

## American Economic Association

---

Nonparametric Regression Techniques in Economics

Author(s): Adonis Yatchew

Source: *Journal of Economic Literature*, Vol. 36, No. 2 (Jun., 1998), pp. 669-721

Published by: [American Economic Association](#)

Stable URL: <http://www.jstor.org/stable/2565120>

Accessed: 19/08/2011 15:48

---

Your use of the JSTOR archive indicates your acceptance of the Terms & Conditions of Use, available at <http://www.jstor.org/page/info/about/policies/terms.jsp>

JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of content in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new forms of scholarship. For more information about JSTOR, please contact support@jstor.org.



American Economic Association is collaborating with JSTOR to digitize, preserve and extend access to *Journal of Economic Literature*.

<http://www.jstor.org>

# Nonparametric Regression Techniques in Economics

ADONIS YATCHEW<sup>1</sup>

## 1. Introduction

### 1.1 *Setting the Stage*

1.1.1 *Benefits of Nonparametric Estimation.* If economics is the dismal science, then econometrics is its ill-fated offspring. The limited number of strong implications derived from economic theory, the all but complete inability to perform controlled experiments, the paucity of durable empirical results, the errant predictions, the bottomless reservoir of difficult to measure but potentially relevant variables—these do not set a pleasant working environment for the empirical analyst in economics.

When one contrasts the overall quality of economic data and empirical inferences with the plethora of sophisticated econometric theory already available, it would appear difficult to justify learning (or teaching) techniques where the regression function depends

on an infinite number of unknown parameters. (Nonparametric regression typically assumes little else about the shape of the regression function beyond some degree of smoothness.) Yet, we will argue that it is because such tools lead to more durable inferences that they will become an enduring—even indispensable—element of every economist's tool kit, in much the same way that linear and nonlinear regression are today. After all, reliable empirical results are essential to the formulation of sound policy.

From the point of view of a pure data analyst, the added value of nonparametric techniques consists in their ability to deliver estimators and inference procedures that are less dependent on functional form assumptions. They are also useful for exploratory data analysis and as a supplement to parametric procedures. If one has reservations about a particular parametric form, specification tests against nonparametric alternatives can provide reassurance.

From the point of view of the economist/econometrician, such techniques possess additional appeal in that most implications of economic theory are nonparametric. Typically, theoretical arguments exclude or include variables, they imply monotonicity, concavity, or homogeneity of various sorts, or they embody more complex structure such as

<sup>1</sup>Department of Economics, University of Toronto. The author is grateful to Andrey Feuerverger, Zvi Griliches, James MacKinnon, Angelo Melino, John Pencavel and particularly to Frank Wolak for their patient reading and insightful comments. The statistics and econometrics literature on nonparametric regression is massive and inevitably many interesting and important results have been given little or no exposure in this introduction to the subject. The author regrets this state of affairs and hopes that his lacunae will be corrected in complementary and more artful reviews by other writers. The author may be contacted at [yatchew@chass.utoronto.ca](mailto:yatchew@chass.utoronto.ca).

the implications of the maximization hypothesis. They almost *never* imply a specific functional form (the pure quantity theory of money equation being one exception). This paper will therefore focus some considerable attention on constrained nonparametric regression estimation and testing.

In some cases, the researcher may feel comfortable with a particular parametric form for one portion of the regression function, but less confident about the shape of another portion. Such varying prior beliefs call for combining parametric and nonparametric techniques to yield semiparametric regression models (these have been studied extensively). Inclusion of the nonparametric component may avoid inconsistent estimation which could result from incorrect parameterization.

1.1.2 *An Appetizer for the Reluctant Palate.* For those who have never considered using a nonparametric regression technique, we suggest the following elementary procedure, much of which can be implemented in *any* standard econometric package. Suppose you are given data  $(y_1, z_1, x_1) \dots (y_T, z_T, x_T)$  on the model  $y = z\beta + f(x) + \varepsilon$  where for simplicity all variables are assumed to be scalars. The  $\varepsilon_t$  are i.i.d. with mean 0 and variance  $\sigma_\varepsilon^2$  given  $(z, x)$ . The  $x$ 's are drawn from a distribution with support, say, the unit interval. Most important, the data are rearranged so that  $x_1 \leq \dots \leq x_T$ . All that is known about  $f$  is that its first derivative is bounded by a constant, say  $L$ . Suppose that we first difference to obtain

$$y_t - y_{t-1} = (z_t - z_{t-1})\beta + (f(x_t) - f(x_{t-1})) + \varepsilon_t - \varepsilon_{t-1} \quad (1.1)$$

As sample size increases—packing the unit interval with  $x$ 's—the typical difference  $x_t - x_{t-1}$  shrinks at a rate of about  $1/T$  so that  $f(x_{t-1})$  tends to cancel  $f(x_t)$ . (This is

because the bound on the first derivative implies that  $|f(x_t) - f(x_{t-1})| \leq L|x_t - x_{t-1}|$ .) As long as  $z$  is not perfectly correlated with  $x$ , the ordinary least squares estimator of  $\beta$  using the differenced data, that is

$$\hat{\beta}_{diff} = \frac{\sum (y_t - y_{t-1})(z_t - z_{t-1})}{\sum (z_t - z_{t-1})^2} \quad (1.2)$$

has the approximate sampling distribution

$$\hat{\beta}_{diff} \sim N\left(\beta, \frac{1}{T} \frac{1.5\sigma_\varepsilon^2}{\sigma_u^2}\right) \quad 1.3$$

where  $\sigma_u^2$  is the conditional variance of  $z$  given  $x$ .<sup>2</sup>

We have applied this estimator to data on the costs of distributing electricity (Figure 1). Factors which can influence distribution costs include customer density (greater distance between customers increases costs), the remaining life of physical plant (newer assets require less maintenance), and wage rates. These are among the “ $z$ ” variables which appear parametrically in the model. However, the nature of scale economies is unclear—the effect on costs could be constant, declining to an asymptote or even  $U$ -shaped. Indeed, important regulatory decisions (such as merger policy) may involve determining the minimum efficient scale.

In the left panel, we estimate the cost function parametrically, incorporating a quadratic term for the scale variable “ $x$ ” which we measure using the number of customers of each utility. In the right panel, we report the coefficient estimates of the  $z$  variables after applying the differencing estimator (1.2), suitably generalized to allow for vector  $z$ . There is moderate change in coefficients.

<sup>2</sup> Throughout the paper, the symbol  $\sim$  will denote that a random variable has the indicated approximate distribution and the symbol  $\cong$  will indicate approximate equality.

Figure 1. Returns to Scale in Electricity Distribution

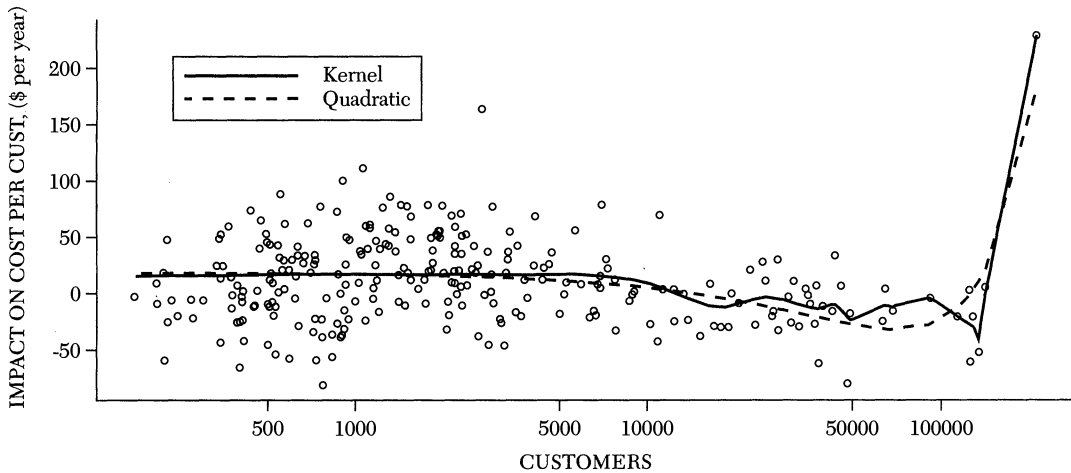
*Model:* Restructuring, vertical unbundling, and deregulation of electricity industries have led to reexamination of scale economies in electricity distribution. The efficiency of small (usually municipal) distributors that exist in the U.S., Norway, New Zealand, Germany and Canada is at issue. The objective is to assess scale economies. The model is given by  $y = z\beta + f(x) + \epsilon$  where  $y$  is variable costs of distribution per customer (COST PER CUST). The vector  $z$  includes: customer density (rural distribution is more expensive) as measured by average distance between customers (DIST); load per customer (LOAD); local generation per customer (GEN); remaining life of assets (LIFE) – older plants require more maintenance; assets per customer (ASSETS); and a proxy for labour wage rates (WAGE). The scale of operation,  $x$ , is measured by the number of customers served (CUST). The scale effect may be nonlinear, (e.g., decreasing to an asymptote or U-shaped), hence the inclusion of a nonparametric component  $f(x)$ .

*Data:* 265 municipal distributors in Ontario, Canada which vary in size from tiny ones serving 200 customers to the largest serving 218,000 (a 'customer' can be a household, or a commercial, industrial or governmental entity).

ESTIMATED MODELS

PARAMETRIC: Quadratic scale effect $y = \alpha + z\beta + \gamma_1x + \gamma_2x^2 + \epsilon$			SEMIPARAMETRIC: Smooth scale effect $y = z\beta + f(x) + \epsilon$		
OLS			Differencing Estimator <sup>1</sup>		
	Coefficient	Standard Error		Coefficient	Standard Error
$\alpha$	15.987	37.002	$\Delta$ DIST	2.568	1.560
DIST	3.719	1.248	$\Delta$ LOAD	.437	.912
LOAD	1.920	.661	$\Delta$ GEN	.0005	.032
GEN	-.051	.023	$\Delta$ LIFE	-4.470	1.070
LIFE	-5.663	.798	$\Delta$ ASSETS	.030	.0072
ASSETS	.037	.005	$\Delta$ WAGE	.003	.00086
WAGE	.003	.0007			
$\gamma_1$	-.00154	.00024			
$\gamma_2$	.1x10 <sup>-7</sup>	.14x10 <sup>-8</sup>			
$R^2$	.45				

ESTIMATED IMPACT OF SCALE ON DOLLAR COSTS PER CUSTOMER PER YEAR



<sup>1</sup> Data reordered so that  $x_1 \leq \dots \leq x_T$ , then all variables differenced  $\Delta w_t = w_t - w_{t-1}$ , followed by ordinary least squares. Standard errors multiplied by  $\sqrt{1.5}$ , as per equation (1.3). Dependent variable is  $\Delta$  COST PER CUST.

Though the “differencing estimator” of  $\beta$  does not require estimation of  $f$ , one can visually assess whether there is a potentially interesting relationship embodied in  $f$  by producing a scatter-plot of  $y_t - z_t \hat{\beta}_{diff}$  against  $x_t$  as we have done in Figure 1. In addition, we plot two curves—the quadratic estimate  $\hat{\gamma}_1 x + \hat{\gamma}_2 x^2$ , (it does not *look* quadratic because  $x$  is scaled logarithmically) and a nonparametric estimate using the “kernel” technique, which we will later discuss in detail. The two estimates are strikingly similar, suggesting that the quadratic specification is adequate (we will later provide formal tests of this proposition). Note further that the standard errors of the differencing estimator are larger than the pure parametric ones, as one would expect from (1.3).

For the most efficient nonparametric estimator, the 1.5 factor in equation (1.3) is replaced by 1, thus the relative efficiency of this differencing estimator is 66.7 percent ( $=1/1.5$ ). Later we will outline how asymptotic efficiency can be achieved by using higher order differences.

Our analysis of this partial linear model has been divided into two parts: first we analyzed the parametric portion of the model, all of which can be done in a standard econometric package; then we estimated the nonparametric portion of the model (estimators for which are widely available in statistical packages such as S-Plus). This modular approach will be a theme of the paper, since it permits one to use existing software and to adapt a variety of parametric or purely nonparametric procedures to this setting.

We make one final observation on the differencing idea which we exploit in this paper because of its simplicity. Nonparametric procedures estimate the value of the regression function at a given point by using neighboring obser-

vations. For the approach described here, the implicit estimate of  $f(x_t)$  is  $f(x_{t-1})$ .

1.1.3 *Objectives of the Paper.* Why have nonparametric regression methods not yet infiltrated the applied literature to the degree one might expect, particularly since other nonparametric techniques are used without hesitation? After all, anyone using the central limit theorem to do inference on a mean from an unknown distribution has engaged in semiparametric inference. The same is true if the mean is a linear function of some explanatory variables, but the distribution of the residuals remains unknown. Even the histogram, which we teach in introductory courses, is a nonparametric estimator of the underlying density.

True, nonparametric regression has not thus far uncovered any particularly startling empirical phenomena that were hitherto inaccessible to parametric practitioners.<sup>3</sup> But this does not explain the absence of greater use in exploratory data analysis, as a confirmatory tool, or as a supplement to the standard parametric fare. Several factors have influenced this relatively slow penetration rate at the same time that there has been an explosion of theoretical results in this area.

First, nonparametric regression techniques are theoretically more complex than the usual tool kit of linear and nonlinear parametric modelling methods.

Second, nonparametric regression techniques are computationally intensive and they require large (in some cases astronomically large) data sets, since relationships are “discovered” by examining nearby observations. (“Nearby” better not be too far away.)

<sup>3</sup> Some may be surprised by the apparent absence of scale economies in electricity distribution (Figure 1). After all, the so-called “wires business” is a natural monopoly.

Third, a *unified* framework for constrained estimation and testing of economic models using nonparametric regression is still in the incipient stage. As a consequence, software which will handle such procedures in an automated fashion does not yet exist.

A central objective of this paper is to demonstrate that these barriers are indeed substantially lower than might first appear. We comment on each in turn.

First, we deal with the issue of theoretical sophistication. Nonparametric regression typically involves either local averaging or some form of least squares estimation. Both ideas are familiar from parametric modelling, and indeed a significant portion of the theory of nonparametric regression involves straightforward extension of results and techniques familiar to parametric practitioners. Unfortunately, nonparametric methods also have critical elements that are not present in the pure parametric setting. Most important among these are the “curse of dimensionality” and the need to select the value of a “smoothing parameter.” Taking as a premise that a technique is unlikely to be used (or worse, unlikely to be used correctly) without a rudimentary understanding and intuitive grasp of the basic theory, we set out, wherever possible, simple arguments supporting the theoretical propositions.

Second is the issue of computation. The precipitous drops in computing costs, data storage, and even data collection (the latter through the proliferation of automated data collection technologies) are effectively eliminating this as a barrier to the use of nonparametric regression techniques. We believe that the forthcoming industry standard (be it a local averaging estimator, a least squares estimator, or a hybrid) will be coupled with computer-intensive inference techniques, such as the bootstrap.

However, for the interested applied economist, there is no need to wait for future software developments. Many of the procedures outlined in this paper can be implemented using off-the-shelf software, in particular S-Plus (see for example, William Venables and Brian Ripley 1994), and XploRe (see Wolfgang Härdle, Sigbert Klinke, and Berwin Turlach 1995).

Third is the issue of constrained estimation and hypothesis testing. In a parametric setting, constraints can be imposed on parameters relatively easily, and many hypotheses can then be tested (for example, by comparing restricted and unrestricted sums of squared residuals).

In a nonparametric setting, imposing constraints on the estimator is often more difficult. However, if a restricted estimator is obtained, one can examine the estimated residuals to see whether they constitute pure error or whether they are related to the explanatory variables. If it is the latter, then this suggests invalidity of the constraints. (Formal testing can proceed by performing a regression of the estimated residuals on all the explanatory variables.)

In summary, our overarching objective is to increase the accessibility of nonparametric techniques to economists. To pursue this main goal, our three subsidiary objectives are to provide implementation details for a few relatively simple nonparametric regression estimation and inference techniques, to summarize central theoretical results which we hope will make the techniques more intelligible, and to outline procedures for constrained estimation and hypothesis testing.

1.1.4 *Charting the Terrain Ahead.* The structure and content of the paper is thus driven by the objectives mentioned above. The remainder of the *Introduction*, entitled *Background and*

*Overview*, categorizes various kinds of regression models, from parametric at the simple extreme, through semiparametric models, to various kinds of nonparametric models which may incorporate additional structure (such as additive separability or monotonicity). The section then discusses the “curse of dimensionality” a phenomenon which has no counterpart in a parametric setting. Following this, two naive nonparametric regression estimators are introduced—the first based on local averaging, the second on least squares. The section closes with a glimpse of essential theoretical results. Upon completion of the *Background and Overview*, the reader should have an appreciation of the range of models under study, the two most basic estimation principles, the theoretical challenges, and a minimal set of theoretical results.

The second section of the paper details two nonparametric regression estimators—the kernel estimator which is based on the principle of local averaging; and nonparametric least squares, which is closely related to spline estimation. The section also summarizes their main statistical properties.

The third section focuses on *The Partial Linear Model* which we believe to be the most likely “entry level” model for economists. This section describes three distinct estimators. In addition to the partial linear model, the class of semiparametric specifications includes other important subclasses such as partial parametric and index models. Despite the prominence of index models in the econometric literature, we have not included them in this paper, essentially because of space limitations. A foothold into that literature may be established by reading Thomas Stoker (1991). See also James L. Powell (1994) and references therein.

The fourth section of the paper dis-

cusses constrained estimation and hypothesis testing.

In the fifth section of the paper, *Extensions and Technical Details*, we collect a variety of topics. The partial linear model is revisited yet again, this time to demonstrate that the parametric and nonparametric portions can be analyzed separately. Thus, a variety of purely nonparametric procedures can be grafted onto the nonparametric portion of the partial linear model. The role of various constraints implied by economic theory is discussed. The section also includes a note on available computer software, and directions for further reading.

Section 6 presents our conclusions.

We will from time to time make use of a number of mathematical ideas dealing with sequences of numbers and of random variables. We have summarized these in Appendix A. Throughout the paper, equations of particular significance or usefulness are indicated by equation numbers set in boldface type.

## 1.2 Background and Overview

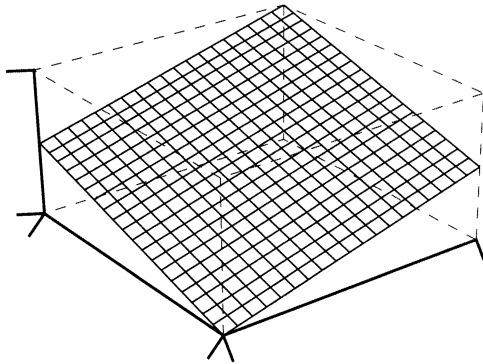
### 1.2.1 Categorization of Models.

Consider the model  $y = f(x) + \varepsilon$  where  $\varepsilon$  is i.i.d. with mean 0 and variance  $\sigma_\varepsilon^2$  given  $x$ . If  $f$  is known only to lie in a family of smooth functions  $\mathfrak{F}$  then the model is nonparametric. If  $f$  satisfies some additional properties (such as monotonicity, concavity, homogeneity or symmetry) and hence lies in  $\mathfrak{F} \subset \mathfrak{F}$ , we will say that the model is constrained nonparametric. If we can partition  $x$  into two subsets  $x_a$  and  $x_b$ , such that  $f$  is of the form  $f_a(x_a) + f_b(x_b)$ , then it is called additively separable.

Next we turn to semiparametric models. In this paper we will focus on the partial linear model already introduced. (See Figure 2 for categorization of various kinds of regression models.)

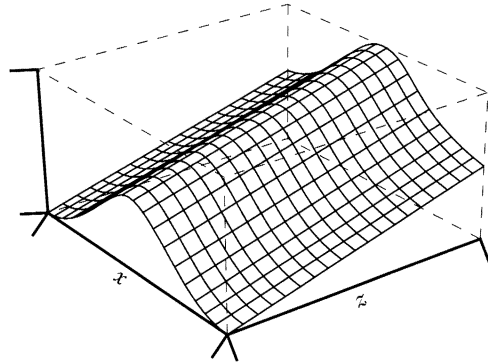
Figure 2. Categorization of Regression Functions

Parametric – Linear



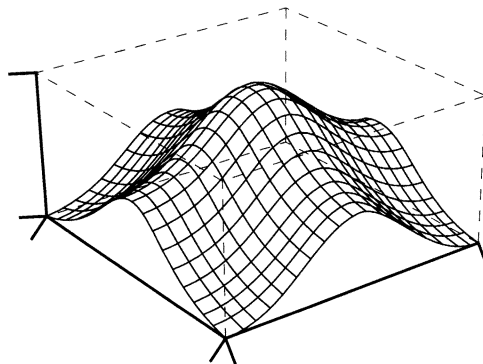
Linear  $y = x\beta + \epsilon$  (pictured above)  
 Nonlinear  $y = g(x;\beta) + \epsilon$ ,  $g$  known

Semiparametric – Partial Linear



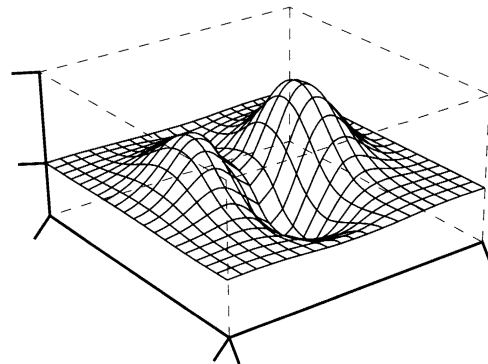
Partial Linear  $y = z\beta + f(x) + \epsilon$ ,  $f \in \mathfrak{F}$  (pictured above)  
 Partial Parametric  $y = g(z;\beta) + f(x) + \epsilon$ ,  $g$  known,  $f \in \mathfrak{F}$   
 Index Models  $y = f(z\beta) + \epsilon$ ,  $f \in \mathfrak{F}$

Additively Separable



Smooth  $y = f_a(x_a) + f_b(x_b) + \epsilon$ ,  $f_a, f_b \in \mathfrak{F}$

Nonparametric



Smooth  $y = f(x) + \epsilon$ ,  $f \in \mathfrak{F}$  (pictured above)  
 Smooth and Constrained  $y = f(x) + \epsilon$ ,  $f \in \bar{\mathfrak{F}}$

$\mathfrak{F}$  is a smooth family of functions.  $\bar{\mathfrak{F}}$  is a smooth family with additional constraints such as monotonicity, concavity, symmetry or other constraints.

1.2.2 *The Curse of Dimensionality and the Need for Large Data Sets.* In comparison to parametric estimation, nonparametric procedures can impose enormous data requirements. To gain an appreciation of the problem as well as remedies for it, we begin with a deterministic framework. Suppose the ob-

jective is to approximate a function  $f$ . If it is known to be linear in one variable, two observations are sufficient to determine the entire function; three are sufficient if  $f$  is linear in two variables. If  $f$  is of the form  $g(x;\beta)$  where  $g$  is known and  $\beta$  is an unknown  $k$ -dimensional vector, then  $k$  judiciously selected points



are usually sufficient to solve for  $\beta$ . No further observations on the function are necessary.

Let us turn to the pure nonparametric case. Suppose  $f$ , defined on the unit interval, is known only to have a first derivative, bounded by  $L$  (i.e.,  $\sup_{x \in [0,1]} |f'| \leq L$ ). If we sample  $f$  at  $T$  equidistant points and approximate  $f$  at any point by the closest point at which we have an evaluation, then our approximation error cannot exceed  $1/2L/T$ . Increasing the density of points reduces approximation error at a rate  $O(1/T)$ .

Now suppose  $f$  is a function on the unit square and that it has derivatives bounded in all directions by  $L$ . In order to approximate the function, we need to sample throughout its domain. If we distribute  $T$  points uniformly on the unit square, each will "occupy" an area  $1/T$  and the typical distance between points will be  $1/T^{1/2}$  so that the approximation error is now  $O(1/T^{1/2})$ . Repeating this argument for functions of  $k$  variables, the typical distance between points becomes  $1/T^{1/k}$  and the approximation error is  $O(1/T^{1/k})$ . In general, this method of approximation yields error proportional to the distance to the nearest observation.

For  $T=100$ , the approximation error is 10 times larger in two dimensions than in one and 40 times larger in five dimensions. Put another way, 100 observations in one dimension would yield the same accuracy as 10,000 observations would in two dimensions and 10 billion would yield in five dimensions. One begins to see the virtues of parametric modelling.

We will consider three types of restrictions which substantially reduce approximation error: a partial linear structure, additive separability, and smoothness assumptions.

Suppose a regression function defined on the unit square has the partial

linear form  $z\beta + f(x)$  (the function  $f$  is unknown except for a derivative bound). In this case, we need two evaluations along the  $z$  axis to completely determine  $\beta$  (see the "Semiparametric" surface in Figure 2). Furthermore,  $T$  equidistant evaluations along the  $x$  axis will ensure that  $f$  can be approximated with error  $O(1/T)$  so that the approximation error for the regression function as a whole is also  $O(1/T)$ , the same as if it were a nonparametric function of one variable.

Next, consider an additively separable function on the unit square:  $f(x_a, x_b) = f_a(x_a) + f_b(x_b)$ , where the functions  $f_a$  and  $f_b$  satisfy a derivative bound ( $f_b(0) = 0$  is imposed as an identification condition). If we take  $2T$  observations,  $T$  along each axis, then  $f_a$  and  $f_b$  can be approximated with error  $O(1/T)$ , so approximation error for  $f$  is also  $O(1/T)$ , once again the same as if  $f$  were a nonparametric function of one variable.

The following proposition should now be plausible—for partially linear or additively separable models, the approximation error depends on the maximum dimension of the pure nonparametric components of the model.

