[3] so one could choose the kernel based on computational considerations, the Gaussian kernel being a popular choice.

Unlike choosing a kernel function, however, choosing an appropriate bandwidth is a crucial aspect of sound nonparametric analysis.

2.3 Bandwidth Selection

The key to sound nonparametric estimation lies in selecting an appropriate bandwidth for the problem at hand. Though the kernel function

[3] See Silverman (1986, p. 43, Table 3.1).

remains important, its main role is to confer differentiability and smoothness properties on the resulting estimate. The bandwidth, on the other hand, drives the finite-sample behavior in a way that the kernel function simply cannot. There are four general approaches to bandwidth selection, (1) reference rules-of-thumb, (2) plug-in methods, (3) cross-validation methods, and (4) bootstrap methods. We would be negligent if we did not emphasize the fact that data-driven bandwidth selection procedures are not guaranteed always to produce good results. For simplicity of exposition, we consider the univariate density estimator for continuous data for what follows. Modification to admit multivariate settings and a mix of different datatypes follows with little modification, and we direct the interested reader to Li and Racine (2003) for further details on the mixed data density estimator.

2.3.1 Reference Rule-of-Thumb

Consider for the moment the estimation of the univariate density function defined in (2.2), whose optimal bandwidth is given in (2.6). A quick peek at (2.6) reveals that the optimal bandwidth depends on the underlying density, which is unknown. The reference rule-of-thumb for choosing the bandwidth uses a standard family of distributions to assign a value to the unknown constant $\int f''(z)^2\,dz$. For instance, for the normal family it can be shown that $\int f''(z)^2\,dz = \frac{3}{8\sqrt{\pi}\sigma^5}$. If you also used the Gaussian kernel, then

$$\int K^2(z)dz = \frac{1}{\sqrt{4\pi}}, \quad \int z^2K(z)dz = 1,$$

so the optimal bandwidth would be

$$h_{\mathrm{opt}} = (4\pi)^{-1/10}\left(\frac{3}{8}\right)^{-1/5}\pi^{1/10}\sigma n^{-1/5} = 1.059\sigma n^{-1/5},$$

hence the "$1.06\sigma n^{-1/5}$" rule-of-thumb. In practice we use $\hat{\sigma}$, the sample standard deviation.

2.3.2 Plug-in

Plug-in methods such as that of Sheather and Jones (1991) involve plugging estimates of the unknown constant $\int f''(z)^2\,dz$ into the optimal bandwidth formula based on an initial estimator of $f''(z)$ that itself is based on a "pilot" bandwidth such as the $1.06\sigma n^{-1/5}$ reference rule-of-thumb. All other constants in h_{opt} are known as we provide the kernel function (i.e., $\int K^2(z)dz$ and $\int z^2K(z)dz$ are known). Though such rules are popular, we direct the interested reader to Loader (1999) for a discussion of the relative merits of plug-in bandwidth selectors versus those discussed below.[4]

2.3.3 Least Squares Cross-Validation

Least squares cross-validation is a fully automatic and data-driven method of selecting the smoothing parameter. This method is based on the principle of selecting a bandwidth that minimizes the IMSE of the resulting estimate. The integrated squared difference between $\hat{f}(x)$ and $f(x)$ is

$$\int \left\{\hat{f}(x) - f(x)\right\}^2 dx = \int \hat{f}(x)^2\, dx - 2\int \hat{f}(x)f(x)\, dx + \int f(x)^2\, dx.$$

We can replace these values with sample counterparts and adjust for bias and obtain an objective function that can be numerically minimized. This approach was proposed by Rudemo (1982) and Bowman (1984).

To appreciate the substance of Loader's (1999) comments, Figure 2.3 plots the bimodal density estimate, the kernel estimate using the plug-in rule, and that using least squares cross-validation.

Figure 2.3 reveals that indeed the plug-in rule is oversmoothing leading to substantial bias for the left peak. Least squares cross-validation rectifies this as Loader (1999) points out, but at the cost of additional variability in the right peak.

One problem with this approach is that it is sensitive to the presence of rounded or discretized data and to small-scale effects in the data.

This example suggests that perhaps the fixed h kernel estimator could be improved on, and there exist "adaptive" kernel estimators

[4] Loader writes "We find the evidence for superior performance of plug-in approaches is far less compelling than previously claimed. In turn, we consider real data examples, simulation studies and asymptotics. Among the findings are that plug-in approaches are tuned by arbitrary specification of pilot estimators and are prone to over-smoothing when presented with difficult smoothing problems."

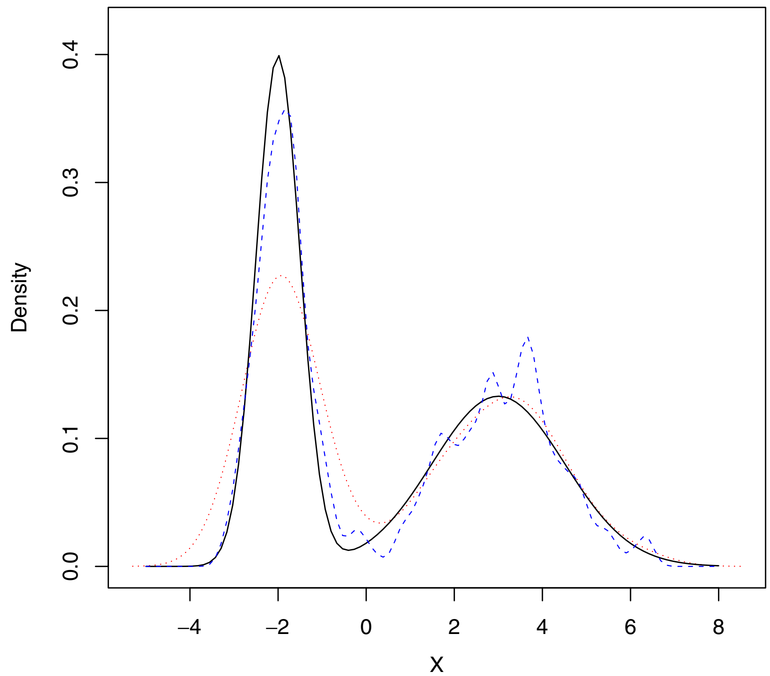

Fig. 2.3 Plug-in versus least squares cross-validation density estimates. The true density is the solid line, the dotted line the plug-in density, and the dashed line the least squares cross-validation density.

that allow h to vary at either the point x or X_i (see Abramson (1982) and Breiman et al. (1977)). These estimators, however, tend to introduce spurious noise in the density estimate. As the fixed h method is dominant in applied work, we proceed with this approach.

2.3.4 Likelihood Cross-Validation

Likelihood cross-validation yields a density estimate which has an entropy interpretation, being that the estimate will be close to the actual density in a Kullback–Leibler sense. Likelihood cross-validation chooses h to maximize the (leave-one-out) log likelihood function given by

$$\mathcal{L} = \log L = \sum_{i=1}^{n} \log \hat{f}_{-i}(x),$$

where $\hat{f}_{-i}(x)$ is the leave-one-out kernel estimator of $f(X_i)$ that uses all points except X_i to construct the density estimate, that is,

$$\hat{f}_{-i}(x) = \frac{1}{(n-1)h} \sum_{j=1, j\neq i}^{n} K\left(\frac{X_j - x}{h}\right).$$

This method is of general applicability, and was proposed by Stone (1974) and Geisser (1975). One drawback of this method is that it can oversmooth for fat-tailed distributions such as the Cauchy.

2.3.5 Bootstrap Methods

Faraway and Jhun (1990) proposed a bootstrap-based method of selecting the bandwidth h by estimating the IMSE defined in (2.5) for any given bandwidth and then minimizing over all bandwidths. The approach uses a smoothed bootstrap method based on an initial density estimate. One drawback with this approach is that the objective function is stochastic which can give rise to numerical minimization issues, while it can also be computationally demanding.

2.4 Frequency and Kernel Probability Estimators

So far we have considered estimating a univariate density function presuming that the underlying data is continuous in nature. Suppose we were interested instead in estimating a univariate *probability* function where the data is discrete in nature. The nonparametric non-smooth approach would construct a frequency estimate, while the nonparametric smooth approach would construct a kernel estimate quite different from that defined in (2.2). For those unfamiliar with the term "frequency" estimate, this is simply the estimator of a probability computed via the sample frequency of occurrence. For example, if a random variable is the result of a Bernoulli trial (i.e., zero or one with fixed probability from trial to trial) then the frequency estimate of the probability of a zero (one) is simply the number of zeros (ones) divided by the number of trials.

First, consider the estimation of a probability function defined for $X_i \in \mathcal{S} = \{0, 1, \ldots, c-1\}$. The non-smooth "frequency" (non-kernel) estimator of $p(x)$ is given by

$$\tilde{p}(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(X_i, x),$$

where $\mathbf{1}(\cdot)$ is again the indicator function defined earlier. It is straightforward to show that

$$\begin{aligned} E\tilde{p}(x) &= p(x), \\ \operatorname{var}\, \tilde{p}(x) &= \frac{p(x)(1-p(x))}{n}, \end{aligned}$$

hence,

$$\operatorname{MSE}(\tilde{p}(x)) = n^{-1}p(x)(1-p(x)) = O(n^{-1}),$$

which implies that

$$\tilde{p}(x) - p(x) = O_p(n^{-1/2})$$

Now, consider the kernel estimator of $p(x)$,

$$\hat{p}(x) = \frac{1}{n}\sum_{i=1}^{n} l(X_i, x), \tag{2.7}$$

where $l(\cdot)$ is a kernel function defined by, say,

$$l(X_i, x) = \begin{cases} 1-\lambda & \text{if } X_i = x \\ \lambda/(c-1) & \text{otherwise,} \end{cases}$$

and where $\lambda \in [0, (c-1)/c]$ is a "smoothing parameter" or "bandwidth." The requirement that λ lie in $[0, (c-1)/c]$ ensures that $\tilde{p}(x)$ is a proper probability estimate lying in $[0, 1]$. It is easy to show that

$$\begin{aligned} E\hat{p}(x) &= p(x) + \lambda\left\{\frac{1-cp(x)}{c-1}\right\}, \\ \operatorname{var}\hat{p}(x) &= \frac{p(x)(1-p(x))}{n}\left(1-\lambda\frac{c}{(c-1)}\right)^2. \end{aligned}$$

This estimator was proposed by Aitchison and Aitken (1976) for discriminant analysis with multivariate binary data. See also Simonoff (1996).

Note that when $\lambda = 0$ this estimator collapses to the frequency estimator $\tilde{p}(x)$, while when λ hits its upper bound, $(c-1)/c$, this estimator is the rectangular (i.e., discrete uniform) estimator which yields equal probabilities across all outcomes.

Using a bandwidth which balances bias and variance, it can be shown that

$$\hat{p}(x) - p(x) = O_p\big(n^{-1/2}\big).$$

Note that, unlike that for the Rosenblatt–Parzen estimator, here we were able to use the exact expressions to obtain our results rather than the approximate expressions used for the former.

2.5 Kernel Density Estimation with Discrete and Continuous Data

Suppose that we were facing a mix of discrete and continuous data and wanted to model the joint density[5] function. When facing a mix of discrete and continuous data, traditionally researchers using kernel methods resorted to a "frequency" approach. This approach involves breaking the continuous data into subsets according to the realizations of the discrete data ("cells"). This of course will produce consistent estimates. However, as the number of subsets increases, the amount of data in each cell falls leading to a "sparse data" problem. In such cases, there may be insufficient data in each subset to deliver sensible density estimates (the estimates will be highly variable).

The approach we consider below uses the concept of "generalized product kernels." For the continuous variables we use standard continuous kernels denoted now by $W(\cdot)$ (Epanechnikov etc.). For an unordered discrete variable $\bar{x}^d$, we could use Aitchison and Aitken's (1976) kernel given by

$$\bar{l}(\bar{X}_i^d, \bar{x}^d) = \begin{cases} 1-\lambda, & \text{if } \bar{X}_i^d = \bar{x}^d, \\ \frac{\lambda}{c-1}, & \text{otherwise.} \end{cases}$$

[5] The term "density" is appropriate for distribution functions defined over mixed discrete and continuous variables. It is the measure defined on the discrete variables in the density function that matters.

For an ordered discrete variable $\tilde{x}^d$, we could use Wang and van Ryzin's (1981) kernel given by

$$\tilde{l}(\tilde{X}_i^d, \tilde{x}^d) = \begin{cases} 1-\lambda, & \text{if } \tilde{X}_i^d = \tilde{x}^d, \\ \frac{(1-\lambda)}{2}\lambda^{|\tilde{X}_i^d - \tilde{x}^d|}, & \text{if } \tilde{X}_i^d \neq \tilde{x}^d. \end{cases}$$

A generalized product kernel for one continuous, one unordered, and one ordered variable would be defined as follows:

$$K(\cdot) = W(\cdot) \times \bar{l}(\cdot) \times \tilde{l}(\cdot). \tag{2.8}$$

Using such product kernels, we can modify any existing kernel-based method to handle the presence of categorical variables, thereby extending the reach of kernel methods.

Estimating a joint probability/density function defined over mixed data follows naturally using these generalized product kernels. For example, for one unordered discrete variable $\bar{x}^d$ and one continuous variable x^c, our kernel estimator of the PDF would be

$$\hat{f}(\bar{x}^d, x^c) = \frac{1}{nh_{x^c}} \sum_{i=1}^{n} \bar{l}(\bar{X}_i^d, \bar{x}^d) W\left(\frac{X_i^c - x^c}{h_{x^c}}\right).$$

This extends naturally to handle a mix of ordered, unordered, and continuous data (i.e., both quantitative and qualitative data). This estimator is particularly well suited to "sparse data" settings. Rather than clutter the page with notation by formally defining the estimator for p continuous, q unordered, and r ordered variables, we presume that the underlying idea of using product kernels is clear, and direct the interested reader to Li and Racine (2003) for details.

2.5.1 Discrete and Continuous Example

We consider Wooldridge's (2002) "wage1" dataset having $n = 526$ observations, and model the joint density of two variables, one continuous ("lwage") and one discrete ("numdep"). "lwage" is the logarithm of average hourly earnings for an individual. "numdep" the number of dependents $(0, 1, \dots)$. We use likelihood cross-validation (see Section 2.3.4) to obtain the bandwidths, and the resulting estimate is presented in Figure 2.4.

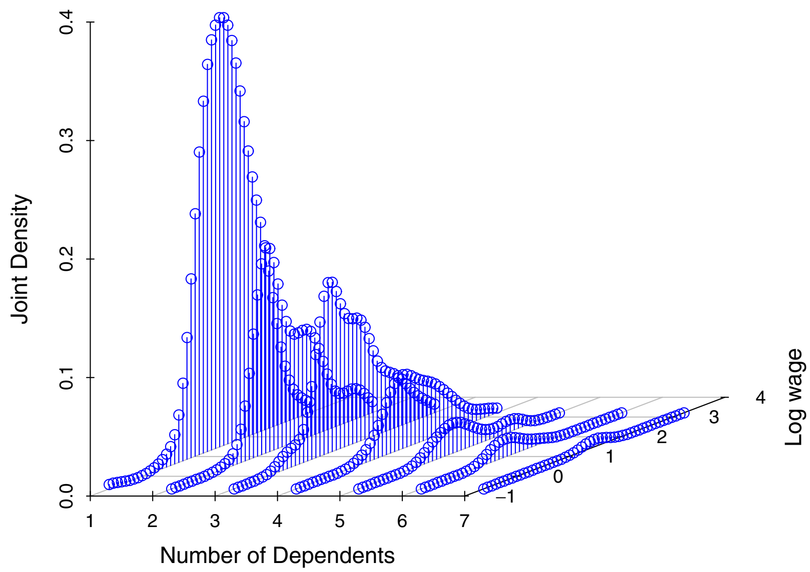

Fig. 2.4 Nonparametric kernel estimate of a joint density defined over one continuous and one discrete variable.

Note that this is indeed a case of "sparse" data for some cells (see Table 2.1), and the traditional approach would require estimation of a nonparametric univariate density function based upon only two observations for the last cell ($c = 6$).

2.6 Constructing Error Bounds

It is possible to construct pointwise and simultaneous confidence intervals for the density estimate, and this is typically done using either the

Table 2.1 Summary of the number of dependents in the Wooldridge (2002) 'wage1' dataset ("numdep") ($c = 0, 1, \ldots, 6$).

c	n_c
0	252
1	105
2	99
3	45
4	16
5	7
6	2

asymptotic formula such as that given in (2.4) in which the unknown components are replaced with their estimates, or using resampling methods such as the bootstrap. Note that the kernel estimator can be shown to be asymptotically normal via application of Liapunov's double array central limit theorem.

