SMOOTHING BIAS IN THE MEASUREMENT OF MARGINAL EFFECTS

by

Thomas M. Stoker
Massachusetts Institute of Technology

WP#3337-91-EFA August 1991

SMOOTHING BIAS IN THE MEASUREMENT OF MARGINAL EFFECTS

by

Thomas M. Stoker

August 1991

* Sloan School of Management, Massachusetts Institute of Technology, Cambridge, MA, 02139, USA. The author wishes to thank R. Carroll, G. Chamberlain, W. Härdle, J. Heckman, D. Jorgenson, S. Marron, W. Newey, J. Powell and A. Zellner for helpful comments.

Abstract

This paper shows how marginal effects or derivatives estimated nonparametrically can contain systematic downward biases. By considering the behavior of kernel regression estimators with finite bandwidths, we indicate how biases in derivatives arise as a natural feature of local averaging.

We indicate how the kernel regression of y on x will measure the regression of y on x + hu, where h is the bandwidth and the distribution of u is given by the kernel density. Thus, smoothing induces a generic "errors-in-variables" problem with finite h. We characterize the impact of this structure on derivatives, including connecting the derivatives to the density of the regressors.

Various results are given, some involving a normal kernel and normal regressors. Bias values associated with a normal linear model are computed using approximately optimal bandwidth values, and are seen to be quite large. A diagnostic statistic is given, and the role of the nonparametric fitting criterion in derivative bias is discussed.

SMOOTHING BIAS IN THE MEASUREMENT OF MARGINAL EFFECTS

by Thomas M. Stoker

1. Introduction

Applications of econometric models either involve full model simulations or partial calculations based on estimated interrelationships among economic variables. For predictor variables that can be changed incrementally, the latter type of application rests on the estimated values of marginal effects or derivatives, often in the form of elasticities.

This role of econometric modeling has likewise affected the types of models chosen for summarizing empirical relationships. For instance, part of the popularity of the standard linear regression model; $E(y|x) = \alpha + \beta^T x$; arises from its parsimonious summary of the predictor marginal effects or derivatives, namely as the values of the coefficients β. Standard simultaneous equations models begin with linear structural equations, again because of the parsimonious representation of interrelationships through coefficients. Intrinsically linear equations arise in much the same fashions; for instance a linear model predicting $y = \ln(Y)$ by $x = \ln(X)$ has coefficients representing the elasticities of Y with respect to X.

Focus on derivatives has guided the evolution of flexible modeling methods in econometrics. The literature of "flexible functional forms" arose from the recognition of substantial restrictions on marginal effects in the popular linear expenditure system used in analyzing consumer demand and the Cobb-Douglas and CES functions used for characterizing production relationships.[1] In particular, a functional form was defined as "flexible" if it was capable of approximating arbitrary derivative structures for a given value of the predictor variables. While this pointwise requirement does not necessarily allow an entire function to be well approximated, it can provide a

reasonable minimum standard for whether estimation results have been contrived by choice of functional form. The shortcoming of this early approach is that the approximation of derivatives at a point, its main criterion, is not upheld for methods of fitting the "flexible" functional forms to data.[2]

Current research on nonparametric methods in econometric modeling is designed to overcome the restrictiveness of assumed functional forms, appealing to functional approximation theory in large samples. It is worthwhile noting that one of the earliest papers advocating nonparametric methods in econometrics, namely Eldawabi, Gallant and Souza (1983), proposed Fourier series approximation as a method for consistently estimating derivatives at a point. Because of the dramatic recent development of nonparametric methods, it is natural to conclude that the problem of derivative measurement has been solved, with errors in results due exclusively to sampling variation. Whether one fits a truncated version of a series expansion (polynomial or Fourier series, for instance), or estimates a relationship using local averages (kernel or nearest neighbor, for instance), the available theory asserts that the estimates will capture arbitrary nonlinearities, at least with a sufficiently large number of observations. Under the same theoretical guidelines, the same can be said of the derivatives of such relationships, namely that they give a consistent (nonparametric) depiction of the effects of the predictors on the response, at each value of the predictor variables.[3]

Yet every one of these methods still involves a degree of functional approximation in any empirical situation, and especially those of small or moderate sample size. Infinite series expansions must be truncated at some point, and it is easy to conceive of situations where the approximation theory is of doubtful relevance. Taking an extreme case, if a linear model is fit to data, the associated polynomial approximation theory has seemingly little to

offer about the quality of derivative estimates at different points. Methods based on smoothing (local averages), may present more assurance of accuracy, because they are better structured to capture bumps, wiggles and other types of nonlinearity. Yet in any empirical application the amount of smoothing (windows for local averaging) must be set, inducing some approximation error.

The purpose of this paper is to show how the results obtained from smoothing estimators can exhibit substantial, predictable biases in small or moderate data samples. In particular, marginal effects or derivatives of such estimators can be systematically downward biased, resulting in measured effects that are too small. While the existence of bias is no surprise, the fact that it can be in one direction, namely downward, is somewhat striking.[4]

Our analysis is based entirely on standard kernel regression estimators, and while the smoothing bias is intrinsic, other methods of smoothing may not exhibit the downward bias to the same degree. Section 2 begins with the basic framework, and some brief motivation of the size of the bias based on simulations of average derivative estimators. Section 3 opens with a characterization of the nature of smoothing, as inducing a generic nonlinear errors-in-variables structure. We then indicate how this induces downward derivative bias, as well as the role played by the density of the regressors in the bias problem. Discussion and examples based on normally distributed regressors are given next, including a general result on downward derivative bias. Section 4 addresses the role of sample size and dimension in derivative bias, by calculating the bias using optimal bandwidth values for a normal linear model. Section 5 completes the main exposition, by giving a simple diagnostic statistic for derivative bias, and then gives a few remarks on the role of the fitting criterion in nonparametric estimation; in particular we point out how global fitting methods can act to mollify mismeasurement of derivatives, at least on average. Section 6 contains some concluding remarks.

The kernel estimator we consider does permit consistent estimation of regression and its derivatives under the currently popular asymptotic theory, that expresses the proper rate of shrinkage of the smoothing parameter or bandwidth.[5] As such, it is useful to point out initially how our analysis differs from this theory. In order to focus on the impact of smoothing, we employ an asymptotic theory that keeps the bandwidth fixed. In particular, since the kernel estimator is a ratio of averages that have the bandwidth as a parameter, we study the limit of that ratio when the bandwidth parameter is set to the value used in a data sample. As such, our approximation theory treats the averages directly, without promising to shrink the bandwidth as sample size expands. The posture of the paper is that this theory may provide a better distributional approximation in realistically sized samples; however whether this is generally true is an practical issue that merits further study.

2. Motivation of the Bias Problem

2.1 Basic Framework and Estimators

We take the observed data $\{(y_i,x_i),\ i=1,\ldots,N\}$ to be an i.i.d. random sample, where y is a response variable of interest and x is a continuously distributed k-vector. The joint density of (y,x) is denoted F(y,x), and the marginal density of x is denoted f(x).

In the spirit of modeling with additive disturbances, we take the economic relationship of interest to be the mean regression of y on x, namely $g(x) = E(y|x)$. The marginal effects of x on y are the derivatives

$$g'(x) \equiv \frac{\partial g(x)}{\partial x} \tag{2.1}$$

When y and x are in log-form, namely $y = \ln Y$ and $x = \ln X$, then the marginal effects are the elasticities of Y with respect to X.

A parametric econometric analysis would specify a g(x) up to a finite number of parameters that would then be estimated. A nonparametric analysis would measure g(x) directly for each x, resulting in a function estimate $\hat{g}(x)$. In either case, the marginal effects are measured by the derivatives of the function obtained through estimation.

Our analysis of smoothing is based on the standard (Nadaraya-Watson) kernel estimator of the regression g(x), defined as

$$\hat{g}(x) = \frac{\hat{c}(x)}{\hat{f}(x)} \tag{2.2}$$

where the numerator is

$$\hat{c}(x) = N^{-1}h^{-k} \sum_{i=1}^{N} \mathcal{K}\left(\frac{x - x_i}{h}\right) y_i \tag{2.3}$$

and the denominator is

$$\hat{f}(x) = N^{-1}h^{-k} \sum_{i=1}^{N} \mathcal{K}\left(\frac{x - x_i}{h}\right) , \tag{2.4}$$

the standard (Rosenblatt-Parzen) kernel estimator of the marginal density f(x) (c.f. Silverman(1986), Härdle(1991)). Here h denotes the bandwidth value that determines the extent of smoothing or local averaging, and $\mathcal{K}(.)$ is a density function that gives local weights for averaging. The marginal effects of y on x are estimated as the derivatives of $\hat{g}(x)$, given formally as

$$\hat{g}'(x) = \frac{\hat{c}'(x)}{\hat{f}(x)} - \frac{\hat{f}'(x)\hat{c}(x)}{\hat{f}(x)^2} \tag{2.5}$$

The standard large sample theory shows how $\hat{g}(x)$ is a (pointwise) consistent estimator of $g(x)$ under conditions governing how the bandwidth h is shrunk with increases in the sample size N, for instance $h^2 \to 0$ and $Nh^k \to \infty$. Analogous theory demonstrates how $\hat{g}'(x)$ is a pointwise consistent estimator of the marginal effects $g'(x)$. Our discussion is concerned with a different issue, namely how well $\hat{g}'(x)$ estimates $g'(x)$ for given bandwidth value h. As such, our focus is on mismeasurement directly involved with smoothing.

The regularity conditions we require are as follows.

Assumption 2.1: The density $f(x)$ has convex (possibly unbounded) support $S_f \subseteq \mathbb{R}^k$, and $f(x) = 0$ for $x \in \partial S_f$, the boundary of its support. $f(x)$ is twice continuously differentiable on $\text{int}(S_f)$. The density $F(y,x)$ is twice continuously differentiable in x. The mean and variance of (y,x) exists, and $g(x) = E(y|x)$ is continuously differentiable on $\text{int}(S_f)$.

