SMOOTHING BIAS IN THE MEASUREMENT OF MARGINAL EFFECTS

by

Thomas M. Stoker

Massachusetts Institute of Technology

WP#3522-93-EFA January 1993
SMOOTHING BIAS IN THE MEASUREMENT OF MARGINAL EFFECTS

by

Thomas M. Stoker

August 1991

Revised: December 1992

* Sloan School of Management, Massachusetts Institute of Technology, Cambridge, MA, 02139, USA. This research was funded by NSF Grant SES 9122886. The author wishes to acknowledge comments from seminar audiences at the University of Chicago, Harvard University, Stanford University and CORE; and thank G. Chamberlain, K. Doksum, W. Härdle, J. Heckman, D. Jorgenson, S. Marron, R. Matzkin, W. Newey, J. Powell, P. Robinson, A. Zellner and the anonymous referees for many helpful suggestions.
Abstract

This paper examines the tendency of kernel regression estimators to underestimate marginal effects, or derivatives. By developing the connection between smoothing and "errors-in-variables" structure, we explain how smoothing itself can lead to attenuation bias in derivatives, or bias toward zero. When the true model is linear and the regressors normal, the attenuation bias is uniform, and can be accurately explained by standard errors-in-variables bias formulae. These formulations also indicate the seriousness of the attenuation bias with different numbers of regressors and different sample sizes. For the situation where the true model is nonlinear, we examine the related errors-in-variables structure and indicate what features affect the size of attenuation bias in derivatives. Finally, we explain the connections between our analysis (which is based on "fixed bandwidth" approximation) and the more familiar nonparametric theory for kernel regression estimators, including brief consideration of technical devices such as higher order kernels.
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. Focus on marginal effects has often provided criteria for the specification of econometric models; for instance, early definitions of "flexible functional forms" rested on whether derivatives of the true relationship could be approximated at a given value of the predictor variables. Current interest in nonparametric methods in econometrics arises from their ability to approximate the true relationship and its derivatives over the full range of predictor variable values.

This paper considers kernel regression, a standard nonparametric method based on smoothing, or local averages. We point out how marginal effects estimated from kernel regression tend to be too small, or more precisely, how derivatives can have a systematic attenuation bias (bias toward zero). We discuss how this tendency arises as a natural consequence of data smoothing, and give some indications of the typical size of the problem in applications.

It is important at the outset to consider the context of our investigation, as well as some of the methods we utilize. There is nothing novel or surprising about the existence of bias in nonparametric methods, as the treatment of bias is a standard element of nonparametric approximation theory. We are concerned here with the proposition that kernel regression, as applied, gives estimates of marginal effects that are far from the true values, and are systematically too small. For instance, if the true regression is linear and the regressors have a unimodal density, the derivatives are biased toward zero.
at every point. As such, we focus on the impact of smoothing in realistically sized samples. We make extensive use of Monte Carlo simulations, and to understand the impact of smoothing in isolation, we develop a "fixed bandwidth" approximation theory for kernel estimators. While most of our remarks are based on simple logic, our analysis differs from that of standard nonparametric approximation theory, and therefore may be somewhat controversial. Consequently, we devote Section 3 of the exposition to explaining the differences between our analysis and standard nonparametric theory.

We illustrate the derivative bias problem after introducing our notation. The observed data \((y_i, x_i), i=1, \ldots, N\) is assumed 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 an additive disturbance, the economic relationship of interest is the mean regression \(g(x) = E(y|x)\) of \(y\) on \(x\). The marginal effects of \(x\) on \(y\) are the derivatives \(g'(x) = \partial g(x)/\partial x\).

We focus on the standard (Nadaraya-Watson) kernel estimator of the regression \(g(x)\), namely

\[
\hat{g}(x) = \frac{\hat{c}(x)}{\hat{f}(x)},
\]

where the numerator is

\[
\hat{c}(x) = N^{-1}h^{-k} \sum_{i=1}^{N} K\left(\frac{x - x_i}{h}\right) y_i
\]

and the denominator is

\[
\hat{f}(x) = N^{-1}h^{-k} \sum_{i=1}^{N} K\left(\frac{x - x_i}{h}\right),
\]

the standard (Rosenblatt-Parzen) kernel estimator of the marginal density \(f(x)\) (c.f. Silverman(1986), Härdle(1991)). Here \(h\) is the bandwidth value that determines the extent of smoothing or local averaging, and \(K(\cdot)\) 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

\begin{equation}
\hat{g}'(x) = \frac{c'(x)}{f(x)} - \frac{f'(x)c(x)}{f(x)^2}
\end{equation}

From standard theory, the estimators $\hat{g}(x)$ and $\hat{g}'(x)$ are capable of measuring $g(x)$ and $g'(x)$ nonparametrically. In particular, under general conditions $\hat{g}(x)$ is a (pointwise) consistent estimator of $g(x)$ if $h \rightarrow 0$ and $Nh^k \rightarrow \infty$, and $\hat{g}'(x)$ is a (pointwise) consistent estimator of $g'(x)$ if $h \rightarrow 0$ and $Nh^{k+2} \rightarrow \infty$.

Our interest is in the accuracy of $\hat{g}'(x)$ as an estimator of $g'(x)$ in practice, where the bandwidth $h$ has been set. I originally encountered the bias problem in results from simulating kernel estimators of the average derivative $\delta = E(g')$, and this provides us with a useful starting point. One estimator is the average of the kernel regression derivatives, or

\begin{equation}
\hat{\delta} = N^{-1} \sum_{i=1}^{N} \hat{g}'(x_i) \hat{l}_i
\end{equation}

where $\hat{l}_i = 1[f(x_i) \geq b]$ is a trimming indicator that drops observations with small estimated density (used in the technical analysis of this estimator). Attenuation bias in $\hat{\delta}$ indicates an average bias in $\hat{g}'(x)$ over $x$ values. For later reference, we also present the (direct) IV estimator of the average derivative

\begin{equation}
\hat{d} = N^{-1} \left[ \sum_{i=1}^{N} \left( x'(x_i) \right)^T \hat{l}_i \right]^{-1} \left[ N^{-1} \sum_{i=1}^{N} \hat{g}'(x_i) \hat{l}_i \right]
\end{equation}

where $\hat{x}'$ is the derivative (matrix) of the kernel estimator of $E(x|x)$; namely (1.1) with $y_i$ replaced by $x_i$. This estimator is the sample analog of the expression $\delta = (E(\delta x/\delta x))^{-1} E(g')$.

Table 1 contains the means and standard deviations of $\hat{\delta}$ and $\hat{d}$ over 400 Monte Carlo
samples of size $N = 100$ for a linear model and a probit model. In each case there are $k = 4$ regressors $x_i$ that are normally distributed, with a normally distributed disturbance. The kernel $K(.)$ is the spherical normal density, the bandwidth value is set to $h = 1$ and the trimming bound is set to drop 1% of the observations. Ordinary least squares estimators are included for comparison.\(^6\)

A stark reflection of the bias problem of interest is given by the displayed means of $\hat{\delta}$, which are 46% of the true values with the linear model and 50% of the true values with the probit model. Since $\hat{\delta}$ is the average of the estimated marginal effects $\hat{g}'(x_i)$, the basic fact is that those effects are way too small.\(^7\) Moreover, the simulation design underlying Table 1 should favor good estimator performance; the regressors for each model are symmetrically distributed, and the linear model has $R^2 = .80$.

To get a clearer view of the kernel regression function and its derivative, in Figure 1.1 we plot the means and approximate 95% probability bands (± 1.96 standard deviation) for kernel regression and its derivative with a univariate linear model with normal regressor.\(^8\) The design has $R^2 = .80$, and again we have set the bandwidth $h = 1$. The flattened slope of the estimated regression is obvious, the mean of the derivative is approximately .50 over the whole range, and nowhere do the probability bands contain the true value $g'(x) = 1$.

Of course, it is natural to argue that $h = 1$ is a very large bandwidth for the univariate design, so bad measurement might be expected.\(^9\) Consequently, we consider the results from using optimal bandwidth values as listed in Tables 2 and 3 (to be discussed later). Figure 1.2 plots the results from using $h = .377$, which minimizes integrated mean square error of $\hat{g}'(x)$. While the true value $g'(x) = 1$ is now contained in the probability bands, the mean derivative is approximately .87 across the entire range of evaluation points. We also give a few results for $h = .249$, the bandwidth value that minimizes integrated mean squared error of $\hat{g}(x)$, again noting how the mean derivative is below the true value. These results illustrate how the attenuation bias in marginal effects can be
TABLE 1: AVERAGE DERIVATIVE SIMULATION RESULTS
(Means and Standard Deviations Over 400 Monte Carlo Samples)

(a) Linear Model: \( y = 1 + x_1 + x_2 + x_3 + x_4 + \epsilon \)

<table>
<thead>
<tr>
<th></th>
<th>( \hat{\delta}_1 )</th>
<th>( \hat{\delta}_2 )</th>
<th>( \hat{\delta}_3 )</th>
<th>( \hat{\delta}_4 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average Kernel</td>
<td>.465</td>
<td>.459</td>
<td>.461</td>
<td>.463</td>
</tr>
<tr>
<td>Derivative ( \hat{\delta} )</td>
<td>(.077)</td>
<td>(.075)</td>
<td>(.073)</td>
<td>(.076)</td>
</tr>
<tr>
<td>Direct IV Estimator ( \hat{d} )</td>
<td>1.001</td>
<td>.996</td>
<td>.997</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td>(.110)</td>
<td>(.108)</td>
<td>(.115)</td>
<td>(.110)</td>
</tr>
<tr>
<td>Ordinary Least Squares</td>
<td>1.001</td>
<td>.998</td>
<td>.997</td>
<td>.997</td>
</tr>
<tr>
<td></td>
<td>(.103)</td>
<td>(.103)</td>
<td>(.110)</td>
<td>(.104)</td>
</tr>
</tbody>
</table>

(b) Probit Model: \( y = 1[ 1 + x_1 + x_2 + x_3 + x_4 + \epsilon > 0 ] \)