Pointwise confidence intervals yield intervals at a given point x and are of the form:

$$P(\hat{f}_l(x) < f(x) < \hat{f}_u(x)) = 1 - \alpha,$$

where α is the probability of a Type I error. Simultaneous confidence intervals, on the other hand, yield intervals of the form:

$$P(\cap_{i=1}^{n}\{\hat{f}_l(X_i) < f(X_i) < \hat{f}_u(X_i)\}) = 1 - \alpha.$$

As construction of the above two types of intervals requires the interval to be centered on $f(x)$, bias correction methods must be used, either via estimation of asymptotic formula such as that given in (2.3) or via resampling methods such as the jackknife or bootstrap.

Alternatively, if interest lies solely in assessing variability of the estimate, error bars can be centered on $\hat{f}(x)$ rather than an unbiased estimate of $f(x)$. Figure 2.5 plots the density estimate in Figure 2.2 along with pointwise 95% variability bounds (i.e., not bias-corrected). One might wonder why bias-corrected intervals are not the norm. One reason is because estimating bias is a notoriously difficult thing to do, and the resulting bias-corrected estimates can be highly variable; see Efron (1982) for further details surrounding bias-corrected estimates.

2.7 Curse-of-Dimensionality

As the dimension of the *continuous* variable space increases, the rates of convergence of kernel methods deteriorate, which is the well known "curse of dimensionality" problem. Letting p denote the number of continuous variables over which the density is defined, it can be shown that

$$\hat{f}(x) - f(x) = O_p\left(n^{-2/(p+4)}\right);$$

see Li and Racine (2003) for a derivation of this results for the mixed data case with least squares cross-validation.

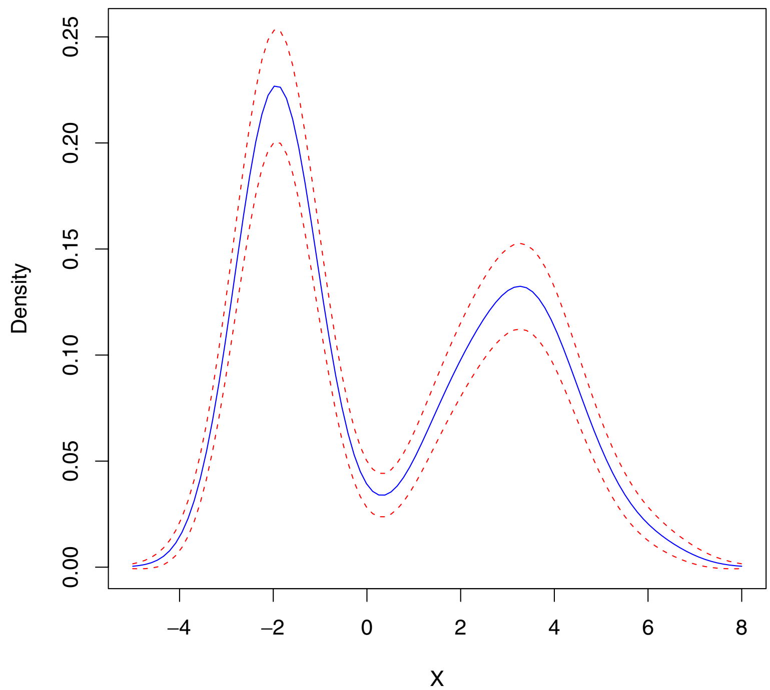

Fig. 2.5 Kernel density estimate $\hat{f}(x) \pm 1.96 \times s$ using the asymptotic standard error, s given in (2.4).

Silverman (1986, p. 94) presents an often cited table that shows the sample size required to ensure that the relative MSE of a *correctly specified parametric estimator* (*multivariate normal*) versus a multivariate kernel density estimator (with continuous datatypes only) is less than 0.1 when evaluated at the multivariate mean, where relative MSE is defined by $E\{\hat{f}(\mu) - f(\mu)\}^2 / f(\mu)^2$, a Gaussian kernel is used, and the optimal point-wise bandwidth is computed. This table is frequently cited by people who have thereby inferred that kernel methods are useless when then dimension exceeds two or three variables.

Though of course Silverman's (1986, p. 94) table is correct, concluding that kernel methods are not going to be of value when the dimension exceeds just a few variables does not follow, for two simple reasons. First, popular parametric models are rarely, if ever, correctly specified.[6] The "horse race" is therefore between *misspecified* and therefore *inconsistent* parametric models and relatively *inefficient* but *consistent* nonparametric models.[7] Second, the curse-of-dimensionality applies only to the number of *continuous* variables involved. In applied settings it is not uncommon to encounter situations involving only a small number of continuous variables or, often, the data is comprised exclusively of categorical variables.

[6] "Normality is a myth; there never was, and never will be, a normal distribution" Geary (1947).

[7] As mentioned earlier, Robinson (1988) refers to parametric models as $\sqrt{n}$-inconsistent (they are typically referred to as $\sqrt{n}$-consistent) to highlight this phenomena.

3

Conditional Density Estimation

Conditional density functions underlie many popular statistical objects of interest, though they are rarely modeled directly in parametric settings and have perhaps received even less attention in kernel settings. Nevertheless, as will be seen, they are extremely useful for a range of tasks, whether directly estimating the conditional density function, modeling count data (see Cameron and Trivedi (1998) for a thorough treatment of count data models), or perhaps modeling conditional quantiles via estimation of a conditional CDF. And, of course, regression analysis (i.e., modeling conditional means) depends directly on the conditional density function, so this statistical object in fact implicitly forms the backbone of many popular statistical methods.

3.1 Kernel Estimation of a Conditional PDF

Let $f(\cdot)$ and $\mu(\cdot)$ denote the joint and marginal densities of (X,Y) and X, respectively, where we allow Y and X to consist of continuous, unordered, and ordered variables. For what follows we shall refer to Y as a dependent variable (i.e., Y is explained), and to X as covariates (i.e., X is the explanatory variable). We use $\hat{f}$ and $\hat{\mu}$ to denote kernel estimators thereof, and we estimate the conditional density $g(y|x) = f(x,y)/f(x)$ by

$$\hat{g}(y|x) = \hat{f}(x,y)/\hat{f}(x). \tag{3.1}$$

The kernel estimators of the joint and marginal densities $f(x,y)$ and $f(x)$ are described in the previous section and are not repeated here; see Hall et al. (2004) for details on the theoretical underpinnings of a data-driven method of bandwidth selection for this method.

3.1.1 The Presence of Irrelevant Covariates

Hall et al. (2004) proposed the estimator defined in (3.1), but choosing appropriate smoothing parameters in this setting can be tricky, not least because plug-in rules take a particularly complex form in the case of mixed data. One difficulty is that there exists no general formula for the optimal smoothing parameters. A much bigger issue is that it can be difficult to determine which components of X are relevant to the problem of conditional inference. For example, if the jth component of X is independent of Y then that component is irrelevant to estimating the density of Y given X, and ideally should be dropped before conducting inference. Hall et al. (2004) show that a version of least-squares cross-validation overcomes these difficulties. It automatically determines which components are relevant and which are not, through assigning large smoothing parameters to the latter and consequently shrinking them toward the uniform distribution on the respective marginals. This effectively removes irrelevant components from contention, by suppressing their contribution to estimator variance; they already have very small bias, a consequence of their independence of Y. Cross-validation also gives us important information about which components are relevant: the relevant components are precisely those which cross-validation has chosen to smooth in a traditional way, by assigning them smoothing parameters of conventional size. Cross-validation produces asymptotically optimal smoothing for relevant components, while eliminating irrelevant components by oversmoothing.

The importance of this result is best appreciated by comparison of the conditions for consistency outlined in Section 2.2.1, where we mentioned standard results for density estimation whereby $h \to 0$ as $n \to \infty$ (bias $\to 0$) and $nh \to \infty$ as $n \to \infty$ (var $\to 0$). Hall et al. (2004) demonstrate that, for irrelevant conditioning variables in X, their bandwidths in fact ought to behave exactly the opposite, namely, $h \to \infty$ as $n \to \infty$ for optimal smoothing. The same has been demonstrated for regression as well; see Hall et al. (forthcoming) for further details.

3.1.2 Modeling an Italian GDP Panel

We consider Giovanni Baiocchi's Italian GDP growth panel for 21 regions covering the period 1951–1998 (millions of Lire, 1990 = base). There are 1,008 observations in total, and two variables, "gdp" and "year." Given their nature, we treat gdp as continuous and year $(1951, 1952, \dots)$ as an ordered discrete variable. We then estimate the density of gdp conditional on year. Figure 3.1 plots the estimated conditional density, $\hat{f}(\text{gdp}|\text{year})$ based upon likelihood

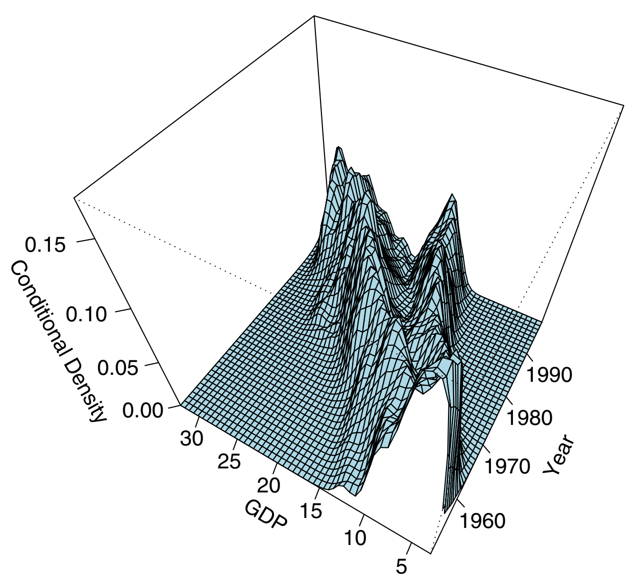

Fig. 3.1 Nonparametric conditional PDF estimate for the Italian gdp panel.

cross-validated bandwidth selection which yielded bandwidths $\hat{h}_{\text{gdp}} = 0.715$ and $\hat{\lambda}_{\text{year}} = 0.671$.

Figure 3.1 reveals that the distribution of income has evolved from a unimodal one in the early 1950s to a markedly bimodal one in the 1990s. This result is robust to bandwidth choice, and is observed whether using simple rules-of-thumb or data-driven methods such as least-squares cross-validation or likelihood cross-validation. The kernel method readily reveals this evolution which might easily be missed were one to use parametric models of the income distribution. For instance, the (unimodal) log-normal distribution is a popular parametric model for income distributions, but is incapable of revealing the multi-modal structure present in this dataset.

3.2 Kernel Estimation of a Conditional CDF

Li and Racine (forthcoming) propose a nonparametric conditional CDF kernel estimator that admits a mix of discrete and categorical data along with an associated nonparametric conditional quantile estimator. Bandwidth selection for kernel quantile regression remains an open topic of research, and they employ a modification of the conditional PDF based bandwidth selector proposed by Hall et al. (2004).

We use $F(y|x)$ to denote the conditional CDF of Y given $X = x$, while $f(x)$ is the marginal density of X. We can estimate $F(y|x)$ by

$$\hat{F}(y|x) = \frac{n^{-1}\sum_{i=1}^{n} G\left(\frac{y-Y_i}{h_0}\right) K_h(X_i, x)}{\hat{f}(x)}, \tag{3.2}$$

where $G(\cdot)$ is a kernel CDF chosen by the researcher, say, the standard normal CDF, h_0 is the smoothing parameter associated with Y, and $K_h(X_i, x)$ is a product kernel such as that defined in (2.8) where each univariate continuous kernel has been divided by its respective bandwidth for notational simplicity.

Figure 3.2 presents this estimator for the Italian GDP panel described in Section 3.1.2.

The conditional CDF presented in Figure 3.2 conveys information presented in Figure 3.1 in a manner better suited to estimating, say, a conditional quantile to which we now turn.

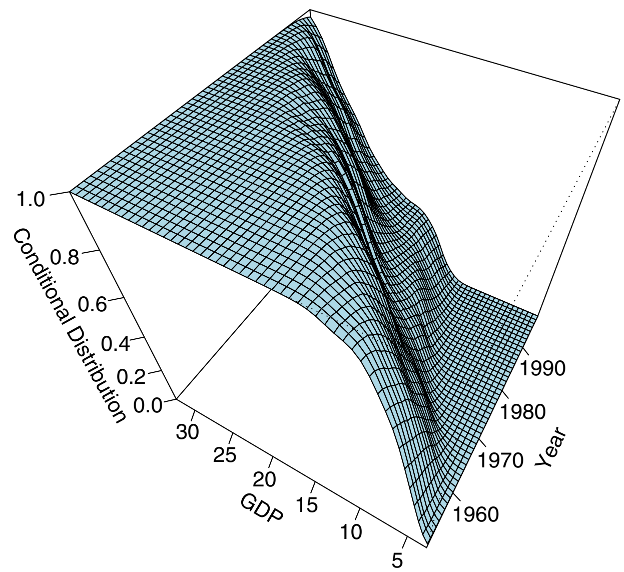

Fig. 3.2 Nonparametric conditional CDF estimate for the Italian GDP panel.

3.3 Kernel Estimation of a Conditional Quantile

Estimating regression functions is a popular activity for practitioners. Sometimes, however, the regression function is not representative of the impact of the covariates on the dependent variable. For example, when the dependent variable is left (or right) censored, the relationship given by the regression function is distorted. In such cases, conditional quantiles above (or below) the censoring point are robust to the presence of censoring. Furthermore, the conditional quantile function provides a more comprehensive picture of the conditional distribution of a dependent variable than the conditional mean function.

Once we can estimate conditional CDFs such as that presented in Figure 3.2, estimating conditional quantiles follows naturally. That is, having estimated the conditional CDF we simply invert it at the desired quantile as described below. A conditional αth quantile of a conditional distribution function $F(\cdot|x)$ is defined by ($\alpha \in (0,1)$)

$$q_\alpha(x) = \inf\{y : F(y|x) \geq \alpha\} = F^{-1}(\alpha|x).$$

Or equivalently, $F(q_\alpha(x)|x) = \alpha$. We can directly estimate the conditional quantile function $q_\alpha(x)$ by inverting the estimated conditional CDF function, i.e.,

$$\hat{q}_\alpha(x) = \inf\{y : \hat{F}(y|x) \geq \alpha\} \equiv \hat{F}^{-1}(\alpha|x).$$

Theoretical details of this method can be found in Li and Racine (forthcoming).

Figure 3.3 presents the 0.25, 0.50 (median), and 0.75 conditional quantiles for the Italian GDP panel described in Section 3.1.2, along with box plots[1] of the raw data. One nice feature of this application is

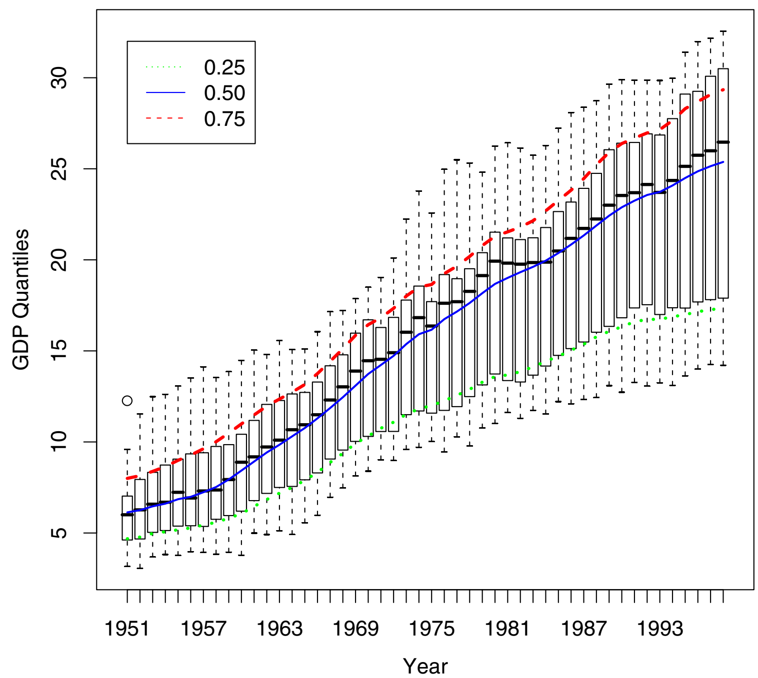

Fig. 3.3 Nonparametric conditional quantile estimates for the Italian GDP panel, $\alpha = (0.25, 0.50, 0.75)$.