Assumption 2.2: The kernel $\mathcal{K}(u)$ has support $S_{\mathcal{K}} \subseteq \mathbb{R}^k$, with $\mathcal{K}(u) > 0$ for $u \in \text{int}(S_{\mathcal{K}})$ and $\mathcal{K}(u) = 0$ for $u \in \partial S_{\mathcal{K}}$, the boundary of $S_{\mathcal{K}}$. The origin $0 \in S_{\mathcal{K}}$, and if $u \in S_{\mathcal{K}}$ then $-u \in S_{\mathcal{K}}$. $\mathcal{K}(u)$ is symmetric, $\mathcal{K}(u) = \mathcal{K}(-u)$, (with $\int u\mathcal{K}(u)du = 0$) and continuously differentiable on $\text{int}(S_{\mathcal{K}})$.

Assumption 2.3: The integrals $\int\mathcal{K}(u)f(x-hu)du$ and $\int\int\mathcal{K}(u)yF(y,x-hu)dudy$ exist for $x \in S_{\mathcal{K}}$ and are differentiable in x, with derivatives $(\int\mathcal{K}(u)f(x-hu)du)' = \int \mathcal{K}(u)f'(x-hu)du$ and $(\int\int\mathcal{K}(u)yF(y,x-hu)dudy)' = \int\int\mathcal{K}(u)y\ (\partial F/\partial x)(y,x-hu)dudy$.

The last condition is stated in the form in which it is used, and could be replaced by various primitive conditions that assure it (see, for example Ibragimov and Has'minskii(1981)).

2.2 Motivation via Average Derivative Estimators

There is little surprise in the observation that there is bias in the estimators $\hat{g}(x)$ and $\hat{g}'(x)$ for a given value of the bandwidth h. The issue that motivates this paper is that the bias in derivatives can be substantial and systematically one-sided, namely that derivative estimates can be uniformly too small. This character of the bias was first noted by the author in the behavior of certain average derivative estimators.[6] For motivation of the potential size of the problem, we now consider some results of this study.

The connection of average derivatives to our inquiry is immediate: the (unweighted) average derivative δ of y on x is defined as the mean of $g'(x)$ over all x values.[7] We consider two estimators below, which are sample analogs of the following expressions of δ,

$$\delta \equiv E(g') = \{E(\partial x/\partial x)\}^{-1} E(g') \quad . \tag{2.6}$$

The first estimator is just the (trimmed) sample average of the estimated derivatives $\hat{g}'(x_i)$, or

$$\hat{\delta} = N^{-1} \sum_{i=1}^{N} \hat{g}'(x_i)\, \hat{1}_i \tag{2.7}$$

where $\hat{1}_i = 1[\hat{f}(x_i) \geq b]$ is a trimming indicator that drops observations with small estimated density (required for the technical analysis of this estimator). The second is a "slope" estimator, defined as

$$\hat{d} = \left[N^{-1} \sum_{i=1}^{N} [\hat{x}'(x_i)]^T \hat{1}_i\right]^{-1} \left[N^{-1} \sum_{i=1}^{N} \hat{g}'(x_i)\, \hat{1}_i\right] \tag{2.8}$$

where $\hat{x}'$ is the derivative (matrix) of the kernel estimator of $E(x|x)$; namely

(2.2) with y_i replaced by x_i. Stoker(1991a) points out how $\hat{d}$ can be written as an instrumental variables estimator of the slope coefficients from estimating the linear equation $y_i = \hat{c} + x_i'\hat{d} + \hat{u}_i$, which explains the terminology.

Under certain shrinking bandwidth conditions, Stoker(1991a) has shown these estimators to be $\sqrt{N}$-consistent for δ, and asymptotically equivalent. These conditions employ "asymptotic undersmoothing," so that the averages above converge to their limits at a faster rate than the estimator $\hat{g}(x)$ converges to $g(x)$ at any point x.[8] Consequently, in a large sample with tiny bandwidths, the distributional properties of the estimators (2.7), (2.8) will be virtually identical.

However, there is no reason to suspect that these estimators will behave identically in small samples, or with substantive bandwidth values. In particular, we can motivate the current study by illustrating how these two estimators can give quite different measures of average derivatives. First consider a linear model with normal regressors

$$(2.9) \qquad y_i = 1 + x_{1i} + x_{2i} + x_{3i} + x_{4i} + \epsilon_i \; ; \quad i = 1,\ldots,N$$

where the $k = 4$ predictors x_{ji}, and the disturbance ϵ_i are (independent) $N(0,1)$ variables. The sample size is $N = 100$, the kernel is the spherical multivariate normal density $K(u) = \Pi\, k(u_j)$ with $k(u_j) = (1/\sqrt{2\pi})\exp(-u_j^2/2)$, the bandwidth is $h = 1$ and the trimming bound b is set to drop 1% of the observations.[9] The average derivative is the vector of coefficients $\delta = (1,1,1,1)'$. Table 1 contains the means and standard errors of each of the average derivative components over 20 Monte Carlo simulations.

The results on the "average" estimator (2.7) illustrate the scope of the problem addressed in this paper. For concreteness, suppose that y and x represented log-output and log-inputs respectively, with the simulated model

TABLE 1: SIMULATION RESULTS - LINEAR MODEL

True Value: $\delta = (1,1,1,1)$

	$\hat{\delta}_1$	$\hat{\delta}_2$	$\hat{\delta}_3$	$\hat{\delta}_4$
Average	.447	.453	.477	.444
(2.7)	(.063)	(.101)	(.059)	(.076)
Slope	1.01	1.02	1.03	.985
(2.8)	(.098)	(.152)	(.115)	(.116)
OLS	1.01	1.01	1.02	.976
	(.078)	(.128)	(.111)	(.107)

(2.9) a Cobb-Douglas production function. Suppose that a log-production function were estimated by the kernel estimator $\hat{g}(x)$, and the output elasticities by $\hat{g}'(x)$. Table 1 (the "average" row), says that on average, the estimates $\hat{g}'(x)$ are 45% of their true values. The OLS estimators are virtually unbiased as they should be, as are the "slope" estimators (2.8).

This simulation design ought to favor good estimator performance. The predictors are symmetrically distributed, independent and have a symmetric impact on y. The R^2 of the true equation is .80, which is not overwhelmingly small for survey applications in economics. One facet of the results which might be guaranteed is the good performance of the slope estimators, because their instrumental variables formulation implies that they are conditionally unbiased for the true coefficients. With this in mind we present simulations of a binary response model in Table 2; namely where the dependent variable is altered to

$$(2.10) \qquad y_i = 1[\ 1 + x_{1i} + x_{2i} + x_{3i} + x_{4i} + \epsilon_i > 0\ ]\ ; \quad i = 1,\ldots,N$$

Now the true average derivative vector is $\delta = .161\ (1,1,1,1)$. The kernel, bandwidth and trimming parameters are the same as for Table 1.

This table displays exactly the same characteristics. In this case the OLS coefficients (of regressing the discrete y on x) are comparable because of the design; with normally distributed regressors, the OLS coefficients consistently estimate the average derivative δ (c.f. Brillinger(1983) and Stoker(1986)).

As indicated above, the results of Stoker(1991a) state that for very large data sets, with tiny bandwidth (and trimming bound) values, the differences seen in Tables 1 and 2 will disappear. In this context, 100 observations is clearly insufficient to constitute a "very large" sample in this sense. Moreover, the bandwidth value h = 1 was not chosen by an

TABLE 2: SIMULATION RESULTS - BINARY RESPONSE MODEL

True Value: $\delta = (.161, .161, .161, .161)$

	$\hat{\delta}_1$	$\hat{\delta}_2$	$\hat{\delta}_3$	$\hat{\delta}_4$
Average	.082	.083	.083	.076
(2.7)	(.021)	(.023)	(.018)	(.014)
Slope	.186	.186	.178	.168
(2.8)	(.039)	(.046)	(.036)	(.034)
OLS	.171	.171	.168	.160
	(.035)	(.033)	(.035)	(.028)

automatic mechanism,[10] and so one might approach this problem by trying to derive "best" bandwidth values for averaging regression derivatives, to see if these differences can be made small.

Our approach is different, namely to study the limiting behavior of the estimators under fixed bandwidth values. This posture focuses on the impact of smoothing, and gives a clear interpretation of how smoothing estimators can differ from procedures based on (global) least squares or other fitting criterion. In fact, exactly the sort of differences observed above follow from this posture, as we now show.

3. Smoothing Bias in Marginal Effects

3.1 Errors-in-Variables Structure

While it would be most preferable to study the behavior of the kernel regression estimator with a fully developed small sample theory, our analysis is based on large sample approximation. This is due to the fact that our underlying regression structure can be quite general, as well as the nonlinearity in the construction of (2.2). In particular, we focus on approximation with the bandwidth h held fixed, which differs from the currently popular theory that shrinks the bandwidth along with increases in sample size. Our posture treats (2.2) as a composition of simple averages, without the promise of shrinking the bandwidth when more data is obtained.

This analysis is quite simple, and gives a natural interpretation. In particular, we shall see that smoothing induces an "errors-in-variables" problem into the results. The bias in derivatives arises in an analogous fashion to the downward bias of a linear regression coefficient when the regressor is measured with independent error. We now make these connections.

Given the bandwidth h, we denote the limit of the estimator $\hat{g}(x)$ as $\gamma_h(x)$. By applying Slutsky's theorem and the weak law of large numbers, this limit is expressed as;

$$(3.1) \qquad \gamma_h(x) \equiv \operatorname{plim} \hat{g}(x) = \frac{\operatorname{plim} \hat{c}(x)}{\operatorname{plim} \hat{f}(x)} = \frac{E[\hat{c}(x)]}{E[\hat{f}(x)]}$$

and so we need to characterize $E[\hat{c}(x)]$ and $E[\hat{f}(x)]$.