<table>
<thead>
<tr>
<th></th>
<th>( \hat{\delta}_1 )</th>
<th>( \hat{\delta}_2 )</th>
<th>( \hat{\delta}_3 )</th>
<th>( \hat{\delta}_4 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average Kernel</td>
<td>.081</td>
<td>.080</td>
<td>.081</td>
<td>.082</td>
</tr>
<tr>
<td>Derivative ( \hat{\delta} )</td>
<td>(.018)</td>
<td>(.019)</td>
<td>(.019)</td>
<td>(.019)</td>
</tr>
<tr>
<td>Direct IV Estimator ( \hat{d} )</td>
<td>.176</td>
<td>.173</td>
<td>.176</td>
<td>.178</td>
</tr>
<tr>
<td></td>
<td>(.037)</td>
<td>(.038)</td>
<td>(.039)</td>
<td>(.037)</td>
</tr>
<tr>
<td>Ordinary Least Squares</td>
<td>.162</td>
<td>.162</td>
<td>.163</td>
<td>.164</td>
</tr>
<tr>
<td></td>
<td>(.033)</td>
<td>(.034)</td>
<td>(.035)</td>
<td>(.033)</td>
</tr>
</tbody>
</table>
FIGURE 1.1: DERIVATIVE BIAS WITH LINEAR MODEL, BANDWIDTH $h = 1$

400 Monte Carlo Samples of Size $N = 100$

Model: $y_i = 1 + x_i + \epsilon_i$, $x_i \sim N(0, 1)$, $\epsilon_i \sim N(0, .25)$, $R^2 = .80$

Selected Evaluation Points:

<table>
<thead>
<tr>
<th>$x$</th>
<th>$g'(x)$</th>
<th>$\hat{E}[g'(x)]$</th>
<th>Probability Band</th>
</tr>
</thead>
<tbody>
<tr>
<td>-2.0</td>
<td>1.0</td>
<td>.50</td>
<td>[.28, .71]</td>
</tr>
<tr>
<td>0.0</td>
<td>1.0</td>
<td>.49</td>
<td>[.36, .61]</td>
</tr>
<tr>
<td>1.0</td>
<td>1.0</td>
<td>.49</td>
<td>[.35, .63]</td>
</tr>
</tbody>
</table>
FIGURE 1.2: DERIVATIVE BIAS WITH LINEAR MODEL, BANDWIDTH $h = .377$

400 Monte Carlo Samples of Size $N = 100$, Design as in Figure 1.1

Selected Evaluation Points:

<table>
<thead>
<tr>
<th>$x$</th>
<th>$g'(x)$</th>
<th>$\hat{g}'(x)$</th>
<th>$\mathbb{E}[g'(x)]$</th>
<th>Probability Band [Lower, Upper]</th>
</tr>
</thead>
<tbody>
<tr>
<td>$h = .377$</td>
<td>-2.0</td>
<td>1.0</td>
<td>.87</td>
<td>[.02, 1.71]</td>
</tr>
<tr>
<td></td>
<td>0.0</td>
<td>1.0</td>
<td>.86</td>
<td>[.54, 1.19]</td>
</tr>
<tr>
<td></td>
<td>1.0</td>
<td>1.0</td>
<td>.87</td>
<td>[.44, 1.30]</td>
</tr>
<tr>
<td>$h = .249$</td>
<td>-2.0</td>
<td>1.0</td>
<td>.89</td>
<td>[-.71, 2.49]</td>
</tr>
<tr>
<td></td>
<td>0.0</td>
<td>1.0</td>
<td>.93</td>
<td>[.36, 1.51]</td>
</tr>
<tr>
<td></td>
<td>1.0</td>
<td>1.0</td>
<td>.93</td>
<td>[.22, 1.64]</td>
</tr>
</tbody>
</table>
uniform across the data range.

With this introduction, we now examine how such a systematic bias could arise. After characterizing kernel regression, we consider the case of a normal linear model, and our conclusions match the results of Figures 1.1 and 1.2 above. We then discuss how the bias arises and is manifested when the true model is nonlinear.

2. **Smoothing Bias in Kernel Regression**

2.1 **Fixed Bandwidth Posture and Some Assumptions**

Our analysis of the bias problem utilizes a large sample approximation to the distribution of the kernel regression \( \hat{g}(x) \) because it is a nonlinear combination of the sample averages \( \hat{c}(x) \) and \( \hat{f}(x) \). As mentioned above, we employ a different posture from the standard nonparametric asymptotic theory, where approximation is based on the bandwidth value shrinking (\( h \to 0 \)) as the sample size is increased. We regard the bandwidth as set to the value used in an application, treating the kernel estimator and its derivative as a simple combination of sample averages. We will see how this posture gives an accurate depiction of the size and nature of the derivative bias problem, and in Section 3 we discuss the relation between our "fixed bandwidth" theory and the standard approximation theory.

All of our results are based on straightforward manipulations of expectations and derivatives of the components of the kernel regression estimator. A sufficient set of regularity conditions for these manipulations are:

**Assumption 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 \( g(x) = \mathbb{E}(y|x) \) is continuously differentiable on \( \text{int}(S_f) \).
Assumption 2: The kernel $K(u)$ has convex support $S_K \subseteq \mathbb{R}^k$, with $K(u) > 0$ for $u \in \text{int}(S_K)$ and $K(u) = 0$ for $u \in \partial S_K$, the boundary of $S_K$. If $u \in S_K$ then $-u \in S_K$ and $K(u)$ is symmetric ($K(u) = K(-u)$) and continuously differentiable on $\text{int}(S_K)$.

Assumption 3: The integrals $\int K(u)f(x-hu)du$ and $\int \int K(u)yF(y,x-hu)dudy$ exist for $x \in S_K$ and are differentiable in $x$, with derivatives ($\int K(u)f(x-hu)du' = \int K(u)f'(x-hu)du$ and $(\int \int K(u)yF(y,x-hu)dudy)' = \int \int K(u)y \frac{\partial F}{\partial x}(y,x-hu)dudy$.

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 Smoothing and Errors-in-Variables

With all of these provisos out of the way, we denote the limit of $g(x)$ given $h$ as $\hat{g}(x)$. From Slutsky's Theorem and the weak law of large numbers,

\begin{equation}
\chi_h(x) = \frac{\text{plim} \hat{g}(x)}{\text{plim} \hat{f}(x)} = \frac{E[\hat{c}(x)]}{E[\hat{f}(x)]},
\end{equation}

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

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

\begin{equation}
E[\hat{c}(x)] = \int y \phi_h(y,x) \, dy
\end{equation}

where

\begin{equation}
\phi_h(y,x) = \int K(u) F(y,x-hu) \, du
\end{equation}

The function $\phi_h$ is clearly a density, in the form of a convolution. Further, note that if $u$ is distributed independently of $(y,x)$, 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 calculation gives

\begin{equation}
E[\hat{f}(x)] = \int K(u) f(x-hu)du \equiv \phi_h(x)
\end{equation}

where $\phi_h(z)$ is easily seen to be the marginal density of $z = x + hu$.\textsuperscript{11}
Combining (2.2) and (2.4) gives

\[ (2.5) \quad \gamma_h(x) = \frac{E[c(x)]}{E[f(x)]} = \frac{\int y \phi_h(y,x) dy}{\phi_h(x)} . \]

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

\[ (2.6) \quad \text{plim } \hat{g}'(x) = \gamma_h'(x) . \]

In words, the estimator \( \hat{g}(x) \) estimates the regression \( \gamma_h \) of \( y \) on \( x + hu \), and its derivative \( \hat{g}'(x) \) estimates the associated derivatives \( \gamma_h' \). As such, smoothing induces an "errors-in-variables" structure, causing the regression of \( y \) conditional on \( x + hu \) to be measured instead of the regression of \( y \) on \( x \). When the bandwidth \( h \) is tiny, these two regressions coincide. However, when \( h \) is finite, the comparison between \( g(x) \) and \( \gamma_h(x) \) gives rise to the derivative bias problem illustrated above.\(^\text{12}\)

We summarize these observations as

**Proposition 1:** Assume that the bandwidth \( h \) is a fixed value, and let \( \gamma_h(z) \) be the expectation of \( y \) given \( z = x + hu \). Given assumptions 1, 2 and 3, as \( N \to \infty \) we have that \( \text{plim } \hat{g}(x) = \gamma_h(x) \) and \( \text{plim } \hat{g}'(x) = \gamma_h'(x) \). Moreover, the limiting distribution of \( \sqrt{N} [\hat{g}(x) - \gamma_h(x)] \) is \( N(0, \sigma_h(x)^2) \), and the limiting distribution of \( \sqrt{N} [\hat{g}'(x) - \gamma_h'(x)] \) is \( N(0, \Sigma_h(x)) \), where \( \sigma_h(x)^2 \) is the variance of

\[ r(x;y_j,x_j) = \int [1/h \phi_h(x)] \|x-x_j\|/h \|y_j-\gamma_h(x)\|, \]

and \( \Sigma_h(x) \) is the covariance matrix of

\[ r^*(x;y_j,x_j) = [1/h^{k+1} \phi_h(x)] \]

\[ \left[ (K'((x-x_j)/h)-h\phi_h'(x)/\phi_h(x) K((x-x_j)/h)) [y_j-\gamma_h(x)] - h\gamma_h'(x) K((x-x_j)/h) \right]. \]
The statements on limiting distribution follow immediately from the Central Limit theorem and the delta method applied to \( \hat{g}(x) \) and \( \hat{g}'(x) \).\(^{13}\) While we do not make further use of the limiting distributions, they give justification for our method of constructing probability bands in our figures.

The "errors-in-variables" intuition is further strengthened by considering what the kernel regression \( \hat{g}(x) \) actually is. Suppose that \((Y,X)\) denotes the random vector distributed with the empirical distribution of the observations \((y_i, x_i), i=1,...,N\). If \( u \) is distributed with density \( K(u) \), independently of \( X \), then the function \( \hat{f}(x) \) is the density of \( X + hu \). Along the line of reasoning of Manski (1988), it is clear that the function \( \hat{g}(x) \) is the regression of \( Y \) conditional on \( X + hu \). Our "fixed bandwidth" approximation yields the analog of this structure in a large sample.

These connections permit us to study the impact of kernel smoothing by studying the associated "errors-in-variables" structure. We now develop the familiar results for linear models in our framework, and indicate how the derivative bias varies with sample size and the number of regressors.