[1] A box-and-whisker plot (sometimes called simply a "box plot") is a histogram-like method of displaying data, invented by J. Tukey. To create a box-and-whisker plot, draw a box with ends at the quartiles Q_1 and Q_3. Draw the statistical median M as a horizontal line in the box. Now extend the "whiskers" to the farthest points that are not outliers (i.e., that are within 3/2 times the interquartile range of Q_1 and Q_3). Then, for every point more than 3/2 times the interquartile range from the end of a box, draw a dot.

that the explanatory variable is ordered and there exist multiple observations per year. The non-smooth quantile estimates generated by the box plot can be directly compared to those obtained via direct estimation of the smooth CDF, and it is clear that they are in agreement.

3.4 Binary Choice and Count Data Models

Another application of kernel estimates of PDFs with mixed data involves the estimation of conditional mode models. By way of example, consider some discrete outcome, say $Y \in \mathcal{S} = \{0, 1, \ldots, c-1\}$, which might denote by way of example the number of successful patent applications by firms. We define a conditional mode of $y|x$ by

$$m(x) = \max_y g(y|x). \tag{3.3}$$

In order to estimate a conditional mode $m(x)$, we need to model the conditional density. Let us call $\hat{m}(x)$ the estimated conditional mode, which is given by

$$\hat{m}(x) = \max_y \hat{g}(y|x), \tag{3.4}$$

where $\hat{g}(y|x)$ is the kernel estimator of $g(y|x)$ defined in (3.1). By way of example, we consider modeling low birthweights (a binary indicator) using this method.

3.4.1 Modeling Low Birthweight (0/1)

For this example, we use data on birthweights taken from the R MASS library (Venables and Ripley (2002)), and compute a parametric Logit model and a nonparametric conditional mode model using (3.4) in which the conditional density was estimated using (3.1) based upon Hall et al.'s (2004) method. We then compare their confusion matrices[2] and assess their classification ability. The outcome y is a binary indicator of low infant birthweight ("low") defined below. The method

[2] A "confusion matrix" is simply a tabulation of the actual outcomes versus those predicted by a model. The diagonal elements contain correctly predicted outcomes while the off-diagonal ones contain incorrectly predicted (confused) outcomes.

Table 3.1 Confusion matrices for the low birthweight data. The table on the left summarizes the parametric logit model, that on the right the kernel model.

	Predicted	
Actual	0	1
0	119	11
1	34	25

	Predicted	
Actual	0	1
0	127	1
1	27	32

can handle unordered and ordered multinomial outcomes without modification. This application has $n = 189$ and 7 explanatory variables in x, "smoke," "race," "ht," "ui," "ftv," "age," and "lwt" defined below.

Variables are defined as follows:

(1) "low" indicator of birth weight less than 2.5 kg
(2) "smoke" smoking status during pregnancy
(3) "race" mother's race ("1" = white, "2" = black, "3" = other)
(4) "ht" history of hypertension
(5) "ui" presence of uterine irritability
(6) "ftv" number of physician visits during the first trimester
(7) "age" mother's age in years
(8) "lwt" mother's weight in pounds at last menstrual period

Note that all variables other than age and lwt are categorical in nature in this example.

We compute the "confusion" matrices for each model using likelihood cross-validation to obtain the bandwidths for the nonparametric conditional mode model. As can be seen, the nonparametric model correctly classifies $(127 + 32)/189 = 84.1\%$ of low/high birthweights while the Logit model correctly classifies only $(119 + 25)/189 = 76.1\%$.

4

Regression

One of the most popular methods for nonparametric kernel regression was proposed by Nadaraya (1965) and Watson (1964) and is known as the "Nadaraya–Watson" estimator though it is also known as the "local constant" estimator for reasons best described when we introduce the "local polynomial" estimator (Fan (1992)). We begin with a brief introduction to the local constant method of estimating regression functions and their derivatives then proceed to the local polynomial method. We remind the reader that we shall rely on many objects outlined in *Density and Probability Function Estimation* and *Conditional Density Estimation* such as generalized product kernels and so forth.

4.1 Local Constant Kernel Regression

We begin by considering the bivariate regression case for notational simplicity.[1]

[1] As will be seen, the multivariate mixed data versions follow naturally, and we will point out the modifications required where appropriate.

4.1.1 The Local Constant Conditional Mean ($\hat{g}(x)$)

By definition, the conditional mean of a continuous random variable Y is given by

$$g(x) = \int y\, g(y|x)\, dy = \int y \frac{f(y,x)}{f(x)}\, dy = \frac{m(x)}{f(x)},$$

where $g(y|x)$ is the conditional PDF defined in *Conditional Density Estimation* and where $m(x) = \int y f(y,x)\, dy$.

The local constant estimator of the conditional mean is obtained by replacing the unknown joint and marginal densities, $f(y,x)$ and $f(x)$, by their kernel estimators defined in *Density and Probability Function Estimation*, which yields

$$\hat{g}(x) = \int y \frac{\hat{f}(y,x)}{\hat{f}(x)}\, dy.$$

With a little algebra the local constant estimator $\hat{g}(x)$ simplifies to

$$\hat{g}(x) = \int y \frac{\hat{f}(y,x)}{\hat{f}(x)}\, dy = \frac{\sum_{i=1}^n Y_i K\left(\frac{X_i - x}{h_x}\right)}{\sum_{i=1}^n K\left(\frac{X_i - x}{h_x}\right)}. \tag{4.1}$$

Note that the integral drops out due to the use of the product kernel function and a change of variables argument.

Note that, under the conditions given in the following section, $\hat{g}(x)$ is a consistent estimate of a conditional mean. In essence, we are locally averaging those values of the dependent variable which are "close" in terms of the values taken on by the regressors. By controlling the amount of local information used to construct the estimate (the "local sample size") and allowing the amount of local averaging to become more informative as the sample size increases, while also decreasing the neighborhood in which the averaging occurs, we can ensure that our estimates are consistent under standard regularity conditions.

4.1.2 Approximate Bias and Variance

Though the local constant estimator is widely used, it suffers from "edge bias" which can be seen by considering its approximate bias which in the bivariate case is given by

$$h^2\left(\frac{1}{2}g''(x) + \frac{g'(x)f'(x)}{f(x)}\right)\kappa_2$$

(see Pagan and Ullah (1999, p. 101) for a derivation). Other things equal, as we approach the boundary of the support of the data, $f(x)$ approaches zero and the bias increases. The class of "local polynomial" estimators described in Section 4.2 do not suffer from edge bias though they are prone to numerical instability issues described shortly. The approximate bias for the local linear estimator introduced shortly is given by

$$\frac{h^2}{2}g''(x)\kappa_2,$$

and it can be seen that the term giving rise to the edge bias in the local constant estimator, namely $g'(x)f'(x)/f(x)$, does not appear in that for the local linear estimator.

In Section 4.2, we describe the local linear estimator for the bivariate case, and at this time point out that the local constant and local linear estimators have identical approximate variance which, for the bivariate case is given by

$$\frac{\sigma^2(x)}{f(x)nh}\int K^2(z)\,dz,$$

where $\sigma^2(x)$ is the conditional variance of y.

4.1.3 Optimal and Data-Driven Bandwidths

The IMSE-optimal bandwidth for the local constant estimator,

$$h_{\text{opt}} = \left[\frac{\sigma^2(x)\int f^{-1}(x)\,dx\int K^2(z)\,dz}{\int\{2g'(x)f'(x)f^{-1}(x) + g''(x)\}^2\,dx\kappa_2^2}\right]^{1/5} n^{-1/5},$$

is obtained in exactly the same manner as was that in Section 2.2.1, and like its density counterpart depends on unknown quantities that depend on the underlying DGP.

Though plug-in methods could be applied, in multivariate settings they are infeasible due to the need to estimate higher order derivatives along with cross-partial derivatives, among others, while in mixed-data settings no general formula exists. Alternative data-driven approaches are used in practice.

Two popular data-driven methods of bandwidth selection that have desirable properties are least-squares cross-validation and the AIC-based method of Hurvich et al. (1998), which is based on minimizing a modified Akaike Information Criterion.

Least-squares cross-validation for regression is based on minimizing

$$\mathrm{CV}(h) = n^{-1} \sum_{i=1}^{n} (Y_i - \hat{g}_{-i}(X_i))^2,$$

where $\hat{g}_{-i}(X_i)$ is the estimator of $g(X_i)$ formed by leaving out the ith observation when generating the prediction for observation i.

Hurvich et al.'s (1998) approach is based on the minimization of

$$\mathrm{AIC}_c = \ln(\hat{\sigma}^2) + \frac{1 + \mathrm{tr}(H)/n}{1 - \{\mathrm{tr}(H) + 2\}/n},$$

where

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} \{Y_i - \hat{g}(X_i)\}^2 = Y'(I - H)'(I - H)Y/n$$

with $\hat{g}(X_i)$ being a nonparametric estimator and H being an $n \times n$ weighting function (i.e., the matrix of kernel weights) with its (i, j)th element given by $H_{ij} = K_h(X_i, X_j)/\sum_{l=1}^{n} K_h(X_i, X_l)$, where $K_h(\cdot)$ is a generalized product kernel.

Both the CV method and the AIC_c method have been shown to be asymptotically equivalent; see Li and Racine (2004) for details.

4.1.4 Relevant and Irrelevant Regressors

For relevant x, conditions for consistency are the same as those outlined for density estimation, namely $h \to 0$ as $n \to \infty$ and $nh \to \infty$ as $n \to \infty$. However, when x is in fact irrelevant, then it can be shown that $h \to \infty$ as $n \to \infty$ will produce optimal smoothing rather than $h \to 0$. It has been shown that the least-squares cross-validation method of bandwidth selection will lead to optimal smoothing for both relevant and irrelevant x; see Hall et al. (forthcoming) for details.

For the local constant estimator of the conditional mean of y, when $h \to \infty$ we observe that

$$\hat{g}(x) = \frac{\sum_{i=1}^{n} Y_i K(0)}{\sum_{i=1}^{n} K(0)} = n^{-1} \sum_{i=1}^{n} Y_i = \bar{y},$$

which is the *unconditional mean* of y. In this instance we say that x has been "smoothed out" of the regression function, which is appropriate when there is no information contained in x that is useful for predicting y.

The intuition underlying the desirability of smoothing out irrelevant regressors is quite simple. The presence of irrelevant x means that the bias of $\hat{g}(x)$ is zero for any h. One could therefore use relatively small values of h, however estimators with relatively small h will necessarily be more variable than those with relatively large h. As cross-validation delivers an approximation to the MSE of the estimator, then MSE is clearly minimized in this case when the variance of $\hat{g}(x)$ is minimized, which occurs when h is such that $\hat{g}(x) = \bar{y}$, i.e., when $h \to \infty$. Again, cross-validation can deliver the appropriate value of h in both relevant and irrelevant settings. Finally, observe that the rate of convergence of the bivariate (i.e., one regressor) local constant kernel estimator using optimal smoothing is (inversely) proportional to $\sqrt{n}$ in the presence of irrelevant regressors, which is the parametric rate, while in the presence of relevant regressors the rate of convergence is proportional to $\sqrt{n^{4/5}}$ using second order kernels, which is slower than the parametric rate. This fact is perhaps not as widely appreciated as it could be and has important implications for automatic dimension reduction in multivariate settings which can mitigate the curse-of-dimensionality in some settings.

The extension to multiple regressors follows naturally, and a mixed-data multivariate version is obtained by simply replacing the kernel with a generalized product kernel defined in *Density and Probability Function Estimation*; see Racine and Li (2004) for theoretical underpinnings of this method.

4.1.5 The Local Constant Response ($\hat{\beta}(x)$)

In addition to estimating the conditional mean, we frequently wish to estimate marginal effects ("derivatives" or "response").

The unknown response $\beta(x)$ for the bivariate case considered above is defined as follows:

$$\begin{aligned}\beta(x) \equiv \frac{d\,g(x)}{dx} &= g'(x) = \frac{f(x)m'(x) - m(x)f'(x)}{f^2(x)} \\ &= \frac{m'(x)}{f(x)} - \frac{m(x)}{f(x)}\frac{f'(x)}{f(x)} = \frac{m'(x)}{f(x)} - g(x)\frac{f'(x)}{f(x)}.\end{aligned}$$

The local constant estimator is obtained by replacing the unknown $f(x)$, $m'(x)$, $g(x)$, and $f'(x)$ with their kernel-based counterparts and is given by

$$\begin{aligned}\hat\beta(x) \equiv \frac{d\,\hat g(x)}{dx} &= \frac{\hat f(x)\hat m'(x) - \hat m(x)\hat f'(x)}{\hat f^2(x)} \\ &= \frac{\hat m'(x)}{\hat f(x)} - \frac{\hat m(x)}{\hat f(x)}\frac{\hat f'(x)}{\hat f(x)} = \frac{\hat m'(x)}{\hat f(x)} - \hat g(x)\frac{\hat f'(x)}{\hat f(x)},\end{aligned}$$

where

$$\begin{aligned}\hat m(x) &= \frac{1}{nh}\sum_i Y_i K\left(\frac{X_i - x}{h}\right) \\ \hat f(x) &= \frac{1}{nh}\sum_i K\left(\frac{X_i - x}{h}\right) \\ \hat m'(x) &= -\frac{1}{nh^2}\sum_i Y_i K'\left(\frac{X_i - x}{h}\right) \\ \hat f'(x) &= -\frac{1}{nh^2}\sum_i K'\left(\frac{X_i - x}{h}\right).\end{aligned}$$

Again, a multivariate version follows naturally, and mixed-data versions follow using the generalized product kernels introduced earlier where of course this estimator is only defined for the continuous regressors.

4.2 Local Polynomial Kernel Regression

The estimator given in (4.1) is called the local constant estimator because it can be seen to be the minimizer of the following:

$$\hat g(x) \equiv \min_a \sum_{i=1}^{n} (Y_i - a)K\left(\frac{X_i - x}{h}\right).$$

We now introduce a popular extension that does not suffer from edge bias, though it does introduce other issues such as potential singularity that often arises in sparse data settings. The most popular local polynomial method is the local linear approach, which we describe below and again consider the bivariate case for notational simplicity.

Assuming that the second derivative of $g(x)$ exists, then in a small neighborhood of a point x, $g(x_0) \approx g(x) + (\partial g(x)/\partial x)(x_0 - x) = a + b(x_0 - x)$. The problem of estimating $g(x)$ is equivalent to the local linear regression problem of estimating the intercept a. The problem of estimating the response $\partial g(x)/\partial x$ is equivalent to the local linear regression problem of estimating the slope b.

We proceed by choosing a and b so as to minimize

$$\begin{aligned}\mathcal{S} &= \sum_{i=1}^{n} (Y_i - a - b(X_i - x))^2 K\left(\frac{X_i - x}{h}\right) \\ &= \sum_{i=1}^{n} (Y_i - a - b(X_i - x))^2 K(Z_i).\end{aligned}$$

The solutions $\hat{a}$ and $\hat{b}$ will be the local linear estimators of $g(x)$ and $\beta(x)$, respectively. Solving we obtain

$$\begin{pmatrix} \hat{g}(x) \\ \hat{\beta}(x) \end{pmatrix} = \left[\sum_{i=1}^{n} \begin{pmatrix} 1 & X_i - x \\ X_i - x & (X_i - x)^2 \end{pmatrix} K(Z_i)\right]^{-1} \sum_{i=1}^{n} \begin{pmatrix} 1 \\ X_i - x \end{pmatrix} K(Z_i) y_i.$$

One feature of this approach is that it directly delivers estimators of the mean and response, which was not the case for the local constant estimator. The approximate bias and variance are given in Section 4.1.2. For the estimation of marginal effects (i.e., $\beta(x)$), it is common to use a higher-order polynomial (i.e., to use a local quadratic regression if you want to estimate first derivatives) as a bias-reduction device (see Fan and Gijbels (1996)).

One problem that often surfaces when using this estimator is that it suffers from singularity problems arising from the presence of sparse data, particularly for small bandwidths, hence various forms of "ridging" have been suggested to overcome these problems. Ridging methods are techniques for solving badly conditioned linear regression problems.

The approach was first proposed by Hoerl and Kennard (1970). For details on the use of ridging methods in a local linear context see Cheng et al. (1997) and Seifert and Gasser (2000).

The behavior of the local linear estimator with regard to h is markedly different from that for the local constant estimator. As $h \to \infty$ the local linear estimator $\hat{g}(x)$ can be shown to approach $\hat{\beta}_0 + \hat{\beta}_1 x$ where $\hat{\beta}_0$ and $\hat{\beta}_1$ are the linear least squares estimators from the regression of y on x. That is, as $h \to \infty$ the locally linear fit approaches the globally linear fit in exactly the same manner as the local constant fit approached the globally constant fit, namely $\bar{y}$. However, while the local constant estimator had the property that irrelevant variables could be totally smoothed out, the same does not hold for the local linear estimator which can lead to excessive variability in the presence of irrelevant regressors.

The bias and variance of this estimator were presented in Section 4.1. A multivariate version of the local linear estimator for mixed data settings follow naturally using generalized product kernels; see Li and Racine (2004) for details.