For $E[\hat{c}(x)]$, a standard change of variables gives

$$(3.2) \qquad E[\hat{c}(x)] = \int y \, \Phi_h(y,x) \, dy$$

where

$$(3.3) \qquad \Phi_h(y,x) = \int K(u) \, F(y, x - hu) \, du$$

The function Φ_h is clearly a density, in the form of a convolution. In particular, if (y,x) is distributed independently of u, with density $K(u)$, then $\Phi_h(y,z)$ is the joint density of y and $z = x + hu$. For $E[\hat{f}(x)]$, a similar standard calculation gives

$$(3.4) \qquad E[\hat{f}(x)] = \int K(u) \, f(x-hu)du \equiv \phi_h(x)$$

With x, u as above, $\phi_h(z)$ is easily seen to be the marginal density of $z = x + hu$,[11] and it is easy to verify that $\phi_h(z) = \int \Phi_h(y,z) \, dy$. For later reference, we define the (translation) score λ_h of ϕ_h as

$$(3.5) \qquad \lambda_h(x) = - \frac{\partial \ln \phi_h}{\partial x} = - \frac{\phi_h'}{\phi_h}$$

$$= - \frac{\int K(u) \, f'(x-hu)du}{\int K(u) \, f(x-hu)du}$$

where the latter equality follows from assumption 2.3.

Combining (3.2) and (3.4) gives

$$(3.6) \qquad \gamma_h(x) \equiv \text{plim } \hat{g}(x) = \frac{E[\hat{c}(x)]}{E[\hat{f}(x)]} = \frac{\int y \, \Phi_h(y,x)dy}{\phi_h(x)} \quad .$$

This expression is easy to interpret. Namely, $\gamma_h(z)$ is the regression function $E(y|z)$, with $z = x + hu$. Moreover, given assumption 2.3, an analogous argument to that above gives

$$(3.7) \qquad \text{plim } \hat{g}'(x) = \gamma_h'(x) \quad .$$

To summarize, for fixed bandwidth h, the regression estimator $\hat{g}$ estimates the regression γ_h of y on x + hu, and the derivative $\hat{g}'$ estimates the associated derivatives γ_h'. As such, local smoothing induces an "errors-in-variables" problem, namely by causing the regression of y conditional on x + hu to be measured instead of the regression of y conditional on x.[12] Of course, if h were allowed to vanish, then so would the difference between these two regression functions. However, for finite h, we show how the difference between γ_h and g generates the derivative bias problem.

3.2 Argument Shifting and Smoothing Bias

It is natural to conjecture that this structure induces a downward bias in $\hat{g}'(x)$ as an estimator of $g'(x)$, in analogy with the downward bias (toward zero) imparted to the least squares coefficient of a linear model when the regressor is measured with error. To develop this further, we first recall the basic linear errors-in-variables structure, with a normal regressor. Suppose that there is a single regressor x, and that the true model is

$$(3.8) \qquad y = \alpha + \beta x + \epsilon$$

where ϵ is distributed with mean 0, independently of x. Suppose further that x is distributed normally with mean μ and variance σ^2, and that $K(u)$ is a normal density, so that u is distributed normally with mean 0 and variance 1. With $z = x + hu$, we have that

$$(3.9) \qquad y = \alpha + \beta(z - hu) + \epsilon$$

$$= \alpha + \beta[z - hE(u|z)] + v$$

where $v = \{\epsilon - h[u - E(u|z)]\}$ has mean zero conditional on z. The standard bias analysis follows from

$$(3.10) \qquad \gamma_h(z) = E(y|z) = \alpha + \beta[z - hE(u|z)]$$

$$= [\alpha + \beta\nu_h\mu] + \beta(1-\nu_h)\, z$$

where $\nu_h = h^2/(\sigma^2+h^2)$, the latter equality using that

$$(3.11) \qquad hE(u|z) = \nu_h\,(z - \mu).$$

Consequently, the OLS coefficient estimates $\beta(1-\nu_h)$, or contains a downward bias (in absolute value) as an estimator of the true coefficient β. The term ν_h is the familiar "noise/total variation" ratio for this problem.

Aside from giving us an immediate example of smoothing bias; here

$$(3.12) \qquad g(x) = \alpha + \beta x\ , \qquad \gamma_h(x) = [\alpha + \beta\nu_h\mu] + \beta(1-\nu_h)\, x$$

so that

$$(3.13) \qquad g'(x) = \beta\ , \qquad \gamma_h'(x) = \beta(1-\nu_h)\ ;$$

this example also serves to indicate how the bias arises. In particular, the regression (3.10) is the true regression function with its argument shifted from z to $z - hE(u|z)$, or toward the mean of z, namely

$$(3.14) \qquad \gamma_h(z) = g[z - hE(u|z)] = g[z - \nu_h\,(z - \mu)]$$

where $g(x) = \alpha + \beta x$. This shifting serves to flatten the slope of γ_h relative to that of g. This point is illustrated in Figure 1.

Returning to the general format, we can see that the shift of the argument is one of two factors affecting the structure of γ_h relative to g. To make this explicit, we first assume that the basic behavioral model has an additive error:

Assumption 3.1: The true model for the response y is of the form

(3.15) $$y = g(x) + \epsilon$$

where ϵ is independent of x.

We then have immediately that

(3.16) $$\gamma_h(z) = E[g(z - hu)|z]$$

$$= E(g[z - hE(u|z) - h(u-E(u|z)]|z)$$

Therefore, γ_h could be constructed from g by i) shifting the central argument from z to $z-hE(u|z)$, and ii) averaging g over the departures $h[u-E(u|z)]$. In general, it is difficult to completely characterize these effects with arbitrary functions g(x), f(x) and $K(u)$. For the remainder of this section, we focus on i), the "argument-shift". The alteration ii) is difficult to characterize generally (aside from examples), although we do present a result for general regression in the next section, where the regressors are assumed to be normally distributed.

The argument-shift depends solely on the distribution of (z,u), making relevant the position of the density of z: for instance, in the univariate, unimodal case, if z is in the right tail of the density it is natural to expect that $E(u|z) > 0$, implying a leftward shift, and if z is in the left

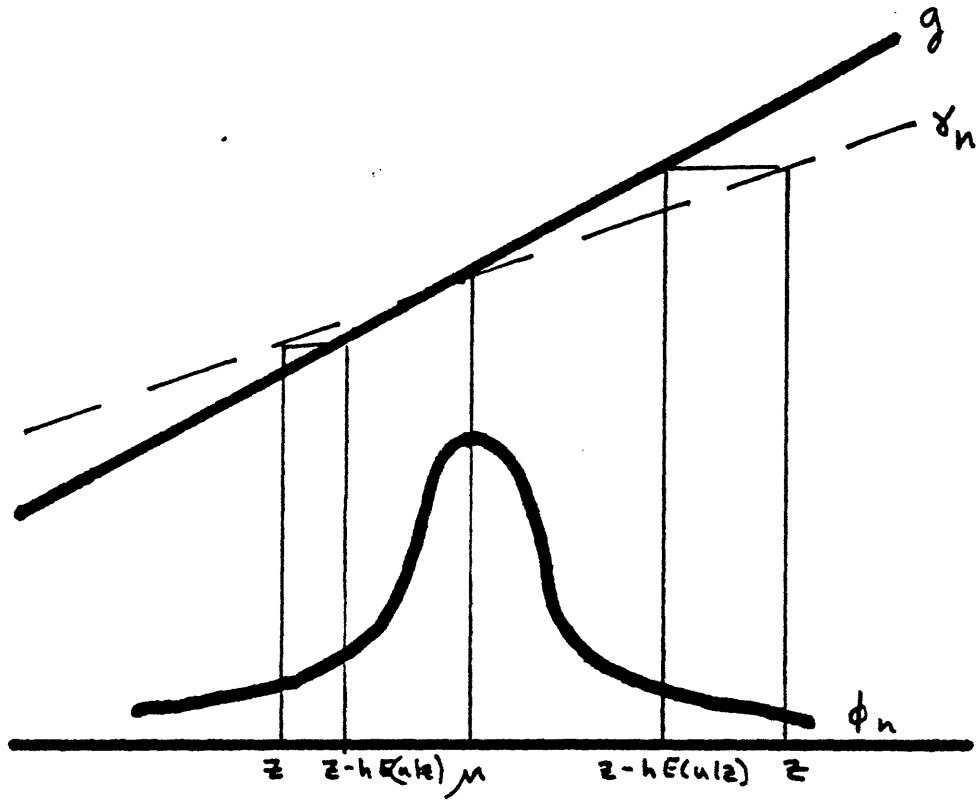

FIGURE 1

tail of the density, it is likewise natural to expect $E(u|z) < 0$, implying a rightward shift. Each of these impacts serves to flatten a curved function $g(x)$, as illustrated in Figures 2a and 2b. If the alteration ii) is minor, the argument-shift gives rise to the downward derivative bias.

To say more, we must characterize the structure of the shift, written as

$$hE(u|z) = h \frac{\int u\mathcal{K}(u)\ f(z-hu)du}{\int \mathcal{K}(u)\ f(z-hu)du} , \tag{3.17}$$

and how it relates to the distribution of x or $z = x + hu$. The connection to the density of z is immediate if we specialize to the case where u is normally distributed; we assume,

Assumption 3.2: The kernel $\mathcal{K}(u)$ is the multivariate normal density, with mean 0 and covariance matrix I, the kxk identity matrix.