Smoothness can also reduce approximation error. Suppose  $f$  is twice differentiable on the unit interval with  $f'$  and  $f''$  bounded by  $L$  and we evaluate  $f$  at  $T$  equidistant values of  $x$ . Consider approximation of  $f$  at  $x_o \in [x_t, x_{t+1}]$ . Using a Taylor expansion we have

$$f(x_o) = f(x_t) + f'(x_t)(x_o - x_t) + \frac{1}{2}f''(x^*)(x_o - x_t)^2 \quad x^* \in [x_t, x_o]. \quad (1.4)$$

If we approximate  $f(x_o)$  using  $f(x_t) + f'(x_t)(x_o - x_t)$  the error is  $O(x_o - x_t)^2 = O(1/T^2)$ . Of course we do not observe  $f'(x_t)$ . However, the bound on the second derivative implies that  $f'(x_t) - [f(x_{t+1}) - f(x_t)]/[x_{t+1} - x_t]$  is  $O(1/T)$  so that

$$f(x_o) = \tag{1.5}$$

$$f(x_t) + \frac{[f(x_{t+1}) - f(x_t)]}{[x_{t+1} - x_t]} (x_o - x_t) + O\left(\frac{1}{T^2}\right).$$

This local linear approximation involves nothing more than joining the observed points with straight lines. If third order ( $k^{\text{th}}$  order) derivatives are bounded, then local quadratic ( $k-1$  order polynomial) approximations will reduce the error further.

In this section, we have used the elementary idea that if a function is smooth, its value at a given point can be approximated reasonably well by using evaluations of the function at neighboring points. This idea is fundamental to nonparametric estimation where of course  $f$  is combined with noise to yield the observed data. All of the results illustrated in this section have analogues in the nonparametric setting. Data requirements grow very rapidly as the dimension of the nonparametric component increases. The rate of convergence (that is, the rate at which we learn about the unknown regression function) can be improved using semiparametric structure (we illustrate the partial linear model but similar results hold for partial parametric or index models), additive separability, and smoothness assumptions. Finally, the curse of dimensionality underscores the paramount importance of procedures which validate models with faster rates of convergence. Among these are specification tests of a parametric null against a nonparametric alternative, and significance tests which may reduce the number of explanatory variables.

**1.2.3 Local Averaging Estimators vs Optimization Estimators.** Our first objective will be to estimate  $y = f(x) + \varepsilon$  given data  $(y_1, x_1) \dots (y_T, x_T)$ . For the moment we will assume that  $x$  is a scalar.

Local averaging estimators are exten-

sions of conventional estimators of location to a nonparametric regression setting. If one divides the scatterplot into vertical bands, then one can compute local means (or medians) as approximations to the regression function. A more appealing alternative is to have the “band” or “window” move along the  $x$  axis, computing a moving average along the way. The wider the band (for the moment we set aside the issue of bandwidth selection), the smoother the estimate, as may be seen in Figure 3 where solid lines depict local averaging estimates. (If one were in a vessel, the “sea” represented by the solid line in the bottom panel would be the most placid.) Suppose then we define the estimator to be

$$\hat{f}(x_o) = \frac{1}{T_o} \sum_{N(x_o)} y_t = \tag{1.6}$$

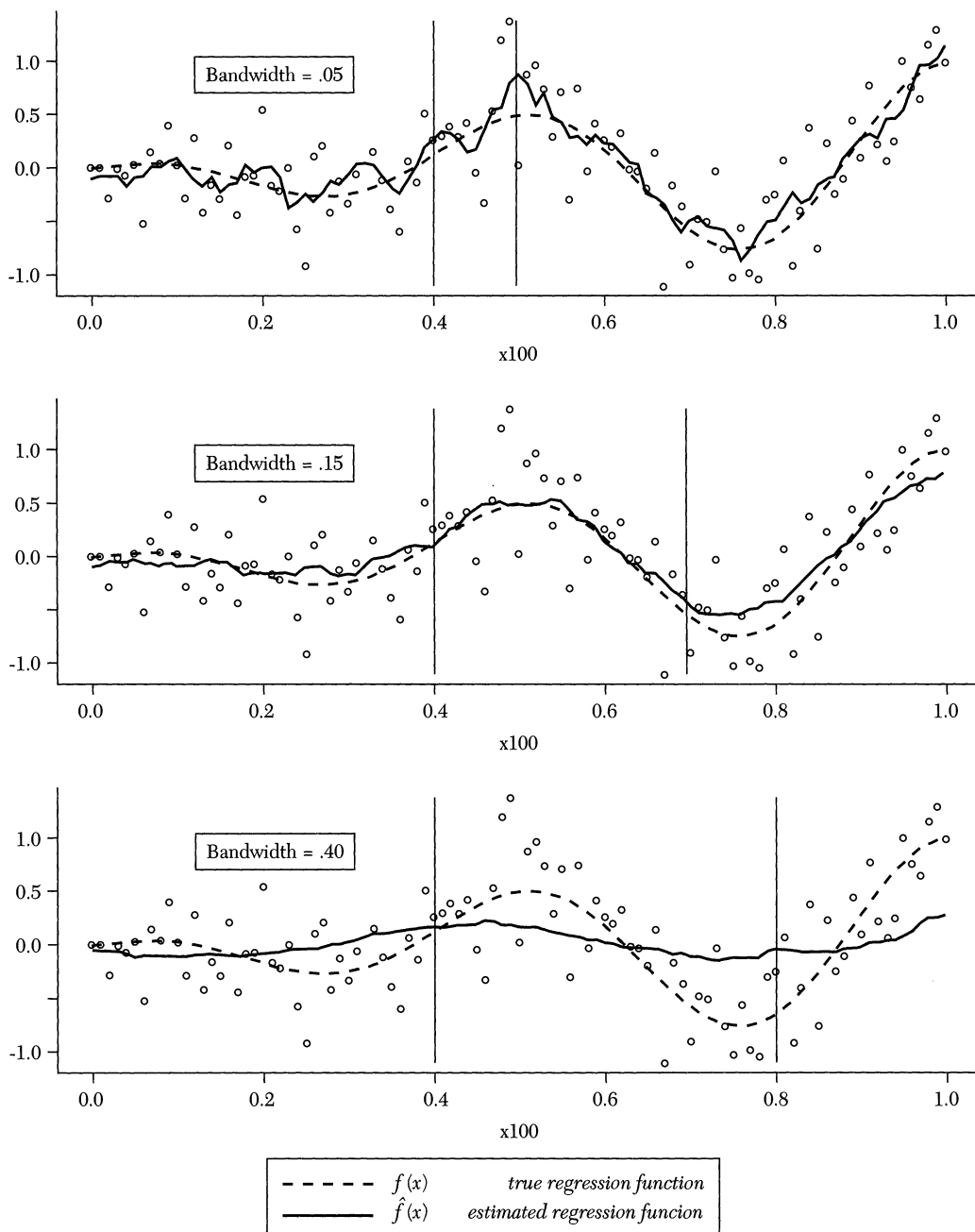
$$f(x_o) + \frac{1}{T_o} \sum (f(x_t) - f(x_o)) + \frac{1}{T_o} \sum \varepsilon_t,$$

where summations are taken over observations in the neighborhood  $N(x_o)$  around  $x_o$  determined by the bandwidth, and  $T_o$  is the number of elements in  $N(x_o)$ . Conditional on the  $x$ 's, the bias of the estimator consists of the second term and the variance is determined by the third term. The mean squared error (that is, the bias squared plus the variance) is given by

$$E[\hat{f}(x_o) - f(x_o)]^2 = \left(\frac{1}{T_o} \sum_{N(x_o)} f(x_t) - f(x_o)\right)^2 + \frac{\sigma_\varepsilon^2}{T_o}. \tag{1.7}$$

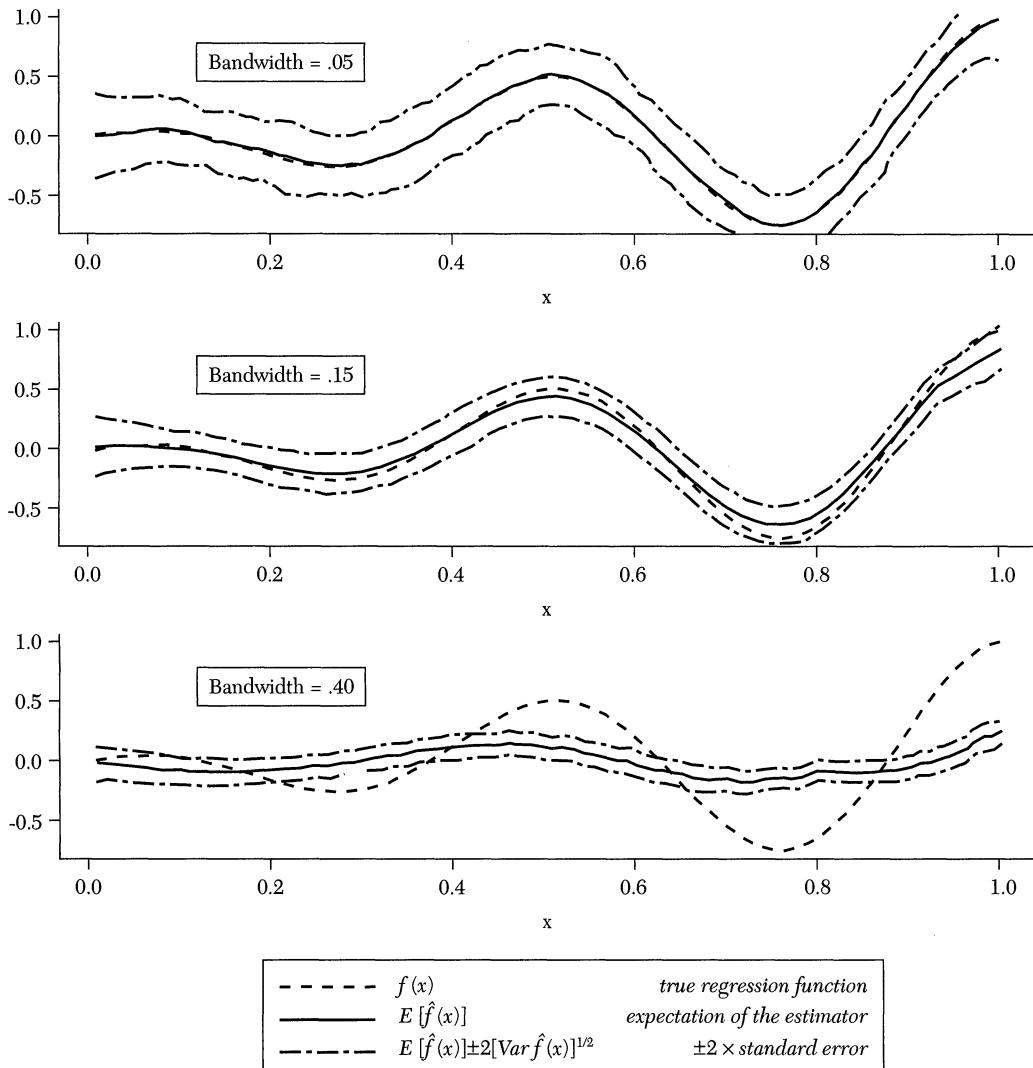
Mean squared error can be minimized by increasing the bandwidth of the neighborhood  $N(x_o)$  until the increase in bias squared is offset by the reduction in variance. (The latter declines since  $T_o$  increases as the bandwidth widens.) This trade-off between bias and variance is illustrated in Figure 4 (which continues

Figure 3 Naive Local Averaging



DATA GENERATING MECHANISM  $y_t = x_t \cos(4\pi x_t) + \varepsilon_t$   $\varepsilon_t \sim N(0, .09)$   $x_t \in [0, 1]$ ,  $T=100$ . Observations are averaged over bands of the indicated width.

Figure 4 Bias-Variance Tradeoff



DATA GENERATING MECHANISM  $y_t = x_t \cos(4\pi x_t) + \varepsilon_t$   $\varepsilon_t \sim N(0, .09)$   $x_t \in [0, 1]$ ,  $T=100$

the example of Figure 3). In the first panel, local averaging is taking place using just 5 percent of the data at each point (since the bandwidth is .05 and 100  $x$ 's are uniformly spaced on the interval  $[0, 1]$ ). The solid line is  $E[\hat{f}(x)]$  and the estimator exhibits little bias—it coin-

cides almost perfectly with the true regression function (the dashed line). The broken lines on either side correspond to two times the standard errors of the estimator at each point— $2(\text{Var}[\hat{f}(x)])^{1/2}$ . In the second panel the bandwidth is substantially broader; we are now averaging

about 15 percent of the data at each point. The standard error curves are tighter but some bias has been introduced. The  $E[\hat{f}(x)]$  no longer coincides perfectly with the true regression curve. In the third panel, averaging is taking place over 40 percent of the data. The standard error curves are even tighter, but now there is substantial bias, particularly at the peaks and valleys of the true regression function. The expectation of the estimator  $E[\hat{f}(x)]$  is fairly flat, while the true regression function undulates widely around it.

A more general formulation of local averaging estimators modifies (1.6) as follows:

$$\hat{f}(x_o) = \sum_1^T w_t(x_o) y_t. \quad (1.8)$$

The estimate of the regression function at  $x_o$  is a weighted sum of the  $y_t$  where the weights  $w_t(x_o)$  depend on  $x_o$ . (A number of local averaging estimators can be put in this form including kernel, nearest neighbor and regressogram.) Since one would expect that observations close to  $x_o$  would have conditional means similar to  $f(x_o)$ , it is natural to assign higher weights to these observations and lower weights to those that are farther away. Local averaging estimators have the advantage that as long as the weights are known, or can be easily calculated,  $\hat{f}$  is also easy to calculate. The disadvantage of such estimators is that it is often difficult to impose additional structure on the estimating function  $\hat{f}$ .

Optimization estimators, on the other hand, are more amenable to incorporating additional structure. As a prelude to our later discussion, consider the following naive estimator. Given data  $(y_1, x_1) \dots (y_T, x_T)$  on  $y_t = f(x_t) + \varepsilon_t$ , where  $x_t \in [0, 1]$ , suppose  $|f'| \leq L$  and we solve

$$\min_{\hat{y}_1, \dots, \hat{y}_T} \frac{1}{T} \sum_t (y_t - \hat{y}_t)^2$$

$$\left| \frac{\hat{y}_t - \hat{y}_s}{x_t - x_s} \right| \leq L \quad s, t = 1, \dots, T. \quad (1.9)$$

Here  $\hat{y}_t$  is the estimate of  $f$  at  $x_t$  and  $\hat{f}$  is a piecewise linear function joining the  $\hat{y}_t$  with slope not exceeding the derivative bound  $L$ . Under general conditions this estimator will be consistent. Furthermore, adding monotonicity or concavity constraints, at least at the points where we have data, is straightforward. As additional structure is imposed, the estimator becomes smoother and its fit to the true regression function improves (see Figure 5).

*1.2.4 A Bird's-Eye View of Important Theoretical Results.* The non/semiparametric literature contains a large number of theoretical results. Here we summarize, in crude form, the main categories of results that are of particular interest to the applied researcher.

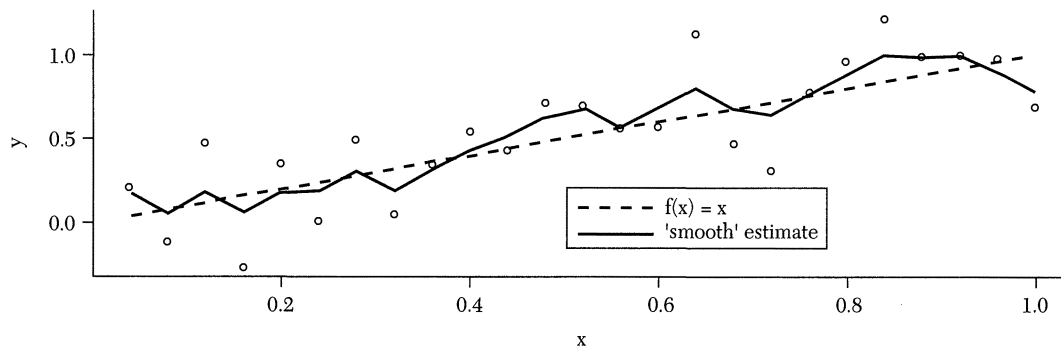
*Computability of Estimators.* Our preliminary exposition of local averaging estimators suggests that their computation is generally straightforward. The naive optimization estimator considered above can also be calculated easily, even with additional constraints on the regression function. What is more surprising is that estimators which minimize the sum of squared residuals over (fairly general) infinite dimensional classes of smooth functions can be obtained by solving finite dimensional (often quadratic) optimization problems. (See Sections 2.1 and 2.2.)

*Consistency.* In nonparametric regression, smoothness conditions (in particular, bounds on derivatives), play a central role in assuring consistency of the estimator. They are also critical in determining the rate of convergence as

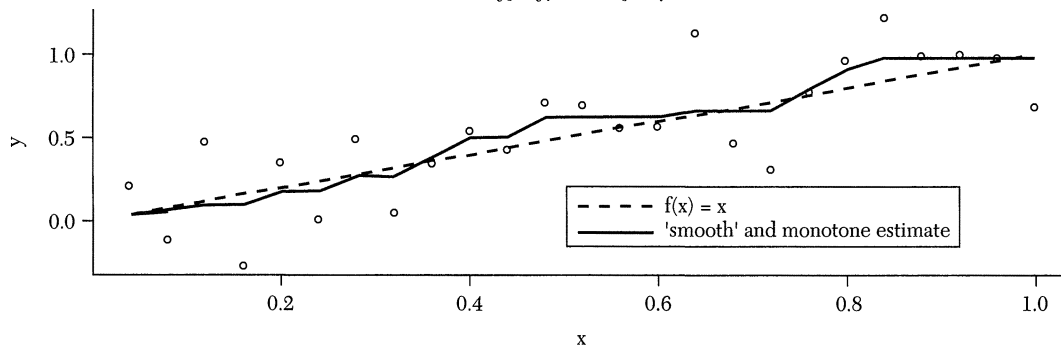
Figure 5 Naive Nonparametric Least Squares

DATA GENERATING MECHANISM  $y_t = x_t + \varepsilon_t$   $\varepsilon_t \sim N(0, 0.04)$   $x_t \in [0, 1]$

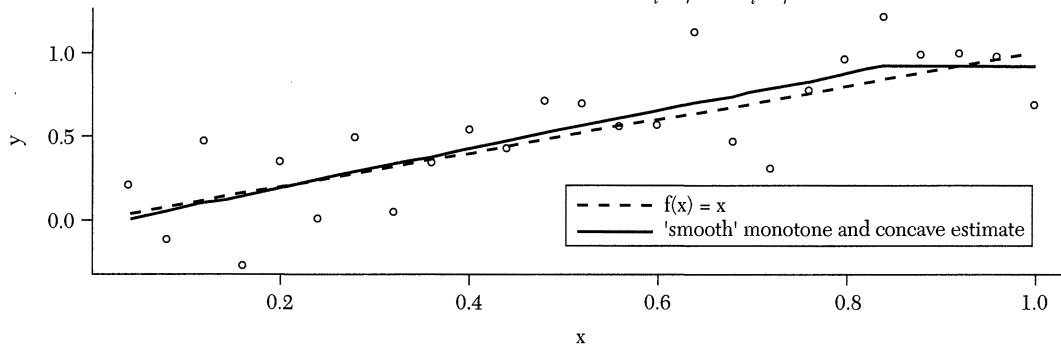
SMOOTHNESS:  $\min_{\hat{y}_1, \dots, \hat{y}_T} \frac{1}{T} \sum (y_t - \hat{y}_t)^2 \quad \left| \frac{\hat{y}_t - \hat{y}_s}{x_t - x_s} \right| \leq 3 \quad s, t = 1, \dots, 25$



SMOOTHNESS AND MONOTONICITY: additional constraints  $\hat{y}_s \leq \hat{y}_t$  for all  $x_s \leq x_t$



SMOOTHNESS, MONOTONICITY AND CONCAVITY: additional constraints  $\hat{y}_s \leq \frac{x_s - x_r}{x_t - x_r} \hat{y}_t + \frac{x_t - x_s}{x_t - x_r} \hat{y}_r$   $r, s, t = 1, \dots, T$ ,  $x_r \leq x_s \leq x_t$



The simulations were performed using GAMS — General Algebraic Modelling System (A. Brooke, D. Kendrick, and A. Meerhaus 1992).

well as certain distributional results.<sup>4</sup> With sufficient smoothness, derivatives of the regression function can be estimated consistently, sometimes by differentiating the estimator of the function itself. (See Sections 2.1 and 2.2.)

*Rate of Convergence.* How quickly does one “discover” the true regression function? In a parametric setting, the rate at which the variance of estimators goes to zero is typically  $1/T$ .<sup>5</sup> It does not depend on the number of explanatory variables. For nonparametric estimators, convergence slows dramatically as the number of explanatory variables increases (recall our earlier discussion of the curse of dimensionality), but this is ameliorated somewhat if the function is differentiable. The optimal rate at which a nonparametric estimator can converge to the true regression function is given by (see Charles J. Stone 1980, 1982)

$$\int [\hat{f}(x) - f(x)]^2 dx = O_P\left(\frac{1}{T^{2m/(2m+d)}}\right), \quad (1.10)$$

where  $m$  equals the degree of differentiability of  $f$  and  $d$  is the dimension of  $x$ . For a twice differentiable function of one variable, (1.10) implies an optimal rate of convergence of  $O_P(T^{-4/5})$  (a case which will recur repeatedly); for a function of two variables it is  $O_P(T^{-2/3})$ .

Local averaging and nonparametric

<sup>4</sup> For example, in proving these results for minimization estimators, smoothness is used to ensure that uniform (over classes of functions) laws of large numbers and uniform central limit theorems apply; see Richard Dudley (1984), David Pollard (1984), and Donald W.K. Andrews (1994a).

<sup>5</sup> In the i.i.d. setting, if  $y = \mu + \varepsilon$ ,  $\text{Var}(\bar{y}) = \sigma_\varepsilon^2/T$  hence  $\mu - \bar{y} = O_P(T^{-1/2})$ . For the linear regression model  $y = \alpha + \beta x + \varepsilon$  we have:

$$\begin{aligned} \int (\alpha + \beta x - \hat{\alpha} - \hat{\beta}x)^2 dx &= (\alpha - \hat{\alpha})^2 \int dx + (\beta - \hat{\beta})^2 \int x^2 dx \\ &+ 2(\alpha - \hat{\alpha})(\beta - \hat{\beta}) \int x dx = O_P(1/T), \end{aligned}$$

since  $\hat{\alpha}, \hat{\beta}$  are unbiased and  $\text{Var}(\hat{\alpha}), \text{Var}(\hat{\beta})$  and  $\text{Cov}(\hat{\alpha}, \hat{\beta})$  converge to 0 at  $1/T$ . The same rate of convergence usually applies to general parametric forms of the regression function.

least squares estimators can be constructed which achieve the optimal rate of convergence (see Sections 2.1 and 2.2.). Rate of convergence also plays an important role in certain test procedures.

If the model is additively separable or partially linear, then the rate of convergence of the optimal estimator depends on the nonparametric component of the model with the highest dimension (Stone 1985, 1986). For example, for the additively separable model  $y = f_a(x_a) + f_b(x_b) + \varepsilon$  where  $x_a, x_b$  are scalars, the convergence rate is the same as if the regression function were a nonparametric function of one variable. The same is true for the partial linear model  $y = z\beta + f(x) + \varepsilon$  where  $x$  and  $z$  are scalars. Estimators of  $\beta$  can be constructed for which the variance shrinks at the parametric rate  $1/T$  and which are asymptotically normal. We have already seen a simple differencing estimator with this property (see Sections 3 and 4.3 for further discussion).

For the hybrid regression function  $f(z, x_a, x_b, x_c) = z\beta + f_a(x_a) + f_b(x_b) + f_c(x_c)$  where  $x_a, x_b, x_c$  are of dimension  $d_a, d_b, d_c$ , respectively, the optimal rate of convergence for the regression as a whole is the same as for a nonparametric regression model with number of variables equal to  $\max\{d_a, d_b, d_c\}$ .

Constraints such as monotonicity or concavity do not enhance the (large sample) rate of convergence if enough smoothness is imposed on the model (see Section 4.4). They can improve performance of the estimator if strong smoothness assumptions are not made or if the dataset is of moderate size (recall Figure 5).

*Bias-Variance Trade-Off.* By increasing the number of observations over which averaging is taking place, one can reduce the variance of a local averaging estimator. But as progressively less similar observations are introduced, the

estimator generally becomes more biased. The objective is to minimize the mean squared error (variance plus bias squared). For nonparametric estimators which achieve optimal rates of convergence, the square of the bias and the variance converge to zero at the same rate (see Section 2.1 below). (In parametric settings the former converges to zero much more quickly than the latter.) Unfortunately, this property complicates the construction of confidence intervals and test procedures.

*Asymptotic Distributions of Estimators.* For a wide variety of nonparametric estimators, the estimate of the regression function at a point is approximately normally distributed. The joint distribution at a collection of points is joint normally distributed, as are various functionals such as the average sum of squared residuals. (See Sections 2.1 and 2.2 below.)

*How Much to Smooth.* Smoothness parameters such as the bandwidth can be selected optimally by choosing the value which minimizes out-of-sample prediction error. The technique, known as cross-validation, will be discussed below (see Section 2.3).

*Testing Procedures.* A variety of specification tests of parametric or semiparametric null hypotheses against nonparametric or semiparametric alternatives are available. Some nonparametric tests of significance are also available. There are also tests of additive separability, monotonicity, homogeneity, concavity and maximization hypotheses. The validity of bootstrap inference procedures has been proved in a number of cases. A fairly unified testing theory can be constructed using conditional moment tests.

### 1.3 Reprise

In this introductory part of the paper, we have argued that nonparametric

regression estimation techniques are based upon principles which should be familiar to the applied economist—in particular, the notion of a local average and the principle of least squares. We have also introduced naive implementations of each of these principles for models with a *single* explanatory variable.

Of course, very few relationships of interest to economists are so simple. Both principles have natural implementations in the presence of multiple explanatory variables. Unfortunately, our ability to accurately estimate the relationship deteriorates as the number of such variables increases. This “curse of dimensionality” can be mitigated somewhat by introducing additional structure such as a semiparametric specification, by assuming higher order differentiability, or by imposing additive separability.