4.2.1 A Simulated Bivariate Example

We consider an example where we simulate a sample of size $n = 50$ where x is uniformly distributed and $y = \sin(2\pi x) + \epsilon$ where ϵ is normally distributed with $\sigma = 0.25$. We first consider the case where least-squares cross-validation is used to select the bandwidths. Figure 4.1 presents the data, the true DGP, and the local constant and local linear estimators of $g(x) = \sin(2\pi x)$.

It can be seen in Figure 4.1 that the local constant estimator displays some apparent edge bias as the estimator flares slightly downwards on the rightmost edge and slightly upwards on the leftmost edge as would be expected when one examines its approximate bias. However, both estimators provide faithful descriptions of the underlying DGP.

Next, we consider the differing behaviors of the local constant and local linear estimators as $h \to \infty$. We set the respective bandwidths at $h = 100{,}000$, and Figure 4.2 presents the data, the true DGP, and the local constant and local linear estimators.

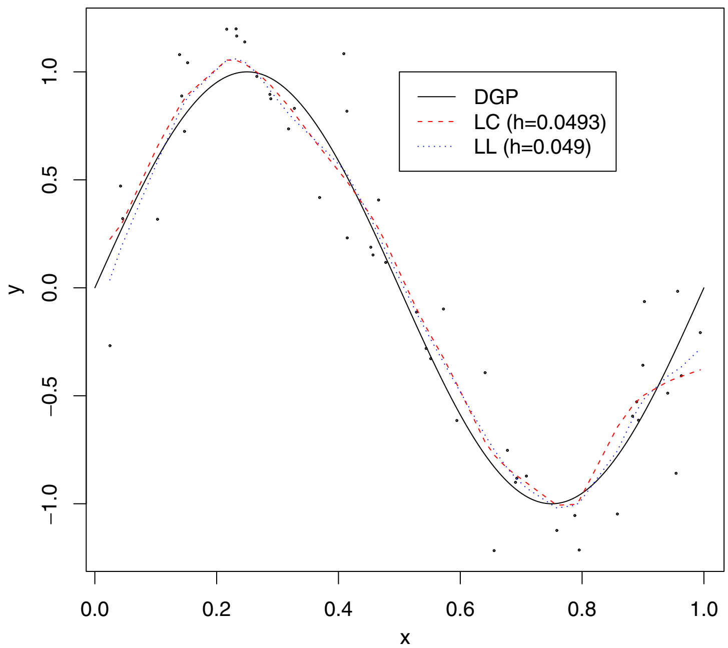

Fig. 4.1 The local constant and local linear estimators using least-squares cross-validation, $n = 50$.

Figure 4.2 clearly illustrates the markedly different properties of each estimator for large h, and underscores the fact that the local linear estimator cannot completely remove a variable by oversmoothing.

Suppose one was interested in marginal effects. In this case you might consider the local constant and local linear estimators of $\beta(x)$. Figure 4.3 plots the resulting estimates of response based upon the cross-validated bandwidths.

Readers may think that these estimators are not all that smooth, and they would of course be correct. Remember that we have a small sample ($n = 50$), are using a stochastic bandwidth, and as n increases the estimates will become progressively smoother. However, this is perhaps a good place to point out that common parametric specifications found in much applied econometric work would completely fail to capture even the simple mean and response considered here. Recall that this is the horse race referred to previously, and though the estimates

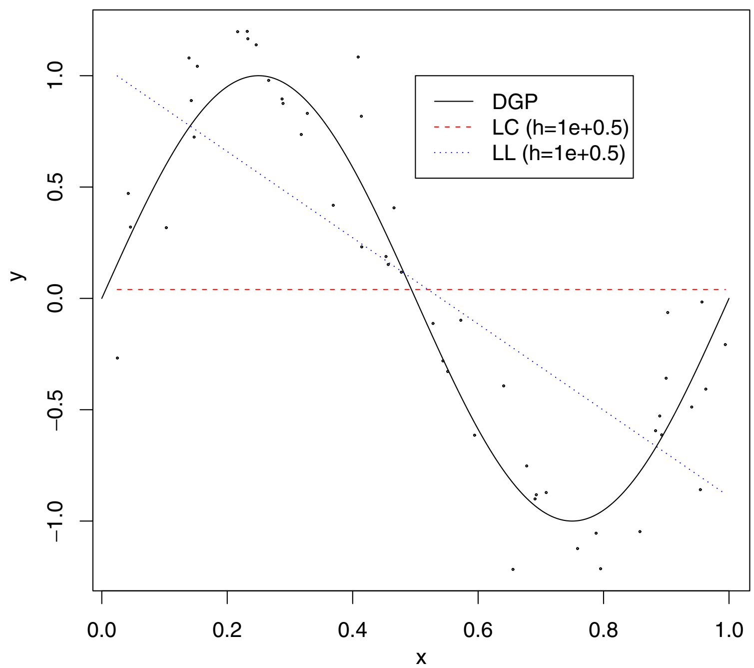

Fig. 4.2 The oversmoothed local constant and local linear estimators using $h = 100{,}000$, $n = 50$.

might not be all that pleasing to some readers, they are indeed highly informative.

4.2.2 An Illustrative Comparison of Bandwidth Selection Methods

To assess how various bandwidth selection methods perform on actual data, we consider the following example using data from Fox's (2002) `car` library in R (R Development Core Team (2007)). The dataset consists of 102 observations, each corresponding to a particular occupation. The dependent variable is the prestige of Canadian occupations, measured by the Pineo–Porter prestige score for occupation taken from a social survey conducted in the mid-1960s. The explanatory variable is average income for each occupation measured in 1971 Canadian dollars. Figure 4.4 plots the data and five local linear regression estimates, each differing in their window widths, the window widths being

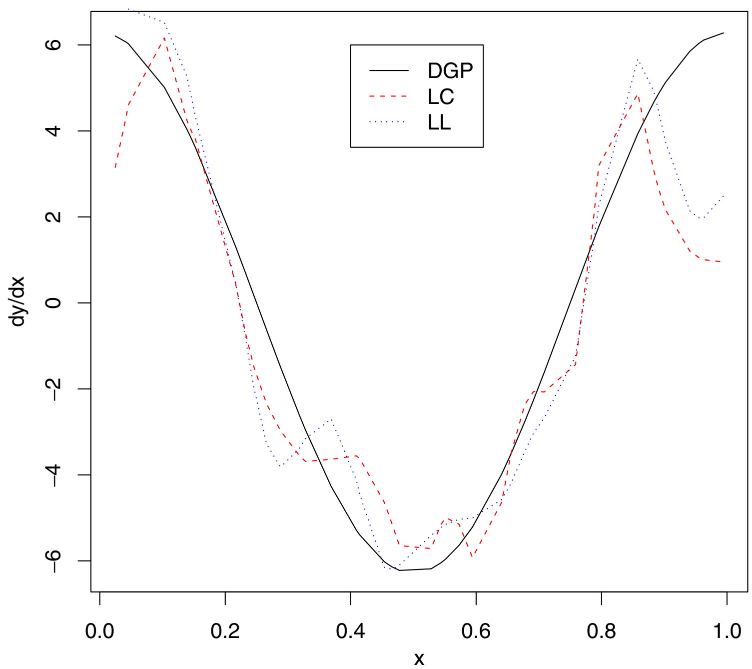

Fig. 4.3 The local constant and local linear estimators of response $\beta(x)$ using least-squares cross-validation, $n = 50$, $dy/dx = 2\pi\cos(2\pi x)$.

undersmoothed, oversmoothed, Ruppert et al.'s (1995) direct plug-in, Hurvich et al.'s (1998) corrected AIC ("AIC$_c$"), and cross-validation. A second order Gaussian kernel was used throughout.

It can be seen that the oversmoothed local linear estimate is globally linear and in fact is exactly a simple linear regression of y on x as expected, while the AIC$_c$ and CV criterion appears to provide the most reasonable fit to this data. As noted, in mixed data settings there do not exist plug-in rules. We have experienced reasonable performance using cross-validation and the AIC$_c$ criterion in a variety of settings.

4.2.3 A Multivariate Mixed-Data Application

For what follows, we consider an application that involves multiple regression analysis with qualitative information. This example is taken from Wooldridge (2003, p. 226).

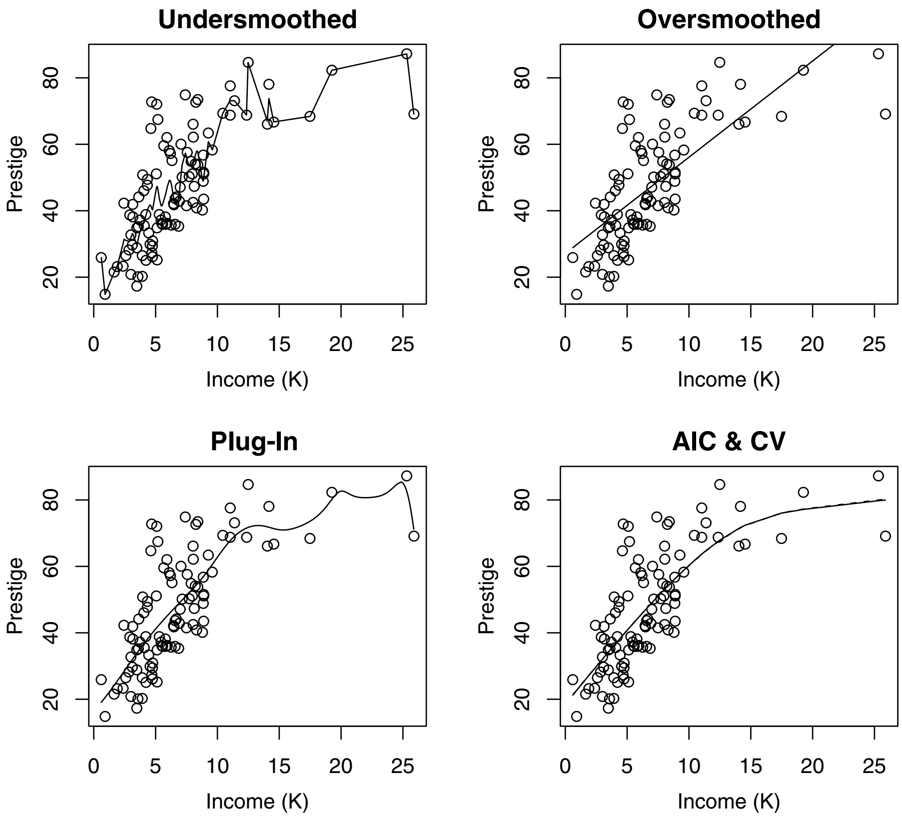

Fig. 4.4 Local linear kernel estimates with varying window widths. Bandwidths are undersmoothed ($0.1\sigma n^{-1/5}$), oversmoothed ($10^3\sigma n^{-1/5}$), AIC_C and CV ($3.54\sigma n^{-1/5}$, $3.45\sigma n^{-1/5}$), and plug-in ($1.08\sigma n^{-1/5}$).

We consider modeling an hourly wage equation for which the dependent variable is log(wage) (lwage) while the explanatory variables include three continuous variables, namely educ (years of education), exper (the number of years of potential experience), and tenure (the number of years with their current employer) along with two qualitative variables, female ("Female"/"Male") and married ("Married"/"Notmarried"). For this example there are $n = 526$ observations. We use Hurvich et al.'s (1998) AIC_c approach for bandwidth selection, which is summarized in Table 4.1.

Table 4.1 Bandwidth summary for the hourly wage equation.

```
Regression Data (526 observations, 5 variable(s)):

Regression Type: Local Linear
Bandwidth Selection Method: Expected Kullback-Leibler Cross-Validation
Formula: lwage ~ factor(female)+factor(married)+educ+exper+tenure
Bandwidth Type: Fixed
Objective Function Value: -0.8570284 (achieved on multistart 5)

factor(female)    Bandwidth:  0.01978275     Lambda Max:  0.500000
factor(married)   Bandwidth:  0.15228887     Lambda Max:  0.500000
educ              Bandwidth:  7.84663015   Scale Factor:  6.937558
exper             Bandwidth:  8.43548175   Scale Factor:  1.521636
tenure            Bandwidth: 41.60546059   Scale Factor: 14.099208

Continuous Kernel Type: Second-Order Gaussian
No. Continuous Explanatory Vars.: 3

Unordered Categorical Kernel Type: Aitchison and Aitken
No. Unordered Categorical Explanatory Vars.: 2
```

We display partial regression plots in Figure 4.5. A “partial regression plot” is simply a 2D plot of the outcome y versus one covariate x_j when all other covariates are held constant at their respective medians/modes. We also plot bootstrapped variability bounds which are often preferable to those obtained via the asymptotic approximations.[2]

Figure 4.6 presents the partial response plots along with their bootstrapped error bounds.

Note that, for the two categorical variables, the gradient is computed as the difference in wages, other variables held constant at their respective medians/modes, when one is, say, married versus not married. Note that for the leftmost value of each attribute (“Female” and “Married”) the difference is zero as we take the difference between each value assumed by the variable and the first level of each; see Racine et al. (2006) for the construction of response for categorical variables.

[2] The asymptotic formula is based on small-h approximations. As noted, sometimes optimal smoothing can appropriately deliver $h \to \infty$. As this cannot be known in advance, the asymptotic approximations will naturally perform poorly when this is the case.

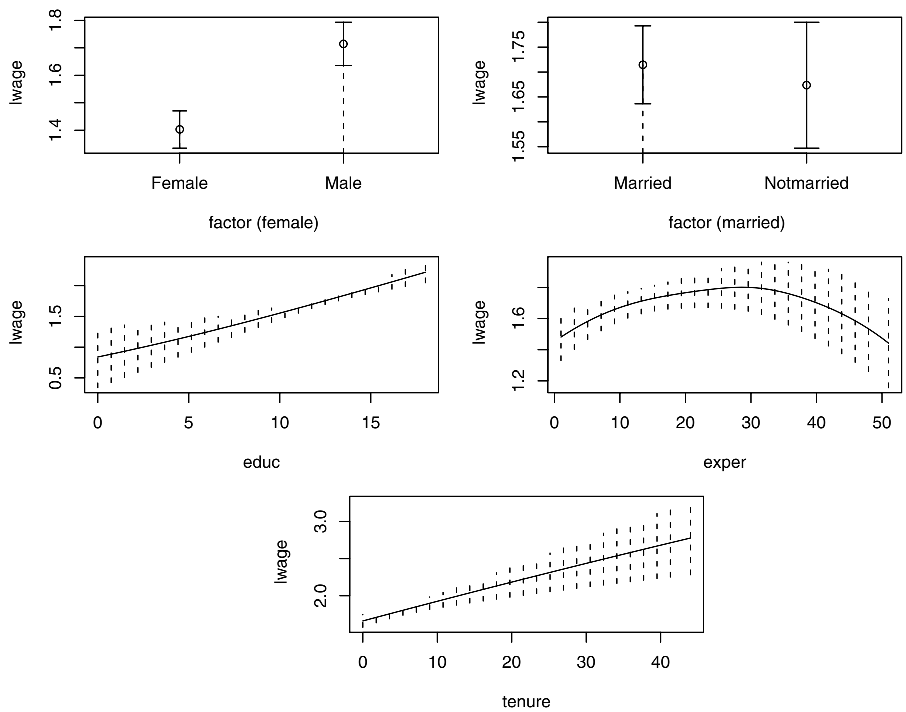

Fig. 4.5 Partial local linear nonparametric regression plots with bootstrapped pointwise error bounds for the Wooldridge (2002) 'wage1' dataset.