### 2.3 Linear Models

For this section, assume that the true model is of the form:

\[
y = \alpha + \beta^T x + \epsilon\tag{2.7}
\]

where \( E(\epsilon|x) = 0 \), so that \( g(x) = \alpha + \beta^T x \) and \( g'(x) = \beta \). With \( z = x + hu \) and \( u \) distributed with density \( K(u) \) (independently of \( x \)), we have that

\[
y = \alpha + \beta^T (z - hu) + \epsilon\tag{2.8}
\]

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

where \( v = (\epsilon - h\beta^T[u - E(u|z)]) \) has mean zero conditional on \( z \). Therefore

\[
E(y|z) = \gamma_h(z) = \alpha + \beta^T [z - hE(u|z)]
\]
and the kernel regression \( \hat{g}(x) \) estimates \( \gamma_h(x) \). To be more specific we need to characterize the term \( hE(u|z) \).

Standard "errors-in-variables" bias formulae arise by assuming

**Assumption NK:** The kernel \( K(u) \) is the normal density, with mean 0 and covariance matrix \( I \).

and \( ^1 \)

**Assumption NR:** The regressors \( \{x_i, i=1,...,N\} \) are a random sample from a normal distribution with mean \( \mu \) and covariance matrix \( \Sigma \).

Under this structure, \( z = x + hu \) and \( u \) are joint normally distributed, which implies that \( hE(u|z) = (I-A_h^T)(z-\mu) \), where

\[
A_h = (\Sigma + h^2 I)^{-1} \Sigma
\]

Consequently, we have

**Proposition 2:** When the true regression model is the linear equation (2.7), under assumptions 1, 3, NK and NR, the kernel regression \( \hat{g}(x) \) estimates

\[
\gamma_h(x) = [\alpha + (\beta - A_h^T \mu)^T] + (A_h \beta)^T x;
\]

and the kernel regression derivative \( \hat{g}'(x) \) estimates

\[
\gamma_h'(x) = A_h \beta.
\]

Therefore, \( \hat{g}'(x) \) is a downward biased estimator of \( g'(x) = \beta \) 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 \( \beta \) and 0, with positive definite weights. In the case where \( \Sigma = \sigma^2 I \), we have \( A_h = (1-\nu_h) I \), where \( \nu_h = h^2/(\sigma^2 + h^2) \) is the familiar "noise/total variation" ratio for this problem.

The design for the average derivatives of Table 1 had \( \sigma^2 = 1 \), \( h = 1 \) and \( \beta = (1,...,1)^T \). The bias factor is \( (1-\nu_h) I = .5 I \), or that each component of \( \hat{g}'(x) \) estimates half the true coefficient, or .5. The same factor applies in the univariate case with \( h = \)

9
1, which coincides quite closely with the results of Figure 1.1. Figure 1.2 has $\sigma^2 = 1$
and $h = .377$, so that $\hat{g}(x)$ estimates a line with slope $1-\nu_h = .876$, which again is exactly
as depicted. For $h = .249$, $\hat{g}(x)$ estimates a line with slope $1-\nu_h = .942$, which is very
close to the results for $x = 0.0$ and $1.0$, although for $x = -2.0$, the mean of the estimated
slope is smaller than the predicted .942.

Proposition 1 stresses how the attenuation bias is determined by the bandwidth value
chosen. Consequently, to get a sense of the typical size of derivative bias, we need to
consider typical bandwidth values, for different sample sizes and different numbers of
regressors. Table 2 gives bandwidth values that minimize approximate (trimmed) integrated
squared error of $\hat{g}(x)$, for the linear model $y = 1 + \sum_j x_j + \epsilon$ with spherically normal
regressors and normal disturbance, and a normal kernel used for kernel regression. To
maintain comparability with different numbers of regressors, the variance of the
disturbance $\epsilon$ is set to maintain the $R^2$ values listed. Table 3 gives bandwidth values for
minimizing approximate mean squared error of the derivative $\hat{g}'(x)$ under the same
guidelines. The calculations underlying these bandwidth values are summarized in the
Appendix.

The depiction of bias given by Tables 2 and 3 is easy to summarize. First, with one
or two regressors, the bias values are fairly small, and decrease steadily with increases
in sample size. When there are more regressors, the bias values are considerably larger
and vanish less quickly with sample size, illustrating the "curse of dimensionality". The
impact of disturbance variance ($R^2$ from .8 to .2) is predictable, with larger bandwidths
for "noisier" designs. Comparing the optimal bandwidths for estimating derivatives in
Table 3 with those in Table 2 gives another reflection of dimensionality problems. In
particular, minimizing integrated mean squared error for derivatives involves larger
bandwidth choices, larger derivative bias, and bias that vanishes less quickly with
increases in sample size.

The values in Tables 2 and 3 depend on the linear design, and are at most suggestive
TABLE 2: DERIVATIVE BIAS: APPROXIMATELY OPTIMAL BANDWIDTHS
FOR ESTIMATING REGRESSION

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

<table>
<thead>
<tr>
<th>Dimension k</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>N = 100</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bandwidth h</td>
<td>0.249</td>
<td>0.367</td>
<td>0.475</td>
<td>0.571</td>
<td>0.657</td>
<td>0.971</td>
</tr>
<tr>
<td>Derivative Bias</td>
<td>5.84%</td>
<td>11.89%</td>
<td>18.38%</td>
<td>24.56%</td>
<td>30.15%</td>
<td>48.55%</td>
</tr>
<tr>
<td>N = 1,000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bandwidth h</td>
<td>0.157</td>
<td>0.250</td>
<td>0.342</td>
<td>0.428</td>
<td>0.509</td>
<td>0.824</td>
</tr>
<tr>
<td>Derivative Bias</td>
<td>2.41%</td>
<td>5.89%</td>
<td>10.45%</td>
<td>15.48%</td>
<td>20.56%</td>
<td>40.44%</td>
</tr>
<tr>
<td>N = 10,000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bandwidth h</td>
<td>0.099</td>
<td>0.170</td>
<td>0.246</td>
<td>0.321</td>
<td>0.394</td>
<td>0.699</td>
</tr>
<tr>
<td>Derivative Bias</td>
<td>0.97%</td>
<td>2.82%</td>
<td>5.70%</td>
<td>9.34%</td>
<td>13.43%</td>
<td>32.83%</td>
</tr>
<tr>
<td>N = 100,000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bandwidth h</td>
<td>0.063</td>
<td>0.116</td>
<td>0.177</td>
<td>0.241</td>
<td>0.305</td>
<td>0.593</td>
</tr>
<tr>
<td>Derivative Bias</td>
<td>0.39%</td>
<td>1.33%</td>
<td>3.03%</td>
<td>5.47%</td>
<td>8.51%</td>
<td>26.02%</td>
</tr>
</tbody>
</table>

Table 2b: $R^2 = .20$

<table>
<thead>
<tr>
<th>Dimension k</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>N = 100</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bandwidth h</td>
<td>0.434</td>
<td>0.583</td>
<td>0.705</td>
<td>0.807</td>
<td>0.894</td>
<td>1.184</td>
</tr>
<tr>
<td>Derivative Bias</td>
<td>15.83%</td>
<td>25.37%</td>
<td>33.21%</td>
<td>39.44%</td>
<td>44.42%</td>
<td>58.37%</td>
</tr>
<tr>
<td>N = 1,000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bandwidth h</td>
<td>0.274</td>
<td>0.397</td>
<td>0.508</td>
<td>0.605</td>
<td>0.692</td>
<td>1.005</td>
</tr>
<tr>
<td>Derivative Bias</td>
<td>6.96%</td>
<td>13.63%</td>
<td>20.48%</td>
<td>26.80%</td>
<td>32.39%</td>
<td>50.23%</td>
</tr>
<tr>
<td>N = 10,000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bandwidth h</td>
<td>0.173</td>
<td>0.271</td>
<td>0.365</td>
<td>0.454</td>
<td>0.536</td>
<td>0.852</td>
</tr>
<tr>
<td>Derivative Bias</td>
<td>2.89%</td>
<td>6.82%</td>
<td>11.77%</td>
<td>17.08%</td>
<td>22.31%</td>
<td>42.07%</td>
</tr>
<tr>
<td>N = 100,000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bandwidth h</td>
<td>0.109</td>
<td>0.184</td>
<td>0.263</td>
<td>0.340</td>
<td>0.415</td>
<td>0.723</td>
</tr>
<tr>
<td>Derivative Bias</td>
<td>1.17%</td>
<td>3.29%</td>
<td>6.46%</td>
<td>10.38%</td>
<td>14.69%</td>
<td>34.33%</td>
</tr>
</tbody>
</table>

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

$x \sim N(0, I)$, $\epsilon \sim N(0, I)$; Constant $R^2: \sigma^2_\epsilon = k(1/R^2 - 1)$

Optimal Bandwidth $h$ Formula in Appendix; Derivative Bias $\nu_h = h^2/(1+h^2)$
### TABLE 3: DERIVATIVE BIAS: APPROXIMATELY OPTIMAL BANDWIDTHS FOR ESTIMATING DERIVATIVES

<table>
<thead>
<tr>
<th>Dimension k</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>N = 100</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bandwidth h</td>
<td>0.377</td>
<td>0.505</td>
<td>0.612</td>
<td>0.706</td>
<td>0.789</td>
<td>1.087</td>
</tr>
<tr>
<td>Derivative Bias</td>
<td>12.47%</td>
<td>20.29%</td>
<td>27.27%</td>
<td>33.26%</td>
<td>38.37%</td>
<td>54.18%</td>
</tr>
<tr>
<td>N = 1,000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bandwidth h</td>
<td>0.272</td>
<td>0.378</td>
<td>0.474</td>
<td>0.561</td>
<td>0.640</td>
<td>0.942</td>
</tr>
<tr>
<td>Derivative Bias</td>
<td>6.87%</td>
<td>12.52%</td>
<td>18.35%</td>
<td>23.92%</td>
<td>29.06%</td>
<td>47.00%</td>
</tr>
<tr>
<td>N = 10,000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bandwidth h</td>
<td>0.195</td>
<td>0.284</td>
<td>0.367</td>
<td>0.445</td>
<td>0.519</td>
<td>0.815</td>
</tr>
<tr>
<td>Derivative Bias</td>
<td>3.68%</td>
<td>7.45%</td>
<td>11.87%</td>
<td>16.56%</td>
<td>21.23%</td>
<td>39.94%</td>
</tr>
<tr>
<td>N = 100,000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bandwidth h</td>
<td>0.141</td>
<td>0.213</td>
<td>0.284</td>
<td>0.354</td>
<td>0.421</td>
<td>0.706</td>
</tr>
<tr>
<td>Derivative Bias</td>
<td>1.94%</td>
<td>4.33%</td>
<td>7.47%</td>
<td>11.13%</td>
<td>15.06%</td>
<td>33.27%</td>
</tr>
</tbody>
</table>