For this kernel, we have that $\mathcal{K}'(u) = - u\mathcal{K}(u)$. Using this, if integration-by-parts is applied to each component of the numerator of (3.17), we have that

$$hE(u|z) = - h^2 \frac{\int \mathcal{K}(u)\ f'(z-hu)du}{\int \mathcal{K}(u)\ f(z-hu)du} = h^2 \left(- \frac{\phi_h'}{\phi_h} \right) = h^2\ \lambda_h(z) \tag{3.18}$$

Therefore, the direction of the argument-shift is determined by the sign of the score λ_h, or equivalently by the sign of the density derivative ϕ_h'. In the univariate, unimodal case, the density derivative is positive to the left of the mode, and negative to the right, giving the shift as depicted in Figures 2a and 2b.[13]

For more general base densities, (3.18) indicates how downward derivative bias will arise in areas of higher density, so that the average of the

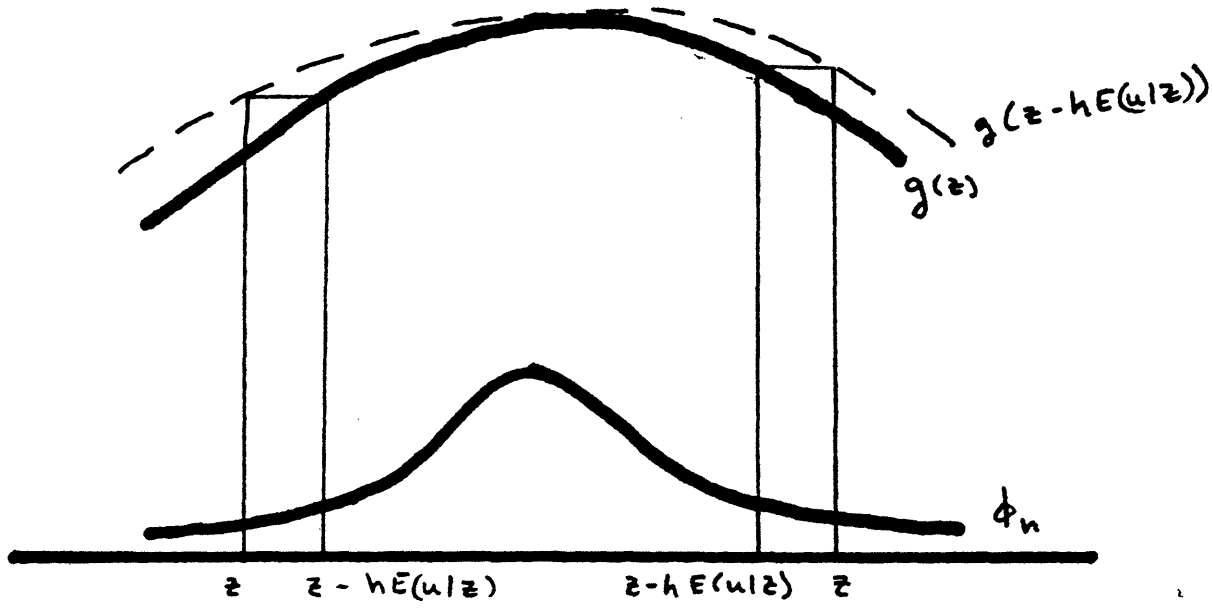

FIGURE 2a

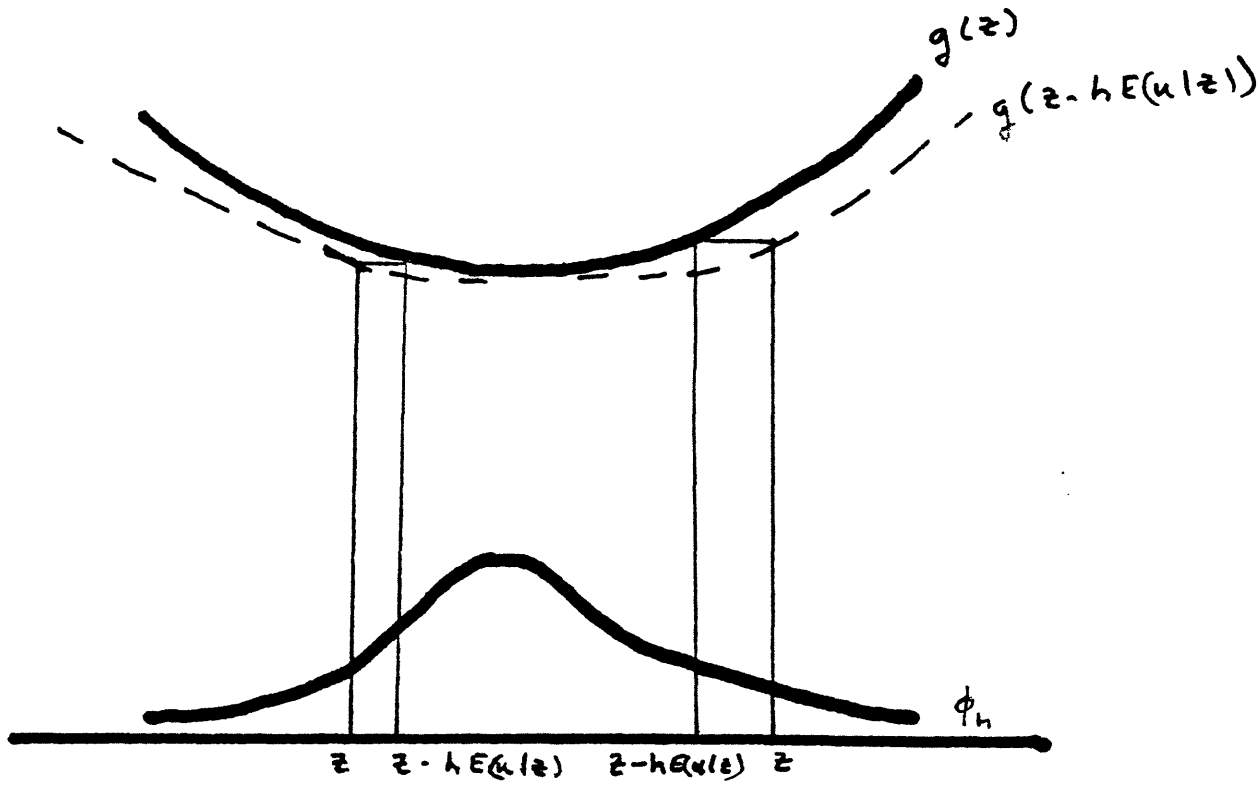

FIGURE 2b

estimated derivatives will be too small. Consider the bimodal density of Figure 3, with the true model linear. Here the slopes are too small over the areas around the models, and too large for the connecting area in the middle. Therefore, the average of the estimated derivatives will be too small, as the sections with downward bias are more heavily weighted than those with upward bias.

With a linear model, $g(x) = \alpha + x^T\beta$, we have that

$$\gamma_h(x) = \alpha + \beta^T[x - h^2\lambda_h(x)] \quad ; \tag{3.18}$$

and

$$\gamma_h'(x) = \beta^T[I - h^2\{-\partial^2 \ln \phi_h/\partial x \partial x^T\}] \quad . \tag{3.19}$$

so that the direction of the derivative bias is determined by the concavity properties of ϕ_h. If ϕ_h is log-concave, $\partial^2 \ln \phi_h/\partial x \partial x^T$ is everywhere negative, and the derivative bias is uniformly downward. Given Assumption 3.2, if the density f is log-concave, then so is ϕ_h (Prekopa(1973, 1980)). The average bias in derivatives from (3.19) is

$$E[\gamma_h'(x)] = \beta^T[I - h^2\ E\{-\partial^2 \ln \phi_h/\partial x \partial x^T\}] \quad ; \tag{3.20}$$

The expectation in the latter expression is $E\{-\partial^2 \ln \phi_h/\partial x \partial x^T\} = \int -\partial^2 \ln \phi_h/\partial x \partial x^T\ f(x)\ dx$, which is not the information matrix of ϕ_h, but nevertheless will be positive if weighting by f does not differ much from weighting by ϕ_h. While we have not established a general result, when this expectation is downward the average derivative bias is downward as expected.

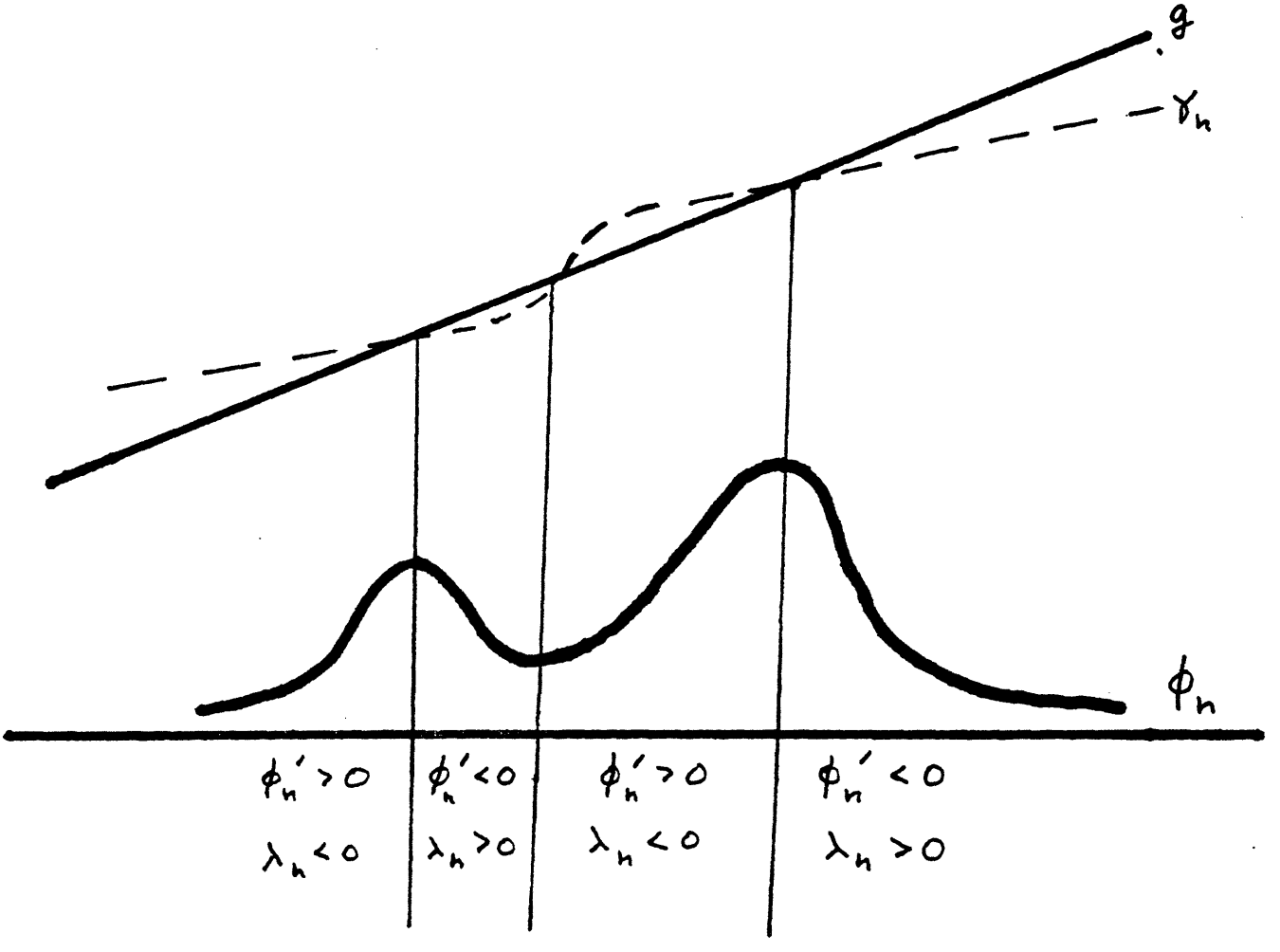
g
γ_n
ϕ_n
$\phi_n' > 0$
$\lambda_n < 0$
$\phi_n' < 0$
$\lambda_n > 0$
$\phi_n' > 0$
$\lambda_n < 0$
$\phi_n' < 0$
$\lambda_n > 0$