4.3 Assessing Goodness-of-Fit

We will require a unit-free measure of goodness-of-fit for nonparametric regression models which is comparable to that used for parametric regression models, namely R^2. Note that this will clearly be a *within-sample* measure of goodness-of-fit. Given the drawbacks of computing R^2 based on the decomposition of the sum of squares (such as possible negative values), there is an alternative definition and method for computing R^2 that can be used that is directly applicable to any model, linear or nonlinear. Letting Y_i denote the outcome and $\hat{Y}_i$ the fitted value for observation i, we may define R^2 as follows:

$$R^2 = \frac{\left[\sum_{i=1}^{n}(Y_i - \bar{y})(\hat{Y}_i - \bar{y})\right]^2}{\sum_{i=1}^{n}(Y_i - \bar{y})^2 \sum_{i=1}^{n}(\hat{Y}_i - \bar{y})^2},$$

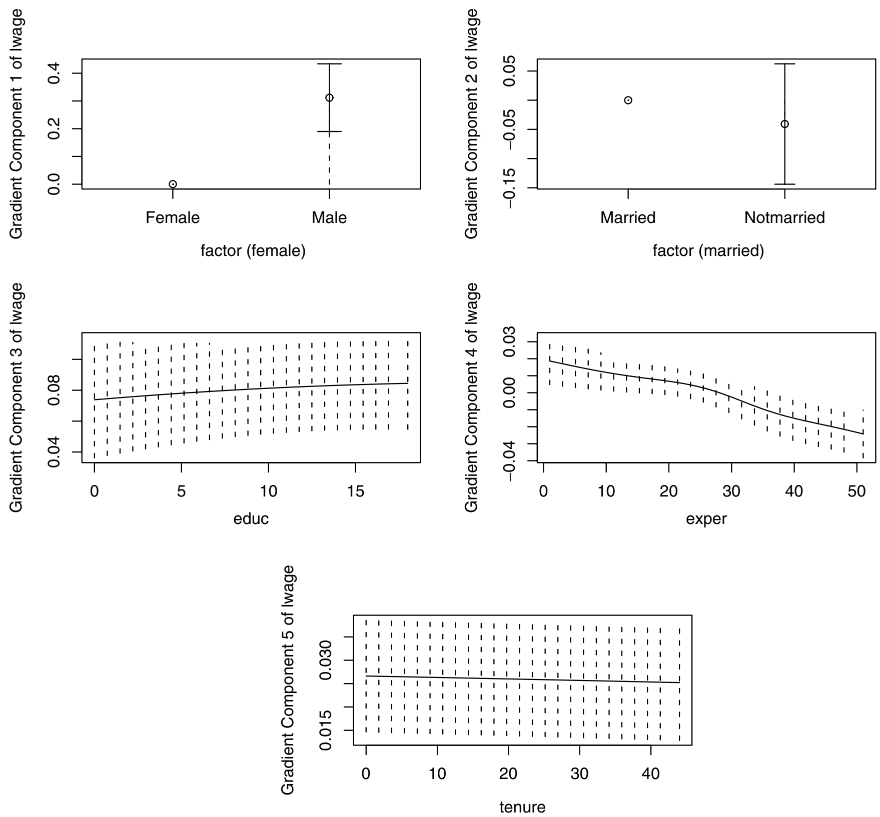

Fig. 4.6 Partial local linear nonparametric response plots with bootstrapped pointwise error bounds for the Wooldridge (2002) 'wage1' dataset.

and this measure will *always* lie in the range $[0, 1]$ with the value 1 denoting a perfect fit to the sample data and 0 denoting no predictive power above that given by the unconditional mean of the target. It can be demonstrated that this method of computing R^2 is identical to the standard measure computed as $\sum_{i=1}^{n}(\hat{Y}_i - \bar{y})^2 / \sum_{i=1}^{n}(Y_i - \bar{y})^2$ *when the model is linear and includes an intercept term and OLS is used to fit it.* This useful measure will permit direct comparison of within-sample goodness-of-fit subject to the obvious qualification that this is by no means a model selection criterion, rather, simply a summary measure that some may wish to report. This measure could, of course, also be computed using out-of-sample predictions and out-of-sample realizations. If we consider models estimated on a randomly selected subset of data and evaluated on an independent sample of hold-out data, this measure computed for the hold-out observations might serve to help assess various models, particularly when averaged over a number of independent hold-out datasets.[3]

By way of example, for the application taken from Wooldridge (2003, p. 226) above, the local linear model had an R^2 of 51.5% using this measure, which is directly comparable to the *unadjusted* R^2 from a parametric model.

4.4 A Resistant Local Constant Method

Nonparametric kernel methods are often (correctly) criticized due to their lack of robustness in the more traditional sense, namely, robust to the presence of contaminated data that can arise due to measurement errors, data entry errors, and the like. Methods that are robust in the more traditional sense are often referred to as "resistant" since they "resist" the presence of a small number of bad data values. Leung (2005) has recently proposed a novel method for resistant robust kernel regression. This is an exciting new development that is deserving of attention.

4.4.1 Leung's (2005) Resistant Kernel Regression Approach

We let $\{X_i, Y_i\}_{i=1}^n$ denote a set of data and consider the regression of Y on X at the n design points $\{X_i\}_{i=1}^n$,

$$Y_i = g(X_i) + \epsilon_i, \quad i = 1, \ldots, n, \tag{4.2}$$

where $g(\cdot)$ is an unknown functional of X and $\{\epsilon_i\}_{i=1}^n$ are *i.i.d.* random errors having distribution $F(\cdot)$.

The local constant kernel smoother of $g(x)$, denoted here as $\hat{g}_h(x)$ is given by

$$\hat{g}_h(x) \equiv \arg\min_a \sum_{i=1}^n (Y_i - a)^2 K\left(\frac{X_i - x}{h}\right), \tag{4.3}$$

[3] There exist a number of alternate measures of goodness of fit that are generated by the package. See the help file (i.e., type ?npreg) for details.

where h is a bandwidth that determines the amount of local smoothing and $K(\cdot)$ is a kernel function such that $\int K(z)\,dz = 1$, $\int zK(z)\,dz = 0$, and $\int z^2K(z)\,dz = k_2 < \infty$, for instance. The main problem in applied settings is how to best choose h.

A resistant local constant kernel smoother, on the other hand, can be obtained via

$$\tilde{g}_{h|c}(x) \equiv \arg\min_a \sum_{i=1}^{n} \rho_c(Y_i - a)K\left(\frac{X_i - x}{h}\right). \tag{4.4}$$

where ρ_c is, for instance, Huber's (1964) ρ_c function underlying M estimators which is given by (Maronna et al. (2006, p. 26))

$$\rho_c(u) = \begin{cases} u^2 & \text{if } |u| \leq c \\ 2c|u| - c^2 & \text{if } |u| > c \end{cases}. \tag{4.5}$$

In order to compute $\tilde{g}_{h|c}(x)$, the resistance parameter c must be specified by the user. One popular rule-of-thumb is $c = 1.345 \times s$ where s is a robust measure of scale such as the median absolute deviation about the median (MAD). This popular rule ensures 95% efficiency relative to the homoskedastic normal model in a location problem. Clearly this approach is more computationally demanding than the methods outlined in *Regression*. However, convincing simulations and applications provided by Leung (2005) indicate that this methods is deserving of attention by those worried about the presence of outliers.

Related work includes Stone (1977) and Cleveland (1979) who consider resistant local polynomial fitting using weighted least squares,[4] Cantoni and Ronchetti (2001) who consider smoothing splines with robust choice of the smoothing parameter along the lines of Leung (2005), Fan and Jiang (2000) who consider robust one-step local polynomial estimators but who did not address the issue of bandwidth selection, and Wang and Scott (1994) who consider locally

[4] Their method "lowess" stands for "locally weighted regression." The robustness follows from iterative fitting where the assigned weights are inversely proportional to the residuals from the previous fit, hence outliers tend to be downweighted.

weighted polynomials fitted via linear programming. See also Čížek and Härdle (2006) who consider robust estimation of dimension-reduction regression models.

The literature on resistant kernel methods is a development that has the potential to refine kernel smoothing along an important dimension leading to a set of truly robust methods.

5

Semiparametric Regression

Semiparametric methods constitute some of the more popular methods for flexible estimation. Semiparametric models are formed by combining parametric and nonparametric models in a particular manner. Such models are useful in settings where fully nonparametric models may not perform well, for instance, when the curse of dimensionality has led to highly variable estimates or when one wishes to use a parametric regression model but the functional form with respect to a subset of regressors or perhaps the density of the errors is not known. We might also envision situations in which some regressors may appear as a linear function (i.e., linear in variables) but the functional form of the parameters with respect to the other variables is not known, or perhaps where the regression function is nonparametric but the structure of the error process is of a parametric form.

Semiparametric models can best be thought of as a compromise between fully nonparametric and fully parametric specifications. They rely on parametric assumptions and can therefore be misspecified and inconsistent, just like their parametric counterparts.

A variety of semiparametric methods have been proposed. For what follows we shall restrict attention to regression-type models, and we consider three popular methods, namely the partially linear, single index, and varying coefficient specifications.

5.1 Partially Linear Models

The partially linear model is one of the simplest semiparametric models used in practice, and was proposed by Robinson (1988) while Racine and Liu (2007) extended the approach to handle the presence of categorical covariates. Many believe that, as the model is apparently simple, its computation ought to also be simple. However, the apparent simplicity hides the perhaps under-appreciated fact that bandwidth selection for partially linear models can be orders of magnitude more computationally burdensome than that for fully nonparametric models, for one simple reason. As will be seen, data-driven bandwidth selection methods such as cross-validation are being used, and the partially linear model involves cross-validation to regress y on Z (Z is multivariate) then *each column of* X on Z, whereas fully nonparametric regression involves cross-validation of y on X only. The computational burden associated with partially linear models is therefore much more demanding than for nonparametric models, so be forewarned.

A semiparametric partially linear model is given by

$$Y_i = X_i'\beta + g(Z_i) + u_i, \quad i = 1, \ldots, n, \tag{5.1}$$

where X_i is a $p \times 1$ vector, β is a $p \times 1$ vector of unknown parameters, and $Z_i \in \mathbb{R}^q$. The functional form of $g(\cdot)$ is not specified. The finite dimensional parameter β constitutes the parametric part of the model and the unknown function $g(\cdot)$ the nonparametric part. The data is assumed to be i.i.d. with $E(u_i|X_i, Z_i) = 0$, and we allow for a conditionally heteroskedastic error process $E(u_i^2|x, z) = \sigma^2(x, z)$ of unknown form. We focus our discussion on how to obtain a $\sqrt{n}$-consistent estimator of β, as once this is done an estimator of $g(\cdot)$ can be easily obtained via the nonparametric regression of $Y_i - X_i\hat{\beta}$ on z.

Taking the expectation of (5.1) conditional on Z_i, we get

$$E(Y_i|Z_i) = E(X_i|Z_i)'\beta + g(Z_i). \tag{5.2}$$

Subtracting (5.2) from (5.1) yields

$$Y_i - E(Y_i|Z_i) = (X_i - E(X_i|Z_i))' \beta + u_i. \tag{5.3}$$

Defining the shorthand notation $\tilde{Y}_i = Y_i - E(Y_i|Z_i)$ and $\tilde{X}_i = X_i - E(X_i|Z_i)$, and applying the least squares method to (5.3), we obtain an estimator of β given by

$$\hat{\beta}_{\text{inf}} = \left[\sum_{i=1}^{n} \tilde{X}_i \tilde{X}_i'\right]^{-1} \sum_{i=1}^{n} \tilde{X}_i \tilde{Y}_i. \tag{5.4}$$

The above estimator $\hat{\beta}_{\text{inf}}$ is not feasible because $E(Y_i|Z_i)$ and $E(X_i|Z_i)$ are unknown. However, we know that these conditional expectations can be consistently estimated using the kernel methods described in *Regression*, so we can replace the unknown conditional expectations that appear in $\hat{\beta}_{\text{inf}}$ with their kernel estimators thereby obtaining a feasible estimator of β. Some identification conditions will be required in order to identify the parameter vector β, and we refer the interested reader to Robinson (1988).

5.1.1 A Partially Linear Example

Suppose that we again consider Wooldridge's (2002) "wage1" dataset, but now assume that the researcher is unwilling to specify the nature of the relationship between exper and lwage, hence relegates exper to the nonparametric part of a semiparametric partially linear model. Table 5.1 presents a summary from the partially linear specification.

It is of interest to compare these results with that for a linear model that is quadratic in experience, which is summarized in Table 5.2 and with the local linear specification outlined in *Regression*. First, we note that the parameter estimates and their respective standard errors are comparable in magnitude with those from the fully parametric specification listed in Table 5.2. Second, in terms of in-sample fit, the semiparametric partially linear specification ($R^2 = 44.9\%$) performs slightly better than the parametric specification ($R^2 = 43.6\%$) while the fully nonparametric specification ($R^2 = 51.5\%$) outperforms both the fully parametric and partially linear specifications.

Table 5.1 Model summary for the partially linear hourly wage equation.

```
Partially Linear Model
Regression data: 526 training points, in 5 variable(s)
With 4 linear parametric regressor(s), 1 nonparametric regressor(s)

                  y(z)
Bandwidth(s): 2.050966

                  x(z)
Bandwidth(s): 4.1943673
              1.3531783
              3.1605552
              0.7646561

                  factor (female) factor (married)     educ       tenure
Coefficient(s):      0.2902499      -0.03722828     0.07879512  0.01662935
Standard error(s):   0.0359527       0.04230253     0.00676465  0.00308927

Kernel Regression Estimator: Local Constant
Bandwidth Type: Fixed

Residual standard error: 0.1553021
R-squared: 0.4493789
```

Table 5.2 Model summary for the fully linear hourly wage equation.

Coefficients:	Estimate	Std. Error
(Intercept)	0.1811615	0.1070747
factor(female)Male	0.2911303	0.0362832
factor(married)Notmarried	-0.0564494	0.0409259
educ	0.0798322	0.0068273
tenure	0.0160739	0.0028801
exper	0.0300995	0.0051931
I(exper^2)	-0.0006012	0.0001099

Multiple R-Squared: 0.4361,   Adjusted R-squared: 0.4296

5.2 Index Models

A semiparametric single index model is of the form:

$$Y = g(X'\beta_0) + u, \tag{5.5}$$

where Y is the dependent variable, $X \in \mathbb{R}^q$ is the vector of explanatory variables, β_0 is the $q \times 1$ vector of unknown parameters, and u is the error satisfying $E(u|X) = 0$. The term $x'\beta_0$ is called a "single index" because it is a scalar (a single index) even though x is a vector. The functional form of $g(\cdot)$ is unknown to the researcher. This model is semiparametric in nature since the functional form of the linear index is specified, while $g(\cdot)$ is left unspecified.

Ichimura (1993), Manski (1988), and Horowitz (1998, pp. 14–20) provide excellent intuitive explanations of the identifiability conditions underlying semiparametric single index models (i.e., the set of conditions under which the unknown parameter vector β_0 and the unknown function $g(\cdot)$ can be sensibly estimated), and we direct the reader to these articles for details.

5.2.1 Ichimura's Method

Consider the case where y is continuous. If the functional form of $g(\cdot)$ were known, we would have a standard nonlinear regression model, and we could use the nonlinear least squares method to estimate β_0 by minimizing

$$\sum_{i=1}(Y_i - g(X_i'\beta))^2 \tag{5.6}$$

with respect to β.

In the case of an unknown function $g(\cdot)$, we first need to estimate $g(\cdot)$. However, the kernel method does not estimate $g(X_i'\beta_0)$ directly because not only is $g(\cdot)$ unknown, but so too is β_0. Nevertheless, for a given value of β we can estimate

$$G(X_i'\beta) \stackrel{\text{def}}{=} E(Y_i|X_i'\beta) = E[g(X_i'\beta_0)|X_i'\beta] \tag{5.7}$$

by the kernel method, where the last equality follows from the fact that $E(u_i|X_i'\beta) = 0$ for all β since $E(u_i|X_i) = 0$.

Note that when $\beta = \beta_0$, $G(X_i'\beta_0) = g(X_i'\beta_0)$, while in general, $G(X_i'\beta) \neq g(X_i'\beta_0)$ if $\beta \neq \beta_0$. Ichimura (1993) suggests estimating $g(X_i'\beta_0)$ by $\hat{G}_{-i}(X_i'\beta)$ and choosing β by (semiparametric) nonlinear least squares, where $\hat{G}_{-i}(X_i'\beta)$ is a leave-one-out nonparametric kernel estimator of $G(X_i'\beta)$.

A Single Index Example for Continuous Y Next, we consider applying Ichimura (1993)'s single index method which is appropriate for contin-

Table 5.3 Model summary for the semiparametric index model of the hourly wage equation.

```
Single Index Model
Regression Data: 526 training points, in 6 variable(s)

      factor (female) factor (married)   educ     exper      expersq    tenure
Beta:       1           -2.783907      9.947963  3.332755  -0.0750266  2.310801
Bandwidth: 2.457583
Kernel Regression Estimator: Local Constant

Residual standard error: 0.1552531
R-squared: 0.4501873
```

uous outcomes, unlike that of Klein and Spady (1993) outlined below. We again make use of Wooldridge's (2002) "wage1" dataset. Table 5.3 presents a summary of the analysis.

It is interesting to compare this model with the parametric and nonparametric models outlined above as it provides an in-sample fit (45.1%) that lies in between the parametric model (43.6%) and the fully nonparametric local linear model (51.5%).

5.2.2 Klein and Spady's Estimator

We now consider the case where y is binary. Under the assumption that ϵ_i and X_i are independent, Klein and Spady (1993) suggested estimating β by maximum likelihood methods. The estimated log-likelihood function is

$$\mathcal{L}(\beta, h) = \sum_i (1 - Y_i) \ln(1 - \hat{g}_{-i}(X_i'\beta)) + \sum_i Y_i \ln(\hat{g}_{-i}(X_i'\beta)), \quad (5.8)$$

where $\hat{g}_{-i}(X_i'\beta)$ is the leave-one-out estimator. Maximizing (5.8) with respect to β and h leads to the semiparametric maximum likelihood estimator of β proposed by Klein and Spady. Like Ichimura's (1993) estimator, maximization must be performed numerically.