## 2. Nonparametric Regression

### 2.1 Kernel Estimators

**2.1.1 Estimation.** We continue with our nonparametric regression model  $y = f(x) + \varepsilon$  where for the time being  $x$  is a scalar. A conceptually convenient way to construct local averaging weights for substitution into (1.8) is to use a unimodal function centered at zero, which declines in either direction at a rate controlled by a scale parameter. Natural candidates for such functions, which are commonly known as kernels, are probability density functions. Let  $K$  be a bounded function which integrates to one and is symmetric around zero. Define the weights to be

$$w_t(x_0) = \frac{\frac{1}{\lambda T} K\left(\frac{x_t - x_0}{\lambda}\right)}{\frac{1}{\lambda T} \sum_1^T K\left(\frac{x_t - x_0}{\lambda}\right)} \quad (2.1)$$



The shape of the weights (which, by construction, sum to one) is determined by  $K$ , while their magnitude is controlled by  $\lambda$  which is known as the bandwidth. A large value of  $\lambda$  results in greater weight being put on observations that are far from  $x_o$ . Using (1.8) the nonparametric regression function estimator (first suggested by E.A. Nadaraya 1964 and G.S. Watson 1964) becomes

$$\hat{f}(x_o) = \frac{\frac{1}{\lambda T} \sum_1^T K\left(\frac{x_t - x_o}{\lambda}\right) y_t}{\frac{1}{\lambda T} \sum_1^T K\left(\frac{x_t - x_o}{\lambda}\right)} \quad (2.2)$$

A variety of other kernels are available (see Figure 6). Generally, selection of the kernel is less important than selection of the bandwidth over which observations are averaged. The simplest is the uniform kernel (a.k.a. rectangular or box kernel), which takes a value of 1 on  $[-1/2, 1/2]$  and 0 elsewhere. We focus on it next as it will provide us with the clearest insights.

2.1.2 *Uniform Kernel,  $x$ 's Uniformly Distributed on  $[0,1]$* . In order to garner some intuition, in addition to working with a uniform kernel, assume for the moment that  $x$  is uniformly distributed on the unit interval. If we draw  $T$  observations on  $x$ , then the *proportion* of observations falling in an interval of width  $\lambda$  will be approximately  $\lambda$  and the *number* will be approximately  $\lambda T$ .<sup>6</sup>

More formally, define the neighborhood around  $x_o$  as  $N(x_o) = \{x_t \mid x_t \in [x_o - \lambda/2, x_o + \lambda/2]\}$  and note that there are roughly  $\lambda T$  observations in  $N(x_o)$ . For

<sup>6</sup> Two conditions are imposed on  $\lambda$ . The first is  $\lambda \rightarrow 0$  which ensures that averaging takes place over a shrinking bandwidth, thus eventually eliminating bias. The second is  $\lambda T \rightarrow \infty$  which ensures that the number of observations being averaged grows, which allows the variance of the estimate to decline to 0.

the uniform kernel, the denominator of (2.2) will be about 1 and the estimator  $\hat{f}(x_o)$  becomes a simple average of the  $y_t$  over the neighborhood  $N(x_o)$ . (For example, if the bandwidth equals .2, one is averaging about .2 $T$  observations or about 20 percent of the data to obtain the estimate.) Thus, we have

$$\begin{aligned} \hat{f}(x_o) &\cong \frac{1}{\lambda T} \sum_{N(x_o)} y_t \\ &= \frac{1}{\lambda T} \sum_{N(x_o)} f(x_t) + \frac{1}{\lambda T} \sum_{N(x_o)} \varepsilon_t \\ &\cong f(x_o) + \frac{f'(x_o)}{\lambda T} \sum_{N(x_o)} (x_t - x_o) \\ &\quad + \frac{f''(x_o)}{2\lambda T} \sum_{N(x_o)} (x_t - x_o)^2 + \frac{1}{\lambda T} \sum_{N(x_o)} \varepsilon_t \\ &\cong f(x_o) + \frac{1}{2} f''(x_o) \frac{1}{\lambda T} \sum_{N(x_o)} (x_t - x_o)^2 \\ &\quad + \frac{1}{\lambda T} \sum_{N(x_o)} \varepsilon_t \end{aligned} \quad (2.3)$$

We have applied a second order Taylor series.<sup>7</sup> Next, we rewrite (2.3) as<sup>8</sup>

<sup>7</sup> In particular,  $f(x_t) = f(x_o) + f'(x_o)(x_t - x_o) + \frac{1}{2} f''(x_o)(x_t - x_o)^2 + o(x_t - x_o)^2$ . We are obviously assuming second order derivatives exist.

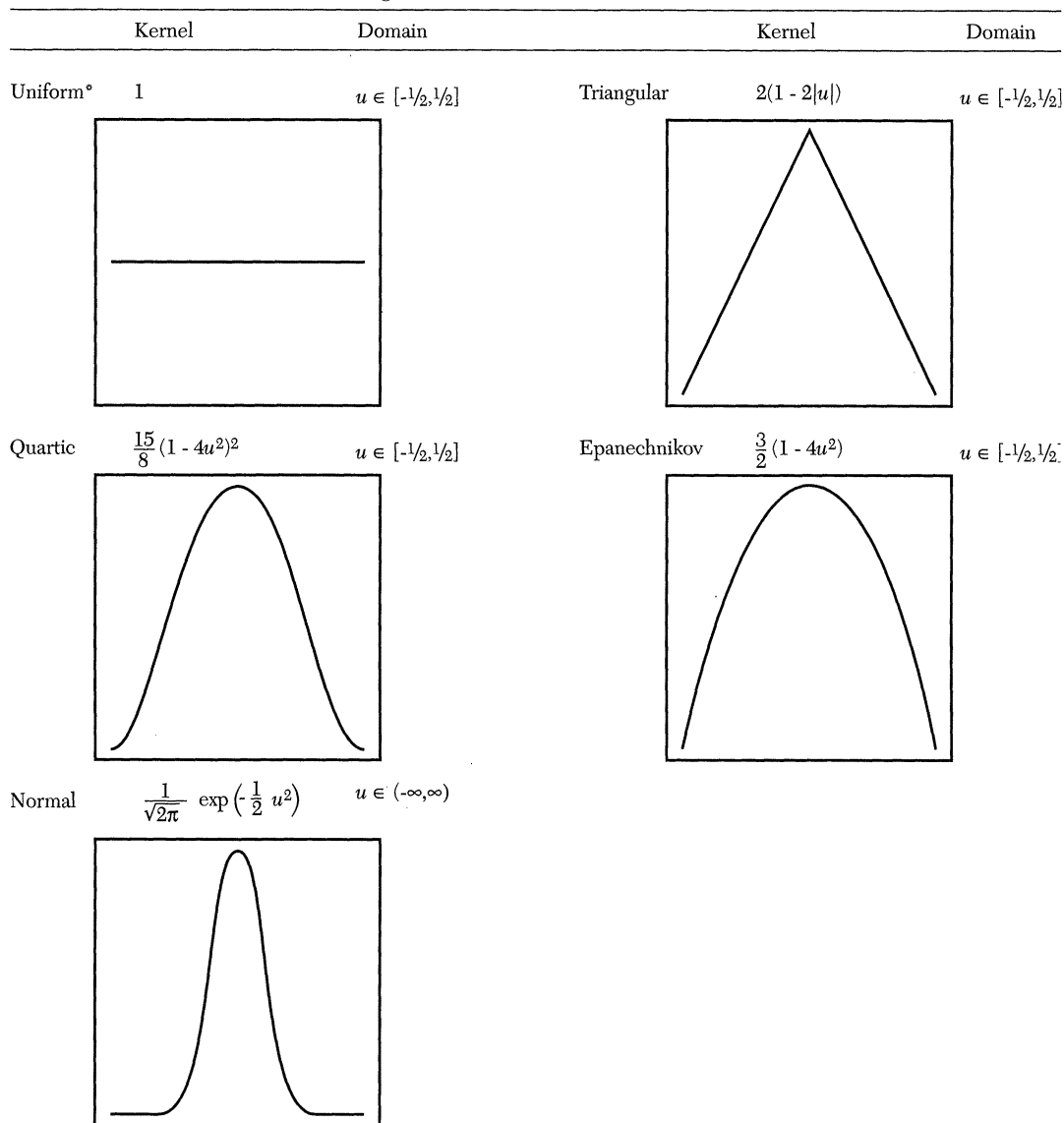
<sup>8</sup> Consider a random variable  $U_\lambda$  which is uniformly distributed on an interval of width  $\lambda$  centered at 0. Then  $\text{Var}(U_\lambda) = \lambda^2/12$ . To obtain (2.4)

from (2.3) it is useful to think of  $\frac{1}{\lambda T} \sum_{x_t \in N(x_o)} (x_t - x_o)$

as an average of  $\lambda T$  such random variables, in which case the average has mean 0 and variance  $\lambda/12T$ . Thus, the second term in the third line of (2.3) converges to 0 fast enough so that for our purposes we can set it to 0. Next, think of  $\frac{1}{\lambda T} \sum_{x_t \in N(x_o)} (x_t - x_o)^2$  as an estimator of the variance of

$U_\lambda$ . The estimator converges quickly enough so that the second term of the last equation of (2.3) can be approximated by  $1/24 f''(x_o) \lambda^2$ .

Figure 6. Alternative Kernel Functions



\*Also known as the 'box' or 'rectangular' kernel.

$$\hat{f}(x_o) \cong f(x_o) + \frac{1}{24} \lambda^2 f''(x_o) + \frac{1}{\lambda T} \sum_{N(x_o)} \varepsilon_i \quad (2.4)$$

The last term is an average of about  $\lambda T$  independent and identical random vari-

ables so that its variance is  $\sigma_\varepsilon^2 / \lambda T$  and we have:

$$\hat{f}(x_o) = f(x_o) + O(\lambda^2) + O_P\left(\frac{1}{(\lambda T)^{1/2}}\right) \quad (2.5)$$

The bias  $E(\hat{f}(x_o) - f(x_o))$  is approximated by the second term of (2.4) and the  $\text{Var}(\hat{f}(x_o))$  is approximately  $\sigma_\varepsilon^2/\lambda T$  so that the mean squared error (the sum of the bias squared and the variance) at a point  $x_o$  is

$$E[\hat{f}(x_o) - f(x_o)]^2 = O(\lambda^4) + O\left(\frac{1}{\lambda T}\right). \quad (2.6)$$

The approximation embodied in (2.4) yields dividends immediately. As long as  $\lambda \rightarrow 0$  and  $\lambda T \rightarrow \infty$ , the second and third terms go to zero and we have a consistent estimator.

The rate at which  $\hat{f}(x_o) - f(x_o) \rightarrow 0$  depends on which of the second or third terms in (2.4) converge to zero more slowly. Optimality is achieved when the bias squared and the variance shrink to zero at the same rate. Using (2.5) one can see that this occurs if  $O(\lambda^2) = O_P((T\lambda)^{-1/2})$  which implies that optimality can be achieved by choosing  $\lambda = O(T^{-1/5})$ . In this case

$$\hat{f}(x_o) = f(x_o) + O\left(\frac{1}{T^{2/5}}\right) + O_P\left(\frac{1}{T^{2/5}}\right). \quad (2.7)$$

Equivalently, we could have solved for the optimal rate using (2.6). Setting  $O(\lambda^4) = O(1/\lambda T)$  and solving we again obtain  $\lambda = O(T^{-1/5})$ . Substituting into (2.6) yields a rate of convergence for the mean squared error at a point  $x_o$  of  $E[\hat{f}(x_o) - f(x_o)]^2 = O(T^{-4/5})$ . This in turn underpins the following:

$$\int [\hat{f}(x) - f(x)]^2 dx = O_P\left(\frac{1}{T^{4/5}}\right) \quad (2.8)$$

a rather pleasant result in that it satisfies Stone's optimal rate of convergence, equation (1.10) above ( $m=2, d=1$ ).

Applying a central limit theorem<sup>9</sup> to the last term of (2.4), we have:

<sup>9</sup>One must be a little bit careful because the number of  $\varepsilon_i$  being summed is random. See, for example, Robert Serfling (1980, p. 32).

$$(\lambda T)^{1/2} \left( \hat{f}(x_o) - f(x_o) - \frac{1}{24} \lambda^2 f''(x_o) \right) \sim N(0, \sigma_\varepsilon^2) \quad (2.9)$$

If we select  $\lambda$  optimally, say,  $\lambda = T^{-1/5}$ , then  $(\lambda T)^{1/2} \lambda^2 = 1$  and the construction of a confidence interval for  $f(x_o)$  is complicated by the presence of the term involving  $f''(x_o)$  (which would need to be estimated). However, if we select  $\lambda$  to go to zero faster than  $T^{-1/5}$  (for example,  $\lambda = T^{-1/4}$ ), then  $(\lambda T)^{1/2} \lambda^2 \rightarrow 0$  and (2.9) becomes  $(\lambda T)^{1/2} (\hat{f}(x_o) - f(x_o)) \sim N(0, \sigma_\varepsilon^2)$ . Intuitively, the bandwidth is shrinking fast enough so that the bias is small relative to the variance (see (2.5) above). In this case, a 95 percent confidence interval for  $f(x_o)$  is approximately  $\hat{f}(x_o) \pm 1.96 \sigma_\varepsilon / (\lambda T)^{1/2}$ .

Let us pause for a moment. In this section, we have illustrated three essential results for a simple kernel estimator: that it is consistent; that by averaging over a neighborhood which shrinks at an appropriate rate it achieves the optimal rate of convergence, and that it is asymptotically normal.

**2.1.3 General Kernel,  $x$ 's Distributed with General Distribution  $p(x)$ .** For a general kernel and assuming that the  $x$ 's are random with probability density  $p(x)$ , the Nadaraya-Watson kernel estimator (2.2) is consistent. The numerator converges to  $f(x_o)p(x_o)$  and the denominator converges to  $p(x_o)$ .

The rate of convergence is optimized if  $\lambda = O(T^{-1/5})$  in which case the integrated squared error converges at the optimal rate  $O_P(T^{-4/5})$  as in (2.8). Confidence intervals may be constructed using:

$$(\lambda T)^{1/2} \left( \hat{f}(x_o) - f(x_o) - \frac{1}{2} a \kappa \lambda^2 \left( f''(x_o) + 2 f'(x_o) \frac{p'(x_o)}{p(x_o)} \right) \right) \sim N \left( 0, \frac{b_k \sigma_\varepsilon^2}{p(x_o)} \right) \quad (2.10)$$

where  $p(\cdot)$  is the density of  $x$  and

$$a_K = \int u^2 K(u) du \quad b_K = \int K^2(u) du \quad (2.11)$$

If the bandwidth converges to zero at the optimal rate, then estimates of the first two derivatives of  $f$ , the density of  $x$  and its first derivative, as well as the variance of the residual must all be obtained in order to construct a confidence interval which in large samples will have the correct coverage probability.<sup>10</sup>

Again, the bias term in (2.10) can be made to disappear asymptotically by permitting the bandwidth to shrink at a rate that is faster than the optimal rate. Averaging over a narrower range reduces the bias but increases the variance of the estimator.

An alternative, which takes account of both bias and variance terms without requiring their calculation, is based on the bootstrap. Figure 7 provides implementation details for both asymptotic and bootstrap confidence intervals.

Thus far we have discussed confidence intervals for  $f$  at a point. A more interesting graphic for nonparametric estimation is a confidence band or ribbon around the estimated function. The plausibility of an alternative specification (such as a parametric estimate, a monotone, or concave estimate) could then be assessed by superimposing the latter on the graph to see if it falls within the band.<sup>11</sup>

Figure 8 provides nonparametric kernel estimates of the Engel curve relating expenditure on food to income. The 90 percent uniform confidence band was constructed using the *reguncb* function in XploRe (Härdle, Klinke and Turlach

1995). Also displayed is an estimated linear regression function which falls within the nonparametric confidence band at all but extreme levels of income.

**2.1.4 Kernel Estimation of Functions of Several Variables.** In economics, it is rarely the case that one is interested in a function of a single variable. Even if we are comfortable incorporating most of our explanatory variables parametrically (for example, within a partial linear model), there may be more than one variable entering nonparametrically. The effect of geographic location (a two-dimensional variable) provides one such example (see Figure 11 below).

Suppose then that  $f$  is a function of two variables. We are given data  $(y_1, x_1), \dots, (y_T, x_T)$  where  $x_t = (x_{t1}, x_{t2})$  on the model  $y_t = f(x_{t1}, x_{t2}) + \epsilon_t$  and we will assume  $f$  is a function on the unit square  $[0, 1]^2$ . We want to estimate  $f(x_o)$  by averaging nearby observations; in particular we will average observations falling in a square of dimension  $\lambda \cdot \lambda$  which is centered at  $x_o$ . If the  $x_t$  are drawn from a uniform distribution on the unit square, there will be (approximately)  $\lambda^2 T$  observations in the neighborhood defined by  $N(x_o) = \{x_t \mid x_{t1} \in x_{o1} \pm \lambda/2, x_{t2} \in x_{o2} \pm \lambda/2\}$ . (For example, any square with sides  $\lambda = .5$  has area .25 and will capture about 25 percent of the observations.) Consider then:

$$\begin{aligned} \hat{f}(x_o) &= \frac{1}{\lambda^2 T} \sum_{N(x_o)} y_t \\ &= \frac{1}{\lambda^2 T} \sum_{N(x_o)} f(x_t) + \frac{1}{\lambda^2 T} \sum_{N(x_o)} \epsilon_t \\ &\cong f(x_o) + O(\lambda^2) + \frac{1}{\lambda^2 T} \sum_{N(x_o)} \epsilon_t \\ &\cong f(x_o) + O(\lambda^2) + O_P\left(\frac{1}{\lambda T^{1/2}}\right). \end{aligned} \quad (2.12)$$

<sup>10</sup> See Härdle and Linton (1994, p. 2310). The normality result in (2.10) extends to the case where one is interested in the joint distribution of estimates at a vector of points.

<sup>11</sup> See Härdle and Oliver Linton (1994, p. 2317), Peter Hall (1993), and Randall Eubank and Paul Speckman (1993).

Figure 7. Confidence Intervals for Kernel Estimators — Implementation

ASYMPTOTIC CONFIDENCE INTERVAL AT  $f(x_0)$

1. Select  $\lambda$  so that  $T^{1/5} \lambda \rightarrow 0$ , e.g.  $\lambda = O(T^{-1/4})$ . This ensures that the bias term does not appear in the limiting distribution (2.10).
2. Select a kernel  $K$  and calculate  $b_K = \int K^2(u)du$ . For the uniform kernel on  $[-1/2, 1/2]$   $b_K = 1$ .
3. Estimate  $f$  using the Nadaraya-Watson estimator (2.2).
4. Calculate  $\hat{\sigma}_\varepsilon^2 = 1/T \sum (y_t - \hat{f}(x_t))^2$ .
5. Estimate  $p(x_0)$  using denominator of (2.2). If the uniform kernel is used,  $\hat{p}(x_0)$  equals the number of  $x_t$  in the interval  $x_0 \pm \lambda/2$  divided by  $\lambda$ .
6. Calculate the confidence interval at  $f(x_0)$  using  $\hat{f}(x_0) \pm 1.96 \sqrt{\frac{b_K \hat{\sigma}_\varepsilon^2}{\hat{p}(x_0) \lambda T}}$

BOOTSTRAP CONFIDENCE INTERVAL\* AT  $f(x_0)$

1. Estimate  $f$  using the optimal bandwidth  $\lambda = O(T^{-1/5})$ , call this estimate  $\hat{f}_\lambda$ , then calculate the residuals  $\hat{\varepsilon}_t = y_t - \hat{f}_\lambda(x_t)$ .
2. Re-estimate  $f$  using a wider bandwidth, say  $\tilde{\lambda}$ , (which will result in some oversmoothing) and call this estimate  $\hat{f}_{\tilde{\lambda}}$ . Resample the estimated residuals  $\hat{\varepsilon}_t$  using the 'wild' bootstrap to obtain bootstrap residuals  $\hat{\varepsilon}_t^B$  and construct a bootstrap dataset  $y_t^B = \hat{f}_{\tilde{\lambda}}(x_t) + \hat{\varepsilon}_t^B, t=1, \dots, T$ .
3. Estimate  $f(x_0)$  using the bootstrap data and  $\lambda = O(T^{-1/5})$  to obtain  $\hat{f}_\lambda^B(x_0)$ . Repeat the resampling many times and obtain the .025 and .975 quantiles of the distribution of  $\hat{f}_\lambda^B(x_0)$ . The result yields a 95% confidence interval for  $f(x_0)$  which has the correct coverage probability in large samples.

\*For the theory underlying this bootstrap methodology see Härdle (1990, Th. 4.2.2, and pp. 106-7, 247). See also Appendix B of this paper.

What we have done here is to mimic the reasoning in equations (2.3)–(2.5) but this time for the bivariate case. (We have assumed that  $f$  is twice differentiable, but have spared the reader the Taylor series expansion.) However, there is a subtle difference. The bias term is still proportional to  $\lambda^2$ , but the variance term is now  $O_p(1/\lambda T^{1/2})$  rather than  $O_p(1/(\lambda T)^{1/2})$  since we are averaging approximately  $\lambda^2 T$  values of  $\varepsilon_t$ .

Hence for consistency, we now need  $\lambda \rightarrow 0$  and  $\lambda T^{1/2} \rightarrow \infty$ . As before, convergence of  $\hat{f}(x_0)$  to  $f(x_0)$  is fastest when the bias and variance terms go to zero at the same rate, that is when  $\lambda = O(T^{-1/6})$ . The second and third terms of the last line of (2.12) are then  $O_p(T^{-1/3})$  and  $\int [\hat{f}(x) - f(x)]^2 dx = O_p(T^{-2/3})$  which is optimal (see (1.10) above).

More generally, if the  $x_t$  are  $d$ -dimensional with probability density  $p(x)$  defined on the unit cube in  $R^d$ , and we are

using a kernel  $K$  then the estimator in (2.2) becomes

$$\hat{f}(x_0) = \frac{\frac{1}{\lambda^d T} \sum_1^T y_t \prod_{i=1}^d K\left(\frac{x_{ti} - x_{oi}}{\lambda}\right)}{\frac{1}{\lambda^d T} \sum_1^T \prod_{i=1}^d K\left(\frac{x_{ti} - x_{oi}}{\lambda}\right)} \quad (2.13)$$

Again, if  $K$  is the uniform kernel on  $[-1/2, 1/2]$ , then the product of the kernels (hence the term product kernel) is one only if  $x_{ti} \in [x_{oi} - \lambda/2, x_{oi} + \lambda/2]$  for  $i = 1, \dots, d$ , that is only if  $x_t$  falls in the  $d$ -dimensional cube centred at  $x_0$  with sides of length  $\lambda$ .

Above we have introduced a simple kernel estimator for functions of several variables which averages observations over a cube centered at  $x_0$ . A multitude of variations and alternatives exists. For example, one could select different

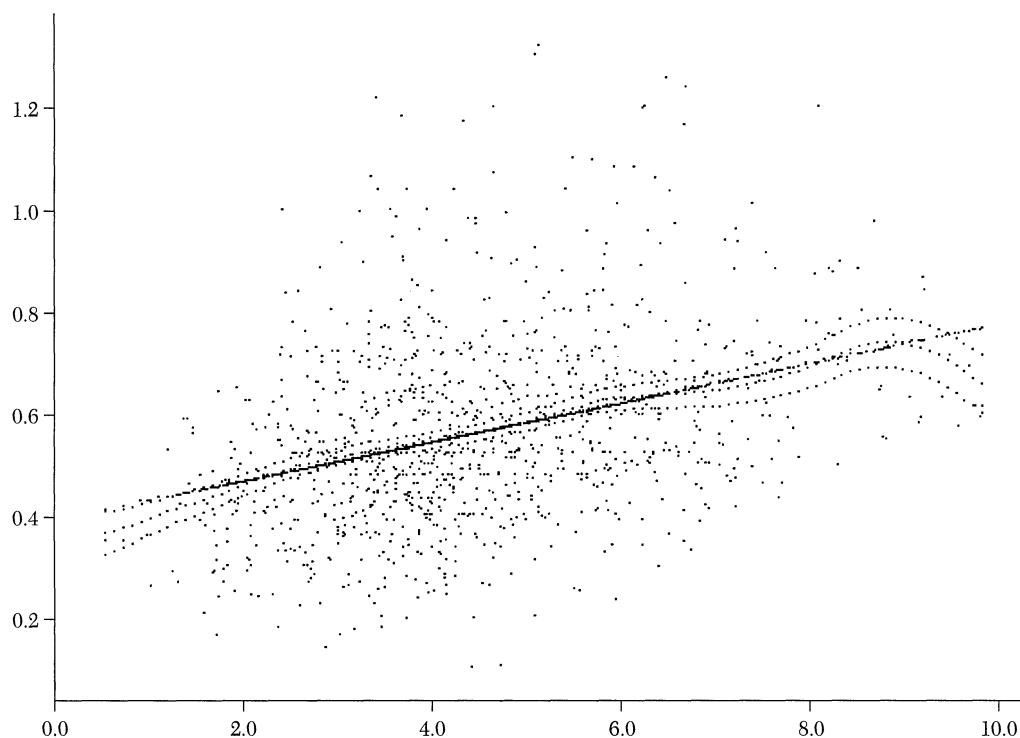
Figure 8. Engel Curve Estimation

MODEL:  $y = f(x) + \epsilon$  where  $y$  is expenditure on food,  $x$  is income (units are in  $10^4$  Canadian dollars).

DATA: the data consist of a sample of 1058 two-parent households with two children below 16 years of age.

Source, 1992 Survey of Family Expenditures, Household Survey Division, Statistics Canada.

## ESTIMATED MODELS



Linear Model and Kernel Estimate with 90% Uniform Confidence Band

Estimation and graphics produced using XploRe (Härdle, Klinke, and Turlach 1995). Uniform confidence band use routine *regunb*.

bandwidths for each dimension so that averaging would take place over rectangular cubes rather than over perfect cubes. Or, one might select different kernels for each dimension. Still more generally, one could average over non-rectangular regions such as spheres or ellipsoids. For details in a multivariate density estimation setting, see for example, David Scott (1992, pp.149–55).