**Table 3b: R^2 = .20**

<table>
<thead>
<tr>
<th>Dimension k</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>N = 100</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bandwidth h</td>
<td>0.561</td>
<td>0.714</td>
<td>0.833</td>
<td>0.932</td>
<td>1.015</td>
<td>1.293</td>
</tr>
<tr>
<td>Derivative Bias</td>
<td>23.93%</td>
<td>33.74%</td>
<td>40.98%</td>
<td>46.46%</td>
<td>50.75%</td>
<td>62.58%</td>
</tr>
<tr>
<td>N = 1,000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bandwidth h</td>
<td>0.404</td>
<td>0.535</td>
<td>0.645</td>
<td>0.740</td>
<td>0.823</td>
<td>1.120</td>
</tr>
<tr>
<td>Derivative Bias</td>
<td>14.01%</td>
<td>22.26%</td>
<td>29.39%</td>
<td>35.38%</td>
<td>40.41%</td>
<td>55.63%</td>
</tr>
<tr>
<td>N = 10,000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bandwidth h</td>
<td>0.291</td>
<td>0.401</td>
<td>0.499</td>
<td>0.588</td>
<td>0.668</td>
<td>0.970</td>
</tr>
<tr>
<td>Derivative Bias</td>
<td>7.78%</td>
<td>13.87%</td>
<td>19.97%</td>
<td>25.68%</td>
<td>30.85%</td>
<td>48.46%</td>
</tr>
<tr>
<td>N = 100,000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bandwidth h</td>
<td>0.209</td>
<td>0.301</td>
<td>0.387</td>
<td>0.467</td>
<td>0.542</td>
<td>0.840</td>
</tr>
<tr>
<td>Derivative Bias</td>
<td>4.19%</td>
<td>8.30%</td>
<td>13.01%</td>
<td>17.90%</td>
<td>22.69%</td>
<td>41.35%</td>
</tr>
</tbody>
</table>

Specification: As in Table 2.

Optimal Bandwidth $h$ Formula in Appendix; Derivative Bias $\nu_h = h^2/(1+h^2)$
of typical biases in more general settings. The effects of dimension, sample size, etc., that we have highlighted above are all in line with intuition. However, one cannot help but notice how many of the bias values are extremely large. As such, we are forced to conclude that derivative bias can be a serious problem in typical empirical settings.

The geometry of smoothing bias with a linear model serves to introduce ideas of value in the study of nonlinear models, as well as the impact of the density of the regressors. Comparing (2.7) and (2.9) shows that impact of smoothing is to replace the argument $x$ of the original model by $x - hE(u|x = x)$; which we refer to as the "argument shift" effect. When the regressor density is normal, the "argument shift" is always toward the mean of $x$, so that the limit of the kernel estimator is a flattened version of the original line as evident from Figures 1.1 and 1.2. If the density of $x$ were unimodal and symmetric, one would expect the "argument shift" to have the flattening effect more generally. In particular, $z = x + hu$ is positively correlated with $u$ by construction, and $E(\mathbb{E}(u|z)(z-\mu)) > 0$ implies that the "shift" is typically toward the mean $\mu$ of $x$.

We can derive precise formulations of the "argument shift" for an arbitrary regressor density when the kernel $k(.)$ is a normal density. Recall that we have denoted the density of $z = x + hu$ as $\phi_h'$, and define the (translation) score $\lambda_h$ of $\phi_h$ as

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

(2.13)

We can then show

**Proposition 3:** Under assumptions 1, 3 and NK, the "argument shift" is

$$hE(u|z) = h^2 \lambda_h(z).$$  

(2.14)

The proof is immediate, as
where the second equality follows from integration-by-parts (noting that the normal kernel implies \( K'(u) = -uK(u) \)) and the third equality follows from assumption 3.

An immediate corollary is

**Proposition 4:** When the true regression model is the linear equation (2.7), under assumptions 1, 3 and NK, the kernel regression \( \hat{g}(x) \) estimates

\[
\gamma_h(x) = \alpha + \beta^T [x - h^2 \lambda_h(x)] ;
\]

and the kernel regression derivative \( \hat{g}'(x) \) estimates

\[
\gamma_h'(x) = \beta^T [1 - h^2 (-\partial^2 \ln \phi_h / \partial x \partial x^T)] .
\]

This reduces to proposition 2 when the regressors are normally distributed, as \( \phi_h \) is then a normal density with mean \( \mu \) and covariance matrix \( \Sigma + h^2 I \), so that \( -\partial^2 \ln \phi_h / \partial x \partial x^T = (\Sigma + h^2 I)^{-1} \). More generally, the direction of the argument shift is determined by the sign of \( \lambda_h' \), or of the density derivative \( \phi_h' \). When \( \phi_h \) is unimodal and symmetric, the shift is always toward the mode (or mean), giving the flattening effect. Further, (2.17) shows how the (pointwise) direction of derivative bias is determined by the concavity properties of \( \phi_h' \). For instance, if \( \phi_h(x) \) is log-concave, then the derivative bias is downward at every point. This structure is implied if \( f(x) \) is log concave when \( K(u) \) is a normal density (c.f. Prekopa(1973, 1980)).

For more general base densities, these formulations indicate how downward derivative bias arises in areas of higher density. For instance, suppose that the regressor density
was a mixture of normals with disparate means, say a 50-50 mixture of $\mathcal{N}(-10,1)$ and $\mathcal{N}(10,1)$. For values near -10, the local normal structure would suggest attenuation bias as above (a "flattened" estimated regression), and likewise for values near 10. Since the kernel regression is continuous, these two "flattened" line segments would be connected by a steeper line segment for range of values strictly between -10 and 10, as predicted by the relation between the argument shift and the derivatives of $\phi_h$. Such structure is illustrated in Figure 2.1, where the simulation is based on a normal mixture density for the regressor, and simulated values of $E[f(x)] = \phi_h(x)$ are displayed. While the means are not as disparate as -10 and 10, the attenuation bias (for derivatives) in areas of higher density is plainly evident. The average bias of derivatives is clearly toward zero, since downward biases occur in higher density areas than where the upward biases occur.

2.4 Nonlinear Models

The connection to errors in variables permits a clear understanding of the source and magnitude of derivative bias when the true model is linear. The paucity of available results on nonlinear models with errors in variables suggests that general results will be difficult to establish, especially since our discussion of nonparametric estimation should include a wide range of possible nonlinear regression structures. However, we can get some insight into the structure of derivative bias by studying the errors-in-variables geometry in this more general setting.

For motivation, consider Figures 2.2 and 2.3, that show simulation results where the true regression $g(x)$ is a quadratic and cubic model respectively, and where the regressor is normally distributed as before. Relative to the previous section, we also graph the results of the regression after the "argument shift;" namely $\tilde{\gamma}(x)$ and $\tilde{\gamma}'(x)$, where

$$\tilde{\gamma}(z) = g[z - hE(u|z)]$$  

For each figure, a substantial downward derivative bias is evident. Further, for each figure the derivative $\tilde{\gamma}'(x)$ is a fairly close representation of the mean derivative $\gamma_h'(x)$. 

13
FIGURE 2.1: DERIVATIVE BIAS WITH LINEAR MODEL, NORMAL MIXTURE DESIGN

400 Monte Carlo Samples of Size N = 100, Bandwidth h = 1.0
Linear Model as in Figure 1.1, \( f(x) = 0.66 \mathcal{N}(-2,1) + 0.34 \mathcal{N}(2,1) \)

Selected Evaluation Points:

<table>
<thead>
<tr>
<th>( x )</th>
<th>( g'(x) )</th>
<th>( E[g'(x)] )</th>
<th>Probability Band [Lower, Upper]</th>
</tr>
</thead>
<tbody>
<tr>
<td>-4.0</td>
<td>1.0</td>
<td>0.49</td>
<td>[0.23, 0.74]</td>
</tr>
<tr>
<td>-2.0</td>
<td>1.0</td>
<td>0.54</td>
<td>[0.38, 0.70]</td>
</tr>
<tr>
<td>0.0</td>
<td>1.0</td>
<td>1.38</td>
<td>[1.02, 1.74]</td>
</tr>
<tr>
<td>0.25</td>
<td>1.0</td>
<td>1.48</td>
<td>[1.10, 1.87]</td>
</tr>
<tr>
<td>2.0</td>
<td>1.0</td>
<td>0.64</td>
<td>[0.40, 0.88]</td>
</tr>
<tr>
<td>4.0</td>
<td>1.0</td>
<td>0.48</td>
<td>[0.13, 0.82]</td>
</tr>
</tbody>
</table>
FIGURE 2.2: DERIVATIVE BIAS WITH QUADRATIC MODEL

400 Monte Carlo Samples of Size N = 100, Bandwidth h = 0.377

Model: \( y_i = x_i^2 + \epsilon_i, x_i \sim N(0,1), \epsilon_i \sim N(0, 0.25) \)

Selected Evaluation Points:

<table>
<thead>
<tr>
<th>x</th>
<th>( g'(x) )</th>
<th>( \tilde{g}'(x) )</th>
<th>( \hat{E}[g'(x)] )</th>
<th>Probability Band</th>
</tr>
</thead>
<tbody>
<tr>
<td>-2.0</td>
<td>-4.00</td>
<td>-3.06</td>
<td>-2.81</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>[-5.11, -0.50]</td>
</tr>
<tr>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>[-0.28, 0.27]</td>
</tr>
<tr>
<td>1.0</td>
<td>2.0</td>
<td>1.53</td>
<td>1.54</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>[0.89, 2.20]</td>
</tr>
</tbody>
</table>
FIGURE 2.3: DERIVATIVE BIAS WITH CUBIC MODEL

400 Monte Carlo Samples of Size N = 100, Bandwidth h = .377

Model: \( y_i = -x_i + .5 x_i^3 + \varepsilon_i, x_i \sim N(0,1), \varepsilon_i \sim N(0, .25) \)

Selected Evaluation Points:

<table>
<thead>
<tr>
<th>x</th>
<th>( g'(x) )</th>
<th>( \tilde{g}'(x) )</th>
<th>( E[g'(x)] )</th>
<th>Probability Band</th>
</tr>
</thead>
<tbody>
<tr>
<td>-2.0</td>
<td>5.00</td>
<td>3.15</td>
<td>3.20</td>
<td>[0.05, 6.35]</td>
</tr>
<tr>
<td>0.0</td>
<td>-1.00</td>
<td>-0.87</td>
<td>-0.71</td>
<td>[-1.03, -0.39]</td>
</tr>
<tr>
<td>1.0</td>
<td>.50</td>
<td>.13</td>
<td>.28</td>
<td>[-.11, .67]</td>
</tr>
</tbody>
</table>
= E[\hat{g}'(x)]. We now consider how these features can arise.