A Single Index Example for Binary Y We again consider data on birthweights taken from the R MASS library (Venables and Ripley (2002)), and compute a single index model (the parametric Logit model and a nonparametric conditional mode model results are reported in *Conditional Density Estimation*). The outcome is an indicator of low

Table 5.4 Confusion matrix for the low birthweight data using the single index model.

	Predicted	
Actual	0	1
0	125	5
1	37	22

infant birthweight (0/1) and so Klein and Spady's (1993) approach is appropriate. The confusion matrix is presented in Table 5.4.

It can be seen that, based on in-sample classification, this model does somewhat better than the parametric logit model when modeling this dataset. The single index model correctly classifies $(125 + 22)/189 = 77.8\%$ of low/high birthweights while the Logit model correctly classifies $(119 + 25)/189 = 76.1\%$.

5.3 Smooth Coefficient (Varying Coefficient) Models

The smooth coefficient model was proposed by Hastie and Tibshirani (1993) and is given by

$$\begin{aligned} Y_i &= \alpha(Z_i) + X_i'\beta(Z_i) + u_i = (1 + X_i') \begin{pmatrix} \alpha(Z_i) \\ \beta(Z_i) \end{pmatrix} + u_i \\ &= W_i'\gamma(Z_i) + u_i, \end{aligned} \tag{5.9}$$

where X_i is a $k \times 1$ vector and where $\beta(z)$ is a vector of unspecified smooth functions of z. Premultiplying by W_i and taking expectations with respect to Z_i yields

$$E[W_iY_i|Z_i] = E[W_iW_i'|Z_i]\gamma(Z_i) + E[W_iu_i|Z_i]. \tag{5.10}$$

We can express $\gamma(\cdot)$ as

$$\gamma(Z_i) = (E[W_iW_i'|Z_i])^{-1}E[W_iY_i|Z_i]. \tag{5.11}$$

Li and Racine (2007b) consider a kernel-based approach that admits both discrete and continuous regressors. They propose using a local constant estimator of the form:

$$\hat{\gamma}(z) = \left[\sum_{j=1}^{n} W_jW_j'K\left(\frac{Z_j - z}{h}\right)\right]^{-1} \sum_{j=1}^{n} W_jY_jK\left(\frac{Z_j - z}{h}\right)$$

Table 5.5 Model summary for the smooth coefficient hourly wage equation.

```
Smooth Coefficient Model
Regression data: 526 training points, in 2 variable(s)

              factor (female) factor (married)
Bandwidth(s):   0.001813091      0.1342957

Bandwidth Type: Fixed

Residual standard error: 0.1470017
R-squared: 0.4787102

Average derivative(s):
   Intercept       educ         tenure         exper       expersq
 0.3402224978  0.0786499683  0.0142981775  0.0300505722 -0.0005950969
```

and propose a variant of cross-validation for bandwidth selection; see Li and Racine (2007b) for details. The fitted model is given by

$$Y_i = \hat{Y}_i + \hat{u}_i = W_i'\hat{\gamma}(Z_i) + \hat{u}_i.$$

5.3.1 A Smooth Coefficient Example

Suppose that we once again consider Wooldridge's (2002) "wage1" dataset, but now assume that the researcher is unwilling to presume that the coefficients associated with the continuous variables do not vary with respect to the categorical variables female and married. Table 5.5 presents a summary from the smooth coefficient specification.

Comparing these results with that for a linear model that is quadratic in experience summarized in Table 5.2, we observe that the average parameter values are comparable in magnitude with those from the fully parametric specification listed in Table 5.2. However, the semiparametric smooth coefficient model performs better than the parametric specification in terms of in-sample fit ($R^2 = 47.8\%$ versus $R^2 = 43.6\%$). This suggests that the additional flexibility offered by allowing all parameters to vary with respect to the continuous variables has resulted in an improved fit.

6

Panel Data Models

The nonparametric and semiparametric estimation of panel data models has received less attention than the estimation of standard regression models. Data panels are samples formed by drawing observations on N cross-sectional units for T consecutive periods yielding a dataset of the form $\{Y_{it}, Z_{it}\}_{i=1,t=1}^{N,T}$. A panel is therefore simply a collection of N individual time series that may be short ("small T") or long ("large T").

The nonparametric estimation of time series models is itself an evolving field. However, when T is large and N is small then there exists a lengthy time series for each individual unit and in such cases one can avoid estimating a panel data model by simply estimating separate nonparametric models for each individual unit using the T individual time series available for each. If this situation applies, we direct the interested reader to Li and Racine (2007a, Chap. 18) for pointers to the literature on nonparametric methods for time series data.

When contemplating the nonparametric estimation of panel data models, one issue that immediately arises is that the standard (parametric) approaches that are often used for panel data models (such as first-differencing to remove the presence of so-called "fixed effects") are no longer valid unless one is willing to presume additively separable effects, which for many defeats the purpose of using nonparametric methods in the first place.

A variety of approaches have been proposed in the literature, including Wang (2003), who proposed a novel method for estimating nonparametric panel data models that utilizes the information contained in the covariance structure of the model's disturbances, Wang et al. (2005) who proposed a partially linear model with random effects, and Henderson et al. (2006) who consider profile likelihood methods for nonparametric estimation of additive fixed effect models which are removed via first differencing. In what follows, we consider direct nonparametric estimation of fixed effects models using the methods outlined in *Regression*.

6.1 Nonparametric Estimation of Fixed Effects Panel Data Models

Consider the following nonparametric panel data regression model,

$$Y_{it} = g(X_{it}) + u_{it}, \quad i = 1, 2, \ldots, N, \;\; t = 1, 2, \ldots, T,$$

where $g(\cdot)$ is an unknown smooth function, $X_{it} = (X_{it,1}, \ldots, X_{it,q})$ is of dimension q, all other variables are scalars, and $E(u_{it}|X_{i1}, \ldots, X_{iT}) = 0$.

We say that panel data is "poolable" if one can "pool" the data, by in effect, ignoring the time series dimension, that is, by summing over both i and t without regard to the time dimension thereby effectively putting all data into the same pool then directly applying the methods in, say, *Regression*. Of course, if the data is not poolable this would obviously not be a wise choice.

However, to allow for the possibility that the data is in fact *potentially* poolable, one can introduce an *unordered* discrete variable, say $\delta_i = i$ for $i = 1, 2, \ldots, N$, and estimate $E(Y_{it}|Z_{it}, \delta_i) = g(Z_{it}, \delta_i)$ nonparametrically using the mixed discrete and continuous kernel approach introduced in *Density and Probability Function Estimation*. The δ_i variable is akin to including cross-sectional dummies (as is done, for instance, in the least-squares dummy variable approach for linear panel data regression models). Letting $\hat{\lambda}$ denote the cross-validated smoothing parameter associated with δ_i, then if $\hat{\lambda}$ is at its upper bound, one gets $g(Z_{it}, \delta_i) = g(Z_{it})$ and the data is thereby pooled in the resulting estimate of $g(\cdot)$. If, on the other hand, $\hat{\lambda} = 0$ (or is close to 0), then this effectively estimates each $g_i(\cdot)$ using only the time series for the

ith individual unit. Finally, if $0 < \hat{\lambda} < 1$, one might interpret this as a case in which the data is partially poolable.

It bears mentioning that, in addition to the issue of poolability, there is also the issue of correcting inference for potential serial correlation in the u_{it} residuals. That is, even if the data is poolable, you cannot blindly apply the asymptotic approach; an appropriate bootstrapping approach is likely best in practice.

6.1.1 Application to a US Airline Cost Panel

We consider a panel of annual observations for six US airlines for the 15 year period 1970 to 1984 taken from the Ecdat R package (Croissant (2006)) as detailed in Greene (2003, Table F7.1, p. 949). The variables in the panel are airline ("airline"), year ("year"), the logarithm of total cost in $1,000 ("lcost"), the logarithm of an output index in revenue passenger miles ("loutput"), the logarithm of the price of fuel ("lpf"), and load factor, i.e., the average capacity utilization of the fleet ("lf"). We treat "airline" as an unordered factor and "year" as an ordered factor and use a local linear estimator with Hurvich et al.'s (1998) AIC$_c$ approach.

Table 6.1 presents a summary of the bandwidths, while Figure 6.1 presents the partial regression plots.

An examination of Table 6.1 reveals that the bandwidth for the unordered variable "airline" is 0.0025 which suggests that the model is not poolable across airlines (i.e., a separate time-series model for each airline is likely appropriate). Figure 6.1 indicates that costs are rising with output and the price of fuel, while they fall with the load factor.

By way of comparison, in Table 6.2 we present results for a linear fixed effects panel data model using the R plm package (Croissant and Millo (2007)).

Table 6.1 Bandwidth summary for the local linear US Airline panel data model.

```
Var.: loutput           Bandwidth: 1020484     Scale Factor: 1.696225e+06
Var.: lpf               Bandwidth: 1417256     Scale Factor: 3.336533e+06
Var.: lf                Bandwidth: 0.0130355   Scale Factor: 0.472229
Var.: ordered(year)     Bandwidth: 0.1107695   Lambda Max:   1.000000
Var.: factor(airline)   Bandwidth: 0.0024963   Lambda Max:   1.000000
```

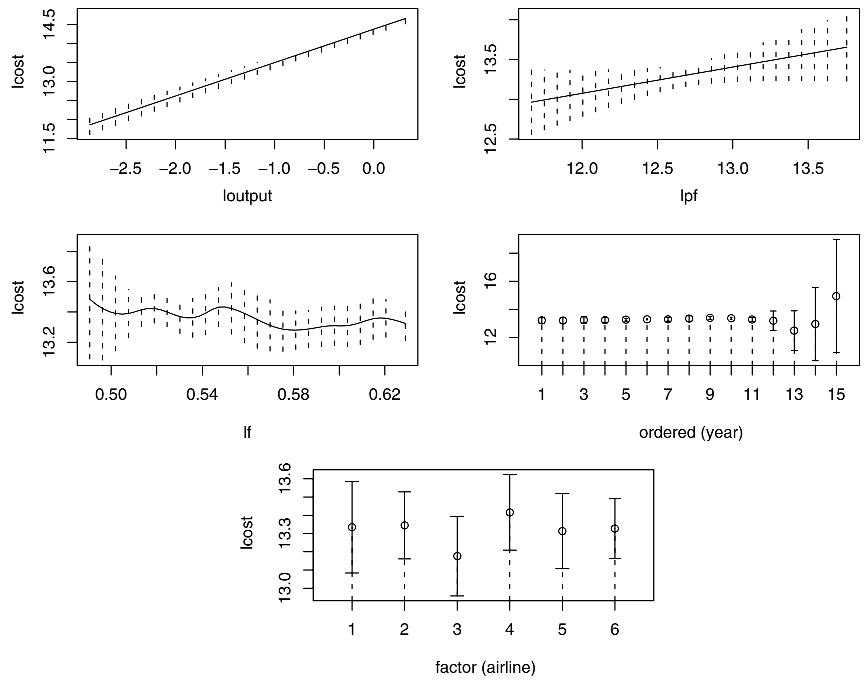

Fig. 6.1 Partial regression plot and bootstrapped error bounds for the US Airline panel.

Table 6.2 Model summary for the parametric fixed effects US Airline panel.

```
Model Description
Oneway (individual) effect
Within model
Model formula          : log(cost) ~ log(output) + log(pf) + lf

Coefficients
             Estimate Std. Error z-value  Pr(>|z|)
log(output)  0.919285   0.028841 31.8743 < 2.2e-16 ***
log(pf)      0.417492   0.014666 28.4673 < 2.2e-16 ***
lf          -1.070396   0.194611 -5.5002 3.794e-08 ***

Signif. codes:  0 ``***'' 0.001 ``**'' 0.01 ``*'' 0.05 ``.'' 0.1 `` '' 1
```

A comparison of Figure 6.1 and Table 6.2 reveals that both the parametric and nonparametric models are in agreement in that costs increase with output and the price of fuel and fall with the load factor, other things equal.

7

Consistent Hypothesis Testing

The literature on the use of nonparametric kernel methods for hypothesis testing has experienced tremendous growth and has spawned a variety of novel approaches for testing a range of hypotheses. There exist nonparametric methods for testing for correct specification of parametric models, tests for equality of distributions and equality of regression functions, among others.

Parametric tests typically require the analyst to specify the set of parametric alternatives for which the null hypothesis will be rejected. If, however, the null is false and yet there exist alternative models that the test cannot detect, then the test is said to be a "inconsistent" since it lacks power in certain directions. Nonparametric methods can be used to construct consistent tests, unlike their parametric counterparts.

To be precise, we define what we mean by a "consistent test." Let H_0 denote a null hypothesis whose validity we wish to test. A test is said to be a *consistent* test if

$$P(\text{Reject } H_0 \mid H_0 \text{ is false}) \to 1 \quad \text{as } n \to \infty.$$

The power of a test is defined as $P(\text{Reject } H_0 \mid H_0 \text{ is false})$. Therefore, a consistent test has asymptotic power equal to one.

For what follows, we consider a handful of tests that might be of use to practitioners.

7.1 Testing Parametric Model Specification

A variety of methods exist for testing for correctly specified parametric regression models including Härdle and Mammen (1993), Horowitz and Härdle (1994), Horowitz and Spokoiny (2001), Hristache et al. (2001) and Hsiao et al. (2007), among others. We briefly describe Hsiao et al.'s (2007) test as it admits the mix of continuous and categorical datatypes often encountered in applied settings.

Suppose one wished to test the correctness of a parametric regression model. We could state the null hypothesis as follows:

$$H_0 : E(Y|x) = m(x, \gamma_0), \text{ for almost all } x \text{ and for some } \gamma_0 \in \mathcal{B} \subset \mathbb{R}^p, \tag{7.1}$$

where $m(x,\gamma)$ is a known function with γ being a $p \times 1$ vector of unknown parameters (which clearly includes a linear regression model as a special case) and where $\mathcal{B}$ is a compact subset of $\mathbb{R}^p$. The alternative hypothesis is the negation of H_0, i.e., H_1: $E(Y|x) \equiv g(x) \neq m(x,\gamma)$ for all $\gamma \in \mathcal{B}$ on a set (of x) with positive measure. If we define $u_i = Y_i - m(X_i, \gamma_0)$, then the null hypothesis can be equivalently written as

$$E(u_i|X_i = x) = 0 \text{ for almost all } x. \tag{7.2}$$

A consistent model specification test can be constructed based on nonparametrically estimating (7.2) and averaging over the u_i in a particular manner, which we briefly describe. First, note that $E(u_i|X_i = x) = 0$ is equivalent to $[E(u_i|X_i = x)]^2 = 0$. Also, since we wish to test the null that $E(u_i|X_i = x) = 0$ for almost all x, we need to consider the expectation $E\{E(u_i|X_i = x)\}$ or equivalently $E\{[E(u_i|X_i = x)]^2\}$. By the law of iterated expectations it can be seen that $E\{[E(u_i|X_i = x)]^2\} = E\{u_i E(u_i|X_i = x)\}$. One can therefore construct a consistent test statistic based on a density weighted version of $E\{u_i E(u_i|X_i = x)\}$, namely $E\{u_i E(u_i|X_i) f(X_i)\}$, where $f(x)$ is the joint PDF of X. Density weighting is used here simply to avoid a random denominator that would otherwise appear in the kernel estimator.

The sample analogue of $E\{u_i E(u_i|X_i) f(X_i)\}$ is given by the formula $n^{-1}\sum_{i=1}^n u_i E(u_i|X_i) f(X_i)$. To obtain a feasible test statistic, we replace u_i by $\hat{u}_i$, where $\hat{u}_i = Y_i - m(X_i, \hat{\gamma})$ is the residual obtained from the parametric null model, and $\hat{\gamma}$ is a $\sqrt{n}$-consistent estimator of γ based on the null model (say, the nonlinear least squares estimator of γ). We estimate $E(u_i|X_i)f(X_i)$ by the leave-one-out kernel estimator $(n-1)^{-1}\sum_{j\neq i}^n \hat{u}_j K_{ij}$. Letting X_i be a vector of mixed discrete and continuous variables and using generalized product kernels, the test statistic is based upon

$$\begin{aligned} I_n &\stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^n \hat{u}_i \left\{ \frac{1}{n-1} \sum_{j=1, j\neq i}^n \hat{u}_j K_{ij} \right\} \\ &= \frac{1}{n(n-1)} \sum_{i=1}^n \sum_{j=1, j\neq i}^n \hat{u}_i \hat{u}_j K_{ij}. \end{aligned} \tag{7.3}$$

The studentized version of this test is denoted by J_n. Bootstrap methods can be used to obtain the distribution of I_n (J_n) under the null which can be used to form a bootstrap p-value; see Hsiao et al. (2007) for details.