## 2.2 Nonparametric Least Squares <sup>12</sup>

**2.2.1 Estimation.** In order to implement nonparametric least squares estimators, we will need a tractable way to impose constraints on various order derivatives. Let  $C^m$  be the set of functions

<sup>12</sup> Spline function estimation, which is closely related to nonparametric least squares, is a widely used technique. We begin with an exposition of the latter, then explain its relationship to the former.

that have continuous derivatives up to order  $m$  (for expositional purposes we restrict these functions to the unit interval). A measure of smoothness that is particularly convenient is given by the Sobolev norm:

$$\|f\|_{sob} = \left[ \int f^2 + (f')^2 + (f'')^2 + \dots + (f^{(m)})^2 \right]^{1/2}, \quad (2.14)$$

where  $(m)$  denotes the  $m^{\text{th}}$  order derivative. A small value of the norm implies that neither the function, nor any of its derivatives up to order  $m$  can be too large over a significant portion of the domain. Indeed, bounding this norm implies that all lower order derivatives are bounded in supnorm.<sup>13</sup> (Recall from Section 1.2 and Figure 5 that even bounding the first derivative produces a consistent nonparametric least squares estimator.)

Suppose we take our estimating set  $\mathfrak{S}$  to be the set of functions in  $C^m$  for which the square of the Sobolev norm is bounded by say  $L$ , that is,  $\mathfrak{S} = \{f \in C^m, \|f\|_{sob}^2 \leq L\}$ . The task of finding the function in  $\mathfrak{S}$  that best fits the data would appear to be daunting. After all,  $\mathfrak{S}$  is an infinite dimensional family. What is remarkable is that the solution  $\hat{f}$  that satisfies

$$s^2 = \min_f \frac{1}{T} \sum_t [y_t - f(x_t)]^2 \quad \text{s.t.} \quad \|f\|_{sob}^2 \leq L \quad (2.15)$$

can be obtained by minimizing a quadratic objective function subject to a quadratic constraint. The solution is of the form  $\hat{f} = \sum_1^T \hat{c}_t r_{x_t}$  where  $r_{x_1}, \dots, r_{x_T}$  are functions computable from  $x_1, \dots, x_T$  and  $\hat{c} = (\hat{c}_1, \dots, \hat{c}_T)$  is obtained by solving

<sup>13</sup> That is, there exist constants  $L_i^*$  such that for all  $f \in \mathfrak{S}$ ,  $\sup_x |f^{(i)}| \leq L_i^*$ ,  $i = 1, \dots, m - 1$ , where  $\mathfrak{S}$  is defined momentarily. This result flows from the Sobolev Imbedding Theorem.

$$\min_c \frac{1}{T} [y - Rc]' [y - Rc] \quad \text{s.t.} \quad c'Rc \leq L \quad (2.16)$$

where  $y$  is the  $T \times 1$  vector of observations on the dependent variable and  $R$  is a  $T \times T$  matrix that is computable from  $x_1, \dots, x_T$ . Note that even though one is estimating  $T$  parameters to fit  $T$  observations, the parameters are constrained so that there is no immediate reason to expect perfect fit.<sup>14</sup>

Repeating the main point: the infinite dimensional nonparametric least squares problem (2.15) may be solved by solving the finite dimensional optimization problem (2.16). Furthermore, if  $x$  is a vector, the Sobolev norm (2.14) generalizes to include various order partial derivatives. The optimization problem has the same quadratic structure as in the one-dimensional case above, and the functions  $r_{x_1}, \dots, r_{x_T}$  as well as the matrix  $R$  are directly computable from the data  $x_1, \dots, x_T$ .

Our Returns to Scale in Electricity Distribution example, which we continue in Figure 10 below, illustrates a nonparametric least squares estimator of  $f$ . We have imposed a fourth order Sobolev norm ( $m=4$  in equation [2.14]), which implies at least three bounded derivatives. As one can see, the nonparametric least squares estimate is very similar to the quadratic estimate of the scale effect.

**2.2.2 Properties of Estimator<sup>15</sup>** The main statistical properties of the procedure are these:  $\hat{f}$  is a consistent estimator of  $f$ , indeed low order derivatives of  $\hat{f}$  consistently estimate the corresponding derivatives of  $f$ . The rate at which  $\hat{f}$

<sup>14</sup> The  $r_{x_t}$  are called representer functions and  $R$ , the matrix of inner products of the  $r_{x_t}$ , the representer matrix. See Grace Wahba (1990) for details. An efficient algorithm for solving the optimization problem in (2.16) may be found in Gene Golub and Charles Van Loan (1989, p. 564).

<sup>15</sup> These results may be proved using empirical processes theory as discussed in Dudley (1984) and Pollard (1984).

converges to  $f$  satisfies equation (1.10). These optimal convergence results are useful in producing consistent tests of a broad range of hypotheses.

The average minimum sum of squared residuals  $s^2$  is a consistent estimator of the residual variance  $\sigma_\varepsilon^2$ . Furthermore, in large samples,  $s^2$  is indistinguishable from the true average sum of squared residuals in the sense that

$$T^{1/2} \left( s^2 - \frac{1}{T} \sum \varepsilon_t^2 \right) \rightarrow 0 \quad (2.17)$$

in probability. Next, since  $T^{1/2}(1/T \sum \varepsilon_t^2 - \sigma_\varepsilon^2) \rightarrow N(0, \text{Var}(\varepsilon^2))$  (just apply an ordinary central limit theorem), equation (2.17) implies

$$T^{1/2}(s^2 - \sigma_\varepsilon^2) \rightarrow N(0, \text{Var}(\varepsilon^2)). \quad (2.18)$$

As we shall see below, this result lies at the heart of demonstrating that nonparametric least squares can be used to produce  $T^{1/2}$  consistent normal estimators in the partial linear model.

Finally, the estimator being considered here is closely related to spline function estimators. Assume for the moment  $\eta > 0$  is a given constant<sup>16</sup> and consider the penalized least squares problem:

$$\min_f \frac{1}{T} \sum_t [y_t - f(x_t)]^2 + \eta \|f\|_{\text{Sob}}^2. \quad (2.19)$$

The criterion function trades off fidelity to the data against smoothness of the function  $f$ . There is a penalty for selecting functions which fit the data extremely well but as a consequence are very rough (recall that the Sobolev norm measures the smoothness of a function and its derivatives). A larger  $\eta$  results in a smoother function being selected.

If one solves (2.15), our nonparametric least squares problem; takes the Lagrangian multiplier, say  $\hat{\eta}$  associated with the smoothness constraint; then

<sup>16</sup> Actually, it is selected using cross-validation, a procedure which we discuss shortly.

uses it in solving (2.19), the resulting  $\hat{f}$  will be identical.<sup>17</sup>

### 2.3 Selection of Smoothing Parameter<sup>18</sup>

**2.3.1 Kernel Estimation.** We now turn to selection of smoothing parameters for kernel estimators. If the bandwidth  $\lambda$  is too large, then oversmoothing will exacerbate bias and eliminate important features of the regression function. Selection of a value of  $\lambda$  that is too small will cause the estimator to track the current data too closely, thus impairing the prediction accuracy of the estimated regression function when applied to new data (recall Figures 3 and 4). To obtain a good estimate of  $f$  we would like to select  $\lambda$  to minimize the mean integrated squared error:

$$\text{MISE}(\lambda) = E \int [\hat{f}(x; \lambda) - f(x)]^2 dx \quad (2.20)$$

where we write  $\hat{f}(x; \lambda)$  to explicitly denote that the kernel estimator depends on the choice of  $\lambda$ . Of course we do not observe  $f$  so the *MISE* cannot be minimized directly. Nor will selecting  $\lambda$  by minimizing the estimate of the residual variance

$$\hat{\sigma}_\varepsilon^2(\lambda) = \frac{1}{T} \sum_{t=1}^T [y_t - \hat{f}(x_t; \lambda)]^2 \quad (2.21)$$

lead to a useful result—the minimum of (2.21) occurs when  $\lambda$  is reduced to the point where the data are fit perfectly. However, this idea *can* be modified to produce useful results. Consider a slight variation on (2.21) known as the cross-validation function:

$$\text{CV}(\lambda) = \frac{1}{T} \sum_{t=1}^T [y_t - \hat{f}_{-t}(x_t; \lambda)]^2. \quad (2.22)$$

<sup>17</sup> In their simplest incarnation, spline estimators use  $\int (f'')^2$  as the measure of smoothness. See Eubank (1988) and Wahba (1990).

<sup>18</sup> Cross-validation was first proposed for the kernel estimator by Clark (1975) and for spline estimation by Wahba and Wold (1975).



The only difference between (2.21) and (2.22) is that the kernel estimator is subscripted with a curious “- $t$ ” which is used to denote that  $\hat{f}_{-t}$  is obtained by *omitting* the  $t^{\text{th}}$  observation. Thus the estimate of  $f$  at each point  $x_t$  is obtained by estimating the regression function using all *other* observations, then predicting the value of  $f$  at the omitted observation. (For a given value of  $\lambda$ ,  $CV(\lambda)$  requires calculation of  $T$  separate kernel estimates.)<sup>19</sup>

This subtle change results in some extremely propitious properties (see, for example, Härdle and James Marron 1985, Härdle, Hall, and Marron 1988). In particular, suppose an optimal  $\lambda$ , say  $\lambda_{OPT}$ , could be chosen to minimize (2.20). Let  $\hat{\lambda}$  be the value that minimizes (2.22). Then  $MISE(\hat{\lambda})/MISE(\lambda_{OPT})$  converges to one; that is, in large samples, selecting  $\lambda$  through cross-validation is as good as knowing the  $\lambda$  that minimizes the integrated mean squared error.

**2.3.2 Nonparametric Least Squares.** The heuristics of selection of the smoothness bound for nonparametric least squares are similar. If we select  $L$  in (2.15) to be much larger than the true norm, then our estimator will be less efficient though it will be consistent. If we select a bound that is smaller than the true norm, then our estimator will generally be inconsistent. The cross-validation function is defined as

$$CV(L) = \frac{1}{T} \sum_{t=1}^T [y_t - \hat{f}_{-t}(x_t)]^2, \quad (2.23)$$

where  $\hat{f}_{-t}$  is obtained by solving

<sup>19</sup> The notion that out-of-sample prediction is a useful criterion for estimation and testing is, of course, quite generally applied in statistics. In the simplest case, one can imagine dividing a sample in two, using one part to estimate the model, and the other to assess its accuracy or validity. This naive approach, however, does not make optimal use of the data, a problem that is resolved through the cross-validation device.

$$\min_f \frac{1}{T} \sum_{i \neq t} [y_i - f(x_i)]^2$$

$$\text{s.t. } \|f\|_{\text{Sob}}^2 \leq L. \quad (2.24)$$

Note the subtle change from 2.15.

The interpretation of the smoothing parameter is somewhat different. In kernel estimation it corresponds to the width of the interval over which averaging takes place; in nonparametric least squares it is the diameter of the set of functions over which estimation takes place.

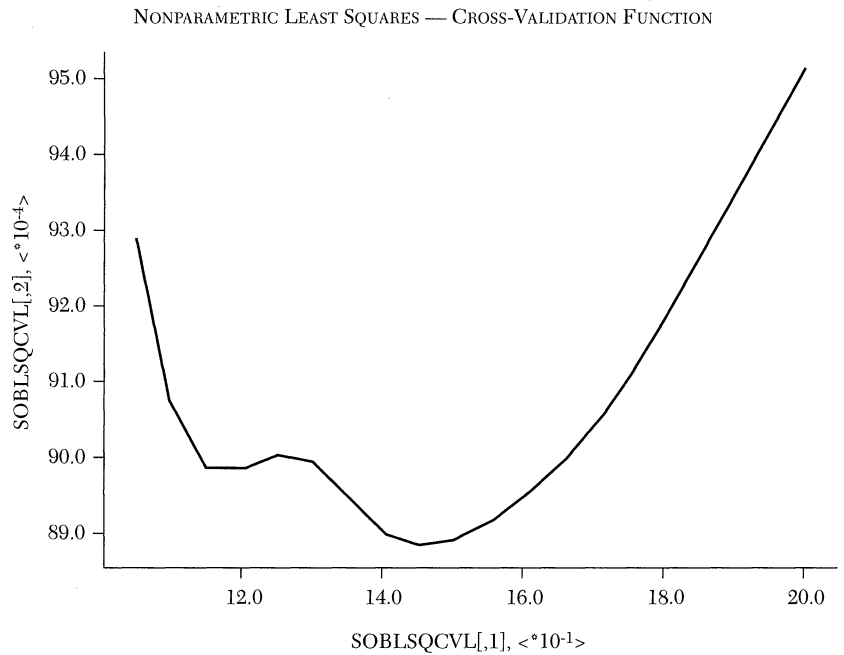
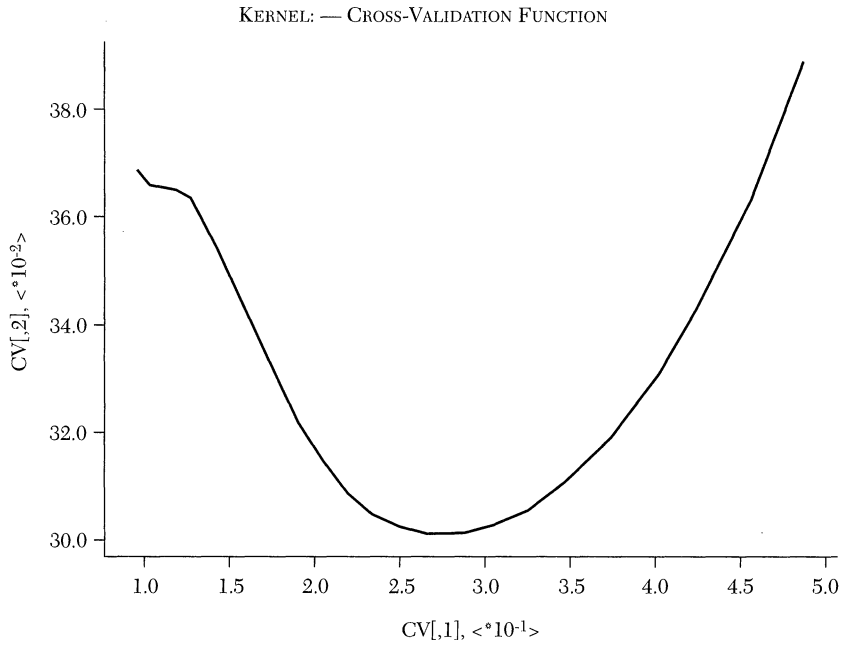
Figure 9 illustrates the behavior of the cross-validation function for both kernel and nonparametric least squares estimators. The data generating mechanism is given by the model  $y_t = x_t + \varepsilon_t$ ,  $\varepsilon_t \sim N(0, 0.1)$ ,  $t = 1, \dots, 25$  where  $x_t$  are equally spaced on the interval  $[0, 1]$ . The minimum of the cross-validation function for the kernel estimator is approximately .25. Thus, at a typical point  $x_0$  the optimal kernel estimator  $\hat{f}(x_0)$  will involve about 25 percent of the observations.

For the nonparametric least squares cross-validation function, note first that the *square* of the Sobolev norm (2.14) of the true regression function is given by  $\int_0^1 x^2 + 1 = 1\frac{1}{3}$ . Thus,  $L = 1\frac{1}{3}$  would be the smallest value which would ensure consistency of the nonparametric least squares problem (2.15). In the simulations (Figure 9), the minimum of the cross-validation function is between 1.4 and 1.5.<sup>20</sup>

**2.3.3 Further Comments.** A number of researchers have investigated alternate procedures for selecting the smoothing parameter. Unlike the case of kernel estimation of density functions, no convenient rules of thumb

<sup>20</sup> For optimality results on cross-validation in a spline setting, see Li (1986, 1987) and Wahba (1990, p. 47).

Figure 9. Selection of Smoothing Parameters



Data generating mechanism:  $y_t = x_t + \varepsilon_t$ ,  $\varepsilon_t \sim N(0, .01)$ ,  $t=1, \dots, 25$ , where  $x_t$  are equally spaced on the interval  $[0,1]$ . Kernel cross-validation performed using XploRe function *regcvl*, see Härdle, Klinke, and Turlach (1995, p. 88). Nonparametric least squares cross-validation performed using Fortran code written by the author.

are available for kernel regression.<sup>21</sup> However, by simply trying different values for the smoothing parameter and visually examining the resulting estimate of the regression function, one can often obtain a useful indication of whether one is over- or under-smoothing.

Furthermore, cross-validation can be automated relatively easily. The kernel cross-validation function in Figure 9 was obtained using the *regcvl* in XploRe. S-Plus uses cross-validation to produce its spline estimates, and other automated procedures are also available.<sup>22</sup>

#### 2.4 Reprise

This portion of the paper focussed on developing the kernel estimator (an example of a local averaging estimator) and on a nonparametric least squares estimator closely related to spline estimation. The statistical properties of the former are relatively easy to derive using basic techniques. For example, it is straightforward to demonstrate that kernel estimators are consistent and approximately normally distributed. If the estimator is to achieve the optimal rate of convergence, then its bias squared and variance must go to zero at the same rate, which complicates the construction of confidence intervals. Despite the somewhat burdensome notation, for example, equation (2.13), kernel estimation of functions of *several* variables is conceptually straightforward—for example, with a uniform kernel, averages are taken over a square (or cube) rather than over an interval.

Derivation of the properties of non-

<sup>21</sup> For “rules of thumb” in a kernel density setting, see Scott (1992, ch. 6). For alternatives to cross-validation in nonparametric regression setting, see for example, Simonoff (1996, p. 197 and references therein).

<sup>22</sup> See Venables and Ripley (1994, p. 250).

parametric least squares or spline estimators is considerably more difficult and we have merely asserted them. Such estimators are consistent and achieve optimal rates of convergence. They also have considerable appeal since it is relatively easy to impose additional structure on such estimators. (We will devote attention to constrained estimation and hypothesis testing in Section 4 below.)

Both local averaging and least squares estimation require the selection of a smoothing parameter. In kernel estimation, it is the neighborhood (or bandwidth or window) over which averaging is to take place. In nonparametric least squares it is the diameter of the ball of functions over which estimation is to take place. Selection of the smoothing parameter is performed by trying different values and selecting the one that minimizes (out-of-sample) prediction error, a technique known as cross-validation.

It should be emphasized that in this section we have introduced but two nonparametric regression estimators. A wide variety of others exists (see for example Härdle 1990, Ch.3 and Jianqing Fan and Irene Gijbels 1996, Ch. 2 for overviews).

### 3. The Partial Linear Model

#### 3.1. Estimation

3.1.1 *Introduction.* Given i.i.d. data  $(y_1, x_1, z_1), \dots, (y_T, x_T, z_T)$  consider the semi-parametric regression model which was discussed in the opening section of the paper:

$$y = z\beta + f(x) + \varepsilon, \quad (3.1)$$

where  $E(y | z, x) = z\beta + f(x)$ ,  $\sigma_\varepsilon^2 = \text{Var}[y | z, x]$ . The function  $f$  is not known to lie in a particular parametric family. An early and important application of this model was that of Robert Engle, Clive Granger,

John Rice, and Andrew Weiss (1986), who used it to study the impact of weather on electricity demand.

Peter Robinson's (1988) influential paper demonstrates that the parameter  $\beta$  can be estimated at parametric rates, that is,  $\hat{\beta} - \beta = O_p(T^{-1/2})$ , despite the presence of the nonparametric function  $f$ . Specifically, Robinson rewrites (3.1) conditioning on  $x$ :

$$y - E(y | x) = y - E(z | x)\beta - f(x) = [z - E(z | x)]\beta + \varepsilon. \quad (3.2)$$

If  $E(y | x)$  and  $E(z | x)$  are known, then ordinary least squares on (3.2) yields an estimate of  $\beta$  which is asymptotically normal with variance  $\sigma_\varepsilon^2/T\sigma_u^2$  where  $\sigma_u^2$  is the variance of the  $z$  conditional on  $x$ . Of course, the regression functions  $E(y | x)$  and  $E(z | x)$  are generally not even known to have particular parametric forms. Robinson then produces nonparametric (kernel) estimators of  $E(y | x)$  and  $E(z | x)$  that converge sufficiently quickly so that their substitution in the OLS estimator does not affect its asymptotic distribution.<sup>23</sup>

**3.1.2 Nonparametric Least Squares.** Returning to (3.1), consider the conditional distribution of  $y, z | x$  where all variables are scalars:

$$\begin{aligned} z &= E(z | x) + u = g(x) + u \\ y &= E(y | x) + v = h(x) + v \\ &= (g(x)\beta + f(x)) + (u\beta + \varepsilon) \end{aligned} \quad (3.3)$$

To simplify exposition we assume both conditional models are homoscedastic so that

$$Cov \begin{pmatrix} u \\ v \end{pmatrix} \equiv \begin{bmatrix} \sigma_u^2 & \sigma_{uv} \\ \dots & \sigma_v^2 \end{bmatrix} = \begin{bmatrix} \sigma_u^2 & \sigma_u^2\beta \\ \dots & \sigma_u^2\beta^2 + \sigma_\varepsilon^2 \end{bmatrix}. \quad (3.4)$$

Under sufficient smoothness assumptions, the nonparametric least squares

<sup>23</sup> For general results of this nature, see Whitney Newey (1994). See also Linton (1995b) who analyzes higher order properties of  $\hat{\beta}$ .

estimator (2.15) can be applied equation by equation. The sample variances  $s_u^2 = \sum \hat{u}_t^2/T$ ,  $s_v^2 = \sum \hat{v}_t^2/T$  are  $T^{1/2}$  consistent asymptotically normal estimators of the corresponding population variances  $\sigma_u^2, \sigma_v^2$  (using equation [2.18]). It can also be demonstrated that  $s_{uv} = \sum \hat{u}_t \hat{v}_t/T$  is a  $T^{1/2}$  consistent asymptotically normal estimator of  $\sigma_{uv}$ . In summary, the *sample* analogue to (3.4), that is, the matrix of estimated variances and covariances, is  $T^{1/2}$  consistent asymptotically normal.

Now  $\beta = \sigma_{uv}/\sigma_u^2$ , so that it is fairly straightforward to show that *its* sample analogue  $\hat{\beta} = s_{uv}/s_u^2$  is also  $T^{1/2}$  consistent asymptotically normal. Furthermore, its variance is given by  $\sigma_\varepsilon^2/T\sigma_u^2$ , the same variance attained by the Robinson estimator.

Thus, inference may be conducted using  $\hat{\beta} \sim N(\beta, \sigma_\varepsilon^2/T\sigma_u^2)$ . Alternatively, the bootstrap may be used to obtain standard errors and critical values (see Mammen and Van de Geer 1995, and Yatchew and Bos 1997).<sup>24</sup>

**3.1.3 The Differencing Estimator Revisited<sup>25</sup>.** In the introduction to this paper, we outlined an estimator of  $\beta$  which involves reordering the data so that the  $x$ 's are in increasing order, then differencing to remove the nonparametric effect. The estimator is given by:

$$\hat{\beta}_{diff} = \frac{\sum (y_t - y_{t-1})(z_t - z_{t-1})}{\sum (z_t - z_{t-1})^2}. \quad (3.5)$$

<sup>24</sup> We note that one can perform semiparametric least squares on the model (3.1) by minimizing the sum of squared residuals with respect to  $\beta$  and  $f$  subject to a smoothness constraint on  $f$ , but the resulting estimator of  $\beta$  would not in general converge at  $T^{-1/2}$ . See Hung Chen (1988).

<sup>25</sup> The idea of differencing to remove a nonparametric effect in pure nonparametric models has been used by John Rice (1984), Yatchew (1988), Peter Hall, J.W. Kay, and D.M. Titterington (1990), and others. Powell (1987) and Hyungtaik Ahn and Powell (1993), among others, use the idea in the partial linear model.

We asserted that  $\hat{\beta}_{diff} \sim N(\beta, 1.5\sigma_\varepsilon^2/T\sigma_u^2)$ , which has 66.7 percent (1/1.5) efficiency relative to the efficient estimator (for example, Robinson 1988). In this section we sketch out the main idea of the proof of this result and provide an estimator based on higher order differencing which is asymptotically efficient.

Using (3.3) to substitute for  $y_t$  and  $z_t$  in  $\hat{\beta}_{diff}$  and then rearranging terms we have

$$\begin{aligned} T^{1/2}(\hat{\beta}_{diff} - \beta) &= T^{1/2} \left[ \frac{1}{T} \sum (f(x_t) - f(x_{t-1}) + \varepsilon_t - \varepsilon_{t-1}) \right. \\ &\quad \left. \cdot (g(x_t) - g(x_{t-1}) + u_t - u_{t-1}) \right] \\ &\quad / \frac{1}{T} \sum (g(x_t) - g(x_{t-1}) + u_t - u_{t-1})^2 \end{aligned} \quad (3.6)$$

For simplicity, consider the case where the  $x_t$  are equally spaced on the unit interval. We assume that first derivatives of  $f$  and  $g$  are bounded by  $L$ . Then:

$$\begin{aligned} T^{1/2} \left| \frac{1}{T} \sum (f(x_t) - f(x_{t-1})) (g(x_t) - g(x_{t-1})) \right| \\ \leq \frac{T^{1/2}}{T} \sum L^2 |x_t - x_{t-1}|^2 = \frac{L^2}{T^{3/2}} \end{aligned} \quad (3.7)$$

and

$$\begin{aligned} \text{Var} \left[ \frac{T^{1/2}}{T} \sum u_t (f(x_t) - f(x_{t-1})) \right] \\ = \frac{\sigma_u^2}{T} \sum (f(x_t) - f(x_{t-1}))^2 \leq \frac{L^2 \sigma_u^2}{T^2}. \end{aligned} \quad (3.8)$$

Thus (3.7) and (3.8) converge to 0. Using similar arguments one can show that all terms in (3.6) which involve  $(f(x_t) - f(x_{t-1}))$  or  $(g(x_t) - g(x_{t-1}))$  converge to 0 sufficiently quickly so that (3.6) is approximately

$$T^{1/2}(\hat{\beta}_{diff} - \beta) \cong \frac{T^{1/2} \frac{1}{T} \sum (\varepsilon_t - \varepsilon_{t-1})(u_t - u_{t-1})}{\frac{1}{T} \sum (u_t - u_{t-1})^2}. \quad (3.9)$$

The denominator converges to  $2\sigma_u^2$  and the numerator has mean zero and vari-

ance  $6\sigma_\varepsilon^2\sigma_u^2$ . Thus, the ratio is asymptotically normal with mean zero and variance given by  $6\sigma_\varepsilon^2\sigma_u^2/(2\sigma_u^2)^2 = 1.5\sigma_\varepsilon^2/\sigma_u^2$ .