In particular, we have

**Proposition 5:** Under assumptions 1, 2 and 3 with a fixed bandwidth \( h \), the kernel regression \( \hat{g}(x) \) consistently estimates the function \( \gamma_h(x) \), where

\[
\gamma_h(z) = g[z-hE(u|z)] + Cg(z) .
\]

and

\[
Cg(z) = E(g[z-hE(u|z) - hw] - g[z-hE(u|z)])|z)
\]

for \( \omega = u - E(u|z) \). The kernel regression derivative \( \hat{g}'(x) \) estimates \( \gamma_h'(x) \), where

\[
\gamma_h'(z) = g'[z-hE(u|z)][1 - h\partial E(u|z)/\partial z] + Cg'(z) .
\]

This result follows from several immediate observations. First, since \( z = x + hu \) and \( g(x) = E(y|x) \), note that \( E(y-g(x)|z) = E_x[z|E(y-g(x)|x,z)] = 0 \), so that \( \gamma_h(z) = E[g(x)|z] \).

Now, if \( \omega = u - E(u|z) \), we can rewrite \( g(x) \) as

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

\[
= g[z - hE(u|z)] + [g[z-hE(u|z) - hw] - g[z-hE(u|z)]] .
\]

Equations (2.19-21) then follow immediately.

Proposition 5 gives a straightforward characterization of the impacts of smoothing. The first term of (2.19) represents the "argument shift" effect, namely to evaluate \( g \) at \( z - hE(u|z) \) instead of \( z \). The second term of (2.19), \( Cg(z) \), is a "curvature adjustment" that reflects averaging over nonlinearity in \( g \) in the vicinity of the shifted argument \( z - hE(u|z) \). "Averaging" refers to averaging over \( \omega \) values, which have mean zero given the value of \( z \). The formulation (2.21) writes the derivatives in terms of these two effects, which we now take up in turn.
In Figures 2.2 and 2.3, the "argument shift" effect works as in the linear case, namely it serves to "flatten" the function (when the regressors have unimodal density) with derivatives made smaller. While it is easy to suspect that this kind of impact would hold in many applications, we now develop the shifting effect further, to learn more about what affects attenuation bias in derivatives.

If we assume that the kernel $K$ is normal, then the structure of proposition 3 is applicable here, which we summarize as

**Proposition 6:** Under assumptions 1, 3 and NK, the "argument shift" components of the regression function and derivatives are expressible as

\[
\hat{y}(z) = g(z - h\mathbb{E}(u|z)) = g(z - h^2\lambda_h(z))
\]

and

\[
\hat{y}'(z) = [I - h^2(-\partial^2\ln \phi_h/\partial x\partial x^T)]g'(z - h^2\lambda_h(z)),
\]

where $\phi_h$ is the density of $z = x + hu$, and $\lambda_h = -\phi_h'/\phi_h$ is the location score of $\phi_h$.

This proposition connects argument shifting to the structure of the regressor density, as in proposition 5. Specializing to the case of normal regressors gives a result comparable to proposition 2, namely

**Proposition 7:** Under assumptions 1, 3, NK, and NR,

\[
\hat{y}(z) = g[A_h z + (I-A_h)\mu]
\]

and

\[
\hat{y}'(z) = A_h g'[A_h z + (I-A_h)\mu],
\]

where $A_h = (\Sigma + h^2I)^{-1}\Sigma$ as in (2.10).

The leading factors of (2.24) and (2.26) reflect the structure discussed for linear models, namely how the argument shift is toward the mean for unimodal designs, causing
"flattening" as in Figures 2.2 and 2.3. More generally, flattening occurs in regions of
greater density, with connecting areas displaying estimated derivatives with positive bias
(as in Figure 2.1).

However, a further difference from the linear case is evident from the second factors
of (2.24) and (2.26), namely how the regression derivative is evaluated at the shifted
argument, and not the original one. Specifically, comparing \( g'(x) \) to \( \tilde{g}'(x) = \lambda_h g'(\lambda_h x + (1-\lambda_h)\mu) \) of (2.26) involves the downweighting by \( \lambda_h \) and the fact that \( g' \) is not evaluated
at \( x \) but rather \( \lambda_h x + (1-\lambda_h)\mu \). When the true model is linear, \( g' \) is constant, and the
change in evaluation point has no impact.

When \( g \) has nonlinear structure, the alteration in the evaluation point of \( g' \) opens up
many possibilities for the direction of bias. For example one could envision a true
regression with many bumps and wiggles, with \( g'(x) \) a different sign from \( \tilde{g}'(x) \) for certain
values. This case could be eliminated for designs where the true marginal effects are
bounded in a positive (or negative) range. However, examples of more practical interest
can be derived where the change in the evaluation point complicates the pointwise
comparison of true derivatives and estimates.

For instance, Figure 2.4 displays the results of estimation when the true model is a
probit model (with \( \mu = 0 \)). The flattening of the estimated regression is clearly evident,
leading to smaller average derivatives (as in Table 1). However, the pointwise comparison
of derivatives is more complicated. Since the true function has small derivatives below
-2.0 and above 0.0, the flattening of the estimated regression mean gives it greater
derivatives for some \( x \) values near these points.

Further, it is possible to devise examples where the change in the evaluation point
effectively cancels the flattening effect of the argument shift over ranges of the data.
Suppose that with a single normal regressor and a normal kernel, the true regression \( g(x) \)
\( \equiv \ln (x) \) for large positive \( x \) values (ignoring the structure of \( g(x) \) for negative values,
or equivalently, ignoring the violation of assumption 1 caused if the normal distribution
FIGURE 2.4: DERIVATIVE BIAS WITH PROBIT MODEL

400 Monte Carlo Samples of Size N = 100, Bandwidth h = .377

Model: \( y_i = 1[1 + x_i + \epsilon_i], \ x_i \sim N(0,1), \ \epsilon_i \sim N(0,.25) \)

Selected Evaluation Points:

<table>
<thead>
<tr>
<th>x</th>
<th>( g'(x) )</th>
<th>( \hat{g}'(x) )</th>
<th>( E[g'(x)] )</th>
<th>Probability Band</th>
</tr>
</thead>
<tbody>
<tr>
<td>-2.0</td>
<td>.11</td>
<td>.23</td>
<td>.27</td>
<td>[0.00, .56]</td>
</tr>
<tr>
<td>-1.5</td>
<td>.48</td>
<td>.57</td>
<td>.50</td>
<td>[.15, .85]</td>
</tr>
<tr>
<td>-1.0</td>
<td>.80</td>
<td>.68</td>
<td>.55</td>
<td>[.25, .85]</td>
</tr>
<tr>
<td>-.5</td>
<td>.48</td>
<td>.37</td>
<td>.37</td>
<td>[.14, .59]</td>
</tr>
<tr>
<td>0.0</td>
<td>.11</td>
<td>.09</td>
<td>.14</td>
<td>[.02, .27]</td>
</tr>
<tr>
<td>1.0</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>[-.01, .02]</td>
</tr>
</tbody>
</table>
were truncated to positive values$^{20}$. From (2.26), such structure would give $\tilde{g}'(x) = A_h g'(A_h x) = A_h/A_h x = 1/x = g'(x)$ or that no bias arises from the argument shift effect. Figure 2.5 illustrates this kind of structure where $g(x) = \ln(x + 4.0)$ and one can see how the attenuation bias vanishes in the right tail.$^{21}$

At any rate, these kinds of examples show how interrelated structure on the regressor density and regression function can preclude a general result about the pointwise derivative bias for all kinds of nonlinear models. It is possible to establish results by artificially bounding the true function (say by parallel lines with positive slope), however, such results do not provide further insight without practical motivation, so we leave them for future research.

Similar remarks apply to the "curvature adjustment" terms $C_g(x)$ and $C_g'(x)$ of (2.20) and (2.21). It is valuable to note that these effects have been small in most of the figures presented (namely $C_g'(x) = \chi_h '(x) - \tilde{g}(x)$), but to rule them out in general at a point $x$ requires $g$ to be suitably smooth around the point $x - hE(u|z=x)$; the "curvature adjustment" reflects the impact of smoothing out bumps and wiggles.$^{22}$

We can isolate the general curvature adjustment in standard asymptotic bias formulae, and we do so in the next section. Here, we close this section by considering the case of normal regressors and a normal kernel, where all the relevant distributions can be specified. With reference to (2.20), $z$ and $hw = hu - hE(u|z)$ are joint normal and uncorrelated, and hence independent. It is easy to verify that the distribution of $hw$ conditional on $z$ is normal with mean 0 and variance $h^2 A_h$. While this states that the averaging in (2.20) is uniform across the sample, it is still not possible to further characterize the impact, without being specific about the structure of the regression $g$ in the vicinity of the point $x - hE(u|z=x)$.