7.1.1 Application — Testing Correct Specification of a Naïve Linear Model

Having estimated, say, a simple parametric wage model that is linear in variables, one might next test the null hypothesis that the parametric model is correctly specified using the approach of Hsiao et al. (2007) described above. Cross-validation is used to select the bandwidths, and a simple bootstrap method is used to compute the distribution of J_n under the null.

Table 7.1 Summary of the model specification test for the parametric hourly wage equation.

```
Consistent Model Specification Test
Parametric null model: lm(formula = lwage ~
                                    factor(female) +
                                    factor(married) +
                                    educ +
                                    exper +
                                    tenure,
                                    data = wage1,
                                    x = TRUE,
                                    y = TRUE)

Number of regressors: 5
IID Bootstrap (399 replications)

Test Statistic 'Jn': 5.542416   P Value: < 2.22e-16
```

Table 7.1 reveals that, not surprisingly, we reject this naïve specification that is linear in all variables, includes no interaction terms, and only allows the intercept to shift with respect to the categorical variables. Note that we are not putting this parametric forward as an ideal candidate, rather, we are simply demonstrating that the test is capable of detecting misspecified parametric models in finite-sample settings.

7.2 A Significance Test for Nonparametric Regression Models

Having estimated a parametric regression model, researchers often then proceed directly to the "test of significance." The significance test is often used to confirm or refute economic theories. However, in a parametric regression framework, sound parametric inference hinges on the correct functional specification of the underlying data generating process, and significance tests for misspecified parametric models will have misleading size and power thereby leading to faulty inference. A variety of approaches have been proposed in the literature including Lavergne and Vuong (1996), who considered the problem of selecting nonparametric regressors in a non-nested regression model framework, Donald (1997), who proposed a nonparametric test for selecting the factors in a multivariate nonparametric relationship, Racine (1997) who considered a consistent significance test for continuous regressors and Racine et al. (2006) who considered a consistent significance test for categorical regressors. See also Delgado and Manteiga (2001) for an alternative nonparametric test of significance of continuous variables in nonparametric regression models.

7.2.1 Categorical Regressors

Suppose we had estimated a nonparametric regression model where some regressors were categorical and some were continuous, and we wished to test whether some of the categorical regressors are irrelevant, i.e., redundant. One might apply the test of Racine et al. (2006), which we briefly describe. Let z denote the categorical explanatory variables that might be redundant, let X denote the remaining explanatory variables in the regression model, and let Y denote the dependent variable. Then the null hypothesis can be written as

$$H_0 : E(Y|x,z) = E(Y|x) \text{ almost everywhere}$$

The alternative hypothesis is the negation of the null hypothesis H_0, i.e., H_1: $E(Y|x,z) \neq E(Y|x)$ on a set with positive measure.

If we let $g(x) = E(Y|x)$ and let $m(x,z) = E(Y|x,z)$, then the null hypothesis is $m(x,z) = g(x)$ almost everywhere. Suppose that the univariate Z assumes c different values, $\{0,1,2,\ldots,c-1\}$. If $c = 2$, then Z is a 0–1 dummy variable, which in practice is probably the most frequently encountered situation.

Note that the null hypothesis H_0 is equivalent to $m(x,z=l) = m(x,z=0)$ for all X and for $l = 1,\ldots,c-1$. The test statistic is an estimator of

$$I = \sum_{l=1}^{c-1} E\left\{[m(x,z=l) - m(x,z=0)]^2\right\}.$$

Obviously $I \geq 0$ and $I = 0$ if and only if H_0 is true. Therefore, I serves as a proper measure for testing H_0. A feasible test statistic is given by

$$I_n = \frac{1}{n}\sum_{i=1}^{n}\sum_{l=1}^{c-1}\left[\hat{m}(X_i,Z_i=l) - \hat{m}(X_i,Z_i=0)\right]^2, \tag{7.4}$$

where $\hat{m}(X_i, Z_i = l)$ is the local constant or local linear regression estimator described in *Regression*.

It is easy to show that I_n is a consistent estimator of I. Therefore, $I_n \to 0$ in probability under H_0 and $I_n \to I > 0$ in probability under H_1. To obtain the null distribution of this statistic or of a studentized version, Racine et al. (2006) proposed two bootstrap procedures, both of which have sound finite-sample properties; see Racine et al. (2006) for details.

7.2.2 Continuous Regressors

Similar to that described above, the null hypothesis when testing for the significance of a continuous regressor can be written

$$H_0 : E(y|x, z) = E(Y|z) \text{ almost everywhere}$$

which is equivalent to

$$H_0 : \frac{\partial E(y|x, z)}{\partial x} = \beta(x) = 0 \text{ almost everywhere}$$

The test statistic is an estimator of

$$I = E\left\{\beta(x)^2\right\}. \tag{7.5}$$

A test statistic can be obtained by forming a sample average of I, replacing the unknown derivatives with their nonparametric estimates $\hat{\beta}(x_i)$ as described in Racine (1997), i.e.,

$$I_n = \frac{1}{n}\sum_{i=1}^{n}\hat{\beta}(X_i)^2, \tag{7.6}$$

where $\hat{\beta}(X_i)$ is the local constant or local linear partial derivative estimator described in *Regression*.

It is easy to show that I_n is a consistent estimator of I. Therefore, $I_n \to 0$ in probability under H_0 and $I_n \to I > 0$ in probability under H_1. To obtain the null distribution of this statistic or of a studentized version, bootstrap procedures can be used; see Racine (1997) for details.

An Illustration We consider Wooldridge's (2002) "wage1" dataset ($n = 526$) that contains a mix of continuous and categorical regressors

Table 7.2 Significance test summaries for the nonparametric local linear hourly wage equation.

```
Kernel Regression Significance Test
Type I Test with IID Bootstrap (399 replications)
Explanatory variables tested for significance:
factor(female) (1), factor(married) (2), educ (3), exper (4), tenure (5)

              factor(female) factor(married)    educ    exper   tenure
Bandwidth(s):     0.01978275       0.1522889 7.84663 8.435482 41.60546

Significance Tests
P Value:
factor(female)  < 2.22e-16 ***
factor(married) 0.0150376 *
educ            < 2.22e-16 ***
exper           < 2.22e-16 ***
tenure          0.0075188 **
---
Signif. codes:  0 ‘‘***’’ 0.001 ‘‘**’’ 0.01 ‘‘*’’ 0.05 ‘‘.’’ 0.1 ‘‘ ’’ 1
```

and apply the tests of significance described above. Results are summarized in Table 7.2, and the p-values indicate that all variables are significant at the conventional 5% level.

8

Computational Considerations

One of the great ironies in this field is that kernel methods, by their nature, are often so computationally burdensome that most researchers hesitate to apply them to those problems for which they are ideally suited, namely, situations involving an embarrassing abundance of data such as large microdata panels, high frequency financial data, and the like.

For small datasets, the computational burden associated with kernel methods is rarely an issue. However, for moderate to large datasets, the computations required to perform data-driven methods of bandwidth selection can easily get out of hand. Interest naturally lies with a general kernel framework, namely, one including unconditional and conditional density estimation, regression, and derivative estimation for both categorical and continuous data, and one that facilitates a range of kernel functions and bandwidth selection methods.

There exist a number of approaches for reducing the computational burden associated with kernel methods. At present we are not aware of methods that provide for a general kernel framework in real-time or close to real-time. We briefly discuss some of the approaches that currently exist and others that hold promise for a breakthrough that will allow real-time kernel estimation on a typical desktop computer or laptop.

8.1 Use Binning Methods

The use of "binning" has produced a number of very computationally attractive estimators. "Binning" refers to an approximate method whereby one first "pre-bins" data on an equally spaced mesh and then applies a suitably modified estimator to the binned data. For instance, binning methods have been proposed by Scott (1985) in which averaged shifted histograms (ASH) are used for smooth nonparametric density estimation, while Scott and Sheather (1985) investigate the accuracy of binning methods for kernel density estimation.

8.2 Use Transforms

Silverman (1986, pp. 61–66) outlines the use of fast Fourier transforms (FFT) for the efficient computation of (univariate) density estimates. This approach restricts estimation to a grid of points (e.g., 512) to further improve computational speed. Elgammal et al. (2003) discuss the use of the fast Gauss transform (FGT) for the efficient computation of kernel density estimates for which the kernel function is Gaussian.

8.3 Exploit Parallelism

Racine (2002) exploits both the parallel nature of most nonparametric methods and the availability of multiple processor computing environments to achieve a substantial reduction in run-time.

8.4 Use Multipole and Tree-Based Methods

Two recent developments in the areas of fast multipole methods and ball-trees hold promise for allowing real-time computation of kernel methods in general. In order to extend these recent developments to a general kernel framework, however, a significant amount of work remains to be done. Multipole and ball-tree methods have been developed only for unconditional density estimation, and for continuous datatypes only.

"Multipole" methods represent a group of approximate methods common in potential field settings where a set of n points interact according to a particular potential function, with the objective being to compute the field at arbitrary points (Greengard (1988), Greengard and Strain (1991)). To speed up calculations, these algorithms exploit the fact that all computations are required to be made only to a certain degree of accuracy.

Trees can be thought of as more powerful generalizations of a grid, being a set of linked grids built at different resolutions. This technique permits application of "divide-and-conquer" approaches that can integrate local information to obtain a global solution having precisely defined point-wise precision (Gray and Moore (2003)). While "kd-trees" share the property of grid representations having complexity that grows exponentially with the dimension q, this is not so for "ball-trees," which have been applied in settings involving literally thousands of dimensions.

Future Challenge The holy grail of applied kernel estimation is the development and implementation of a library that would serve as the basis for a package having, say, the capabilities of the np package (Hayfield and Racine (2007)) but which provides the computational benefits of the best of the methods listed above. This is, however, a rather formidable project but one which would be warmly received by the community.

9

Software

There exist a range of options for those who wish to undertake nonparametric modeling. No one package we are aware of will suit all users, and all are lacking in functionality. Many implement one and two dimensional density and regression methods but do not allow for higher dimensions, while others allow for higher dimensions but are otherwise narrow in scope. The list below is not in any way meant to be an endorsement nor is it meant in any way to be exhaustive. Rather, it is included merely to provide a starting point for the interested reader.

- EasyReg (econ.la.psu.edu/∼hbierens/EASYREG.HTM) is a Microsoft Windows program for regression which contains a module for nonparametric kernel regression with one or two explanatory variables.
- Limdep (www.limdep.com) has modules for kernel density estimation, among others.
- R (www.r-project.org) has a range of libraries offering a range of kernel methods including the base "stats" library, the "KernSmooth" library (orginal by Matt Wand. R port by

Ripley, B. (2007)), and the "np" library (Hayfield and Racine (2007)).

- SAS (www.sas.com) has modules for kernel regression, locally weighted smoothing, and kernel density estimation.
- Stata (www.stata.com) contains some modules for univariate and bivariate kernel density estimation and some modules for local constant kernel regression.
- TSP (www.tspintl.com) has some routines for univariate kernel density estimation and simple kernel regression.

Conclusions

Nonparametric kernel smoothing methods have experienced tremendous growth in recent years, and are being adopted by applied researchers across a range of disciplines. Nonparametric kernel approaches offer a set of potentially useful methods to those who must confront the vexing issue of parametric model misspecification. The appeal of nonparametric approaches stems mainly from their robustness to functional misspecification, in contrast to their parametric counterparts. Though the underlying theory for many of these methods can be daunting for some practitioners, we have attempted to demonstrate how a range of nonparametric methods can in fact be applied in a fairly straightforward manner. We have explicitly avoided any attempt at encyclopedic coverage of the field, rather we have tried to direct the interested reader to the textbooks mentioned in the introduction and, of course, the original journal articles themselves. By presenting a range of semiparametric and nonparametric models spanning a variety of application areas we hope that we have encouraged interested readers to attempt some of these methods in their particular problem domains.

We have tried to emphasize the fact that nonparametric kernel methods can be computationally burdensome, particularly when dealing with large datasets. This arises from the fact that data-driven methods of bandwidth selection must be deployed in applied settings, and these algorithms have run-times that tend to increase exponentially with the amount of data available. However, as noted in *Computational Considerations*, there exist approximate methods having the potential to dramatically reduce the amount of computation required, and those who are willing to contribute in this area are warmly encouraged to do so as their efforts would be particularly helpful to practitioners.

Acknowledgments

I would like to thank Bill Greene, Zac Rolnick, and an anonymous referee for their exemplary editorial assistance and guidance. I would also like to express my indebtedness to numerous co-authors and, in particular, to my friend and co-author Qi Li, with whom I have had the pleasure to work and interact in ways that have been profoundly gratifying.

Finally, I would also like to thank the Shared Hierarchical Academic Research Computing Network (SHARCNET:www.sharcnet.ca) for their ongoing support. I would also like to gratefully acknowledge financial support from the Natural Sciences and Engineering Research Council of Canada (NSERC:www.nserc.ca) and from the Social Sciences and Humanities Research Council of Canada (SSHRC:www.sshrc.ca)

Background Material

Though we have attempted to avoid unnecessary theoretical detail, kernel methods often require dealing with approximations and you will encounter statements such as "$\hat{f}(x) - f(x) = O_p(n^{-2/5})$" in various places in this review. Perhaps the best approach is to first consider the notation for nonstochastic sequences ($O(\cdot)$ and $o(\cdot)$) and then consider a similar notation for stochastic sequences ($O_p(\cdot)$ and $o_p(\cdot)$). In essence, we will need to determine the "order of magnitude" of a "sequence," and the "magnitude" is determined by considering the behavior of the sequence as the sample size n increases.

Order: Big $O(\cdot)$ and Small $o(\cdot)$

For a positive integer n, we write $a_n = O(1)$ if, as $n \to \infty$, a_n remains bounded, i.e., $|a_n| \leq C$ for some constant C and for all large values of n (a_n is a bounded sequence). We write $a_n = o(1)$ if $a_n \to 0$ as $n \to \infty$. Similarly, we write $a_n = O(b_n)$ if $a_n/b_n = O(1)$, or equivalently $a_n \leq Cb_n$ for some constant C and for all n sufficiently large. We write $a_n = o(b_n)$ if $(a_n/b_n) \to 0$ as $n \to \infty$.

By way of example, if $a_n = n/(n+1)$, then $a_n = O(1)$ since $a_n \leq 1$ for all n. Again, by way of example, if $a_n = 10/(n+1)$, then $a_n = o(1)$ because $a_n \to 0$ as $n \to \infty$.

Order in Probability: Big $O_p(\cdot)$ and Small $o_p(\cdot)$

A sequence of real (possibly vector-valued) random variables $\{\mathcal{X}_n\}_{n=1}^{\infty}$ is said to be bounded in probability if, for every $\epsilon > 0$, there exists a constant M and a positive integer N (usually $M = M_\epsilon$ and $N = N_\epsilon$), such that

$$P\left[||\mathcal{X}_n|| > M\right] \leq \epsilon \tag{9.1}$$

for all $n \geq N$. Note that for scalar a the notation $||a||$ is taken to mean the absolute value of a, while for vector a it is taken to mean $\sqrt{a'a}$, i.e., the square root of the sum of squares of the elements in a.

We say that $\mathcal{X}_n$ is bounded in probability if, for any arbitrarily small positive number ϵ, we can always find a positive constant M such that the probability of the absolute value (or "norm") of $\mathcal{X}_n$ being larger than M is less than ϵ. Obviously, if $\mathcal{X}_n = O(1)$ (bounded), then $\mathcal{X}_n = O_p(1)$; however the converse is not true. Letting $\{\mathcal{X}_n\}_{n=1}^{\infty}$ denote i.i.d. random draws from an $N(0,1)$ distribution, then $\mathcal{X}_n \neq O(1)$, however, $\mathcal{X}_n = O_p(1)$. We write $\mathcal{X}_n = O_p(1)$ to indicate that $\mathcal{X}_n$ is bounded in probability. We write $\mathcal{X}_n = o_p(1)$ if $\mathcal{X}_n \xrightarrow{p} 0$. Similarly, we write $\mathcal{X}_n = O_p(\mathcal{Y}_n)$ if $(\mathcal{X}_n/\mathcal{Y}_n) = O_p(1)$, and $\mathcal{X}_n = o_p(\mathcal{Y}_n)$ if $(\mathcal{X}_n/\mathcal{Y}_n) = o_p(1)$. Note that if $\mathcal{X}_n = o_p(1)$, then it must be true that $\mathcal{X}_n = O_p(1)$. However, when $\mathcal{X}_n = O_p(1)$, $\mathcal{X}_n$ may not be $o_p(1)$.