We now have an explanation for the 1.5 factor in the introduction to the paper (which comes as no surprise to time series practitioners). By differencing we have introduced an MA(1) process in the residuals. Applying ordinary least squares to the differenced data is, as a result, less efficient. (But beware, do not attempt to apply generalized least squares for then you will reverse the beneficial effects of differencing.) Thus the simplicity of differencing is purchased at the price of some lost efficiency relative to the kernel-based procedures or the nonparametric least squares procedures above.

Fortunately, efficiency can be improved substantially by using higher order differences (see Yatchew 1997). Fix the order of differencing  $m$ . Consider now the following generalization of the estimator in (3.5):

$$\hat{\beta}_{diff} = \frac{\sum_t \left( \sum_{j=0}^m d_j y_{t-j} \right) \left( \sum_{j=0}^m d_j z_{t-j} \right)}{\sum \left( \sum_{j=0}^m d_j z_{t-j} \right)^2}, \quad (3.10)$$

where  $d_0, \dots, d_m$  are differencing weights satisfying the conditions  $\sum_0^m d_i = 0$  and  $\sum_0^m d_i^2 = 1$ . The first condition ensures that differencing removes the nonparametric effect in large samples. The second ensures that the variance of the residual in the transformed model is the same as in the original model. (Thus, the weights for the simplest differencing estimator (3.5) which were  $d_0 = 1, d_1 = -1$ , would be normalized to  $d_0 = .7071, d_1 = -.7071$ .)

TABLE 1  
OPTIMAL DIFFERENCING WEIGHTS

1	(0.7071, -0.7071)
2	(0.8090, -0.5000, -0.3090)
3	(0.8582, -0.3832, -0.2809, -0.1942)
4	(0.8873, -0.3099, -0.2464, -0.1901, -0.1409)
5	(0.9064, -0.2600, -0.2167, -0.1774, -0.1420, -0.1103)
6	(0.9200, -0.2238, -0.1925, -0.1635, -0.1369, -0.1126, -0.0906)
7	(0.9302, -0.1965, -0.1728, -0.1506, -0.1299, -0.1107, -0.0930, 0.0768)
8	(0.9380, -0.1751, -0.1565, -0.1389, -0.1224, -0.1069, -0.0925, -0.0791, -0.0666)
9	(0.9443, -0.1578, -0.1429, -0.1287, -0.1152, -0.1025, -0.0905, -0.0792, -0.0687, -0.0588)
10	(0.9494, -0.1437, -0.1314, -0.1197, -0.1085, -0.0978, -0.0877, -0.0782, -0.0691, -0.0606, -0.0527)

In contrast to those in Hall et al (1990), the above optimal weight sequences decline in absolute value towards zero.

If the weights  $d_0, \dots, d_m$  are chosen optimally, it can be shown that

$$\hat{\beta}_{diff} \sim N\left(\beta, \frac{1}{T} \left(1 + \frac{1}{2m}\right) \frac{\sigma_\epsilon^2}{\sigma_u^2}\right). \quad (3.11)$$

By increasing the order of differencing from 1 to 2 to 3, the efficiency of the estimator relative to the Robinson procedure improves from 66.7 percent (=1/1.5) to 80 percent (=1/1.25) to 85.7 percent (=1/1.167). Optimal differencing weights do not have analytic expressions but are tabulated (up to  $m = 10$ ) in Table 1.

Suppose  $z$  the parametric variable is a vector, but  $x$  the nonparametric variable continues to be a scalar. Fix the order of differencing  $m$ . Select the differencing weights  $d_0, \dots, d_m$  optimally (as indicated above). Define  $\Delta y$  to be the  $(T-m) \times 1$  vector whose elements are  $[\Delta y]_t = \sum_{j=0}^m d_j y_{t-j}$  and  $\Delta Z$  to be the  $(T-m) \times p$  matrix with entries  $[\Delta Z]_{ti} = \sum_{j=0}^m d_j z_{t-j,i}$ . Then,

$$\hat{\beta}_{diff} = [\Delta Z' \Delta Z]^{-1} \Delta Z' \Delta y$$

$$\sim N\left(\beta, \left(1 + \frac{1}{2m}\right) \frac{\sigma_\epsilon^2}{T} \Sigma_u^{-1}\right). \quad (3.12)$$

Define the following:

$$s_{diff}^2 = \frac{1}{T} \sum (\Delta y_t - \Delta z_t \hat{\beta}_{diff})^2 \quad (3.13)$$

$$\hat{\Sigma}_{u,diff} = \frac{1}{T} \Delta Z' \Delta Z. \quad (3.14)$$

Then  $s_{diff}^2$  converges to  $\sigma_\epsilon^2$  and  $\hat{\Sigma}_{u,diff}$  converges to  $\Sigma_u$ , the conditional covariance matrix of  $z$  given  $x$ . If  $x$  is a vector, then re-ordering so that  $x_t$  and  $x_{t-1}$  are close is not unique. Nevertheless, for reasonable ordering rules, the procedure works as long as the dimension of  $x$  does not exceed 3.

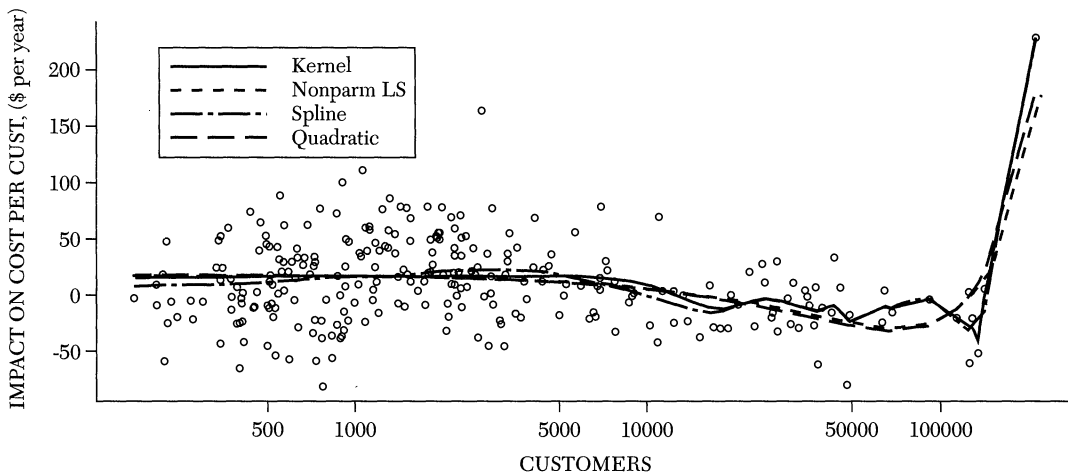
3.1.4 *Examples.* We have already introduced an example on Returns to Scale in Electricity Distribution (Figure 1). We extend it here to include second order optimal differencing and Hal White's (1985) heteroscedasticity consistent standard errors (Figure 10).

As an additional example with a two dimensional nonparametric variable  $x$ , we consider Hedonic Prices of Housing Attributes (Figure 11). Parametric variables include lot size, area of living space and presence of various amenities. The location effect, which has no natural parametric specification, is incorporated nonparametrically.

Figure 10. Returns to Scale in Electricity Distribution (Continued)

		ESTIMATED MODELS							
PARAMETRIC: Quadratic scale effect $y = \alpha + z\beta + \gamma_1x + \gamma_2x^2 + \epsilon$			SEMIPARAMETRIC: Smooth scale effect $y = \alpha + z\beta + f(x) + \epsilon$						
OLS			First Order Differencing <sup>2</sup>			Second Order Optimal Differencing <sup>3</sup>			
	Coeff	SE		Coeff	SE	HCSE		Coeff	SE
$\alpha$	15.987	37.002							
DIST	3.719	1.248	$\Delta$ DIST	2.568	1.560	1.890	$\Delta$ DIST	3.998	1.377
LOAD	1.920	.661	$\Delta$ LOAD	.437	.912	.854	$\Delta$ LOAD	.562	.814
GEN	-.051	.023	$\Delta$ GEN	.0005	.032	.030	$\Delta$ GEN	-.004	.028
LIFE	-5.663	.798	$\Delta$ LIFE	-4.470	1.072	.899	$\Delta$ LIFE	-5.107	.955
ASSETS	.037	.005	$\Delta$ ASSETS	.030	.0072	.006	$\Delta$ ASSETS	.030	.0064
WAGE	.003	.0007	$\Delta$ WAGE	.003	.00086	.00074	$\Delta$ WAGE	.003	.00078
$\gamma_1$	-.00154	.00024							
$\gamma_2$	.1x10 <sup>-7</sup>	.14x10 <sup>-8</sup>							
R <sup>2</sup>	.45								
s <sup>2</sup> <sub>res</sub>	1259.75 <sup>1</sup>		s <sup>2</sup> <sub>diff</sub>	1249.7			s <sup>2</sup> <sub>diff</sub>	1255.4	

ESTIMATED IMPACT OF SCALE ON DOLLAR COSTS PER CUSTOMER PER YEAR<sup>4</sup>



<sup>1</sup> Under the null hypothesis that scale has no effect ( $f$  is constant),  $s_{res}^2 = 1508.5$ . <sup>2</sup>Data reordered so that  $x_1 \leq \dots \leq x_T$ , ( $T=265$ ). Then all variables differenced  $\Delta w_t = w_t - w_{t-1}$ , followed by ordinary least squares. SE's obtained by multiplying OLS standard errors by  $\sqrt{1.5}$ , (set  $m=1$  in equation [3.12]). HCSE's are heteroscedasticity consistent standard errors. Dependent variable is  $\Delta$  COST PER CUST. <sup>3</sup>Data reordered so that  $x_1 \leq \dots \leq x_T$ , then all variables differenced  $\Delta w_t = .809w_{t-1} - .5w_{t-2} - .309w_{t-3}$ , followed by ordinary least squares. OLS standard errors multiplied by  $\sqrt{1.25}$  (set  $m = 2$  in equation [3.12]). Dependent variable is  $\Delta$  COST PER CUST. <sup>4</sup>Kernel and spline estimates obtained using *ksmooth* and *smooth.spline* functions in *S-Plus*. Dependent variable is  $y_t - z_t \hat{\beta}_{diff}$  where  $\hat{\beta}_{diff}$  is calculated using first order differencing. Nonparametric least squares applied using fourth order Sobolev norm. The quadratic estimate is from the parametric model above.

### 3.2 Reprise

In this portion of the paper, we have focussed on the partial linear model

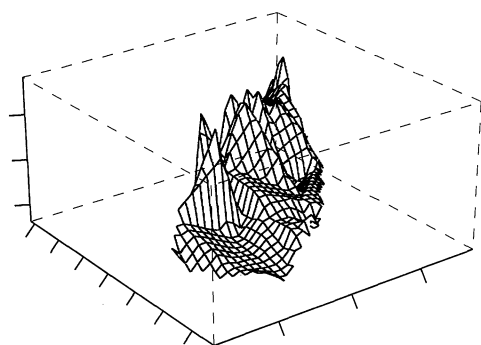
$y = z\beta + f(x) + \epsilon$ . Three distinct estimators of  $\beta$  have been discussed—the first two involve initial nonparametric estimation using either kernel or nonpara-

Figure 11. Hedonic Prices of Housing Attributes

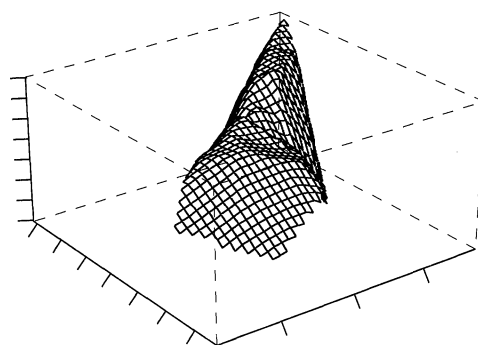
The semiparametric model was estimated by Michael Ho (1995) using semiparametric least squares; the dependent variable  $y$  is SALE PRICE;  $z$  variables in clude lot size (LOTAREA), square footage of housing (USESPC), number of bedrooms (NRBED), average neighborhood income (AVGINC), distance to highway (DHWY), presence of garage (GRGE), fireplace (FRPLC), or luxury appointments (LUX). A critical determinant of price is location which has no natural parametric specification (indeed most urban areas have multi-modal location premia), thus the inclusion of a nonparametric function  $f(x_1, x_2)$  where  $x_1, x_2$  are location coordinates. The data consist of 92 detached homes in the Ottawa area which sold during 1987.

ESTIMATED MODELS									
PARAMETRIC: Linear location effects $y = \alpha + z\beta + \gamma_1x_1 + \gamma_2x_2 + \epsilon$			SEMIPARAMETRIC: Smooth location effect $y = z\beta + f(x_1, x_2) + \epsilon$						
OLS			First Order Differencing <sup>2</sup>			Second Order Optimal Differencing <sup>3</sup>			
	Coeff	S E		Coeff	S E	HCSE		Coeff	S E
$\alpha$	74.0	18.0							
FRPLC	11.7	6.2	$\Delta$ FRPLC	11.3	6.0	5.7	$\Delta$ FRPLC	8.4	6.0
GRGE	11.8	5.1	$\Delta$ GRGE	2.5	5.5	4.1	$\Delta$ GRGE	7.3	5.1
LUX	60.7	10.5	$\Delta$ LUX	55.3	10.8	15.4	$\Delta$ LUX	52.6	10.3
AVGINC	.478	.22	$\Delta$ AVGINC	.152	.35	.22	$\Delta$ AVGINC	.10	.29
DHWY	-15.3	6.7	$\Delta$ DHWY	-5.0	15.9	9.2	$\Delta$ DHWY	-1.4	11.2
LOTAREA	3.2	2.3	$\Delta$ LOTAREA	4.1	2.3	2.0	$\Delta$ LOTAREA	5.6	2.2
NRBED	6.6	4.9	$\Delta$ NRBED	3.3	4.9	3.6	$\Delta$ NRBED	2.5	4.5
USESPC	21.1	11.0	$\Delta$ USESPC	36.5	12.0	9.3	$\Delta$ USESPC	36.5	11.1
$\gamma_1$	7.5	2.2							
$\gamma_2$	-3.2	2.5							
$R^2$	.62								
$s_{res}^2$	424.3 <sup>1</sup>		$s_{diff}^2$	309.2			$s_{diff}^2$	324.8	

DATA WITH PARAMETRIC EFFECT REMOVED



ESTIMATED LOCATION EFFECTS<sup>4</sup>



<sup>1</sup> Under the null that location has no effect, ( $f$  is constant),  $s_{res}^2 = 507.4$ . <sup>2</sup>Data reordered using nearest neighbor algorithm for  $(x_{1t}, x_{2t})$ ,  $t=1, \dots, T$  ( $T=92$ ). Then all variables differenced  $\Delta w_t = w_t - w_{t-1}$ , followed by ordinary least squares. Standard errors multiplied by  $\sqrt{1.5}$ , (set  $m=1$  in equation [3.12]). HCSE's are heteroscedasticity consistent standard errors. Dependent variable is  $\Delta$  SALE PRICE. <sup>3</sup>Data reordered according to nearest neighbor algorithm, then all variables differenced  $\Delta w_t = .809w_t - .5w_{t-1} - .309w_{t-2}$ , followed by ordinary least squares. OLS standard errors multiplied by  $\sqrt{1.25}$  (set  $m = 2$  in equation [3.12]). Dependent variable is  $\Delta$  SALE PRICE. <sup>4</sup>Smoothed estimate obtained using *loess* function in S-Plus. Dependent variable is  $y_t - z_t \beta_{diff}$  where  $\beta_{diff}$  is calculated using first order differencing.



metric least squares methods. The third estimator, which is based on differencing, circumvents this initial step. Each estimator has a variance that shrinks to zero at  $1/T$  (indeed, each is  $T^{1/2}$ -consistent) as well as being asymptotically normal. The relative efficiency of the differencing estimator improves with higher order of differencing, if "optimal" differencing weights are used.

The partial linear model discussed above is part of the much larger class of semiparametric models. The most closely related are partial parametric models,  $y = \mu(z; \beta) + f(x) + \varepsilon$ , where  $\mu$  is a known function but  $f$  and  $\beta$  remain unknown. Each of the techniques described above may be used to obtain  $T^{1/2}$  consistent estimators in partial parametric models.

An important class of semiparametric models is index models which are of the form  $y = f(x\beta) + \varepsilon$ ;  $x$  and  $\beta$  are vectors. Both  $f$  and  $\beta$  are unknown ( $x\beta$  is the 'index', in this case linear);  $T^{1/2}$  consistent estimation of  $\beta$  is also possible (see Hidehiko Ichimura 1993, as well as Roger Klein and Richard Spady 1993). Other important references on semiparametric models include Stoker (1991), Gary Chamberlain (1992), Andrews (1994b), Newey (1994), Linton (1995a), and Joel Horowitz (1996).

#### 4. *Constrained Estimation and Hypothesis Testing*

##### 4.1 *Introduction*

Economic theory rarely dictates a specific functional form. Instead, it typically specifies a collection of potentially related variables and general functional properties of the relationship. For example, economic theory may imply that the impact of a given variable is positive or negative (monotonicity), that doubling of prices and incomes should not alter consumption patterns (homo-

geneity of degree zero), that a proportionate increase in all inputs will increase output by the same proportion (constant returns to scale or, equivalently, homogeneity of degree one), that the effect of one explanatory variable does not depend on the level of another (additive separability), that the relationship possesses certain curvature properties such as concavity or convexity or that observed consumption patterns result from optimization of utility subject to a budget constraint (the maximization hypothesis).

Empirical investigation is then required to assess whether one or another variable is significant or whether a particular property holds. In parametric regression modelling, a functional form is selected and properties are tested by imposing restrictions on the parameters. However, rejection of a hypothesis may be a consequence of the specific functional form that has been selected (but not implied by economic theory). Thus, while the translog production function is richer and more flexible than the Cobb-Douglas, it may not capture all the interesting features of the production process and may indeed lead to incorrect rejection of restrictions. Nonparametric procedures, on the other hand, provide both richer families of functions and more robust tests for assessing the implications of economic theory. Within this framework it is also possible to test whether a specific parametric form is adequate.

In the following sections, we therefore focus on the imposition of additional constraints on nonparametric regression estimation, such as separability and monotonicity, and on testing of hypotheses, particularly specification and significance.<sup>26</sup>

<sup>26</sup>We will not consider tests on the stochastic structure of the residuals, such as heteroscedasticity or autocorrelation.

However, before proceeding, we provide some standardized notation (the ideas are illustrated graphically in Figure 12). Begin with the true model:

$$y = f(x) + \varepsilon. \tag{4.1}$$

We will maintain that  $f$  lies in the set  $\mathfrak{F}$  which is a smooth set of functions. We will want to estimate  $f$  subject to constraints of the form  $f \in \bar{\mathfrak{F}} \subset \mathfrak{F}$  where the set  $\bar{\mathfrak{F}}$  combines smoothness with additional functional properties. We denote the restricted estimator as  $\hat{f}_{res}$  with corresponding estimated residual variance:

$$s_{res}^2 = \frac{1}{T} \sum (y_t - \hat{f}_{res}(x_t))^2. \tag{4.2}$$

Our general null and alternative hypotheses will be of the form:

$$\begin{aligned} H_0: f &\in \bar{\mathfrak{F}} \\ H_1: f &\in \mathfrak{F} \end{aligned} \tag{4.3}$$

We will assume that  $\hat{f}_{res}$  converges to some function  $\bar{f}$  in the restricted set  $\bar{\mathfrak{F}}$ .<sup>27</sup> When the null hypothesis is true,  $\bar{f}$  and  $f$  are identical (since  $f \in \bar{\mathfrak{F}}$ ), and the restricted estimator converges to  $\bar{f} = f$ .

One final important notational convention. Since certain tests will depend on the *difference* between the true regression function and the closest function in  $\bar{\mathfrak{F}}$ , we will reserve special notation for it. In particular:

$$f_{\Delta} = f - \bar{f} \tag{4.4}$$

If the null hypothesis is true,  $f_{\Delta} = 0$ .

### 4.2 Specification Tests

To fix an objective, let us return momentarily to Figure 10, Returns to Scale in Electricity Distribution. Four estimates of the scale effect are illustrated: kernel, nonparametric least squares, spline, and a simple quadratic model. Inspection suggests that the quadratic

provides a fit strikingly similar to all three nonparametric estimates (indeed, it tracks the nonparametric least squares estimate very closely). This is reinforced by the observation that the estimate of the residual variance does not decrease markedly as one moves from the completely parametric model ( $s_{res}^2 = 1259.75$ ) to the differencing estimator ( $s_{diff}^2 = 1249.7$ ). Our objective will be to parlay these casual observations into a formal test of the quadratic specification.<sup>28</sup>

**4.2.1 A Simple Differencing Test of Specification.** To motivate our first and simplest test we return to the idea of differencing, but this time in a purely nonparametric setting. Suppose one is given data  $(y_1, x_1) \dots (y_T, x_T)$  on the model  $y = f(x) + \varepsilon$  where all variables are scalars,  $\varepsilon_t$  are i.i.d. with mean 0, variance  $\sigma_{\varepsilon}^2$ , and independent of  $x$ . The  $x$ 's are drawn from a distribution with support, say the unit interval, and once again we have rearranged the data so that  $x_1 \leq \dots \leq x_T$ . If we first difference to obtain  $y_t - y_{t-1} = f(x_t) - f(x_{t-1}) + \varepsilon_t - \varepsilon_{t-1}$  and define

$$s_{diff}^2 = \frac{1}{2T} \sum (y_t - y_{t-1})^2, \tag{4.5}$$

then as the typical distance between  $x_t$  and  $x_{t-1}$  goes to zero,  $s_{diff}^2$  converges to  $\sigma_{\varepsilon}^2$ . Suppose under the null, the regression function is hypothesized to be quadratic and we obtain an estimate of the variance

$$s_{res}^2 = \frac{1}{T} \sum (y_t - \hat{\gamma}_0 - \hat{\gamma}_1 x_t - \hat{\gamma}_2 x_t^2)^2. \tag{4.6}$$

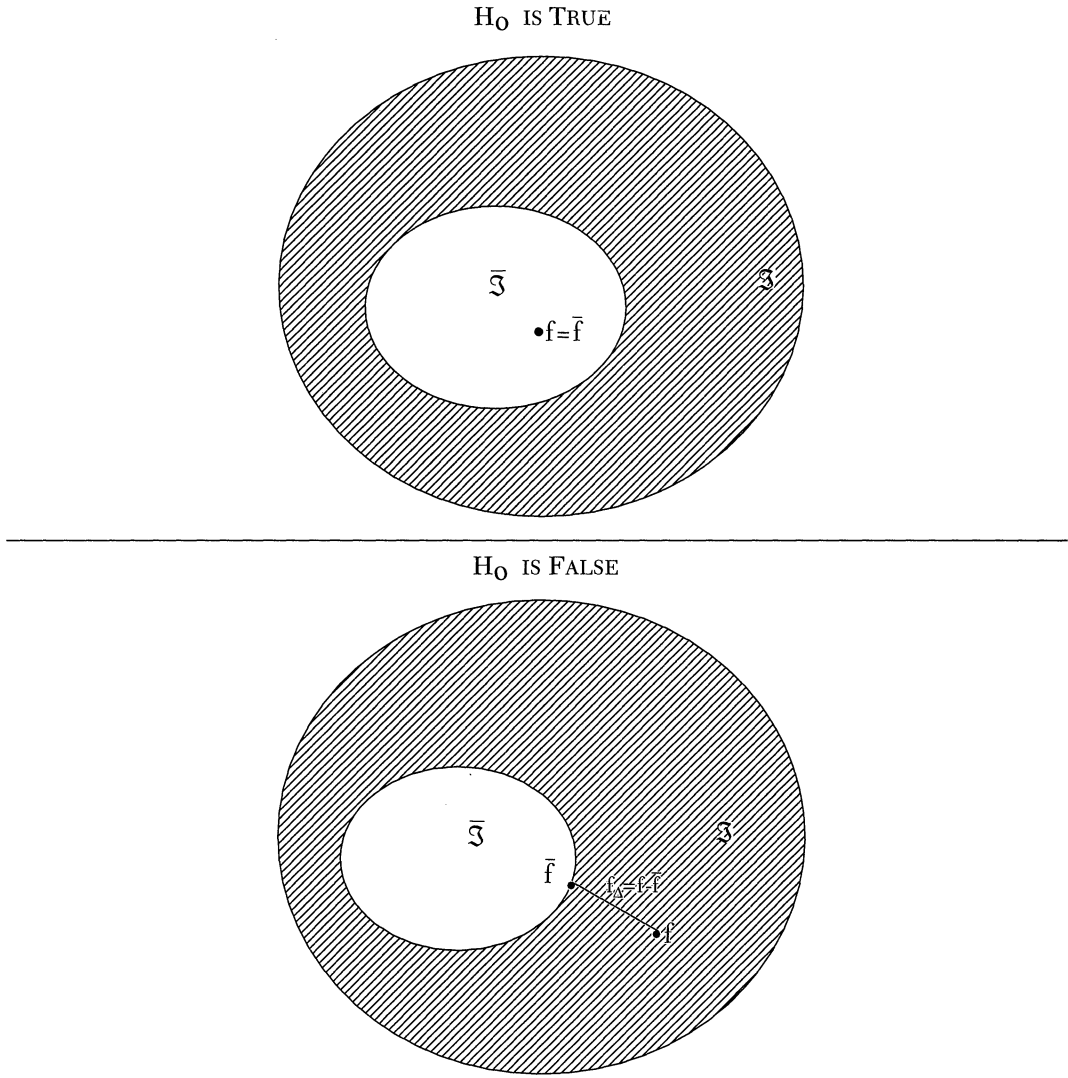
If the null hypothesis is true, then

$$V = \frac{T^{1/2}(s_{res}^2 - s_{diff}^2)}{s_{diff}^2} \sim N(0,1). \tag{4.7}$$

<sup>28</sup> There is a huge literature on specification testing. See James MacKinnon (1992) and White (1994). In this section we focus specifically on tests where the alternative involves a nonparametric component to the regression function.