This much structure does permit us to give a concrete characterization of the average smoothing bias in derivatives, as follows. Add one weak regularity condition
FIGURE 2.5: DERIVATIVE BIAS WITH LOGARITHMIC MODEL

400 Monte Carlo Samples of Size N = 100, Bandwidth h = .377

Model: \( y_1 = \ln(x_1 + 4) + \epsilon_1, x_1 \sim N(0,1), \epsilon_1 \sim N(0, .25) \)

Selected Evaluation Points:

<table>
<thead>
<tr>
<th>x</th>
<th>( g'(x) )</th>
<th>( \gamma'(x) )</th>
<th>( E[g'(x)] )</th>
<th>Probability Band [Lower, Upper]</th>
</tr>
</thead>
<tbody>
<tr>
<td>-2.0</td>
<td>.50</td>
<td>.39</td>
<td>.40</td>
<td>[-.28, 1.10]</td>
</tr>
<tr>
<td>-1.0</td>
<td>.33</td>
<td>.28</td>
<td>.27</td>
<td>[-.06, .62]</td>
</tr>
<tr>
<td>0.0</td>
<td>.25</td>
<td>.22</td>
<td>.22</td>
<td>[-.04, .49]</td>
</tr>
<tr>
<td>1.0</td>
<td>.20</td>
<td>.18</td>
<td>.16</td>
<td>[-.18, .51]</td>
</tr>
<tr>
<td>2.0</td>
<td>.17</td>
<td>.15</td>
<td>.18</td>
<td>[-.48, .85]</td>
</tr>
</tbody>
</table>
**Assumption ND:** For any normally distributed variable $v$ and vector $\zeta$, we have
\[
\frac{\partial}{\partial \zeta}(E_v[g(v+\zeta)]) = E_v[g'(v+\zeta)].
\]

Our result is then

**Proposition 8:** Under assumption 1, 3, NR, NK and ND, we have that

\[
(2.27) \quad E_x[\gamma'_h(x)] = A_h E_w[g'(w)]
\]

where $w \sim N(\mu, \Sigma - h^2 A_h^T A_h)$.

This proposition follows directly from the normality and independence of $z$ and $hw = hu - hE(u|z)$ discussed above. The decomposition (2.22) is written as

\[
(2.28) \quad \gamma_h(z) = E_\omega[g(z - E(hu|z) - hw)] = E_\omega[g(A_h^T z + (I-A_h^T)\mu - hw)]
\]

so that assumption ND implies $\gamma'_h(z) = A_h E_\omega[g'(A_h^T z + (I-A_h^T)\mu - hw)]$. Since $hw$ is independent of the argument $z$, by evaluating at $z = x$ and taking expectations we have

\[
(2.29) \quad E_x[\gamma'_h(x)] = A_h E_x E_\omega[g'(A_h^T x + (I-A_h^T)\mu - hw)] = A_h E[g'(w)].
\]

where $w = A_h^T x + (I-A_h^T)\mu - hw$. $w$ is normally distributed with mean $\mu$ and covariance matrix $A_h^T \Sigma A_h + h^2 A_h^T A_h = \Sigma - h^2 A_h^T A_h$. This verifies (2.27).

Consequently, the average bias in derivatives is characterized by proportional downweighting by $A_h$, and averaging the true derivative $g'(x)$ over a more compact normal design ($N(\mu, \Sigma - h^2 A_h^T A_h)$ instead of $N(\mu, \Sigma)$). The demonstration above indicates how to isolate the average impacts of the "argument shift" and "curvature adjustment" effects.

In particular, the "argument shift" gives average derivative bias of

\[
(2.30) \quad E_x[\tilde{\gamma}'(x)] = A_h E_{\tilde{w}}[g'(\tilde{w})]
\]

where $\tilde{w} \sim N(\mu, A_h^T \Sigma A_h)$, and the "curvature adjustment" is the remainder.
\[(2.31) \quad E_x(Cg') = A_h (E_w[g'(\tilde{w})] - E_w[g'(w)]) \]

where \( w \sim N(\mu, A_h T A_h + h^2 A_h) \) as above. Consequently, each effect involves downweighting by \( A_h \), and the effects of the spread of the data on the averaging of derivatives.\(^{23}\)

A couple more specific examples illustrate these points.

**Example 1 (Quadratic Model):** Consider a univariate problem where \( x \sim N(\mu, \sigma^2) \), and the true regression is quadratic

\[(2.32) \quad g(x) = \alpha + \beta x + \beta_1 x^2 \]

Let \( \mathcal{K} \) be the standard normal density, and recall that \( A_h = 1 - \nu_h, \nu_h = h^2/(\sigma^2 + h^2) \), the "noise/total variation" ratio. The limit \( \gamma_h(x) \) of \( \hat{g}(x) \) is

\[(2.33) \quad \gamma_h(x) = [\alpha + \beta \nu_h \mu + \beta_1 \nu_h^2 \mu^2 + \beta_1 h^2 (1 - \nu_h)]
+ [\beta (1 - \nu_h) + 2 \beta_1 \nu_h (1 - \nu_h) \mu] x + [\beta_1 (1 - \nu_h)^2] x^2 . \]

The limit \( \gamma_h(x) \) is comprised of the "argument shift" term \( \tilde{\gamma}(z) = g((1 - \nu_h)z + \nu_h \mu) \) evaluated at \( z = x \), or

\[(2.34) \quad \tilde{\gamma}(x) = [\alpha + \beta \nu_h \mu + \beta_1 \nu_h^2 \mu^2] + [\beta (1 - \nu_h) + 2 \beta_1 \nu_h (1 - \nu_h) \mu] z + [\beta_1 (1 - \nu_h)^2] z^2 \]

and the "curvature adjustment" term \( Cg(x) \), namely

\[(2.35) \quad Cg(x) = \beta_1 h^2 (1 - \nu_h) , \]

which is constant over \( x \). In this case \( \hat{g}'(x) \) estimates

\[(2.36) \quad \gamma_h'(x) = \beta (1 - \nu_h) + 2 \beta_1 \nu_h (1 - \nu_h) \mu + 2 [\beta_1 (1 - \nu_h)^2] x = \tilde{\gamma}'(x) \]

which is determined solely by the "argument shift" term \( \tilde{\gamma}(x) \), and differs from \( g'(x) = \beta + \)
2β₁x by terms that depend on the relative position of the x density (through μ) and the relative amount of smoothing (through νₗ). To verify Proposition 8, we see that

(2.37) \( E[γₗ'(x)] = β(1-νₗ) + 2β₁(1-νₗ) = (1-νₗ)[β + 2β₁μ] \)

\[ = (1-νₗ)[g'(w)] = (1-νₗ)[g'(x)] \]

Here, since the derivative of g is linear in x, the fact that (2.27) involves averaging over a more compact normal distribution is inconsequential, and we have that the derivatives are biased downward on average with the same factor that applies to a linear model, namely \(1-νₗ\). The same conclusion applies in multivariate settings: it is easy to verify that \(E[γₗ'(x)] = AₗE[g'(x)]\) when the true model is quadratic.

Example 2 (Probit Model): Consider the univariate normal setting as in Example 1, but where the true model is a probit model, or

(2.38) \( y = I[ε < α + βx] \)

where \(I[\ ]\) is the indicator function, and \(ε \sim N(0,1)\), independent of x. If \(ψ\) denotes the standard normal c.d.f. and \(ψ\) the standard normal density, then we have

(2.39) \( g(x) = ψ(α + βx) \)

\( g'(x) = ψ(α + βx) \).

To derive \(γₗ(z) = E(y|z)\), we insert \(z - hu\) for x into \(g(x)\), giving

(2.40) \( y = I[ ε + βh(u-E(u|z)) < α + β(z - hE(u|z))] \)

\[ = I[ η < α + βνₗμ + β(1-νₗ) z] , \]

where \(η = ε + bh(u-E(u|z))\), again using that \(hE(u|z) = νₗ(z-μ)\). The variable \(η\) is distributed normally, independently of \(z\), with variance \(1 + β²ₗ²(1-νₗ)\). Therefore, if
(2.41) \[ c_1 = [1 + \beta^2 \nu \sigma^2]^{1/2} \]

then we have that \( \hat{g}(x) \) estimates

(2.41) \[ \gamma_h(x) = \psi( c_1^{-1} [\alpha + \beta \nu \mu + \beta(1-\nu) x] ) \]

and that \( \hat{g}'(x) \) estimates

(2.42) \[ \gamma_h'(x) = \beta (1-\nu_h) c_1^{-1} \psi( c_1^{-1} [\alpha + \beta \nu \mu + \beta(1-\nu) x] ) \]

Consequently, the bias in derivatives arises from the downweighting appropriate for a linear model \((1-\nu_h)\), together with the scaling by \(c_1\).

The "argument shift" terms omit \(c_1\): namely

(2.43) \[ \tilde{\gamma}(x) = \psi( \alpha + \beta \nu \mu + \beta(1-\nu) x \) \]

(2.44) \[ \tilde{\gamma}'(x) = \beta (1-\nu_h) \psi( \alpha + \beta \nu \mu + \beta(1-\nu_h) x \) \]

and the "smoothing adjustment" terms are just the differences \(Cg_h = \gamma_h - \tilde{\gamma}\) and \(Cg' = \gamma_h' - \tilde{\gamma}'\).

The average derivatives reflect the same downweighting and rescaling: \( \psi \) Define

(2.44) \[ C_0 = [1 + \beta^2 \sigma^2]^{1/2} \]

(2.44) \[ C_1 = [1 + \beta^2 \nu \sigma^2]^{1/2} \]

(2.44) \[ C_2 = [1 + \beta^2 \nu \sigma^2]^{1/2} \]

then we have that

(2.45) \[ E[g'(x)] = \beta C_0^{-1} \psi( C_0^{-1} [\alpha + \beta \mu] ) \]

(2.45) \[ E[\gamma_h'(x)] = \beta (1-\nu_h) C_1^{-1} \psi( C_1^{-1} [\alpha + \beta \mu] ) \]
\[ E[\tilde{\gamma}'(x)] = \beta (1-\nu_h) \ C_2^{-1} \psi( \ C_2^{-1} [\alpha + \beta\mu] ) \]

It is easy to verify identical formulations for a multivariate design: namely if \( x \) is a \( k \)-variate normal vector, then \( \beta (1-\nu_h) \) is replaced by the downweighted coefficient vector \( A_h\beta \) as in a linear design, with scalar factors \( c_1, \ C_0, \ C_1 \) and \( C_2 \) defined analogously to (2.44).