FIGURE 3

3.3 Normal Predictors and Smoothing Bias

For nonlinear models, we can obtain more explicit formulations for the derivative bias when the predictor variables are normally distributed (and a normal kernel is used). Consequently, we now add the assumption

Assumption 3.3: The density $f(x)$ is the multivariate normal density, with mean μ and nonsingular covariance matrix Σ.

With assumption 3.2, the density ϕ_h of $z = x + hu$ is a normal density with mean μ and covariance matrix $\Sigma + h^2 I$, so that

$$\lambda_h(z) = -\partial \ln \phi_h / \partial z = (\Sigma + h^2 I)^{-1}(z - \mu) \tag{3.21}$$

and

$$-\partial^2 \ln \phi_h / \partial z \partial z^T = (\Sigma + h^2 I)^{-1} \quad . \tag{3.22}$$

Further, we have that $z = x + hu$ and u are joint normally distributed, which implies that z and $u - E(u|z)$ are independently distributed normal variables. Defining

$$A_h = (\Sigma + h^2 I)^{-1}\Sigma \tag{3.23}$$

then (3.18) implies that $hE(u|z) = (I - A_h^T)(z - \mu)$.

The multivariate linear model is cast appears in this notation as follows. Suppose now that y follows a normal regression model; $y = \alpha + \beta^T x + \epsilon$, where ϵ is distributed with mean 0, independently of x. Here the true regression is $g(x) = \alpha + \beta^T x$, with derivatives $g'(x) = \beta$, and the regression of y on $z = x + hu$ is

$$\gamma_h(z) \equiv E(y|z) = [\alpha + (\beta - A_h\beta)^T \mu] + (A_h\beta)^T z \quad , \tag{3.24}$$

with derivatives $\gamma_h'(z) = A_h\beta$. Therefore, $\hat{g}'(x)$ will estimate $A_h\beta$ for each

x, and is a downward biased estimator of $g'(x) = b$ in the sense of Chamberlain and Leamer(1976), namely $A_h\beta = A_h\beta + (I-A_h)0$ is a matrix weighted average of β and 0, with positive definite weights. In the univariate case we have $A_h = 1-\nu_h = \sigma^2/(\sigma^2+h^2)$ as before, and in the multivariate case where $\Sigma = \sigma^2 I$, we have $A_h = (1-\nu_h) I$. The design of Table 1 had $h = 1$ and $\sigma^2 = 1$, so that the downward derivative bias factor predicted by this analysis is $1 - \nu_h = 1/2$ for each component. While not exactly matching the results of Table 1, this depiction does explain a substantial amount of the difference observed between the "average" row and the true values.

Because of the "argument-shift" structure, it is natural to conjecture that the downweighting matrix $A_h = (\Sigma+h^2I)^{-1}\Sigma$ plays a role in the derivative bias associated with the estimation of a nonlinear regression function g(x). In particular, with the additional regularity condition

Assumption 3.4: Suppose that for any normally distributed variable v and scalar ζ, we have $\partial/\partial\zeta\ (E_v[g(v+\zeta)]) = E_v[g'(v+\zeta)]$,

we can give a general characterization of the average derivative bias with normal predictors. In particular, we can show

$$E_x[\gamma_h'(x)] = A_h E_w[g'(w)] \tag{3.25}$$

where $w \sim N(\mu_x, \Sigma_x[I - A_h(I-A_h)])$. For the case where $\Sigma = \sigma^2 I$, with $\nu_h \equiv h^2/(\sigma^2+h^2)$, this result specializes to

$$E_x[\gamma_h'(x)] = (1-\nu_h)E_w[g'(w)] \tag{3.26}$$

where $w \sim N(\mu, \sigma^2 I\ [1 - \nu_h(1-\nu_h)])$. As such, the average impact of smoothing bias in derivatives is to downweight by A_h (or $1-\nu_h$), and alter the average

derivative value $E[g'(x)]$ to $E[g'(w)]$, where w has a more compact normal distribution than x; with both distributions centered at μ.

The demonstration of (3.25) follows directly from the normality and independence of z and $hu - hE(u|z) = e$ implied by assumptions 3.2 and 3.3. The decomposition (3.16) is written as

$$(3.27) \qquad \gamma_h(z) = E_e[g(z - E(hu|z) - e)] = E_e\{g[A_h^T z + (I-A_h^T)\mu - e]\}$$

so that assumption 3.4 implies

$$(3.28) \qquad \gamma_h'(z) = A_h E_e\{g'[A_h^T z + (I-A_h^T)\mu - e]\} .$$

Since e is independent of the argument z, by evaluating at $z = x$ and taking expectations we have

$$(3.29) \qquad E_x[\gamma_h'(x)] = A_h E_x E_e\{g'[A_h^T x + (I-A_h^T)\mu - e]\} = A_h E[g'(w)] .$$

where $w = A_h^T x + (I-A_h^T)\mu - e$. w is normally distributed with mean μ_x and covariance matrix $A_h^T \Sigma A_h + h^2 A_h^T = \Sigma[I - A_h(1-A_h)]$. This verifies (3.25).

The derivative bias formulation (3.25) has the striking implication that on average, the same downweighting applies to derivatives when the true model is quadratic or linear; for either $g(x) = \alpha + \beta^T x$ or $g(x) = \alpha + \beta^T x + x^T B x$, we have that

$$(3.30) \qquad E_x[\gamma_h'(x)] = A_h E[g'(x)]$$

This is because the derivatives of the quadratic function are linear in x, so that the spread change (from x to w in (3.25)) is inconsequential.

It is important to stress "on average" above, as one does not have pointwise proportional downweighting, or $\gamma_h'(x) = A_h g'(x)$ for each x for a quadratic function g(x). A useful exercise (details left for the reader) is to verify this in the univariate normal case. We sketch this as

Example 3.1: Let $x \sim N(\mu,\sigma^2)$ and

$$g(x) = \alpha + \beta x + \beta_1 x^2$$

then $\gamma_h(z)$ is

$$(3.31) \quad \gamma_h(z) = [\alpha + \beta\nu_h\mu + \beta_1\nu_h{}^2\mu^2 + \beta_1 h^2\sigma_{u|z}{}^2]$$

$$+ [\beta(1-\nu_h) + 2\beta_1\nu_h\mu - 2\beta_1\nu_h{}^2\mu]\ z \ + \ [\beta_1(1 - \nu_h)^2]\ z^2$$

with $\nu_h = h^2/(\sigma_x{}^2+h^2)$ as before, and $\sigma_{u|z}{}^2$ a (positive) constant. Here γ_h is a shifted quadratic function, and γ_h is concave (convex) if and only if g is concave (convex). Since

$$(3.32) \quad \gamma_h'(z) = [(\beta + 2\beta_1\mu\nu_h)(1-\nu_h)] + 2\ [\beta_1(1 - \nu_h)^2]\ z$$

$$g'(z) = \beta + 2\beta_1\ z$$

the relative values of $\gamma_h'(z)$ and $g'(z)$ depend on the position of the predictor density (through μ), h (through ν_h) and z, matching only at $z = \mu$. Nevertheless, we can verify (3.26,30) as

$$(3.33) \quad E[\gamma_h'(x)] = \beta(1-\nu_h) + 2\beta_1(1-\nu_h)\mu$$

$$= (1-\nu_h)[\beta + 2\beta_1\mu] = (1-\nu_h)E[g'(x)]$$

so that the mean of γ_h' is biased downward by the same factor $(1-\nu_h)$ that applies to the coefficient of a univariate linear model with errors-in-variables.

We close this section with a further example, namely one that gives similar results without the additive disturbance structure of assumption 3.1.