<sup>27</sup> For example,  $\bar{f}$  could be the closest function in  $\bar{\mathfrak{F}}$  to  $f$  in the sense that  $\bar{f}$  satisfies  $\min_{g \in \bar{\mathfrak{F}}} \int (g - f)^2 dx$ .

Figure 12. Constrained and Unconstrained Estimation and Testing



$\mathfrak{S}$  is the *unrestricted* set of functions,  $\bar{\mathfrak{S}}$  is the *restricted* set of functions. Let  $\bar{f}$  be the closest function in  $\bar{\mathfrak{S}}$  to the true regression function  $f$ . If  $H_0$  is true, then  $f$  lies in  $\bar{\mathfrak{S}}$  and  $f = \bar{f}$ . If  $H_0$  is false, then the difference  $f_\Delta = f - \bar{f} \neq 0$ .

If the null is false,  $V$  grows large.<sup>29</sup> The test procedure may be applied not

<sup>29</sup> Higher order differencing may also be used to obtain  $s_{diff}^2$ . See also Plosser, Schwert, and White (1982) who propose specification tests that are based on differencing but that do not involve reordering of the data.

only in a pure nonparametric setting, but also to the partial linear model  $y = z\beta + f(x) + \varepsilon$ . To test the quadratic specification for the scale effect in electricity distribution costs, obtain  $s_{res}^2$  by regressing  $y$  on  $z$ ,  $x$  and  $x^2$ . Obtain  $s_{diff}^2$

Figure 13. Differencing Test of Specification — Implementation

HYPOTHESES: given data  $(y_1, x_1) \dots (y_T, x_T)$  on the model  $y = f(x) + \epsilon$ , test parametric null against nonparametric alternative. Below, the parametric null is a quadratic function.

$$\text{STATISTIC: } V = T^{1/2} \frac{(s_{res}^2 - s_{diff}^2)}{s_{diff}^2} \sim N(0,1) \text{ under } H_0$$

TEST PROCEDURE

1. Reorder data so that  $x_1 \leq \dots \leq x_T$
2. Calculate  $s_{diff}^2 = \frac{1}{2T} \sum (y_t - y_{t-1})^2$
3. Perform restricted regression of  $y$  on  $x$  and  $x^2$  to obtain  $\hat{\gamma}_0 + \hat{\gamma}_1 x_t + \hat{\gamma}_2 x_t^2$ .
4. Calculate  $s_{res}^2 = \frac{1}{T} \sum (y_t - \hat{\gamma}_0 - \hat{\gamma}_1 x_t - \hat{\gamma}_2 x_t^2)^2$ .
5. Calculate  $V$  and perform one-sided test comparing to critical value from a  $N(0,1)$

Application to Returns to Scale in Electricity Distribution (See Figure 10). The model is  $y = z\beta + f(x) + \epsilon$  and we wish to test a quadratic specification for  $f$ . First, obtain a  $T^{1/2}$  consistent estimator of  $\beta$  such as the differencing estimator of equation (3.12). Calculate  $y_t - z_t \beta$  and apply the above procedures to this newly defined dependent variable. Using S-Plus we obtain  $V = .131$  which supports a quadratic model for the scale effect. To test significance of the nonparametric variable specify a constant function for  $f$ . In this case we obtain  $V = 3.37$  so that the scale variable is significant.

Notes: The derivation of the distribution of  $V$  is straightforward. In large samples,  $s_{diff}^2 \cong \frac{1}{2T} \sum (\epsilon_t - \epsilon_{t-1})^2 \cong \frac{1}{T} \sum \epsilon_t^2 - \frac{1}{T} \sum \epsilon_t \epsilon_{t-1}$  and  $s_{res}^2 \cong \frac{1}{T} \sum \epsilon_t^2$ . Hence, the numerator of  $V$  is approximately  $T^{1/2} \frac{1}{T} \sum \epsilon_t \epsilon_{t-1}$  which using a central limit theorem is  $N(0, \sigma_\epsilon^4)$ . The denominator converges to  $\sigma_\epsilon^2$ . Test may be generalized using higher order differencing to obtain  $s_{diff}^2$ .

by applying equation (3.13). The value of the test statistic is  $V = .131$  in which case the quadratic model appears to be adequate. (See Figure 13 for an implementation summary.)

A test of significance of the nonparametric variable  $x$  is a special case of the above procedure—in this case the null hypothesis is that  $f$  is a constant function. Regressing  $y$  only on  $z$  we obtain  $s_{res}^2 = 1508.5$  in which case  $V = 3.37$  so that scale is a statistically significant factor in the costs of distributing electricity.

The procedure described above may be applied if  $x$  is two- or three-dimensional, but more sophisticated rules for reordering the data so that the  $x$ 's are close, are required. The test is generally not valid if the dimension of  $x$  exceeds 3.

4.2.2 A Conditional Moment Test of Specification. Qi Li (1994) and John Xu Zheng (1996) propose a conditional moment test of specification against a

parametric null. This test may be motivated by writing

$$y - \bar{f}(x) = f(x) - \bar{f}(x) + \epsilon = f_\Delta(x) + \epsilon. \tag{4.8}$$

We assume that the restricted regression estimator  $\hat{f}_{res}$  estimates  $\bar{f}$  consistently and note that if the null hypothesis is true, that is, if  $f \in \bar{\mathfrak{F}}$  then  $f_\Delta = 0$ . Thus, if we do an “auxiliary” regression of the estimated residuals  $y_t - \hat{f}_{res}(x_t)$  on  $x_t$  to estimate  $f_\Delta$  and perform a significance test, then we will have a test of the null hypothesis  $f \in \bar{\mathfrak{F}}$ .<sup>30</sup>

In fact, the moment condition that we will consider involves a slight variation on this idea. Consider the following:

$$E_{\epsilon, x}[(y - \bar{f}(x))f_\Delta(x)p_x(x)] = E_x f_\Delta^2(x)p_x(x) \geq 0 \tag{4.9}$$

<sup>30</sup> This idea, of course, is not new and has been exploited extensively for purposes of specification testing in parametric regression models (see for example, MacKinnon 1992).

where  $p_x(x)$  is the density of  $x$ . Recall that the *numerator* of the kernel estimator in equation (2.2) consistently estimates the product of the regression function and the density of  $x$ . Thus, a sample analogue for the left expression of (4.9) may be obtained by calculating

$$U = \frac{1}{T} \sum_t (y_t - \hat{f}_{res}(x_t)) \left[ \frac{1}{\lambda T \sum_{s \neq t} p^2(x_s)} \sum_{s \neq t} (y_s - \hat{f}_{res}(x_s)) K \left( \frac{x_s - x_t}{\lambda} \right) \right]. \quad (4.10)$$

To interpret this expression, note that the term in square brackets may be thought of as an estimator of  $f_{\Delta}(x_t)p_x(x_t)$ . If the null hypothesis is true, then

$$T\lambda^{1/2}U \sim N(0, 2\sigma_{\varepsilon}^4 \int p^2(x) K^2(u)), \quad (4.11)$$

where  $\sigma_U^2 = Var(U) = 2\sigma_{\varepsilon}^4 \int p^2(x) K^2(u) / \lambda T^2$  may be estimated using

$$\hat{\sigma}_U^2 = \frac{2}{\lambda^2 T^4} \sum_t \sum_{s \neq t} (y_t - \hat{f}_{res}(x_t))^2 (y_s - \hat{f}_{res}(x_s))^2 K^2 \left( \frac{x_s - x_t}{\lambda} \right) \quad (4.12)$$

so that  $U/\hat{\sigma}_U \sim N(0,1)$ . Both  $U$  and  $\hat{\sigma}_U^2$  are relatively straightforward quadratic forms which involve the residuals from the parametric regression.

Returning to our example on Returns to Scale in Electricity Distribution, Figure 14 plots the estimated residuals from the regression of  $y$  on  $z$ ,  $x$  and  $x^2$ . Casual observation suggests that there is no further relationship between  $y$  and  $x$ .

Testing the quadratic specification formally, we obtain  $U/\hat{\sigma}_U = .168$  so that the null hypothesis is not rejected. As with the differencing test, we can perform a nonparametric test of significance of the  $x$  variable by inserting a constant function for  $f$  in the restricted model. In this case,  $U/\hat{\sigma}_U = 2.82$ , indicating significance of the scale variable.

The test generalizes readily to the case where  $x$  is a vector and the bootstrap may be used to obtain critical values.

4.2.3 *Other Specification Tests*. Bierens (1990) considers a moment condition of the form

$$E_{\varepsilon, x}[e^{\tau x}(y - \bar{f}(x))] = E_x[e^{\tau x}(f(x) - \bar{f}(x))] \quad (4.13)$$

which for (almost) any real number  $\tau$  does not equal 0 if the null hypothesis is false. (The expressions equal 0 if the null is true since  $f = \bar{f}$ .) Bierens proposes a test based on a sample analogue of the left expression.

Härdle and Mammen (1993) base their specification test on the integrated squared difference  $I_T = \int (\hat{f}_{res}(x) - \hat{f}_{unr}(x))^2 dx$  where  $\hat{f}_{unr}$ , the unrestricted estimator, is a kernel estimator of  $f$ . The restricted estimator  $\hat{f}_{res}$  is (for technical reasons a smoothed version of) the parametric estimator of  $f$ . In simulations, Härdle and Mammen find that the normal approximation to the distribution of their test statistic is substantially inferior to bootstrapping the critical values of the test statistic. They demonstrate that the “wild” bootstrap (see Appendix B) yields a test procedure that has correct asymptotic size under the null and is consistent under the alternative. (They also demonstrate that conventional bootstrap procedures fail.) The test can be applied to circumstances where  $x$  is a vector and where  $\varepsilon$  is heteroscedastic.

Yongmiao Hong and White (1995) propose tests based on series expansions, in particular the flexible Fourier form (Ronald Gallant 1981). The unrestricted regression model is given by

$$f(x) = \delta_0 + \delta_1 x + \delta_2 x^2 + \sum_{j=1}^{n_f} \gamma_{1j} \cos(jx) + \gamma_{2j} \sin(jx) \quad (4.14)$$

Figure 14. Conditional Moment Test of Specification — Implementation

HYPOTHESES: given data  $(y_1, x_1) \dots (y_T, x_T)$  on the model  $y = f(x) + \varepsilon$ , test parametric null against nonparametric alternative. Below, the parametric null is a quadratic function.

STATISTIC: We implement using the uniform kernel so that  $\int K^2 = 1$ . With only slight abuse of notation, let  $K_{st}$  be the  $s$ -th entry of the kernel matrix defined below.

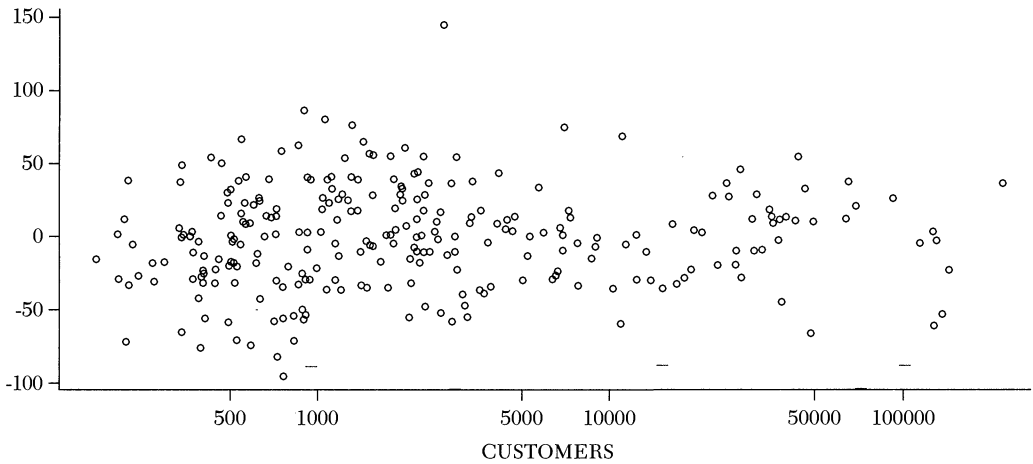
$$U = \frac{1}{T} \sum_t (y_t - \hat{\gamma}_0 - \hat{\gamma}_1 x_t - \hat{\gamma}_2 x_t^2) \left[ \frac{1}{\lambda T} \sum_{s \neq t} (y_s - \hat{\gamma}_0 - \hat{\gamma}_1 x_s - \hat{\gamma}_2 x_s^2) K_{st} \right] \sim N \left( 0, \frac{2\sigma^4 \int p^2(x)}{\lambda T^2} \right)$$

TEST PROCEDURE

1. Perform (restricted) regression  $y$  on  $x$  and  $x^2$  to obtain  $\hat{\gamma}_0 - \hat{\gamma}_1 x_t - \hat{\gamma}_2 x_t^2$ .
2. Calculate the kernel matrix  $K_{st}$  as follows:  
 $K_{st} = 1$  if  $|x_s - x_t| \leq \lambda/2$   $s \neq t$  (note that diagonal elements  $K_{tt} = 0$ )  
 $K_{st} = 0$  otherwise
3. Calculate  $U$
4. Define  $\sigma_U^2 = \text{Var}(U) = 2 \sigma^4 \int p^2(x) / \lambda T^2$  and estimate it using  
 $\hat{\sigma}_U^2 = \frac{2}{T^4 \lambda^2} \sum_t \sum_{s \neq t} (y_t - \hat{\gamma}_0 - \hat{\gamma}_1 x_t - \hat{\gamma}_2 x_t^2) (y_s - \hat{\gamma}_0 - \hat{\gamma}_1 x_s - \hat{\gamma}_2 x_s^2) K_{st}^2$   
 (Note that since we are using the uniform kernel,  $K_{st}^2 = K_{st}$ .)
5. Perform a one sided test comparing  $U/\hat{\sigma}_U$  to the critical value from the  $N(0,1)$ .

APPLICATION TO RETURNS TO SCALE IN ELECTRICITY DISTRIBUTION: (See Figure 10). Under the alternative, the model is  $y = z\beta + f(x) + \varepsilon$  which we wish to test against  $y = z\beta + \gamma_0 + \gamma_1 x + \gamma_2 x^2 + \varepsilon$ . Heuristically, we want to perform a nonparametric regression of the estimated residuals from the latter on  $x$ . Visual examination of these residuals (below) suggests no relationship with  $x$ . Implementing in S-Plus we obtain  $U/\hat{\sigma}_U = .168$  indicating acceptance of a quadratic model for the scale effect. To test significance, specify a constant function for  $f$ . In this case we obtain  $U/\hat{\sigma}_U = 2.82$  so that the scale variable is significant.

ESTIMATED RESIDUALS FROM PARAMETRIC REGRESSION:  $\hat{\varepsilon}_t = y_t - z_t \hat{\beta} - \hat{\gamma}_0 - \hat{\gamma}_1 x_t - \hat{\gamma}_2 x_t^2$



where the number of unknown coefficients  $n_T = 3 + 2n_T^*$  increases with sample size. The rate at which  $n_T$  is permitted to grow depends on the null hypothesis be-

ing tested. The test statistic is based on the difference between  $s_{unr}^2$ , the estimate of the residual variance obtained from (4.14), and  $s_{res}^2$ , which is obtained by esti-

mating the parametric model under the null.

Other procedures not discussed here but worthy of note include those of A. Azzalini, Adrian Bowman and Härdle (1989), Eubank and Clifford Spiegelman (1990), B.J. Lee (1991), Jeffrey Wooldridge (1992), Pedro Gozalo (1993), Yoon-Jae Whang and Andrews (1993), and Horowitz and Härdle (1994).

4.2.4 *Significance Tests*. The simplest nonparametric significance tests are those where the null hypothesis is of the form  $f(x) = \mu$ ;  $\mu$  is a constant. In this case, the null model is entirely parametric and so *all* specification testing methodologies described above immediately yield tests of significance of this kind. If  $x$  is a vector, then this null hypothesis corresponds to testing the *joint* significance of all the explanatory variables.<sup>31</sup>

What is more challenging (and much more useful) is the derivation of tests of significance for a *subset* of the explanatory variables. In this case, the null hypothesis involves a nonparametric component and so the above tests can no longer be directly applied. (For example, suppose one is estimating the model  $y = f(x_1, x_2) + \varepsilon$  and the null hypothesis is that  $y = f(x_1) + \varepsilon$ .) However, the conditional moment test may be extended to this case. (See, for example, Fan and Li 1996.)

### 4.3 Additive Separability

The curse of dimensionality which haunts nonparametric estimation has focussed attention on improving the rate of convergence of nonparametric estimators. Additive models which in the simplest case are of the form  $f(x_a, x_b) = f_a(x_a) + f_b(x_b)$  provide a useful compromise

<sup>31</sup> See Stoker (1989) who, in addition to proposing tests of significance, develops tests of symmetry and homogeneity, and general tests of additive derivative constraints.

between a more rapid rate of convergence and a loss in the richness of the set of functions that can be approximated.<sup>32</sup> For example, if  $x_a$  and  $x_b$  are scalars, then the optimal rate of convergence that a nonparametric estimator can achieve is the same as if  $f$  were a function of only one variable. On the other hand, the additive model does not encompass such relatively common specifications as the multiplicative model  $f(x_a, x_b) = x_a \cdot x_b$ .

An important and useful feature of additive models is the ease with which the effects of different variables may be summarized. Consider, for example, per capita gasoline consumption as a function of per capita disposable income and the price of gasoline. If the two effects are indeed additive, then they may be represented on separate graphs. As may be seen in Figure 15, price has a strong negative effect on demand, while income a strong positive effect. (The data are from Jeffrey Simonoff 1996, Appendix A.)

4.3.1 *A General Estimation Algorithm—Backfitting*. A powerful and general algorithm used to estimate additively separable models is motivated by the observation that

$$E[y - f_a(x_a) | x_b] = f_b(x_b) \\ \text{and } E[y - f_b(x_b) | x_a] = f_a(x_a). \quad (4.15)$$

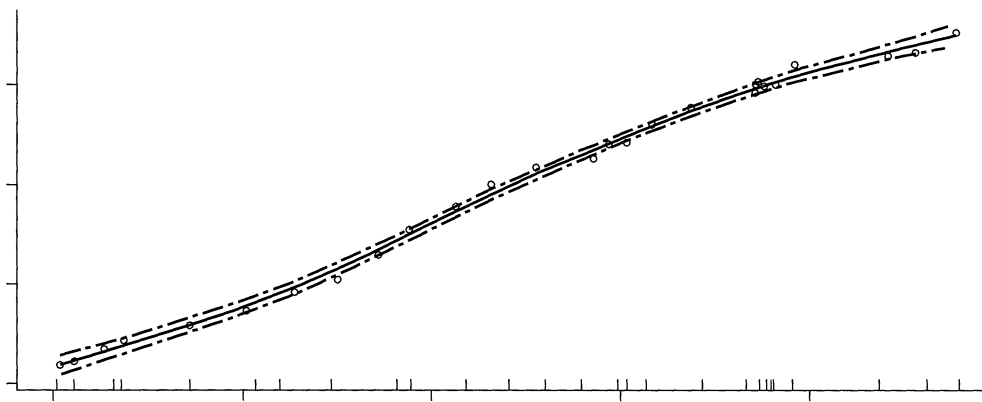
If  $\hat{f}_a$  is a good estimate of  $f_a$  then we may estimate  $f_b$  by nonparametric regression of  $y - \hat{f}_a(x_a)$  on  $x_b$ . (A parallel argument holds for estimation of  $f_a$ .) Beginning with these observations, the algorithm in Table 2 has been widely studied. The initial estimates  $f_a^0, f_b^0$  may be set to zero or to the estimates from a parametric procedure (such as a linear regression).

<sup>32</sup> See for example, Stone (1985, 1986); Andreas Buja, Trevor Hastie, and Robert Tibshirani (1987); and Hastie and Tibshirani (1987, 1990).

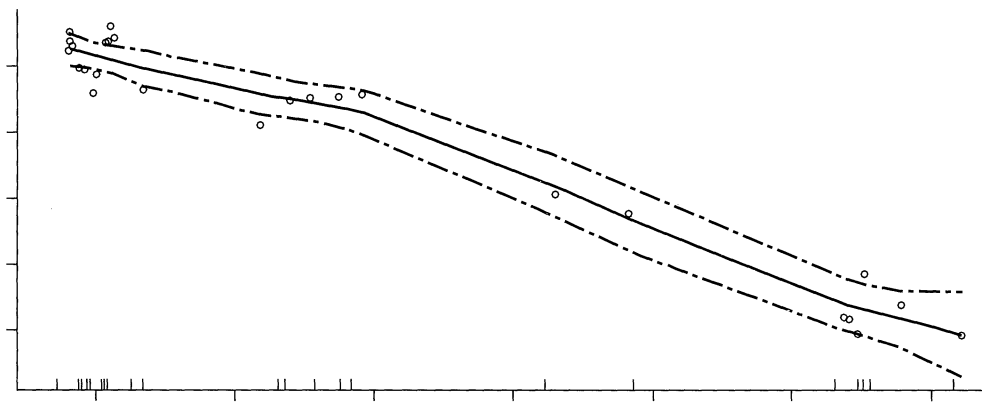
Figure 15. Gasoline Consumption

MODEL: given the model  $y = f_a(x_a) + f_b(x_b) + \epsilon$ , where  $y$  is per capita gasoline consumption,  $x_a$ ,  $x_b$  are per capita income and the real price of gasoline. Data are from Simonoff (1996). The procedure uses the back-fitting algorithm applying spline estimation to each component. Estimation is performed using the *gam* function in S-Plus. Partial residuals are given by  $y_t - \hat{f}_b(x_{bt})$  and  $y_t - \hat{f}_a(x_{at})$  respectively. Bands represent point-wise standard errors.

ESTIMATED EFFECT OF PER CAPITA INCOME ON GASOLINE CONSUMPTION



ESTIMATED EFFECT OF REAL PRICE OF GASOLINE ON CONSUMPTION



The procedure may be generalized in the obvious fashion to additively separable models with more than two additive terms where each term may be a function of several variables. Parametric components may also be included. Assuming that optimal nonparametric estimators are applied to each component, the rate of convergence of the esti-

mated regression function equals the rate of convergence of the component with the largest number of explanatory variables.

In performing the component nonparametric regressions, a variety of techniques may be used, including kernel and spline estimation. Indeed, the algorithm is particularly versatile in that



TABLE 2  
THE BACKFITTING ALGORITHM

Initialization	Select initial estimates $f_a^0, f_b^0$
Iteration	Obtain $\hat{f}_a^i$ by nonparametric regression of $y - \hat{f}_b^{i-1}(x_b)$ on $x_a$ Obtain $\hat{f}_b^i$ by nonparametric regression of $y - \hat{f}_a^{i-1}(x_a)$ on $x_b$
Convergence	Continue iteration until there is little change in individual function estimates

Note: See Hastie and Tibshirani (1990, Ch. 4 and references therein).

different techniques may be selected for different components. For example,  $f_a$  may be estimated using kernel regression and  $f_b$  using nonparametric least squares (or even nonparametric least squares subject to constraints). The algorithm is available in S-Plus using the function *gam* (generalized additive model) which we have applied in Figure 15.

4.3.2 *Further Comments.* It is also possible to estimate the component functions simultaneously using nonparametric least squares or spline procedures. For example, the infinite dimensional optimization problem

$$\min_{f_a, f_b} \frac{1}{T} \sum_t [y_t - f_a(x_{at}) - f_b(x_{bt})]^2$$

$$\text{s.t. } \|f_a + f_b\|_{\mathfrak{S}_{ob}} \leq L \quad (4.16)$$

can be rewritten as a finite dimensional optimization problem with a quadratic objective function and a quadratic constraint. A similar procedure is available if the model is multiplicatively separable, for example,  $f(x_a, x_b) = f_a(x_a) \cdot f_b(x_b)$ . Note that this restriction is useful in imposing homotheticity.<sup>33</sup>

<sup>33</sup> For related work see Daniel Barry (1993), Eubank et al, (1995), and Gozalo and Linton (1997).

#### 4.4 Monotonicity

Economic theory often leads to predictions that a certain effect should be positive or negative. (Indeed, one of the most common complaints expressed by students in econometrics is that despite all their efforts, coefficient estimates are of the wrong sign.) The imposition and testing of inequality constraints on coefficients of parametric models is a well-studied problem (see Frank Wolak 1989 and references therein). The imposition of monotonicity restrictions in nonparametric regression has also been studied extensively. The isotonic regression literature, in the simplest case, considers least squares regression subject only to monotonicity constraints: that is, given data  $(y_1, x_1) \dots (y_T, x_T)$  on the model  $y_t = f(x_t) + \varepsilon_t$ , solve

$$\min_{\hat{y}_1, \dots, \hat{y}_T} \frac{1}{T} \sum_t (y_t - \hat{y}_t)^2$$

$$\hat{y}_s \leq \hat{y}_t \quad \text{for } x_s \leq x_t. \quad (4.17)$$

This literature goes back several decades (see for example, Richard Barlow et al 1972, and Tim Robertson et al 1988).

Monotonicity combined with smoothness assumptions has been studied by Hari Mukarjee (1988), Mammen (1991), and Mukarjee and Steven Stern (1994). See also Ian Wright and Edward Wegman (1980), Florencio Utreras (1984), Wing Wong (1984), Miguel Villalobas and Wahba (1987), John Ramsay (1988), and Steven Goldman and Paul Ruud (1992).