3. **Relation to Standard Asymptotic Theory**

Our arguments are based on a "fixed bandwidth" analysis of bias of nonparametric regression estimators. This posture differs from standard nonparametric asymptotic theory, which is based on limits as the approximation is sharpened (\( h \to 0 \)) and as sample size increases (\( N \to \infty \)). Since the latter formulation has attracted much recent attention, we now clarify some differences between our analysis and the more familiar theory.\(^{25}\)

First off, the standard theory is the proper one for answering questions of how flexible a technique is, which is an important question for choosing a method as part of an approach to empirical analysis. In particular, the fact that \( \hat{g}(x) \) and \( \hat{g}'(x) \) are pointwise consistent estimators of \( g(x) \) and \( g'(x) \) as \( N \to \infty \) and \( h \to 0 \) at appropriate rates, states that \( \hat{g}(x) \) and \( \hat{g}'(x) \) are capable of arbitrarily fine accuracy in the measurement of \( g(x) \) and \( g'(x) \) in sufficiently large samples. This is the central issue of a nonparametric approach, namely to choose methods that do not impose restrictions on statistical measurement at the outset. Rates of convergence and other depictions of nonparametric accuracy give refined answers to the questions of inherent flexibility of a particular nonparametric method.

Our analysis takes the position that a "fixed bandwidth" posture is more accurate than the standard theory for a different set of questions, namely how to study the distribution of a kernel regression estimator in an application. In particular, once a bandwidth value has been set in an application, the regression estimator \( \hat{g}(x) \) will have
characteristics that are determined by that bandwidth value, and our approach has been to focus on those characteristics. This is true regardless of whether the bandwidth value has been set optimally (say minimizing mean squared error) or not. Moreover, if the bandwidth value were set as a function of the data, a full analysis would involve discussing bias as we have done, together with the impact of sampling variation in the chosen bandwidth value. Our posture is that the special bandwidth value \( h = 0 \) (the limit in standard theory) does not give an especially informative analysis relative to one based on the bandwidth value that has been actually set in computing \( \hat{g}(x) \). When the chosen bandwidth value is, in fact, tiny, then "fixed" and "shrinking" bandwidth approximations coincide; otherwise the relevance of the standard "shrinking" theory is open to question.

Because the kernel regression \( \hat{g}(x) \) is a nonlinear combination of sample averages; 
\[
\hat{g}(x) = \frac{c(x)}{f(x)} \text{ of } (1.1)-(1.3),
\]
we have used large sample (\( N \to \infty \)) results to study its bias properties, holding the bandwidth \( h \) fixed. It is useful to spell out how this large sample approximation works, to clarify the differences with the standard theory, and explain the connection to standard asymptotic bias results. In particular, we have

\[
(3.1) \quad \hat{g}(x) - g_h(x) = \frac{1}{E[f(x)]} (c(x) - E[c(x)]) - \frac{\gamma_h(x)}{E[f(x)]} (\hat{f}(x) - E[\hat{f}(x)]) + R_1
\]

with 
\[
R_1 = \frac{(\hat{g}(x) - g_h(x)) (\hat{f}(x) - E[\hat{f}(x)])}{E[\hat{f}(x)]}
\]

Our large sample posture omits \( R_1 \), on the grounds that products of deviations of sample averages from their means are of smaller order than the deviations themselves, as commonplace in central limit theory for sample averages. Standard nonparametric theory considers the deviations

\[
(3.2) \quad \hat{g}(x) - g(x) = \frac{1}{f(x)} (\hat{c}(x) - c(x)) - \frac{g(x)}{f(x)} (\hat{f}(x) - f(x)) + R_2
\]
and omits $R_2$ on the grounds that products of deviations of $\hat{c}(x)$ and $\hat{f}(x)$ from their nonparametric limits are of smaller order than the deviations themselves. This illustrates the difference in posture, and we remark only that there is no reason from theory alone to expect that $R_2$ will be smaller than $R_1$. In terms of bias, our analysis is based on studying $\gamma_h(x)$ directly, noting that the deviations in the leading terms of (3.1) have mean 0. Standard asymptotic bias analysis focuses on the expectations of the leading terms ($\hat{c}(x)-c(x)$ and $\hat{f}(x)-f(x)$) of (3.2). In particular, these terms are approximated by a Taylor series in $h$ around $h = 0$, and the leading terms of those series constitute the bias approximation of the standard theory. Again, we only remark that there is no reason from theory to regard this approximation as more accurate than one based on the chosen bandwidth value (and we have presented several examples where the opposite is true). In addition, it is important to note that this entire discussion could have been based the regression derivative estimator $\hat{g}'(x)$, which involves more complicated formulae but exactly the same issues.

There is certainly nothing objectionable about the standard mathematics of asymptotic bias, given that its proper context is understood. As such, we examine the standard bias in a bit more detail, to shed light on our analysis. Suppose that $x$ is a scalar random variable, then by deriving the leading term of the Taylor series at $h = 0$, it is easy to see that the asymptotic bias of $\hat{g}(x)$ is

\begin{equation}
\text{AB}[\hat{g}(x)] = h^2 \left[ \int u^2 K(u) du \left[ - g'(x) \ell(x) + \frac{g''(x)}{2} \right] \right]
\end{equation}

where $\ell(x) = -f''(x)/f(x)$ is the score of the density of $x$. This term reflects the division of bias into "argument shift" and "curvature adjustment" terms coinciding with our analysis above. Perhaps the easiest way to see this is to assume the kernel is standard normal (so $\int u^2 K(u) du = 1$) and examine the "argument shift" (2.23) of
Proposition 6. From the mean value theorem, we have that

\[ (3.4) \quad \tilde{g}(x) - g(x) = g(x - h^2 \lambda_h(x)) - g(x) = -(\lambda_h(x) g'(\tilde{x})) h^2 \]

where \( \tilde{x} \) is between \( x \) and \( x - h^2 \lambda_h(x) \). Evaluating the coefficient of \( h^2 \) at \( h = 0 \) implies \( \lambda_h(x) \equiv \ell(x) \) and \( g'(\tilde{x}) \equiv g'(x) \), giving the first term of the asymptotic bias above. The "curvature adjustment" term is associated with \( h^2 g''(x)/2 \): for tiny bandwidths, the relevant nonlinearity for this adjustment is given by the second derivative of \( g(x) \).

Similarly, the standard asymptotic bias of the derivative \( \hat{g}'(x) \) is easily seen to be

\[ (3.5) \quad AB[\hat{g}'(x)] = \frac{h^2}{2} \left[ \int (u^2 \mathcal{K}(u) du \right] \left[ - \frac{\partial [g'(x) \ell(x)]}{\partial x} + \frac{g''(x)}{2} \right] \]

We have again split the terms into "argument shift" and "curvature adjustment" effects, noting how the nonlinearity relevant for the curvature adjustment is given by the third derivative of \( g \). Several further connections with our earlier formulae can be illustrated here; for instance, suppose that the true model is linear, \( g(x) = \alpha + \beta x \), with a normal regressor and a normal kernel. Here \( g'(x) = \beta \), \( g''(x) = 0 \) and \( \ell(x) = \sigma^2 (x - \mu) \).

Therefore, the asymptotic bias \( AB[\hat{g}'(x)] \) is \(- (h^2/\sigma^2) \beta \). Our exact bias for this case is \(- h \beta = - [h^2/(\sigma^2 + h^2)] \beta \), and if we expand it in \( h \) at \( h = 0 \), then the leading term is \(- (h^2/\sigma^2) \beta \), as expected. Here the asymptotic bias formula overstates the true downward derivative bias.

As the leading term of a series approximation, the asymptotic bias formulae can be expected to give accurate depictions of the derivative bias with fairly small bandwidths. For instance, in the above example with \( \sigma^2 = 1 \), \( h = 1 \) implies a true derivative bias of \(- .5 \beta \) and asymptotic bias approximation of \(- \beta \), whereas \( h = .377 \) implies \(- .124 \beta \) and \(- .142 \beta \) respectively, and \( h = .249 \) implies \(- .058 \beta \) and \(- .062 \beta \).

It is possible that the asymptotic bias formulae will be accurate for larger bandwidths if the order of the leading term is increased, say to \( h^4 \) or \( h^8 \). It is well known that this kind of increase can occur by using positive and negative local weighting,
namely a higher order kernel. In particular, if a higher order kernel is used; say $K^*$ where $\int u^j K^*(u) du = 0$ for $j < J$, $\int u^J K^*(u) du \neq 0$ for $j = J$, then the leading term of the series expansion is of order $h^J$. For $J > 2$, as $h \to 0$ this term converges more quickly to zero than $h^2$, and under standard methods will permit faster rates of convergence of $\hat{g}(x)$ to $g(x)$, provided $g(x)$ is sufficiently smooth. While this is no assurance that the finite sample bias will be smaller for a given value of $h$, we briefly consider the results of using higher order kernels, deferring a more detailed study to future research.

There are many ways to construct higher order kernels, and we use the method of taking differences of normal densities recently discussed by Robinson (1987). In particular, for $J = 2$ and $j \leq J$, let $\psi_j$ denote a $N(0, \sigma_j^2)$ density, with $\sigma_1 = 1$ and $\sigma_j \neq \sigma_j'$ for $j \neq j'$. Then a kernel of order $2J$ is given by $K^* = [\psi_1 + \sum \tau_j \psi_j] / [1 + \sum \tau_j]$, where $\tau_2$, $\ldots$, $\tau_J$ are solved from the equations $\int u^j K^*(u) du = 0$, for $j' < 2J$. For our calculations, we use $J = 4$ normal densities to construct a kernel of order 8, namely

$$K^* = 1.600 [\psi_1 - .5 \psi_2 + .14286 \psi_3 - .01786 \psi_4]$$

where $\psi_2$, $\psi_3$, $\psi_4$ have standard deviations 2, 3, 4 respectively. We also display some summary results from the kernel of order 4 given as $K^{**} = 1.333 [\psi_1 - .25 \psi_2]$, constructed in a similar fashion.