Notations and Acronyms

Here are some notation and associated definitions used in this review.

- $f(x)$ — The (unconditional) probability density function (PDF) of the random vector X.
- $g(y|x)$ — The PDF of the random variable Y conditional upon the realization of the random vector X.
- $F(x)$ — The (unconditional) cumulative distribution function (CDF) of the random vector X.
- $F(y|x)$ — The cumulative distribution function (CDF) of the random vector Y conditional upon the realization of the random vector X.
- $g(x) = E[Y|x] \equiv \int y f(y|x)\, dy$ — The expectation of the random variable Y conditional upon the realization of the random vector X.
- $q_\alpha(x)$ — The conditional αth quantile of a conditional CDF $F(y|x)$.
- h — A vector of smoothing parameters known as "bandwidths" for continuous variables.

- λ — A vector of smoothing parameters known as "bandwidths" for categorical variables.
- $K(\cdot)$ — A kernel ("weight") function.
- n — The number of observations.

Here are some terms used in this review along with a brief description.

- Cross-validation — A computationally demanding data driven method for bandwidth selection.
- Generalized Product Kernel — A kernel function obtained from the product of different kernels, each different kernel being appropriate for a particular datatype.
- Multistarting — A numerical technique used to minimize/maximize an objective function which is *less* likely to become trapped in local minima but is more computationally burdensome and therefore more time-intensive.

  Multistarting simply involves restarting the search algorithm a number of times from *different* random initial values for the parameters of interest and automatically saving those which are found to minimize/maximize the function of interest over all restarts of the search algorithm.

  Multistarting is highly recommended for serious data analysis.
- Partial Regression Plot — A 2D plot of the outcome y versus one covariate x_j when all other covariates are held constant at their respective medians/modes.
- Resampling — A method which can be used for obtaining the empirical distribution of an object of interest (e.g., bootstrapping).

References

Abramson, I. S. (1982), 'On bandwidth variation in kernel estimates — a square root law'. *The Annals of Statistics* **10**, 1217–1223.

Aitchison, J. and C. G. G. Aitken (1976), 'Multivariate binary discrimination by the kernel method'. *Biometrika* **63**(3), 413–420.

Azzalini, A. and A. W. Bowman (1997), *Applied Smoothing Techniques for Data Analysis: The Kernel Approach with S-plus Illustrations*. New York: Oxford University Press.

Bickel, P. J., C. A. J. Klaassen, Y. Ritov, and J. A. Wellner (1993), *Efficient and Adaptive Estimation for Semiparametric Models*. Baltimore: Johns Hopkins University Press.

Bowman, A. W. (1984), 'An alternative method of cross-validation for the smoothing of density estimates'. *Biometrika* **71**, 353–360.

Bowman, A. W., P. Hall, and T. Prvan (1998), 'Bandwidth selection for the smoothing of distribution functions'. *Biometrika* **85**, 799–808.

Breiman, L., W. Meisel, and E. Purcell (1977), 'Variable kernel estimates of multivariate densities'. *Technometrics* **19**, 135–144.

Brian, R. S. (2007), *KernSmooth: Functions for Kernel Smoothing for Wand and Jones (1995)*. R package version 2.22-21. URL: http://www.maths.unsw.edu.au/wand (original by Matt Wand. R port).

Cameron, A. C. and P. K. Trivedi (1998), *Regression Analysis of Count Data*. New York: Cambridge University Press.

Cantoni, E. and E. Ronchetti (2001), 'Resistant selection of the smoothing parameter for smoothing splines'. *Statistics and Computing* **11**, 141–146.

Cheng, M.-Y., P. Hall, and D. Titterington (1997), 'On the shrinkage of local linear curve estimators'. *Statistics and Computing* **7**, 11–17.

Čížek, P. and W. Härdle (2006), 'Robust estimation of dimension reduction space'. *Computational Statistics and Data Analysis* **51**, 545–555.

Cleveland, W. S. (1979), 'Robust locally weighted regression and smoothing scatterplots'. *Journal of the American Statistical Association* **74**, 829–836.

Croissant, Y. (2006), *Ecdat: Data Sets for Econometrics*. R package version 0.1-5. URL: http://www.r-project.org.

Croissant, Y. and G. Millo (2007), *Plm: Linear Models for Panel Data*. R package version 0.2-2. URL: http://www.r-project.org.

Delgado, M. A. and W. G. Manteiga (2001), 'Significance testing in nonparametric regression based on the bootstrap'. *Annals of Statistics* **29**, 1469–1507.

Devroye, L. and L. Györfi (1985), *Nonparametric Density Estimation: The L^1 View*. New York: Wiley.

Donald, S. G. (1997), 'Inference concerning the number of factors in a multivariate nonparametric relationship'. *Econometrica* **65**, 103–131.

Draper, N. R. (1987), *Empirical Model-Building and Response Surfaces*. Wiley.

Efromovich, S. (1999), *Nonparametric Curve Estimation: Methods, Theory and Applications*. New York: Springer Verlag.

Efron, B. (1982), *The Jackknife, the Bootstrap, and Other Resampling Plans*. Society for Industrial and Applied Mathematics, Philadelphia, Pennsylvania 19103.

Elgammal, A., R. Duraiswami, and L. Davis (2003), 'The fast gauss transform for efficient kernel density evaluation with applications in computer vision'. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Epanechnikov, V. A. (1969), 'Nonparametric estimation of a multidimensional probability density'. *Theory of Applied Probability* **14**, 153–158.

Eubank, R. L. (1999), *Nonparametric Regression and Spline Smoothing*. Second edition. New York: Marcel Dekker.

Fan, J. (1992), 'Design-adaptive nonparametric regression'. *Journal of the American Statistical Association* **87**, 998–1004.

Fan, J. and I. Gijbels (1996), *Local Polynomial Modelling and Its Applications*. London: Chapman and Hall.

Fan, J. and J. Jiang (2000), 'Variable bandwidth and one-step local m-estimator'. *Science in China* (*Series A*) **43**, 65–81.

Fan, J. and Q. W. Yao (2005), *Nonlinear Time Series: Nonparametric and Parametric Methods*. Springer.

Faraway, J. J. and M. Jhun (1990), 'Bootstrap choice of bandwidth for density estimation'. *Journal of the American Statistical Association* **85**, 1119–1122.

Fix, E. and J. L. Hodges (1951), 'Discriminatory analysis. nonparametric estimation: Consistency properties'. Technical Report 4, Randolph Field, Texas, USAF School of Aviation Medicine. Project No. 21–49–004.

Fox, J. (2002), *An R and S-PLUS Companion to Applied Regression*. Thousand Oaks, CA: Sage.

Geary, R. C. (1947), 'Testing for normality'. *Biometrika* **34**, 209–242.

Geisser, S. (1975), 'A predictive sample reuse method with application'. *Journal of the American Statistical Association* **70**, 320–328.

Gray, A. and A. W. Moore (2003), 'Very fast multivariate kernel density estimation via computational geometry'. Joint Statistical Meeting.

Greene, W. H. (2003), *Econometric Analysis*. Fifth edition. Upper Saddle River, NJ: Prentice Hall.

Greengard, L. (1988), *The Rapid Evaluation of Potential Fields in Particle Systems*. MIT Press.

Greengard, L. and J. Strain (1991), 'The fast gauss transform'. *Society for Industrial and Applied Mathematics: Journal of Science and Computation* **12**(1), 79–94.

Härdle, W. (1990), *Applied Nonparametric Regression*. New York: Cambridge University Press.

Härdle, W., M. Müller, S. Sperlich, and A. Werwatz (2004), *Nonparametric and Semiparametric Models*. Berlin: Springer Series in Statistics.

Härdle, W. and E. Mammen (1993), 'Comparing nonparametric versus parametric regression fits'. *Annals of Statistics* **21**, 1926–1947.

Hall, P., Q. Li, and J. S. Racine (2007), 'Nonparametric estimation of regression functions in the presence of irrelevant regressors'. *The Review of Economics and Statistics* **89**(4), 784–789.

Hall, P., J. S. Racine, and Q. Li (2004), 'Cross-validation and the estimation of conditional probability densities'. *Journal of the American Statistical Association* **99**(468), 1015–1026.

Hart, J. D. (1997), *Nonparametric Smoothing and Lack-of-Fit Tests*. New York: Springer Verlag.

Hastie, T. and R. Tibshirani (1993), 'Varying-coefficient models'. *Journal of the Royal Statistical Society Series B* **55**, 757–796.

Hayfield, T. and J. S. Racine (2007), Np: *Nonparametric Kernel Smoothing Methods for Mixed Datatypes*. R package version 0.13-1.

Henderson, D., R. J. Carroll, and Q. Li (2006), 'Nonparametric estimation and testing of fixed effects panel data models'. Unpublished manuscript, Texas A & M University.

Hodges, J. L. and E. L. Lehmann (1956), 'The efficiency of some nonparametric competitors of the t-test'. *Annals of Mathematical Statistics* **27**, 324–335.

Hoerl, A. E. and R. W. Kennard (1970), 'Ridge regression: Biased estimation for nonorthogonal problems'. *Technometrics* **12**, 55–67.

Horowitz, J. L. (1998), *Semiparametric Methods in Econometrics*. New York: Springer-Verlag.

Horowitz, J. L. and W. Härdle (1994), 'Testing a parametric model against a semiparametric alternative'. *Econometric Theory* **10**, 821–848.

Horowitz, J. L. and V. G. Spokoiny (2001), 'An adaptive, rate-optimal test of a parametric mean-regression model against a nonparametric alternative'. *Econometrica* **69**(3), 599–631.

Hristache, M., A. Juditsky, and V. Spokoiny (2001), 'Direct estimation of the index coefficient in a single-index model'. *Annals of Statistics* **29**, 595–623.

Hsiao, C., Q. Li, and J. S. Racine (2007), 'A consistent model specification test with mixed categorical and continuous data'. *Journal of Econometrics* **140**, 802–826.

Huber, P. J. (1964), 'Robust estimation of a location parameter'. *The Annals of Statistics* **35**, 73–101.

Hurvich, C. M., J. S. Simonoff, and C. L. Tsai (1998), 'Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion'. *Journal of the Royal Statistical Society Series B* **60**, 271–293.

Ichimura, H. (1993), 'Semiparametric least squares (SLS) and weighted SLS estimation of single-index models'. *Journal of Econometrics* **58**, 71–120.

Johnston, J. and J. DiNardo (1997), *Econometric Methods*. Fourth edition. McGraw-Hill.

Klein, R. W. and R. H. Spady (1993), 'An efficient semiparametric estimator for binary response models'. *Econometrica* **61**, 387–421.

Lavergne, P. and Q. Vuong (1996), 'Nonparametric selection of regressors: The nonnested case'. *Econometrica* **64**, 207–219.

Leung, D. (2005), 'Cross-validation in nonparametric regression with outliers'. *The Annals of Statistics* **33**, 2291–2310.

Li, Q. and J. S. Racine (2003), 'Nonparametric estimation of distributions with categorical and continuous data'. *Journal of Multivariate Analysis* **86**, 266–292.

Li, Q. and J. S. Racine (2004), 'Cross-validated local linear nonparametric regression'. *Statistica Sinica* **14**(2), 485–512.

Li, Q. and J. S. Racine (2007a), *Nonparametric Econometrics: Theory and Practice*. Princeton University Press.

Li, Q. and J. S. Racine (2007b), 'Smooth varying-coefficient nonparametric models for qualitative and quantitative data'. Unpublished manuscript, Department of Economics, Texas A&M University.

Li, Q. and J. S. Racine (forthcoming), 'Nonparametric estimation of conditional CDF and quantile functions with mixed categorical and continuous data'. *Journal of Business and Economic Statistics*.

Loader, C. R. (1999), 'Bandwidth selection: Classical or plug-in?'. *Annals of Statistics* **27**(2), 415–438.

Manski, C. F. (1988), 'Identification of binary response models'. *Journal of the American Statistical Association* **83**(403), 729–738.

Maronna, A., R. D. Martin, and V. J. Yohai (2006), *Robust Statistics: Theory and Methods*. Wiley.

Nadaraya, E. A. (1965), 'On nonparametric estimates of density functions and regression curves'. *Theory of Applied Probability* **10**, 186–190.

Pagan, A. and A. Ullah (1999), *Nonparametric Econometrics*. New York: Cambridge University Press.

Parzen, E. (1962), 'On estimation of a probability density function and mode'. *The Annals of Mathematical Statistics* **33**, 1065–1076.

Prakasa Rao, B. L. S. (1983), *Nonparametric Functional Estimation*. Orlando, FL: Academic Press.

R Development Core Team (2007), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing. Vienna, Austria. (ISBN 3-900051-07-0. URL: http://www.R-project.org).

Racine, J. S. (1997), 'Consistent significance testing for nonparametric regression'. *Journal of Business and Economic Statistics* **15**(3), 369–379.

Racine, J. S. (2002), 'Parallel distributed kernel estimation'. *Computational Statistics and Data Analysis* **40**, 293–302.

Racine, J. S., J. D. Hart, and Q. Li (2006), 'Testing the significance of categorical predictor variables in nonparametric regression models'. *Econometric Reviews* **25**, 523–544.

Racine, J. S. and Q. Li (2004), 'Nonparametric estimation of regression functions with both categorical and continuous data'. *Journal of Econometrics* **119**(1), 99–130.

Racine, J. S. and L. Liu (2007), 'A partially linear kernel estimator for categorical data'. Unpublished manuscript, McMaster University.

Robinson, P. M. (1988), 'Root-n consistent semiparametric regression'. *Econometrica* **56**, 931–954.

Rosenblatt, M. (1956), 'Remarks on some nonparametric estimates of a density function'. *The Annals of Mathematical Statistics* **27**, 832–837.

Rudemo, M. (1982), 'Empirical choice of histograms and kernel density estimators'. *Scandinavian Journal of Statistics* **9**, 65–78.

Ruppert, D., R. J. Carroll, and M. P. Wand (2003), *Semiparametric Regression Modeling*. New York: Cambridge University Press.

Ruppert, D., S. J. Sheather, and M. P. Wand (1995), 'An effective bandwidth selector for local least squares regression'. *Journal of the American Statistical Association* **90**, 1257–1270.

Scott, D. W. (1985), 'Averaged shifted histograms: Effective nonparametric density estimators in several dimensions'. *Annals of Statistics* **13**, 1024–1040.

Scott, D. W. (1992), *Multivariate Density Estimation: Theory, Practice, and Visualization*. New York: Wiley.

Scott, D. W. and S. J. Sheather (1985), 'Kernel density estimation with binned data'. *Communication in Statistics: Theory and Methods* **14**, 1353–1359.

Seifert, B. and T. Gasser (2000), 'Data adaptive ridging in local polynomial regression'. *Journal of Computational and Graphical Statistics* **9**, 338–360.

Sheather, S. J. and M. C. Jones (1991), 'A reliable data-based bandwidth selection method for kernel density estimation'. *Journal of the Royal Statistal Society, Series B* **53**, 683–690.

Silverman, B. W. (1986), *Density Estimation for Statistics and Data Analysis*. New York: Chapman and Hall.

Simonoff, J. S. (1996), *Smoothing Methods in Statistics*. New York: Springer Series in Statistics.

Stone, C. J. (1974), 'Cross-validatory choice and assessment of statistical predictions (with discussion)'. *Journal of the Royal Statistical Society* **36**, 111–147.

Stone, C. J. (1977), 'Consistent nonparametric regression'. *Annals of Statistics* **5**, 595–645.

Venables, W. N. and B. D. Ripley (2002), *Modern Applied Statistics with S*. Fourth edition. New York: Springer.

Wand, M. P. and M. C. Jones (1995), *Kernel Smoothing*. London: Chapman and Hall.

Wang, F. T. and D. W. Scott (1994), 'The l_1 method for robust nonparametric regression'. *Journal of the American Statistical Association* **89**, 65–76.

Wang, M. C. and J. van Ryzin (1981), 'A class of smooth estimators for discrete distributions'. *Biometrika* **68**, 301–309.

Wang, N. (2003), 'Marginal nonparametric kernel regression accounting for within-subject correlation'. *Biometrika* **90**, 43–52.

Wang, N., R. J. Carroll, and X. Lin (2005), 'Efficient semiparametric marginal estimation for longitudinal/clustered data'. *Journal of the American Statistical Association* **100**, 147–157.

Watson, G. S. (1964), 'Smooth regression analysis'. *Sankhya* **26**(15), 359–372.

Wooldridge, J. M. (2002), *Econometric Analysis of Cross Section and Panel Data*. Cambridge: MIT Press.

Wooldridge, J. M. (2003), *Introductory Econometrics*. Thompson South-Western.

Yatchew, A. J. (2003), *Semiparametric Regression for the Applied Econometrician*. New York: Cambridge University Press.