4.4.1 *Why Monotonicity May Not Enhance the Rate of Convergence.* Suppose we are interested in imposing and testing monotonicity *while maintaining smoothness*. That is, the unrestricted set  $\mathfrak{S}$  consists of smooth functions while the restricted set  $\mathfrak{S}$  imposes in addition monotonicity (recall Figure 12). If the

true regression function is strictly monotone, then monotonicity constraints will not improve the rate of convergence.<sup>34</sup>

To see why this is the case, consider the following example in a simplified parametric setting. Suppose we are estimating the model  $y = \mu + \varepsilon$  subject to the constraint  $\mu \leq 2$ . The usual (unconstrained) estimator of  $\mu$  is the sample mean  $\bar{y}$ . An estimator  $\hat{\mu}$  which incorporates the inequality constraint would set  $\hat{\mu} = \bar{y}$  if  $\bar{y} \leq 2$ , and  $\hat{\mu} = 2$  if  $\bar{y} > 2$ . If the true mean is, say, 1.5, then as sample size increases, the probability that the unconstrained estimator equals the constrained estimator goes to one. Thus, the constraint becomes nonbinding.

In nonparametric regression, an analogous result holds. If the true function is strictly monotone (that is, if the first derivative is bounded away from zero), then with sufficient smoothness assumptions, the monotonicity restrictions become nonbinding as sample size increases. (Thus if the first derivative is estimated consistently, as sample size increases, the derivative estimate will also be bounded away from zero with probability going to 1.) The constrained estimator then has the same convergence rate as the unconstrained estimator.<sup>35</sup> This negative finding, however, does not imply that monotonicity will be uninformative in small samples or in the absence of smoothness assumptions. (Nor does it preclude testing for the presence of this property.) Indeed, one could argue that given the paucity of a priori information present in nonparametric estimation, *any* additional constraints should be exploited as far as

possible, particularly in moderately sized samples. (Recall Figure 5 where the imposition of monotonicity results in better fit.)

4.4.2 *Procedures for Combining Smoothness with Monotonicity.* Mammen (1991) analyzes two estimators which combine smoothing with monotonicity constraints in estimation. The first estimator consists of two steps: smoothing of the data by applying a kernel estimator, followed by determination of the closest set of monotonic points to the smoothed points. That is, given data  $(y_1, x_1), \dots, (y_T, x_T)$ , let  $(\tilde{y}_1, x_1), \dots, (\tilde{y}_T, x_T)$  be the set of points obtained by applying a kernel estimator, then solve

$$\min_{\hat{y}_1, \dots, \hat{y}_T} \frac{1}{T} \sum_t (\tilde{y}_t - \hat{y}_t)^2$$

$$\hat{y}_s \leq \hat{y}_t \quad \text{for } x_s \leq x_t. \quad (4.18)$$

The second estimator examined by Mammen reverses the two steps.<sup>36</sup> An alternative approach involves augmenting the nonparametric least squares optimization problem (2.15) or (2.16) with monotonicity constraints.

4.4.3 *Further Comments.* Concavity restrictions can be imposed on nonparametric least squares. If sufficient derivatives are bounded and if the true regression function is strictly concave, then the concavity constraints will not enhance the large sample rate of convergence.

Considerably more complex systems of constraints, such as the inequalities of Sydney Afriat (1967) which embody the implications of demand theory, can also be imposed on nonparametric least squares estimators.

Finally, techniques for estimating sev-

<sup>34</sup> Van de Geer (1990) demonstrates that if one imposes monotonicity only, as in (4.17), then  $\int (\hat{f} - f)^2 = O_p(T^{-2/3}(\log T)^{2/3})$ .

<sup>35</sup> Utreras (1984), and Mammen (1991) find this result for different estimators.

<sup>36</sup> The optimization problem (4.18) may be implemented in GAMS (Brooke et al. 1992). Implementation is also available in XploRe (Härdle et al 1995), using the function *monreg*.

eral nonparametric regressions (which may be "similar in shape") have been developed. Applications include Engel curve modelling and panel data. (See Coenraad Pinkse and Robinson 1995; Ramsay and Bernard Silverman 1997; Yatchew 1998.)

#### 4.5 Reprise

Section 4 of the paper has focussed on constrained estimation and hypothesis testing. Because of the curse of dimensionality and the consequences for convergence rates, it is extremely desirable to improve the accuracy of estimates by validating parametric specifications. Accordingly, we have provided implementation details for two specification tests where the alternative is nonparametric. Reducing the number of explanatory variables or imposing a separable structure also enhances convergence rates. For this purpose, tests of significance and separability are useful.

The discussion of estimation subject to monotonicity constraints underlines one of the advantages of the nonparametric least squares estimator: such constraints can be imposed relatively easily. Indeed, estimation subject to concavity constraints and other implications of economic theory can also be incorporated into nonparametric least squares or the related spline estimation procedures with little difficulty.

As in parametric approaches, a general methodology for testing hypotheses can be based upon an examination of the residuals from the *constrained* regression. If the null hypothesis is true, these residuals should be unrelated to the explanatory variables. Thus, the procedure involves a nonparametric regression of the constrained residuals on all explanatory variables. The resulting test, which can be applied in a wide variety of circumstances, is

based on a conditional moment procedure.<sup>37</sup>

### 5. Extensions And Technical Details

#### 5.1 Modular Analysis of the Partial Linear Model

The watchful reader may have noticed that our applications of the partial linear model  $y = z\beta + f(x) + \varepsilon$  leave some untidy loose ends. Typically our analysis is divided into two components: first we obtain a  $T^{1/2}$ -consistent estimate of  $\beta$  and undertake inference procedures on  $\beta$  as if  $f$  were not present in the model. Then we analyze  $f$  by performing nonparametric estimation and inference on the newly constructed data  $(y_t^*, x_t) = (y_t - z_t\hat{\beta}, x_t)$  as if  $\beta$  were known. Is such a modular approach valid? Separate analysis of the *parametric* portion is justified by virtue of results like (1.3) or (3.12). However, we have to this point not commented on the appropriateness of our analysis of the *nonparametric* part.

Suspicion mounts if one reflects for a moment on the case where  $f$  is parametric; for example,  $y = z\beta + x\delta + \varepsilon$ . Suppose one obtains a  $T^{1/2}$ -consistent estimate of  $\beta$  (say by doing ordinary least squares on the whole model), then one applies ordinary least squares to the constructed data set  $(y_t - z_t\hat{\beta}, x_t)$  to obtain  $\hat{\delta}$ . Assume further that  $(y_t, z_t, x_t)$  are sampled randomly from a joint distribution and that  $x$  and  $z$  are correlated. The estimate of the variance of  $\hat{\delta}$  produced by a standard software package ( $\hat{\sigma}_\varepsilon^2 / \Sigma x_t^2$ ) will be incorrect. To see this, rewrite the estimator as

<sup>37</sup>For related papers on nonparametric estimation and testing with constraints such as additive separability, monotonicity, concavity or demand theory, see also Larry Epstein and Yatchew (1985), Hal Varian (1985, 1990), Rosa Matzkin (1994, and references therein), Jerry Hausman and Newey (1995), and Yatchew and Bos (1997).

$$\begin{aligned} \hat{\delta} &= \frac{\frac{1}{T} \sum (y_t - z_t \hat{\beta}) x_t}{\frac{1}{T} \sum x_t^2} \\ &= \delta + \frac{\frac{1}{T} \sum x_t \varepsilon_t}{\frac{1}{T} \sum x_t^2} + (\beta - \hat{\beta}) \frac{\frac{1}{T} \sum x_t z_t}{\frac{1}{T} \sum x_t^2} \\ &= \delta + \frac{O_P(T^{-1/2})}{O_P(1)} + O_P(T^{-1/2}) \frac{O_P(1)}{O_P(1)}. \end{aligned} \tag{5.1}$$

There are two random terms, each of the same order (the first is due to the variation in  $\varepsilon$ ; the second results from using  $\hat{\beta}$  instead of  $\beta$ ) and if one estimates  $\text{Var}(\hat{\delta})$  using  $\hat{\sigma}_\varepsilon^2 / \sum x_t^2$  then one is incorporating only the first one. Indeed, two-stage estimation in purely *parametric* models often involves adjusting second stage standard errors based on a first-stage procedure.

The partial linear model does not suffer from this additional burden. Essentially, the reason is that the parametric portion of the model converges more quickly than the nonparametric portion so that analysis of the nonparametric portion can proceed as if the parametric portion were known. To see this a little more precisely, suppose we perform a kernel regression of  $y_t - z_t \hat{\beta}$  on  $x_t$  where we assume uniformly distributed  $x$ 's and the uniform kernel. Then, recalling (2.3) we have

$$\begin{aligned} \hat{f}(x_o) &\equiv \frac{1}{\lambda T} \sum_{N(x_o)} y_t - z_t \hat{\beta} \\ &= \frac{1}{\lambda T} \sum_{N(x_o)} f(x_t) + \frac{1}{\lambda T} \sum_{N(x_o)} \varepsilon_t + (\beta - \hat{\beta}) \frac{1}{\lambda T} \sum_{N(x_o)} z_t \\ &\equiv f(x_o) + \frac{1}{2} f''(x_o) \frac{1}{\lambda T} \sum_{N(x_o)} (x_t - x_o)^2 \\ &\quad + \frac{1}{\lambda T} \sum_{N(x_o)} \varepsilon_t + (\beta - \hat{\beta}) \frac{1}{\lambda T} \sum_{N(x_o)} z_t. \end{aligned} \tag{5.2}$$

Each summation will have approximately  $\lambda T$  terms, so that in each case we are calculating a simple average. Recall that the term involving the second derivative corresponds to the bias (see Section 2.1), the next corresponds to the variance term, and the last is the term arising out of the two-stage estimation procedure. Following (2.5),

$$\begin{aligned} \hat{f}(x_o) - f(x_o) &= O(\lambda^2) + O_P((\lambda T)^{-1/2}) + O_P(T^{-1/2}) O_P(1) \\ &= O(T^{-2/5}) + O_P(T^{-2/5}) + O_P(T^{-1/2}) O_P(1) \\ &\quad \text{if } \lambda = O(T^{-1/5}). \end{aligned} \tag{5.3}$$

Consistency of the kernel estimator is unaffected, since all three terms converge to zero (we continue to require  $\lambda \rightarrow 0, T\lambda \rightarrow \infty$ ). The optimal rate of convergence is unaffected— $\lambda = O(T^{-1/5})$  still minimizes the rate at which the (sum of the) three terms converge to zero. The order of each of the first two terms is  $O_P(T^{-2/5})$ , while the third term converges to zero more quickly and independently of  $\lambda$ . Confidence intervals may also be constructed as before (see 2.9), since

$$\begin{aligned} (\lambda T)^{1/2} (\hat{f}(x_o) - f(x_o)) &= O((\lambda T)^{1/2} \lambda^2) + O_P(1) + O_P(\lambda^{1/2}) \\ &= O(1) + O_P(1) + O_P(T^{-1/10}) \text{ if } \lambda = O(T^{-1/5}) \end{aligned} \tag{5.4}$$

and the third term goes to zero, albeit slowly. If the optimal bandwidth  $\lambda = O(T^{-1/5})$  is selected, then confidence intervals must correct for a bias term.

Returning to nonparametric least squares estimation (Section 2.2), if we regress  $y_t^* = y_t - z_t \hat{\beta}$  on  $x_t$ , the estimator  $\hat{f}$  remains consistent and its rate of convergence is unchanged.

The practical point is that we can separate the analysis of the parametric portion of the model from the analysis of the nonparametric portion. Given a

$T^{1/2}$ -consistent estimate of  $\beta$ , we may construct the new dependent variable,  $y_t^* = y_t - z_t \hat{\beta}$ , set aside the original  $y_t$ , and analyze the data  $(y_t^*, x_t)$  as if they came from the pure nonparametric model  $y_t^* = f(x_t) + \varepsilon_t$ . None of the large sample properties that we have discussed will be affected. This holds true regardless of the dimension of the parametric variable  $z$ .

Where does this leave us? Since  $\hat{f}$  depends on  $\hat{\beta}$ , variation in the latter *will* affect variation in the former. The arguments in this section merely state that in large samples, the impact of the variation in  $\hat{\beta}$  is small relative to the impact of the other terms in equation (5.2). If we want to obtain better approximations to our sampling distributions (for example by using the bootstrap), then we may not want to ignore the term associated with prior estimation of  $\hat{\beta}$ , the last term in (5.2) and (5.3).

### 5.2 The Value of Constrained Estimation

Since economic theory rarely provides parametric functional forms, exploratory data analysis and testing which rationalizes a specific parametric regression function is particularly beneficial. In this connection we have outlined two specification tests and referenced a variety of others.

Even though parametric specification is not its forte, economic theory does play a role in producing other valuable restrictions on the regression function. By specifying which variables are potentially relevant to an equation, and excluding myriad others from consideration, rate of convergence is improved. (Exclusion restrictions may come disguised, for example, as homogeneity of degree zero.) The imposition of exclusion restrictions on either local averaging or minimization estimators is straightforward—one simply reduces the dimensionality of the regression func-

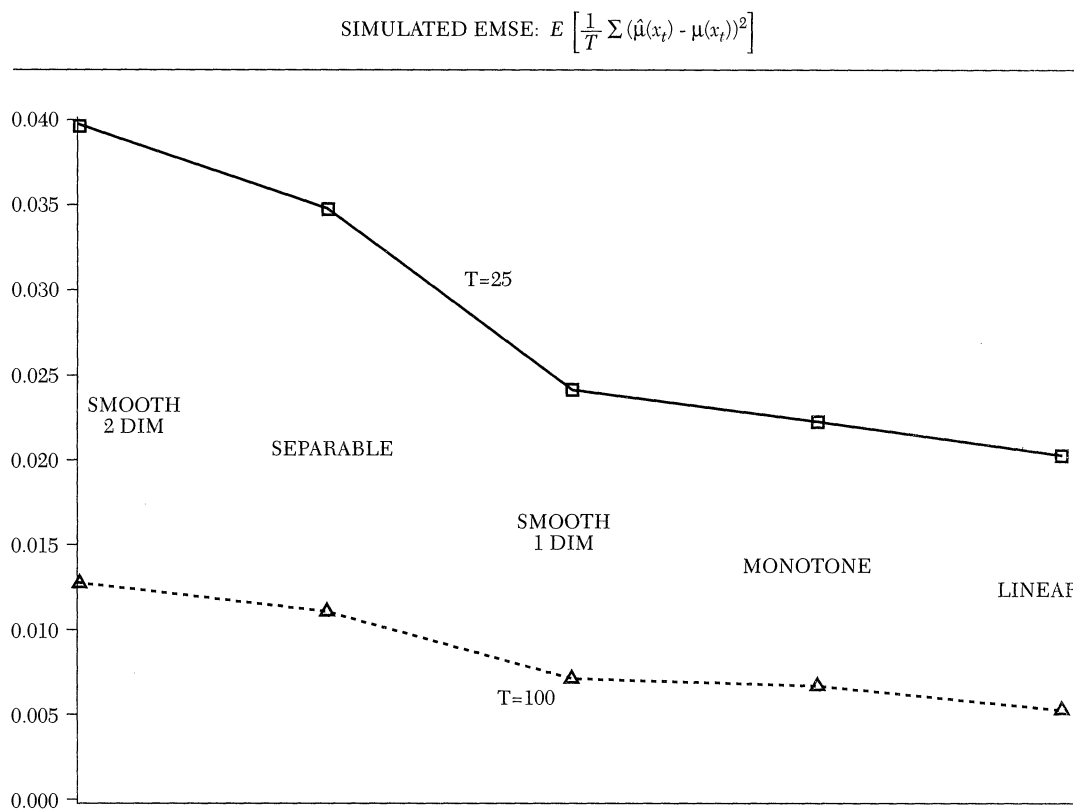
tion. Other restrictions which may be driven by considerations of economic theory and which enhance convergence rates are additive separability and semiparametric modelling. Monotonicity and concavity restrictions do not enhance the (large sample) rate of convergence if sufficient smoothness is imposed, but are beneficial in small samples. Alternatively, their presence can be used to reduce the dependency on smoothness assumptions.

Figure 16 illustrates the consequences of imposing progressively more stringent restrictions on a model which, unbeknownst to the investigator, is linear in one variable. The benefits of learning that the model is a function of one variable rather than two are evident. The expected mean squared error (EMSE), *given fixed sample size*, declines by more than 40 percent as one moves from the smooth two-dimensional to a smooth one-dimensional model. This observation underscores the importance of powerful significance tests for nonparametric models. As expected, separability and partial linearity can also substantially improve the accuracy of the estimator.

### 5.3 Computer Software and Further Reading

Extensive software already in existence can perform many nonparametric procedures automatically. Most of the kernel estimation in this paper was performed in S-Plus, a statistical programming language with extensive graphical capabilities. S-Plus also performs a variety of other nonparametric techniques, including spline and nearest neighbor estimation, and the estimation of additively separable models. Within S-Plus, it is straightforward to implement the differencing estimator, confidence intervals for nonparametric function estimators, and bootstrapping procedures,

Figure 16. Constrained Estimation — Simulated Expected Mean Squared Error



DATA GENERATING MECHANISM:  $y = x_1 + \varepsilon$ ,  $x_1 \in [1,2]$ ,  $\varepsilon \sim N(0, \sigma^2 = .25)$ ,  $\|\cdot\|_{Sob}^2 = \int_1^2 (x_1^2 + 1) dx_1 = 3.33$

ESTIMATED MODELS  $y = \mu(\cdot) + \varepsilon$ ,  $\|\cdot\|_{Sob}^2 \leq 10.0$

SMOOTH - 2 DIM	$\mu(\cdot) = f(x_1, x_2)$
SEPARABLE	$\mu(\cdot) = f_1(x_1) + f_2(x_2)$
SMOOTH - 1 DIM	$\mu(\cdot) = f(x_1)$
MONOTONE	$\mu(\cdot) = f(x_1), f(x_{1\tau}) \leq f(x_{1\tau}), x_{1\tau} \leq x_{1\tau}$
LINEAR	$\mu(\cdot) = \beta_0 + \beta_1 x_1$

Sobolev smoothness norms are of fourth order. In each case, 1000 replications were performed

as well as the tests we have discussed. An excellent reference is Venables and Ripley (1994).<sup>38</sup> Other useful references include Phil Spector (1994) and John Chambers and Hastie (1993). (Complete reference manuals for the

<sup>38</sup> The volume of Härdle (1990), which focusses specifically on nonparametric techniques, contains numerous implementation algorithms.

software are of course available, but this is not the place to start.)

XploRe is a relatively newer package with considerable nonparametric and semiparametric capabilities and good graphical functions. Härdle, Klinke, and Turlach (1995) provide a fine introduction to its capabilities. XploRe performs a variety of nonparametric regression

procedures such as kernel, nearest neighbor, and isotonic regression. The logical learning sequence is to begin with S-Plus prior to attempting to use XploRe.

The nonparametric least squares procedures described in this paper require two steps. First, one must calculate the representor matrices that enter into the optimization problem. (The author uses Fortran for this purpose.) The resulting constrained optimization problems themselves may be solved in GAMS—General Algebraic Modelling System (Brooke, Kendrick and Meerhaus 1992). This software is designed to solve a wide range of linear, nonlinear and integer programming problems and is sufficiently versatile so that bootstrap resampling can be fully automated. It is not specifically designed for statistical use and presently has no graphic capabilities.

We note that the computational burden of bootstrap techniques will not limit their use in nonparametric estimation, since the bootstrap is particularly amenable to parallel computation. Thus, estimation could proceed simultaneously on individual bootstrap samples.

There exist a number of papers and monographs that survey various aspects of the nonparametric/semiparametric literatures. Härdle's (1990) book provides a good introductory treatment of nonparametric regression. Wahba's (1990) monograph is a standard reference on spline estimation. Miguel Delgado and Robinson (1992) provide a valuable survey of semiparametric regression. See also Stoker (1991). Aman Ullah and Hrishikesh Vinod (1993) and Theo Gasser, Joachim Engel and Burkhardt Seifert (1993) provide useful surveys of nonparametric regression in the *Handbook of Statistics* series. Papers by Andrews (1994a), Härdle and Linton (1994), Matzkin (1994), and Powell

(1994) in the *Handbook of Econometrics* series provide fine overviews of various aspects related to nonparametric/semiparametric estimation and testing. See also Linton (1995a) who reviews semiparametric estimation. Fan and Gijbels (1996) provide a fine introduction to local polynomial modelling, a technique which has recently been gaining popularity.

### 6. Conclusions

The averaging of identically distributed objects to estimate the mean of an unknown distribution involves working in a nonparametric setting. So does the use of sample moments to estimate population moments (method of moment estimation), the application of the central limit theorem to do inference on a characteristic of a population, or the construction of a histogram. All these are familiar to most applied economists. Parametric estimation such as ordinary least squares is also widely applied. And we routinely examine residuals, often subjecting them to diagnostic tests or auxiliary regressions to see if anything else is going on that is not being captured by the model.

Nonparametric regression techniques draw upon many of these ideas. When identical objects are not available, one looks at similar objects to draw inferences. Thus kernel estimation averages similarly distributed objects to estimate the collection of conditional means known as the regression function. Nonparametric least squares (which is closely related to spline estimation) minimizes the sum of squared residuals but searches over a richer set of functions (typically a set of smooth functions) than the family of straight lines. Tests on residuals (for example, the conditional moment specification test

we have described) provide a means of assessing a broad range of hypotheses such as whether the sign of the slope of a relationship changes or whether the relationship is additive, concave, or homothetic.

But there are also profound differences between conventional and nonparametric regression techniques. A common question asked in introductory econometrics courses is how one would distribute  $x$ 's on the unit interval if one wanted to learn as much as possible about a linear relationship. The answer is that one wants to maximize the variation of the  $x$ 's and this is done by dividing the  $x$ 's between the end points. The further apart they are, the better.<sup>39</sup> In *parametric* regression, knowing the relationship accurately at a relatively small number of points allows one to infer the parameters and then to use the model equation to interpolate or extrapolate elsewhere. Not so for *nonparametric* regression. Knowing the conditional means at the end points reveals little of what transpires in between. Accurate knowledge of the relationship must be acquired in small neighborhoods covering the whole domain, and so the  $x$ 's are best distributed throughout the unit interval. (In this business, one learns about an individual by looking at close neighbors.)

Estimation problems emerge as the number of explanatory variables increases (the so-called curse of dimensionality). For a fixed number of observations, "neighbors" are much further apart in higher dimensions than in lower. For example, 100  $x$ 's dispersed in the unit cube are more than 20 times farther apart than on the unit interval (Figure 17). If similarity declines proportionately with distance, then averag-

ing over "neighbors" will produce much less accurate estimates in higher dimensions.

In a sense, nonparametric regression is truly empirical in spirit because data must be experienced pretty much everywhere. (The word empiricism descends from a word which means experience.) Interpolation is only deemed reliable among close neighbors, and extrapolation outside the observed domain is considered entirely speculative.

Averaging similar (but not identically distributed) observations introduces bias. One is tempted to increase the number of observations being averaged in order to reduce variance, but this increases bias as progressively less similar observations are added to the mix. The balance is struck when the increase in bias (squared) is offset by the reduction in variance.

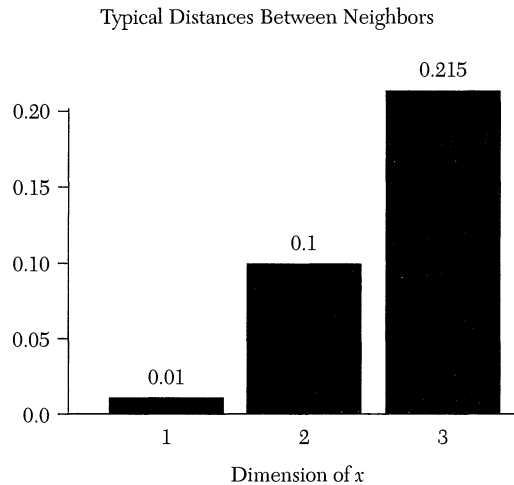
A true sceptic would claim that since economic theory yields little in the way of specific functional forms, all relationships and all variables should be investigated nonparametrically. (Symmetric treatment of various  $x$ 's also has an aesthetic, even egalitarian, appeal.) For applied economists this is neither reasonable nor practical. The potential relevance of numerous variables would conspire with the curse of dimensionality to lead to a state of nihilism—nothing of *significance* could be said about anything. Fortunately, we often have much stronger priors about the influence of some variables than others, and so semiparametric specifications such as the partial linear model provide a promising route. The "curse" also increases the importance of procedures which would confirm a parametric model (specification tests) or reduce the number of explanatory variables (significance tests).

This paper introduces procedures that are relatively simple. In some

<sup>39</sup> For the model  $y = \alpha + x\beta + \varepsilon$ ,  $Var(\hat{\beta}) = \sigma_\varepsilon^2 / \sum (x_i - \bar{x})^2$ .



Figure 17. The Curse of Dimensionality



On the unit interval, 100 equally spaced  $x$ 's will be a distance of only .01 from each other. Allow the  $x$ 's to spread out on the unit square and the distance to the nearest neighbor increases to  $.1(=1/100^{1/2})$ . Distribute them uniformly in the three-dimensional unit cube and the distance increases further to  $.215(=1/100^{1/3})$ .

cases, more powerful but more complex techniques are available. But increased use of even simple nonparametric techniques is part of an endogenous process where greater use will cause software developers to add more of them to econometric packages. This in turn will drive further use and further software development. However, the ultimate test of nonparametric regression techniques resides in their ability to discover new and unusual relationships.