Example 3.2: Consider a normal probit model, where

$$y = 1[\epsilon < \alpha + \beta^T x] \tag{3.34}$$

where 1[ ] is the indicator function, and $\epsilon \sim N(0,1)$, independently of x. Here $g(x) = \Psi(\alpha + \beta^T x)$, where Ψ is the normal c.d.f., and $g'(x) = \beta\, \psi(\alpha+\beta^T x)$, where ψ is the standard normal density function. With $z = x + hu$, to derive $\gamma_h(z) = E(y|z)$ we note that

$$\begin{aligned} y &= 1[\, \epsilon + \beta^T h(u-E(u|z)) < \alpha + \beta^T (z - hE(u|z))] \\ &= 1[\, \eta < \alpha + (\beta - A_h\beta)^T \mu + (A_h\beta)^T z \,] , \end{aligned} \tag{3.35}$$

where $\eta = \epsilon + b^T h(u-E(u|z))$, using that $hE(u|z) = (I-A_h^T)(z-\mu)$. The variable η is distributed independently of z, and $\eta \sim N(0,\sigma_\eta^2)$, where $\sigma_\eta^2 = 1 + h^2\beta^T [I + (h^4-2h^2)(\Sigma+h^2 I)^{-1}] \beta$. Therefore

$$\gamma_h(Z) = \Psi[(\alpha + (\beta-A_h\beta)^T \mu + (A_h\beta)^T z)/\sigma_\eta] \tag{3.36}$$

and

$$\gamma_h'(Z) = (A_h\beta)/\sigma_\eta \;\; \psi[(\alpha + (\beta-A_h\beta)^T \mu + (A_h\beta)^T z)/\sigma_\eta] \tag{3.37}$$

involving A_h from the argument-shift, as well as other differences. Finally, by exploiting the McFadden-Reid(1975) formulae (applied to average derivatives as in Stoker(1986a,b)), we have that

$$E[g'(x)] = c_1\, \beta \tag{3.38}$$

$$E[\gamma_h'(x)] = c_2\, A_h\beta = (c_2/c_1)\, A_h\, E[g'(x)] \tag{3.39}$$

where the constants c_1 and c_2 are given as

$$\begin{aligned} c_1 &= C_1^{-1} \psi[(\alpha + \beta^T \mu)/C_1] \\ C_1 &= [1 + \beta^T \Sigma \beta]^{1/2} \end{aligned} \tag{3.40}$$

$$c_2 = C_2^{-1}\psi[(\alpha + \beta^T\mu)/C_2]$$

$$C_2 = [1 + \beta^T\Sigma(\Sigma + h^2I)^{-1}\Sigma(\Sigma + h^2I)^{-1}\Sigma\beta]^{1/2}$$

Therefore, the average downweighting is similar to that before, involving a scaling of the linear factor A_h.

4. Size of the Derivative Bias: Dimension and Sample Size

Our discussion so far has indicated how derivative bias arises, and has given formulations of the bias in terms of the bandwidth value h. Aside from the simulation results of Tables 2 and 3, where the bandwidth was set in an ad hoc fashion, we still have no sense of the "typical" size of the bias. For applications where bandwidth values are small, there will be little derivative bias; for those where the bandwidth is large, the bias can be a substantive problem. Moreover, since different bandwidth values are appropriate for different sample sizes, and for different dimensions (number of x's), the derivative bias may arise only in limited settings; for instance, small samples of moderate dimension.

We address this issue by considering the bias implied by bandwidth values set in an approximately optimal fashion. In particular, we consider where the true model is

$$y_i = \alpha + x_{1i} + \ldots + x_{ki} + \sigma_k\epsilon_i, \quad i = 1,\ldots,N \tag{4.1}$$

where x_j, $j = 1,\ldots,k$ and ϵ are independent, univariate normal random variables, and α is a constant. For comparability across dimensions, we fix the true R^2 value of the equation by setting $\sigma_k^2 = k(1 - R^2)$. From (3.24), for a given bandwidth value h, the bias factor A_h for this design is $(1 - \nu_h)$ I, where $\nu_h = h^2/(1 + h^2)$. In words, each derivative will estimate $(1 - \nu_h)$ of its true value.

We compute the bandwidth values by minimizing weighted mean integrated squared error, approximated by its leading terms of the bias and variance. Weighting is uniform over the centered (rectangular) region of probability .95 for each dimension. The calculations and specific formulae are given in the Appendix.

The bandwidth values h and the derivative bias ν_h are computed for R^2 values of .8, .5 and .2 respectively in Tables 3a, b and c. Aside from the case of one dimension, the bias values are strikingly large. For instance, with $R^2 = .8$, dimension $k = 4$ and $N = 100$, the optimal bandwidth value is $h = 1.085$, with a downward derivative bias of 54%. In comparison, Table 1 used $h = 1$, and therefore slightly understated the bias problem. The "curse of dimensionality" appears here in spades, namely the derivative bias greatly increases with dimension, and does not decrease rapidly with sample size. The impact of the true noise in the equation is predictable but not terribly large, as the bias increases when the R^2 is lowered, but the differences are not as substantial as those seen for the other features.

Of course, Table 3 is dependent upon the normal linear model, and different bandwidth values would arise if any aspect of the design were altered. Nevertheless, these results suggest strongly that derivative bias can be a serious problem in typical empirical problems.

5. Bias in Marginal Effects and Fitting Criteria

We close our discussion by presenting a diagnostic statistic for bias in marginal effects, and a few remarks on the structure of the bias with different kinds of nonparametric regression estimators. The fact that derivative bias arises with a linear model is symptomatic of certain fitting problems of the kernel regression estimator. In particular, if the kernel estimator is computed with $y = x$, the result is a biased depiction of the true

TABLE 3: DERIVATIVE BIAS - APPROXIMATELY OPTIMAL BANDWIDTHS

Specification:

Linear Model: $y_i = a + \sum x_{ji} + \sigma\epsilon_i$, $i = 1,\ldots,N$

Normal Regressors: $x \sim N(0,I)$

Normal Disturbance: $\epsilon \sim N(0,1)$; Constant R^2: $\sigma^2 = k(1 - R^2)$

Optimal Bandwidth: $h = A(k)\ N^{-1/(k+4)}$, A(k) formula given in Appendix.

Derivative Bias: $\nu_h = h^2/(1+h^2)$

Table 3a: $R^2 = .80$

A(k)	0.719	1.193	1.589	1.930	2.199	3.111
Dimension k	1	2	3	4	5	10
N = 25						
Bandwidth h	0.378	0.698	1.003	1.291	1.538	2.472
Derivative Bias	12.48%	32.74%	50.17%	62.49%	70.29%	85.94%
N = 50						
Bandwidth h	0.329	0.622	0.909	1.184	1.424	2.353
Derivative Bias	9.76%	27.87%	45.23%	58.35%	66.97%	84.70%
N = 100						
Bandwidth h	0.286	0.554	0.823	1.085	1.318	2.239
Derivative Bias	7.57%	23.47%	40.39%	54.08%	63.48%	83.37%
N = 500						
Bandwidth h	0.207	0.423	0.654	0.888	1.103	1.996
Derivative Bias	4.13%	15.20%	29.96%	44.06%	54.87%	79.93%
N = 1000						
Bandwidth h	0.181	0.377	0.592	0.814	1.021	1.899
Derivative Bias	3.16%	12.46%	25.98%	39.84%	51.03%	78.30%
N = 5000						
Bandwidth h	0.131	0.288	0.471	0.666	0.854	1.693
Derivative Bias	1.68%	7.68%	18.14%	30.70%	42.16%	74.14%
N = 10000						
Bandwidth h	0.114	0.257	0.426	0.610	0.790	1.611
Derivative Bias	1.28%	6.20%	15.38%	27.14%	38.45%	72.19%
N = 100000						
Bandwidth h	0.072	0.175	0.307	0.458	0.612	1.367
Derivative Bias	0.51%	2.97%	8.60%	17.32%	27.25%	65.14%

Table 3b: $R^2 = .50$

A(k)	0.864	1.390	1.811	2.164	2.435	3.321
Dimension k	1	2	3	4	5	10
N = 25						
Bandwidth h	0.454	0.813	1.144	1.447	1.703	2.639
Derivative Bias	17.07%	39.78%	56.67%	67.68%	74.36%	87.45%
N = 50						
Bandwidth h	0.395	0.724	1.036	1.327	1.577	2.512
Derivative Bias	13.49%	34.40%	51.76%	63.78%	71.31%	86.32%
N = 100						
Bandwidth h	0.344	0.645	0.938	1.217	1.460	2.390
Derivative Bias	10.57%	29.39%	46.82%	59.69%	68.06%	85.11%
N = 500						
Bandwidth h	0.249	0.493	0.746	0.995	1.221	2.131
Derivative Bias	5.85%	19.57%	35.72%	49.76%	59.84%	81.95%
N = 1000						
Bandwidth h	0.217	0.439	0.675	0.913	1.130	2.028
Derivative Bias	4.49%	16.19%	31.32%	45.44%	56.09%	80.44%
N = 5000						
Bandwidth h	0.157	0.336	0.537	0.746	0.945	1.808
Derivative Bias	2.41%	10.15%	22.35%	35.77%	47.18%	76.57%
N = 10000						
Bandwidth h	0.137	0.299	0.486	0.684	0.875	1.720
Derivative Bias	1.84%	8.23%	19.10%	31.90%	43.37%	74.74%
N = 100000						
Bandwidth h	0.086	0.204	0.350	0.513	0.678	1.459
Derivative Bias	0.74%	4.00%	10.90%	20.85%	31.46%	68.05%