## REFERENCES

- AHN, HYUNGTAIK AND POWELL, JAMES. "Semiparametric Estimation of Censored Selection Models with a Nonparametric Selection Mechanism," *J. Econometrics*, 1993, 58, pp. 2-29.
- AFRIAT, SYDNEY. "The Construction of a Utility Function from Expenditure Data," *Int. Econ. Rev.*, 1967, 8, pp. 66-77.
- ANDREWS, DONALD W.K. "Empirical Process Methods in Econometrics," in *Handbook of econometrics*, Vol. IV. Eds.: R. ENGLE AND D. MCFADDEN. Amsterdam: North Holland, 1994a, pp. 2247-94.
- . "Asymptotics for Semiparametric Econometric Models via Stochastic Equicontinuity," *Econometrica*, 1994b, 62, pp. 43-72.
- AZZALINI A.; BOWMAN, ADRIAN AND HÄRDLE, WOLFGANG. "On the Use of Nonparametric Regression for Model Checking," *Biometrika*, 1989, 76, pp. 1-11.
- BARLOW, RICHARD E.; BARTHOLOMEW, D.J., BREMNER, J.M. AND BRUNK, H.D. *Statistical inference under order restrictions*, New York: John Wiley, 1972.
- BARRY, DANIEL. "Testing for Additivity of a Regression Function," *Annals Statist.*, 1993, 21, pp. 235-54.
- BERAN, RUDY AND DUCHARME, GILLES. *Asymptotic theory for bootstrap methods in statistics*, Centre de Recherches Mathématiques, Université de Montréal, 1991.
- BIERENS, HERMAN. "A Consistent Conditional Moment Test of Functional Form," *Econometrica*, 1990, 58, pp. 1443-58.
- BROOKE, A.; KENDRICK, D. AND MEERAUS, A. GAMS, Redwood City, California: Scientific Press, 1992.
- BUJA, ANDREAS; HASTIE, TREVOR AND TIBSHIRANI, ROBERT. "Linear Smoothers and Additive Models," *Annals Statist.*, 1989, 17, pp. 453-555 (with discussion).
- CHAMBERLAIN, GARY. "Efficiency Bounds for Semiparametric Regression," *Econometrica*, 1992, 60, pp. 576-96.
- CHAMBERS, JOHN M. AND HASTIE, TREVOR. *Statistical models in S*, New York: Chapman and Hall, 1993.
- CHEN HUNG. "Convergence Rates for Parametric Components in a Partly Linear Model," *Annals Statist.*, 1988, 16, pp. 136-46.

- CLARK, R. "A Calibration Curve for Radiocarbon Dates," *Antiquity*, 1975, 49, pp. 251-66.
- DELGADO, MIGUEL AND ROBINSON, PETER M. "Nonparametric and Semiparametric Methods for Economic Research," *J. Econ. Surveys*, 1992, 6(3).
- DUDLEY, RICHARD M. "A Course on Empirical Processes," *Lecture notes in mathematics*, Ecole d'Été de Probabilités de Saint-Flour XII-1982, New York: Springer-Verlag, 1984.
- EFRON, BRADLEY. "Bootstrap Methods: Another Look at the Jackknife," *Annals Statist.*, 1979, 7, pp. 1-26.
- EFRON, BRADLEY AND TIBSHIRANI, ROBERT J. *An introduction to the bootstrap*, New York/London: Chapman & Hall, 1993.
- ENGLE, ROBERT; GRANGER, CLIVE W.J., RICE, JOHN AND WEISS, ANDREW. "Semiparametric Estimates of the Relation between Weather and Electricity Sales," *J. Amer. Statist. Assoc.*, 1986, 81, pp. 310-20.
- EPSTEIN, LARRY AND YATCHEW, ADONIS. "Nonparametric Hypothesis Testing Procedures and Applications to Demand Analysis," *J. Econometrics*, 1985, 30, pp. 150-69.
- EUBANK, RANDALL. *Spline smoothing and nonparametric regression*, New York: Marcel Dekker, 1988.
- EUBANK, RANDALL; HART, JEFFREY D., SIMPSON, D.G. AND STEFANSKI, LEONARD A. "Testing for Additivity in Nonparametric Regression," *Annals Statist.*, 1995, 23, pp. 1896-920.
- EUBANK, RANDALL AND SPECKMAN, PAUL. "Confidence Bands in Nonparametric Regression," *J. Amer. Statist. Assoc.*, 1993, 88, pp. 1287-301.
- EUBANK, RANDALL AND SPIEGELMAN, CLIFFORD H. "Testing the Goodness of Fit of a Linear Model Via Nonparametric Regression Techniques," *J. Amer. Statist. Assoc.*, 1990, 85, pp. 387-92.
- FAN, JIANQING AND GIJBELS, IRENE. *Local polynomial modelling and its applications*, New York/London: Chapman & Hall, 1996.
- FAN YANQIN AND LI, QI. "Consistent Model Specification Tests: Omitted Variables and Semiparametric Functional Forms," *Econometrica*, 1996, 64, pp. 865-90.
- GALLANT, A. RONALD. "Unbiased Determination of Production Technologies," *J. Econometrics*, 1981, 20, pp. 285-323.
- GASSER, THEO; ENGEL, JOACHIM AND SEIFERT, BURKHARDT. "Nonparametric Function Estimation," in *Handbook of statistics*, Vol. 11, Ed.: C.R. RAO. 1993, pp. 423-65.
- GOLDMAN, STEVEN M. AND RUUD, PAUL. "Nonparametric Multivariate Regression Subject to Monotonicity and Convexity Constraints," manuscript, University of California, Berkeley, 1992.
- GOLUB, GENE H. AND VAN LOAN, CHARLES. *Matrix computations*, Baltimore: Johns Hopkins University Press, 1989.
- GOZALO, PEDRO L. "A Consistent Specification Test for Nonparametric Estimation of Regression Function Models," *Econometric Theory*, 1993, 9, pp. 451-77.
- GOZALO PEDRO L. AND LINTON, OLIVER. "Testing Additivity in Generalized Nonparametric Regression Models," manuscript, 1997.
- HALL, PETER. *The bootstrap and edgeworth expansion*, New York: Springer Verlag, 1992.
- . "On Edgeworth Expansions and Bootstrap Confidence Bands in Nonparametric Curve Estimation," *J. Royal Statist. Soc. B*, 1993, 55, pp. 291-304.
- HALL, PETER; KAY, J.W. AND TITTERINGTON, D.M. "Asymptotically Optimal Difference-based Estimation of Variance in Nonparametric Regression," *Biometrika*, 1990, 77, pp. 521-28.
- HÄRDLE, WOLFGANG. *Applied nonparametric regression*, Econometric Society Monograph Series, 19, Cambridge University Press, 1990.
- HÄRDLE WOLFGANG; HALL, PETER AND MARRON, JAMES S. "How Far Are Automatically Chosen Regression Smoothing Parameters from Their Optimum?" *J. Amer. Statist. Assoc.*, 1988, 83, pp. 86-99 (with discussion).
- HÄRDLE, WOLFGANG; KLINKE, SIGBERT AND TURLACH, BERWIN. *XploRe: An interactive statistical computing environment*, New York: Springer-Verlag, 1995.
- HÄRDLE, WOLFGANG AND LINTON, OLIVER. "Applied Nonparametric Methods," in the *Handbook of econometrics*, Vol. IV. Eds.: R. ENGLE AND D. MCFADDEN. Amsterdam: North Holland, 1994, pp. 2297-334.
- HÄRDLE, WOLFGANG AND MAMMEN, ENNO. "Comparing Nonparametric vs Parametric Regression Fits," *Annals Statist.*, 1993, 21, pp. 1926-47.
- HÄRDLE, WOLFGANG AND MARRON, JAMES S. "Optimal Bandwidth Selection in Nonparametric Regression Estimation," *Annals Statist.*, 1985, 13, pp. 1465-81.
- HART, JEFFREY D. *Nonparametric smoothing and lack-of-fit tests*. New York: Springer, 1997.
- HASTIE, TREVOR J. AND TIBSHIRANI, ROBERT. "Generalized Additive Models: Some Applications," *J. Amer. Statist. Assoc.*, 1987, 82, pp. 371-86.
- . *Generalized additive models*, London: Chapman and Hall, 1990.
- HAUSMAN, JERRY AND NEWEY, WHITNEY. "Nonparametric Estimation of Exact Consumer Surplus and Deadweight Loss," *Econometrica*, 1995, 63, pp. 1445-76.
- HO, MICHAEL. *Essays on the housing market*, unpublished Ph.D. dissertation, University of Toronto, 1995.
- HONG, YONGMIAO AND WHITE, HALBERT. "Consistent Specification Testing via Nonparametric Series Regression," *Econometrica*, 1995, 63, pp. 1133-60.
- HOROWITZ, JOEL. "Semiparametric Estimation of a Regression Model with an Unknown Transformation of the Dependent Variable," *Econometrica*, 1996, 64, pp. 103-38.
- HOROWITZ, JOEL AND HÄRDLE, WOLFGANG.

- "Testing a Parametric Model against a Semiparametric Alternative," *Econometric Theory*, 1994, 10, 821-48.
- ICHIMURA, HIDEHIKO. "Semiparametric Least Squares (SLS) and Weighted SLS Estimation of Single-Index Models," *J. Econometrics*, 1993, 58, pp. 71-120.
- KLEIN, ROGER AND SPADY, RICHARD. "An Efficient Semiparametric Estimator for Binary Response Models," *Econometrica*, 1993, 61, pp. 387-422.
- LEE, B.J. "A Model Specification Test against the Nonparametric Alternative," Department of Economics, University of Colorado, manuscript, 1991.
- LEPAGE, RAOUL AND BILLARD, LYNNE. *Exploring the limits of the bootstrap*, New York: John Wiley, 1992.
- LI, KER-CHAU. "Asymptotic Optimality of  $CL$  and Generalized Cross-Validation in Ridge Regression with Application to Spline Smoothing," *Annals Statist.*, 1986, 14, pp. 1101-12.
- . "Asymptotic Optimality for  $C_p$ ,  $CL$ , Cross-Validation and Generalized Cross-Validation: Discrete Index Set," *Annals Statist.*, 1987, 15, pp. 958-75.
- LI, QI. "Some Simple Consistent Tests for a Parametric Regression Function versus Semiparametric or Nonparametric Alternatives," Department of Economics, University of Guelph, manuscript, 1994.
- LINTON, OLIVER. "Estimation in Semiparametric Models: A Review," in *Advances in econometrics and quantitative economics, essays in honor of Professor C.R. Rao*. Eds.: G.S. MADDALA, P.C.B. PHILLIPS, T.N. SRINIVASAN. 1995a, pp. 146-71.
- . "Second Order Approximation in the Partially Linear Regression Model," *Econometrica*, 1995b, 63, pp. 1079-112.
- MACKINNON, JAMES. "Model Specification Tests and Artificial Regressions," *J. Econ. Lit.*, 1992, 30, pp. 102-46.
- MAMMEN, ENNO. "Estimating a Smooth Monotone Regression Function," *Annals Statist.*, 1991, 19, pp. 724-40.
- . *When does bootstrap work?* New York: Springer Verlag, 1992.
- MAMMEN, ENNO AND VAN DE GEER, SARA. "Penalized Quasi-Likelihood Estimation in Partial Linear Models," manuscript, Humboldt-Universität Berlin and University of Leiden, 1995.
- MATZKIN, ROSA L. "Restrictions of Economic Theory in Nonparametric Methods," in *Handbook of econometrics*, Vol. IV. Eds.: R. ENGLE AND D. MCFADDEN. Amsterdam: North Holland, 1994, pp. 2524-59.
- MUKARJEE, HARI. "Monotone Nonparametric Regression," *Annals Statist.*, 1988, 16, pp. 741-50.
- MUKARJEE, HARI AND STERN, STEVEN. "Feasible Nonparametric Estimation of Multi-argument Monotone Functions," *J. Amer. Statist. Assoc.*, 1994, 89, pp. 77-80.
- NADARAYA, E.A. "On Estimating Regression," *Theory Prob. Applic.*, 1964, 10, pp. 186-90.
- NEWWEY, WHITNEY. "The Asymptotic Variance of Semiparametric Estimators," *Econometrica*, 1994, 62, pp. 1349-82.
- PINKSE, COENRAAD AND ROBINSON, PETER M. "Pooling Nonparametric Estimates of Regression Functions with Similar Shape," in *Advances in econometrics and quantitative economics*. Eds.: G.S. MADDALA, P.C.B. PHILLIPS, AND T.N. SRINIVASAN. Oxford: Blackwell, 1995, pp. 172-97.
- PLOSSER, CHARLES G.; SCHWERT, WILLIAM AND WHITE, HALBERT. "Differencing as a Test of Specification," *Int. Econ. Rev.* 1982, 23, pp. 535-52.
- POLLARD, DAVID. *Convergence of stochastic processes*, New York: Springer-Verlag, 1984.
- POWELL, JAMES L. "Semiparametric Estimation of Bivariate Latent Variable Models," working paper 8704, Social Systems Research Institute of Wisconsin, University of Wisconsin, Madison, 1987.
- . "Estimation of Semiparametric Models," in *Handbook of econometrics*, Vol. IV. Eds.: R. ENGLE AND D. MCFADDEN. Amsterdam: North Holland, 1994, pp. 2444-523.
- RAMSAY, JAMES O. "Monotone Regression Splines in Action," *Statist. Sci.* 1988, 3, pp. 425-61.
- RAMSAY, JAMES O. AND SILVERMAN, BERNARD. *Functional data analysis*, New York: Springer, 1997.
- RICE JOHN. "Bandwidth Choice for Nonparametric Regression," *Annals Statist.*, 1984, 12, pp. 1215-30.
- ROBERTSON, TIM; WRIGHT, F.T. AND DYKSTRA, R.L. *Order restricted statistical inference*, New York: John Wiley, 1988.
- ROBINSON, PETER M. "Root-N-Consistent Semiparametric Regression," *Econometrica*, 1988, 56, pp. 931-54.
- SCOTT, DAVID W. *Multivariate density estimation*, New York: John Wiley, 1992.
- SERFLING, ROBERT. *Approximation theorems of mathematical statistics*, New York: John Wiley, 1980.
- SHAO, JUN AND TU, DONGSHENG. *The jackknife and bootstrap*, New York: Springer-Verlag, 1995.
- SIMONOFF, JEFFREY. *Smoothing methods in statistics*, New York: Springer-Verlag, 1996.
- SPECTOR, PHIL. *An introduction to S and S-plus*, Belmont, California: Duxbury Press, 1994.
- STOKER, THOMAS M. "Tests of Additive Derivative Constraints," *Rev. Econ. Stud.* 1989, 56, pp. 535-52.
- . *Lectures on semiparametric econometrics*, CORE Foundation Louvain-La-Neuve, 1991.
- STONE, CHARLES J. "Optimal Rates of Convergence for Nonparametric Estimators," *Annals Statist.* 1980, 8, pp. 1348-60.
- . "Optimal Global Rates of Convergence for Nonparametric Regression," *Annals Statist.*, 1982, 10, pp. 1040-53.
- . "Additive Regression and Other Nonpara-

- metric Models," *Annals Statist.*, 1985, 13, pp. 689–705.
- . "The Dimensionality Reduction Principle for Generalized Additive Models," *Annals of Statistics*, 1986, 14, pp. 590–606.
- ULLAH, AMAN AND VINOD, HRISHIKESH D. "General Nonparametric Regression Estimation and Testing in Econometrics," in *Handbook of statistics*, Vol. 11. Ed.: C.R. RAO. 1993, pp. 85–116.
- UTRERAS, FLORENCIO. "Smoothing Noisy Data under Monotonicity Constraints: Existence, Characterization and Convergence Rates," *Numer. Math.*, 1984, 47, pp. 611–25.
- VAN DE GEER, SARA. "Estimating a Regression Function," *Annals Statist.* 1990, 18, pp. 907–924.
- VARIAN, HAL R. "Nonparametric Analysis of Optimizing Behavior with Measurement Error," *J. Econometrics*, 1985, 30, pp. 445–58.
- . "Goodness of Fit in Optimizing Models," *J. Econometrics*, 1990, 46, pp. 125–40.
- VENABLES, WILLIAM AND RIPLEY, BRIAN. *Modern applied statistics with S-Plus*, New York: Springer-Verlag, 1994.
- VILLALOBOS, MIGUEL AND WAHBA, GRACE. "Inequality-Constrained Multivariate Smoothing Splines with Application to the Estimation of Posterior Probabilities," *J. Am. Statist. Assoc.*, 1987, 82, pp. 239–48.
- WAHBA, GRACE. *Spline models for observational data*, CBMS-NSF Regional Conference Series in Applied Mathematics, #59, Society for Industrial and Applied Mathematics, 1990.
- WAHBA, GRACE AND WOLD, S. "A Completely Automatic French Curve: Fitting Spline Functions by Cross-Validation," *Communications in statistics, Series A*, 1995, 4, pp. 1–17.
- WATSON, G. S. "Smooth Regression Analysis," *Sankhya, Series A*, 1964, 26, pp. 359–72.
- WHANG, YOON-JAE AND ANDREWS, DONALD W.K. "Tests of Specification for Parametric and Semiparametric Models," *J. Econometrics*, 1993, 57, pp. 277–318.
- WHITE, HALBERT. *Asymptotic theory for econometricians*, New York: Academic Press, 1985.
- . *Estimation, inference and specification analysis*, Cambridge University Press, 1994.
- WOLAK, FRANK A. "Testing Inequality Constraints in Linear Econometric Models," *J. Econometrics*, 1989, 41, pp. 205–235.
- WONG, WING HUNG. "On Constrained Multivariate Splines and Their Approximations," *Numer. Math.*, 1984, 43, pp. 141–52.
- WOOLDRIDGE, JEFFREY M. "A Test for Functional Form against Nonparametric Alternatives," *Econometric Theory*, 1992, 8, pp. 452–75.
- WRIGHT, IAN AND WEGMAN, EDWARD. (1980): "Isotonic, Convex and Related Splines," *Annals Statist.*, 1980, 8, pp. 1023–35.
- WU, CHIEN-FU JEFF. "Jackknife, Bootstrap and Other Resampling Methods in Regression Analysis" *Annals Statist.*, 1986, 14, pp. 1261–350 (with discussion).
- YATCHEW, ADONIS. "Some Tests of Nonparametric Regressions Models," in *Dynamic econometric modelling, Proceedings of the third international symposium in economic theory and econometrics*. Eds.: W. BARNETT, E. BERNDT AND H. WHITE. Cambridge University Press, 1988, pp. 121–135.
- . "Nonparametric Regression Model Tests Based on Least Squares," *Econometric Theory*, 1992, 8, pp. 435–451.
- . "An Elementary Estimator of the Partial Linear Model," *Econ. Lett.*, 1997, 57, pp. 135–43.
- . "A Nonparametric Differencing Test of Equality of Regression Functions," manuscript, 1998.
- YATCHEW, ADONIS AND BOS, LEN. "Nonparametric Regression and Testing in Economic Models," *J. Quant. Econ.*, 1997, 13, pp. 81–131.
- ZHENG, JOHN XU. "A Consistent Test of Functional Form via Nonparametric Estimation Techniques," *J. Econometrics*, 1996, 75, pp. 263–89.

## APPENDIX A

### Some Useful Mathematical Background

Suppose  $a_T, T = 1, \dots, \infty$  is a sequence of numbers. Then the sequence  $a_T$  is of smaller order than the sequence  $T^{-r}$ , written  $a_T = o(T^{-r})$  if  $T^r a_T$  converges to zero. For example, if  $a_T = T^{-1/4}$  then  $a_T = o(T^{-1/5})$  since  $T^{1/5} \cdot T^{-1/4} \rightarrow 0$ . A sequence is  $o(1)$  if it converges to 0.

The sequence  $a_T$  is the same order as the sequence  $T^{-r}$ , written  $a_T = O(T^{-r})$  if  $T^r a_T$  is a bounded sequence. For example, the sequence  $a_T = 7 \cdot T^{-1/4} +$

$3 \cdot T^{-1/5} = O(T^{-1/5})$  since  $T^{1/5} a_T$  converges to 3 and hence is a bounded sequence. A sequence is  $O(1)$  if it is bounded.

Now suppose  $a_T, T = 1, \dots, \infty$  is a sequence of random variables. Then,  $a_T = o_p(T^{-r})$  if  $T^r a_T$  converges in probability to zero. For example, let  $a_T = \bar{\epsilon}_T = 1/T \sum_{t=1}^T \epsilon_t$  where  $\epsilon_t$  are independently and identically distributed with mean zero and variance  $\sigma_\epsilon^2$ . Then  $E(\bar{\epsilon}_T) = 0$ ,  $\text{Var}(\bar{\epsilon}_T) = \sigma_\epsilon^2/T$ . Since the mean is 0 and the variance converges to 0, the sequence  $\bar{\epsilon}_T$  converges in probability to 0

and is  $o_P(1)$ . Furthermore, for any  $r < 1/2$ ,  $\bar{\varepsilon}_T = o_P(T^{-r})$  since  $\text{Var}(T^r \bar{\varepsilon}_T) = \sigma_\varepsilon^2 / T^{1-2r}$  converges to 0.

Next, write  $a_T = O_P(T^{-r})$  if  $T^r a_T$  is bounded in probability.<sup>1</sup> For example, suppose  $T^r a_T$  converges to a random variable with finite mean and variance, then  $a_T = O_P(T^{-r})$ . Thus, using the central limit theorem  $T^{1/2} \bar{\varepsilon}_T$  converges to a  $N(0, \sigma_\varepsilon^2)$  in which case  $\bar{\varepsilon}_T = O_P(T^{-1/2})$  and  $T^{1/2} \bar{\varepsilon}_T = O_P(1)$ .

Suppose  $y_t = \mu_y + \varepsilon_t$  where  $\mu_y$  is a constant and define the sample mean based

on  $T$  observations  $\bar{y}_T$ . Then  $\bar{y}_T = O_P(\mu_y + \bar{\varepsilon}_T) = \mu_y + O_P(\bar{\varepsilon}_T) = O(1) + O_P(T^{-1/2})$  and  $T^{1/2}(\bar{y}_T - \mu_y) = T^{1/2} \bar{\varepsilon}_T = O_P(1)$ .

Let  $\lambda_T$  be a sequence of real numbers converging to zero.<sup>2</sup> Typically we consider sequences of the form  $\lambda_T = T^{-r}$  where  $0 < r < 1$ . Let  $a_T = \sum_{t=1}^{\lambda_T T} \varepsilon_t / \lambda_T T$  be the average of the first  $\lambda_T T = T^{1-r}$  values of  $\varepsilon_t$ . For example, if  $T = 100$ ,  $r = 1/5$  then we are averaging the first  $39.8 \cong 40$  observations. Then,  $E[a_T] = 0$  and  $\text{Var}[a_T] = \sigma_\varepsilon^2 / \lambda_T T = \sigma_\varepsilon^2 / T^{1-r}$ . Hence,  $a_T = O_P((\lambda_T T)^{-1/2}) = O_P(T^{-1/2(1-r)})$ .

## APPENDIX B

### *Nonparametric Regression and the Bootstrap*

Bootstrap procedures, which were introduced by Efron (1979), are simulation-based techniques which provide, among other things, estimates of variability, confidence intervals, and critical values for test procedures. In part, because the simulated sampling distributions are constructed from correctly sized samples, bootstrap techniques are often more accurate than those based on asymptotic distribution theory. In many circumstances they are also simpler to implement.

The fundamental idea is to create replications by treating the existing data set (say of size  $T$ ) as a population from which samples (of size  $T$ ) are obtained. In the bootstrap world, sampling from the original data becomes the data-generating mechanism. Variation in estimates results from the fact that upon selection, each data point is replaced within the population. Although the

bootstrap requires resampling many times, calculations need not be done serially, but can be done contemporaneously, making the bootstrap particularly suitable to parallel processing.

The monograph by Efron and Tibshirani (1993) contains a readable introduction to the bootstrap. Beran and Ducharme (1991) provide an approachable treatment of the large sample validity of the bootstrap. Hall (1992) demonstrates that the bootstrap often yields superior finite sample performance than the usual asymptotic distribution theory. Shao and Tu (1995) provide a survey of recent developments.

Suppose we have data  $(y_1, x_1) \dots (y_T, x_T)$  on the model  $y = f(x) + \varepsilon$  where  $f$  may or may not lie in a parametric family. A joint resampling methodology involves drawing i.i.d. observations with replacement from the original collection of ordered pairs.

Residual resampling, on the other hand, proceeds as follows. First,  $f$  is estimated using, say,  $\hat{f}$ . The estimated residuals are assembled and recentered by subtracting off their mean to produce  $\hat{\varepsilon}_t = y_t - \hat{f}(x_t) - \sum_{s=1}^T (y_s - \hat{f}(x_s)) / T$ . One then samples independently from these to construct a bootstrap data set:  $(y_1^B, x_1), \dots, (y_T^B, x_T)$  where  $y_t^B = \hat{f}(x_t) + \hat{\varepsilon}_t^B$ . The "B" su-

<sup>1</sup> That is, for any  $\delta > 0$ , no matter how small, there exists a constant  $A_\delta$  and a point in the sequence  $T_\delta$ , such that for all  $T > T_\delta$ ,  $\text{Prob}[|a_T| > A_\delta] < \delta$ .

<sup>2</sup> Although throughout the paper the bandwidth  $\lambda$  depends on  $T$ , we have suppressed the  $T$  subscript.

perscript signifies a bootstrap observation. Statistics that are of interest are then computed using the bootstrap dataset.

An alternative residual resampling methodology, known as the “wild” bootstrap, is useful in heteroscedastic and in certain nonparametric regression settings. In this case, for each residual  $\hat{\epsilon}_t = y_t - \hat{f}(x_t)$  estimated from the original data one creates a two-point distribution for a random variable, say  $\omega_t$ , with the probability distribution

$\omega_t$	$Prob(\omega_t)$
$\hat{\epsilon} (1 - \sqrt{5})/2$	$(5 + \sqrt{5})/10$
$\hat{\epsilon} (1 + \sqrt{5})/2$	$(5 - \sqrt{5})/10$

The random variable  $\omega_t$  has the properties  $E(\omega_t) = 0$ ,  $E(\omega_t^2) = \hat{\epsilon}_t^2$ ,  $E(\omega_t^3) = \hat{\epsilon}_t^3$ . One then draws from this distribution to obtain  $\hat{\epsilon}_t^B$  for  $t = 1, \dots, T$ . The bootstrap data set  $(y_1^B, x_1), \dots, (y_T^B, x_T)$  is then constructed, where  $y_t^B = \hat{f}(x_t) + \hat{\epsilon}_t^B$ , and statistics of interest are calculated. See Wu (1986) and Härdle (1990, pp. 106–8, 247).

Härdle (1990) discusses various applications of the bootstrap in a nonparametric setting and provides algorithms. See also Hall (1992, pp. 224–34), Hall (1993), Mammen (1992), LePage and Billard (1992), Härdle and Mammen (1993), Li (1995), and Yatchew and Bos (1997).