Table 3c: $R^2 = .20$

A(k)	0.949	1.503	1.937	2.295	2.566	3.435
Dimension k	1	2	3	4	5	10
N = 25						
Bandwidth h	0.498	0.879	1.223	1.535	1.794	2.729
Derivative Bias	19.89%	43.59%	59.94%	70.20%	76.30%	88.16%
N = 50						
Bandwidth h	0.434	0.783	1.108	1.407	1.661	2.597
Derivative Bias	15.84%	38.01%	55.10%	66.45%	73.40%	87.09%
N = 100						
Bandwidth h	0.378	0.698	1.003	1.291	1.538	2.472
Derivative Bias	12.48%	32.74%	50.17%	62.49%	70.29%	85.94%
N = 500						
Bandwidth h	0.274	0.534	0.797	1.055	1.286	2.204
Derivative Bias	6.97%	22.16%	38.86%	52.69%	62.32%	82.92%
N = 1000						
Bandwidth h	0.238	0.475	0.722	0.968	1.191	2.097
Derivative Bias	5.37%	18.43%	34.27%	48.37%	58.65%	81.47%
N = 5000						
Bandwidth h	0.173	0.363	0.574	0.791	0.996	1.869
Derivative Bias	2.90%	11.67%	24.77%	38.51%	49.79%	77.75%
N = 10000						
Bandwidth h	0.150	0.324	0.520	0.726	0.922	1.779
Derivative Bias	2.21%	9.49%	21.27%	34.50%	45.95%	75.99%
N = 100000						
Bandwidth h	0.095	0.221	0.374	0.544	0.714	1.509
Derivative Bias	0.89%	4.64%	12.27%	22.85%	33.76%	69.49%

regression $E(x|x) = x$, along the lines discussed above. This feature motivates the discussion of this section.[14]

The simulations of Section 2.2 showed how the "slope estimator" $\hat{d}$ gave essentially unbiased results for average derivatives. This estimator multiplied the average of regression derivatives by the inverse of the factor

$$M_{hx} = N^{-1}\sum \hat{x}'(x_i)\,\hat{1}_i \quad . \tag{5.1}$$

From our discussion above, it is clear that M_{hx} acts to correct for level derivative bias in the estimation of the function $x = E(x|x)$ by the kernel regression estimator. Moreover, for the normal examples above, if the trimming is ignored (or allowed to vanish with h fixed), we have that M_{hx} estimates the downweighting factor A_h. M_{hx} can then be computed as a diagnostic statistic, and compared to the identity matrix. Of course, if the normal design is called for, $A_h = (\Sigma+h^2I)^{-1}\Sigma$ could be estimated directly to indicate the extent of the problem(or with standardized data $\nu_h = h^2/(1 + h^2)$). M_{hx} would be preferable in nonnormal settings, although it only measures the average proportion of derivative bias in estimating a linear model. Example 3.1 showed how M_{hx} may not give an accurate correction to the derivatives $\hat{g}'(x)$ at different evaluation points x.

Given that the issue arises in the estimation of the regression $E(x|x)$, it is natural to conjecture that the derivative bias might be smaller for a estimation method that reproduced x exactly. While several methods exhibit this property, here we consider some generic features of approximating g(x) by global least squares, such as fitting truncated polynomial or Fourier series expansions.

From Härdle (1991), the kernel regression estimator arises from the local least squares problem

$$(5.2) \qquad \hat{g}(x) = \underset{\mu}{\operatorname{argmin}} \sum_{i=1}^{N} \mathcal{K}\left(\frac{x - x_i}{h}\right) [y_i - \mu]^2$$

As above, with h held fixed, plim $\hat{g}'(x) = \gamma_h'(x)$. From integrating by parts, the mean difference between g' and γ_h' can always be written as

$$(5.3) \qquad E(g') - E(\gamma_h') = E\{[g(x) - \gamma_h(x)]^T \ell(x)\}$$

where $\ell(x) = -f'(x)/f(x)$ is the translation score of the true density. Mean bias arises from the level differences $g(x) - \gamma_h(x)$ being systematically correlated with the score $\ell(x)$.

With global least squares, this interaction with the density can be refined. In particular, suppose that the regression was estimated by

$$(5.4) \qquad \tilde{g}(x) = \underset{\mu(x)\in\mathcal{P}_q}{\operatorname{argmin}} N^{-1} \sum_{i=1}^{N} [y_i - \mu(x_i)]^2$$

where $\mathcal{P}_q$ is a class of functions, for example polynomials of degree q. A nonparametric theory for this estimation would be based on changing q with N, so as to broaden $\mathcal{P}_q$ so that its closure is sure to contain g(x). However, consider the implications of such a method when N increases, for fixed q. Suppose that the class $\mathcal{P}_q$ were sufficiently regular to allow us to demonstrate that plim $\tilde{g}(x) = \tilde{\gamma}(x)$ where $\tilde{\gamma}$ solved the least squares problem

$$(5.5) \qquad \tilde{\gamma}(x) = \underset{\mu(x)\in\mathcal{P}_q}{\operatorname{argmin}} E\{[g(x) - \mu(x)]^2\} .$$

Moreover, the least squares criterion would imply that $g(x) - \tilde{\gamma}(x)$ is uncorrelated (orthogonal) to any element of $\mathcal{P}_q$; for instance, the solution $\tilde{\ell}(x)$ of the (population) problem of fitting the density score $\ell(x)$, or

$$(5.6) \qquad \tilde{\ell}(x) = \underset{\mu(x)\in P_q}{\operatorname{argmin}} \; E([\ell(x) - \mu(x)]^2) \quad .$$

The fact that $g(x) - \tilde{\gamma}(x)$ and $\tilde{\ell}(x)$ are uncorrelated allows us to refine (5.3) to

$$(5.7) \qquad E(g') - E(\tilde{\gamma}') = E([g(x) - \tilde{\gamma}(x)]^T[\ell(x) - \tilde{\ell}(x)]) \quad .$$

The mean derivative bias is now the product of differences in fitting the function $g(x)$ and the score $\ell(x)$.[15] While general results are left for future research, we can see how our earlier results are affected when $x \in \mathcal{P}_q$, with $\mathcal{P}_q$ a linear space (for instance, if $\mathcal{P}_q$ denotes polynomials of degree $q \geq 1$). Obviously, there is no derivative bias for linear models, as $\tilde{\gamma}(x) = g(x)$ in that case. Moreover, there is no mean bias when the regression is nonlinear but the predictors are normal,[16] as $\ell(x)$ is a linear function of x, so $\tilde{\ell}(x) = \ell(x)$ here. Since this is not a pointwise justification, it does not recommend series estimators generally (it is not clear how multimodal or skewed design will impact then). However, it does illustrate how the fitting criterion can alter the dependence of the bias on the design of the regressors.

6. Concluding Remarks

The intention of this paper is to be thought-provoking. While the analysis is clearly critical about the incautious use of marginal effects estimated by local smoothing, this does not serve as an argument for or against the use of local smoothing estimators for examining the basic structure of regression relationships. Such estimators are well equipped for detecting bumps, troughs and other kinds of nonlinear structure, that other estimators (such as truncated series) may miss. The main point of the paper is that smoothing tends to dampen the size of the bumps, etc., leading to

mismeasurement of marginal effects. Moreover, when the bias is roughly proportional, a simple correction could remove the bias, as with the "slope" version of the average derivative estimator given above. Similarly, while the last section indicated how estimators based on global fitting may have less derivative bias that kernel estimators when the design is normal, the same property is not given for more general distributions of the regressors.

The focus of the paper on pointwise bias was originally motivated by simulation results such as those of Table 1 and 2, which showed substantial average mismeasurement. However, it is important to recall that pointwise variance plays an equal role in the accuracy of nonparametric estimators. While this feature is incorporated in the optimal bandwidth calculations of Table 3, it gives further reason for not jumping immediately to the conclusion that alternative nonparametric estimators dominate those using local smoothing. For instance, if there are $k = 5$ regressors and $N = 100$ data points, fitting a cubic polynomial (the lowest degree permitting non-concave or convex relations) involves estimating 56 coefficients, which can impart substantive variance to the estimated function (fitted values) from such an analysis.

These are practical issues, which should be seen as distinct from the theoretical results on nonparametric and semiparametric methods. The myriad of results that now exist for optimal rates of convergence for function estimation, or $\sqrt{N}$ consistency for semiparametric methods, do not address these implementation issues directly. For instance, for estimating average derivatives, the bandwidth conditions of Powell, Stock and Stoker(1989) and Härdle and Stoker(1989) indicate that the bandwidth should converge to zero more quickly than for pointwise estimation, giving rise to "asymptotic undersmoothing". While tempting, this does not establish that bandwidths should be set to smaller values for a given finite sample size, but just how

they would shrink in with increases in sample size.[17] These papers also make use of bias reducing "higher-order" kernels, which are not considered in this paper, but could affect the mismeasurement of marginal effects.

The point of this is that the myriad of theoretical results now available on optimal nonparametric and semiparametric results are not informative about how these procedures will work practically. This paper has presented a large sample analysis, but used a fixed bandwidth theory to try to accurately depict the impacts of smoothing in realistically sized samples. Which analysis is better depends on the quality of approximation, which is a practical issue. We have taken the stance that when a bandwidth of $h = 1$ is set, the distribution of the estimator will be better approximated but fixing $h = 1$ in the approximation, than by considering h to be tiny, as in the limit. The same issue exists for every nonparametric estimator; are the precision properties of a function fit by a cubic polynomial better described by assuming the polynomial degree is large, as in the limit, or by recognizing that a cubic formula has been fit to an unknown function.

Econometric theory has undergone spectacular development over the past decade, coming to an equal position with mathematical statistics in terms of understanding flexible measurement methods. This development has not been accompanied by a large number of empirical applications, in part because of the high degree of technical prowess now required for an applied researchers to follow and assess the literature. The spirit of this paper is to suggest that the practical issues be given much more weight in this econometric research program. Without such practical validation, the impact of the theoretical progress to date will be limited.

Appendix: Approximate Optimal Bandwidth Formulae

The bandwidth values of Table 3 are picked by criterion discussed in Härdle(1991), following closely the calculations in Hausman and Newey(1990). For kernel estimator $\hat{g}_h(x)$ of a regression $g(x) = E(y|x)$, we choose the bandwidth h to minimize the approximate weighted integrated mean squared error, or

$$IMSE(h) = \int w(x) \, [\tilde{Var}(\hat{g}_h(x) + \tilde{Bias}(\hat{g}_h(x))^2] \, f(x) \, dx$$

where w(x) is a weighting function, and pointwise variance and bias are approximated by their leading terms in the bandwidth h, namely

$$\tilde{Var}(\hat{g}_h(x)) = N^{-1}h^{-k} \int \mathcal{K}(u)^2 du \; \sigma^2(x)/f(x)$$

$$\tilde{Bias}(\hat{g}_h(x)) = (h^2/2) \, Trace\{\partial^2 g/\partial x \partial x^T + 2 \, f'(x)/f(x) g'(x)^T\} \int uu^T \mathcal{K}(u) du$$

where $\sigma^2(x) = Var(y|x)$. For the linear model (4.1) with standard normal regressors, and a normal kernel, IMSE(h) specializes to

$$IMSE(h) = C_1 N^{-1} h^{-k} + C_2 h^4$$

where

$$C_1 = \int \mathcal{K}(u)^2 du \int w(x) dx \; \sigma^2 = 2^{-k/2} \int w(x) dx \; \sigma^2$$

$$C_2 = \int (\sum x_j)^2 \, w(x) \, f(x) dx$$

The optimal bandwidth value is then given as

$$h = A \, N^{-1/(k+4)} \quad ; \quad A = (kC_1/4C_2)^{1/(k+4)}$$

We utilize uniform weighting on 95% of the sample; namely

$$w(x) = 1[-c_k < x_j < c_k; \; j = 1, \ldots, k]$$

where c_k is set such that the (marginal) probability of $-c_k < x_j < c_k$ is $(.95)^{1/k}$, so that $E(w) = .95$ for every dimension value k. Recalling that we set $\sigma^2 = k(1 - R^2)$, the constants C_1 and C_2 are then solved for as

$$C_1 = 2^{-k/2} \int w(x)dx \; \sigma^2 = 2^{-k/2}[2c_k]^k \; \sigma^2 = 2^{k/2} c_k{}^k \; k(1 - R^2)$$

$$C_2 = \left(\int_{-c_k}^{c_k} x_j{}^2 \, \psi(x_j)dx_j \right)^k = [(.95)^{1/k} - 2c_k\psi(c_k)]^k$$

where $\psi(.)$ is the standard normal density, and the latter equality follows from twice differentiating and evaluating the truncated normal moment generating function. Therefore, the optimal bandwidth value is determined as a function of k and N by

$$h = A(k) \; N^{-1/(k+4)} \quad ; \qquad \text{where}$$

$$A(k) = \{k^2 2^{k/2} c_k{}^k / 4[(.95)^{1/k} - 2c_k\psi(c_k)]^k\}^{1/(k+4)} (1 - R^2)^{1/(k+4)}$$

Notes

[1] See Lau(1986), Barnett and Lee(1985) and Elbadawi, Gallant and Souza(1983) for references to this literature.

[2] For instance, a linear equation is capable of measuring first derivatives at a point, but OLS coefficient estimates will not necessarily equal the derivatives at a point of interest, such as the sample mean of the predictors (c.f. White(1980), among others).

[3] See Prakasa-Rao(1983) and Härdle(1991) for references to the relevant statistical literature, and Delgado and Robinson (1991) for an extensive survey of recent work in econometrics.

[4] Stoker(1991b) discusses the estimation of density derivatives with a similar conclusion. However, reasons for the downward bias in kernel density derivatives are quite different than for the kernel regression case as studied here. Very preliminary versions of some of the results of this paper and Stoker (1991b) were contained in a draft entitled "Smoothing Bias in Derivative Estimation," revised July 1990.

[5] Härdle (1991) gives a good summary of this theory and the relevant literature.

[6] Extensive versions of the simulation results below will be reported in Stoker and Villas-Boas(1991).

[7] See Stoker(1986), Härdle and Stoker(1989) and Powell, Stock and Stoker(1989) among others. A semiparametric motivation for the estimation of δ derives from models with the index structure $E(y|x) = G(\beta^T x)$, which is commonly available in models of limited dependent variables. Without loss of generality, the coefficients in a model of this type can be scaled so that they represent the average effect on the observed response; namely one can set $\beta = \delta$ defined above. This imparts a concrete interpretation to the empirical effects that are represented by the values of the coefficients β.

[8] The conditions also employ "higher-order" kernels, and many other smoothness conditions that are not used in the analysis of this paper.

[9] While the positive kernel is not consistent with the asymptotic normality results above, Powell, Stock and Stoker(1989) and Stoker and Villas-Boas(1991) find that a positive kernel gives superior small sample performance. Moreover, a positive kernel is used in the remainder of the exposition, so that the results reported are relevant.

[10] Rather it was chosen as roughly the smallest value for which stable performance was seen for the "average" estimator.

[11] Silverman (1986) notes this structure for density estimators, and it is exploited to study bias in Stoker(1991b).

[12] To avoid possible confusion in the following, we note that z denotes the random variable $z = x + hu$ as well as the argument for evaluating γ_h. In particular, for constructing $\gamma_h(z)$, we recognize that $z = x + hu$. However, our point here is that $\hat{g}'(x)$ estimates $\gamma_h'(x)$, namely the function $\gamma_h'(z)$ evaluated by setting the argument $z = x$.

[13] Stoker(1991b) shows some elementary connections between the true density f and the convolution ϕ_h. For instance, for a small enough bandwidth, it is easy to show that the modal structure of ϕ_h will be analogous to that of f.

[14] A recent paper by Gasser and Engel(1990) criticizes the use of Nadaraya-Watson weights in regression estimation, as in (2.2-4), for different reasons than discussed here. However, some aspects of their analysis are similar, such as noting the sensitivity of $\hat{g}(x)$ to the density of x.

[15] This argument, for polynomials, was used by Newey(1991) to study asymptotic bias.

[16] This feature of polynomials is shown in Florens, Ivaldi and Larribeau-Nori(1991).

[17] Goldstein and Messer (1990) demonstrate the "undersmoothing" argument for estimation of general functionals, and make the suggestion of undersmoothing in finite samples.

References

Barnett, W.A. and Y.W. Lee (1985), "The Global Properties of the Minflex Laurent, Generalized Leontief, and Translog Flexible Functional Forms," *Econometrica*, 53, 1421-1437.

Brillinger, D.R. (1983), "A Generalized Linear Model with 'Gaussian' Regressor Variables," in P.J. Bickel, K.A. Doksum and J.l. Hodges, eds., *A Festschrift for Erich L. Lehmann*, Woodsworth International Group, Belmont, CA.

Chamberlain, G. and E. Leamer(1976), "Matrix Weighted Averages and Posterior Bounds," *Journal of the Royal Statistical Society, B*, 73-84.

Delgado, M.A. and P.M. Robinson (1991), "Nonparametric Methods and Semiparametric Methods for Economic Research," draft, London School of Economics.

Elbadawi, I., A. R. Gallant and G. Souza (1983), "An Elasticity Can Be Estimated Consistently Without A Priori Knowledge of Functional Form,: *Econometrica*, 51, 1731-1752

Florens, J-P., M. Ivaldi and S. Larribeau (1991), "Sobolev Estimation of Approximate Regressions," draft, July, Universite' des Sciences Sociales, Toulouse.

Gasser, T. and J. Engel(1990), "The Choice of Weights in Kernel Regression Estimation," *Biometrika*, 77, 377-381.

Goldstein, L. and K. Messer (1990), "Optimal Plug-in Estimators for Nonparametric Functional Estimation," Technical Report No. 277, Stanford University.

Härdle, W. (1991), *Applied Nonparametric Regression*, Cambridge, Cambridge University Press, Econometric Society Monographs.

Härdle, W. and T.M Stoker (1989), "Investigating Smooth Multiple Regression by the Method of Average Derivatives," *Journal of the American Statistical Association*, 84, 986-995.

Hausman, J.A. and W.K. Newey (1990), "Nonparametric Estimation of Exact Consumer Surplus and Deadweight Loss," MIT Department of Economics Working Paper, October.

Ibragimov, I.A. and R.Z. Has'minskii (1981), *Statistical Estimation: Asymptotic Theory*, New York: Springer-Verlag.

Lau, L.J.(1986), "Functional Forms in Econometric Model Building," in Z. Griliches and M.D. Intrilligator, eds., *Handbook of Econometrics, Volume 3*, North Holland, Amsterdam.

McFadden, D. and F. Reid (1975), "Aggregate Travel Demand Forecasting From Disaggregated Behavioral Models," *Transportation Research Record*, No. 534.

Newey, W.K. (1991), "The Asymptotic Variance of Semiparametric Estimators," Working Paper, MIT Department of Economics, August..

Prakasa-Rao, B.L.S.(1983), *Nonparametric Functional Estimation*, Academic Press, New York.

Prekopa, A.(1973), "On Logarithmic Concave Measures and Functions," *Acta. Sci. Math.*, 34, 355-343.

Prekopa, A.(1980), "Logarithmic Concave Measures and Related Measures," in M.A.H. Dempster, ed., *Stochastic Programming*, Academic Press, New York.

Powell, J.L., J.H. Stock and T.M. Stoker(1989), "Semiparametric Estimation of Index Coefficients," *Econometrica*, 57, 1403-1430.

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

Stoker, T.M. (1986a), "Aggregation, Efficiency and Cross Section Regression," *Econometrica*, 54, 171-188.

Stoker, T.M. (1986b), "Consistent Estimation of Scaled Coefficients," *Econometrica*, 54, 1461-1482.

Stoker, T.M. (1991a), "Equivalence of Direct, Indirect and Slope Estimators of Average Derivatives," in *Nonparametric and Semiparametric Methods in Econometrics and Statistics*, ed. by W.A. Barnett, J.L. Powell and G. Tauchen, Cambridge University Press.

Stoker, T.M. (1991b), "Smoothing Bias in the Estimation of Density Derivatives," MIT Sloan School of Management Working Paper, August.

Stoker, T.M. and J.M. Villas-Boas (1991), "Monte Carlo Simulation of Average Derivative Estimators", draft.

White, H. (1980), "Using Least Squares to Approximate Unknown Regression Functions," *International Economic Review*, 21, 149-170.