Figures 3.1 and 3.2 illustrate the results from using the higher order kernel for bandwidth values $h = 1$ and $h = .377$, and so they are comparable to Figures 1.1 and 1.2. Taking Figure 3.2 first, we see a finite sample reflection of the asymptotic bias-variance tradeoff associated with higher order kernels; namely there is no derivative bias but substantially larger variance. For the larger bandwidth $h = 1$ of Figure 3.1, there is a considerably larger variance but the downward derivative bias is still very much in evidence. On the relevance of $h = 1$, it can be regarded as indicative of bias in higher dimensional problems, namely where $h = 1$ is set using the product kernel $\Pi K^*(u_j)$, which is of (multivariate) order 8. The summary results for the kernel of order 4 illustrate
**FIGURE 3.1:** DERIVATIVE BIAS WITH ORDER 8 KERNEL, BANDWIDTH h = 1.00

400 Monte Carlo Samples of Size N = 100, Design as in Figure 1.1

Selected Evaluation Points:

<table>
<thead>
<tr>
<th>x</th>
<th>g'(x)</th>
<th>E[g'(x)]</th>
<th>Probability Band [Lower, Upper]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kernel Order 8</td>
<td>-2.0</td>
<td>1.0</td>
<td>1.18 [0.66, 1.70]</td>
</tr>
<tr>
<td></td>
<td>-1.5</td>
<td>1.0</td>
<td>0.79 [0.52, 1.06]</td>
</tr>
<tr>
<td></td>
<td>0.0</td>
<td>1.0</td>
<td>0.59 [0.44, 0.74]</td>
</tr>
<tr>
<td></td>
<td>1.0</td>
<td>1.0</td>
<td>0.66 [0.47, 0.84]</td>
</tr>
<tr>
<td>Kernel Order 4</td>
<td>-2.0</td>
<td>1.0</td>
<td>0.82 [0.49, 1.16]</td>
</tr>
<tr>
<td></td>
<td>-1.5</td>
<td>1.0</td>
<td>0.65 [0.44, 0.87]</td>
</tr>
<tr>
<td></td>
<td>0.0</td>
<td>1.0</td>
<td>0.56 [0.42, 0.70]</td>
</tr>
<tr>
<td></td>
<td>1.0</td>
<td>1.0</td>
<td>0.59 [0.43, 0.75]</td>
</tr>
</tbody>
</table>
FIGURE 3.2: DERIVATIVE BIAS WITH ORDER 8 KERNEL, BANDWIDTH $h = 0.377$

400 Monte Carlo Samples of Size $N = 100$, Design as in Figure 1.1

<table>
<thead>
<tr>
<th>$x$</th>
<th>$g'(x)$</th>
<th>$E[g'(x)]$</th>
<th>Probability Band [Lower, Upper]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kernel</td>
<td>-2.0</td>
<td>1.0</td>
<td>1.81</td>
</tr>
<tr>
<td>Order 8</td>
<td>-1.5</td>
<td>1.0</td>
<td>1.06</td>
</tr>
<tr>
<td></td>
<td>0.0</td>
<td>1.0</td>
<td>0.98</td>
</tr>
<tr>
<td></td>
<td>1.0</td>
<td>1.0</td>
<td>1.01</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$x$</th>
<th>$g'(x)$</th>
<th>$E[g'(x)]$</th>
<th>Probability Band [Lower, Upper]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kernel</td>
<td>-2.0</td>
<td>1.0</td>
<td>1.73</td>
</tr>
<tr>
<td>Order 4</td>
<td>-1.5</td>
<td>1.0</td>
<td>1.05</td>
</tr>
<tr>
<td></td>
<td>0.0</td>
<td>1.0</td>
<td>0.94</td>
</tr>
<tr>
<td></td>
<td>1.0</td>
<td>1.0</td>
<td>0.98</td>
</tr>
</tbody>
</table>
similar features, although the bias is not as effectively eliminated as with the kernel of order 8.

While we have only touched on the use of higher order kernels for bias reduction, a preliminary conclusion is that such estimators are not immune to smoothing bias problems. While the derivative bias is small for fairly small bandwidths, the cost in pointwise variance is absolutely enormous. Consequently, while one can see a reflection of the asymptotic structure in finite sample results, higher order kernels do not offer a panacea here, without substantively further analysis.

4. Concluding Remarks

In this paper we have discussed various aspects of the finite sample behavior of kernel regression estimators. Our results should not be regarded as an overt criticism of kernel estimators. On the contrary, estimators based on local averages are well designed for detecting bumps, troughs and other qualitative kinds of nonlinear structure, that other nonparametric estimators (such as truncated polynomial series) can miss. However, we have argued against the incautious use of marginal effects or derivatives estimated by local smoothing. In several leading cases, such as a linear model with regressors drawn from a unimodal density, marginal effects contain a systematic downward bias. That bias can be substantial when "typical" bandwidth values are used, even in realistically large samples.

We have departed from the standard nonparametric asymptotic theory for kernel estimators to study finite sample bias. In particular, we have used a "fixed bandwidth" approximation, focusing on bias for the bandwidth value actually set to compute the estimator. The structure of kernel regression is immediately interpretable under this posture, establishing a connection between smoothing and errors-in-variables structure. Much of our analysis is equally applicable to nonlinear errors-in-variables problems as to the impacts of regression smoothing.

We have focused on kernel regression in part because its structure is simple enough
to obtain some informative results on finite sample properties. Our focus should not be construed as suggesting that kernel regression has mismeasurement problems that are "worse" than other nonparametric estimators. On the contrary, to implement any nonparametric method in a finite sample involves approximation, and every nonparametric method will miss certain types of structure. Future research should be directed to cataloging what features are well measured, and which are missed, by the myriad of nonparametric methods that have been proposed. For instance, what structure is missed by a polynomial regression that only includes terms up to degree 3?

Over the past decade the theory of nonparametric and semiparametric methods has undergone spectacular development. 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 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

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

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

\[
\text{AV} \hat{g}_h(x) = N^{-1} h^{-k} \int K(u)^2 du \sigma^2(x) / f(x)
\]

\[
\text{AB}(\hat{g}_h(x)) = (h^2/2) \text{Trace}(\sigma^2 g / \partial x \partial x^T + 2 f'(x)/f(x)g'(x)^T) \int uu^T K(u) du
\]

where \( \sigma^2(x) = \text{Var}(y|x) \). For the linear model (2.7) 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

\[
C_1 = \int K(u)^2 du \int w(x) dx \sigma^2 = (4\pi)^{-k/2} \int w(x) dx \sigma^2
\]

\[
C_2 = \int (\sum x_j)^2 w(x) f(x) dx
\]

The optimal bandwidth value is then given as

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

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

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

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 \( \sigma^2 = k(1/R^2 - 1) \), the constants \( C_1 \) and \( C_2 \) are then solved for as

\[
C_1 = \pi^{-k/2} c_k^k k(1/R^2 - 1); \quad C_2 = k(0.95)^{(k-1)/k} \left( \int_{-c_k}^{c_k} x_j^2 \psi(x_j)dx_j \right)^k
\]

where \( \psi(.) \) is the standard normal density. These expressions are inserted into (A.1) for the bandwidths of Table 2.

A similar calculation gives the optimal bandwidth for derivative estimation, where we again use \( .95 \) weighting as above. The resulting formula is

(A.2) \[ h = \left[ (k+2)\pi^{-k/2} c_k^k k(1/R^2 - 1)/7.61 \right]^{1/(k+6)} N^{-1/(k+6)}, \]

which is used to calculate the bandwidth values in presented in Table 3.
Notes

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

2 See Prakasa-Rao(1983) and Härdle(1991) for references to the relevant statistical literature on regression estimation, and Bierens (1987) and Delgado and Robinson (1991) for an extensive survey of recent work in econometrics.

3 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 $X$.

4 Average derivatives coincide with coefficient parameters in semiparametric index models, c.f. Stoker (1992a) for discussion and references.

5 Stoker(1991) points out how $d$ can be written as an instrumental variables estimator of the slope coefficients from the equation $y_i = c + x_i d + u_i$, which explains the terminology. This same reference shows the first-order equivalence of $\delta$ and $d$ under standard theory. The conditions also employ "higher-order" kernels, and many other smoothness conditions that are not the main focus in the analysis of this paper.

6 Brillinger (1983) and Stoker (1986b) point out how the OLS estimators consistently estimate the average derivative when the regressors are normal distributed for any model, such as the probit model.

7 If $x$ and $y$ represented log-inputs and log-outputs, the linear design for $g(x)$ represents a Cobb-Douglas model, and the estimated elasticities $g'(x_i)$ are 46% of their true values on average. Our calculations in Section 2.3 predict values of 49.5% here (pointwise value of 50% combined with the 1% trimming).

8 Specifically, for each Monte Carlo sample, we compute the kernel estimator and its derivative for a grid of evaluation points. The "means" of these estimators are the sample averages across the 400 Monte Carlo samples for each evaluation point. The probability bands are constructed using the sample standard deviations across Monte Carlo samples, again for each evaluation point. Our figures plot these results for the grid of evaluation points.

9 It is well known that $\hat{g}(x)$ estimates the constant $\bar{y} = N^{-1} \sum y_i$ when $h \to \infty$. 
As discussed in Section 3, we require the "fixed bandwidth" posture only because of the nonlinearity of \( g(x) \) in \( c(x) \) and \( f(x) \). Since \( c(x) \) and \( f(x) \) are just ordinary sample averages, there is little question that their central tendency will be toward their means \( E[c(x)] \) and \( E[f(x)] \) (which happen to coincide with the "fixed bandwidth" limits as \( N \to \infty \)).

It is easy to verify that \( \phi_h(z) = \int \phi_h(y,z) \, dy \). Silverman (1986) notes this structure for kernel density estimators, and it is exploited to study density derivative bias in Stoker (1992b).

Our posture can be explained by analogy to the use of a trimmed mean to minimize mean square error, as in the study of ridge regression and Stein-James estimators. In particular, suppose that \( y \) is distributed with mean \( \mu_y \) and variance \( \sigma_y^2 \), and that we propose to estimate \( \mu_y \) by \( \hat{y} = A \bar{y} \). The constant \( A \) that minimizes mean square error is \( A = \frac{\mu_y^2}{\mu_y^2 + \sigma_y^2 / N} \), where obviously this optimal choice obeys \( A \to 1 \) as \( N \to \infty \). Now, suppose that our application has \( A = .8 \), and we are interested in how one should best approximate the distribution of \( \hat{y} = A \bar{y} \). Analogous to our "fixed bandwidth" approach is to regard \( \bar{y} \) as distributed with mean \( A \mu_y = .8 \mu_y \) and variance \( A^2 \sigma_y^2 / N = .64 \sigma_y^2 / N \). Analogous to the standard "shrinking bandwidth" theory is to regard \( y \) as distributed with mean \( \mu_y \) and variance \( \sigma_y^2 / N \), as implied by letting \( A \to 1 \).

The same reasoning is used in Bierens (1987) to establish the distribution of kernel regression in the case where \( h \to 0 \).

Note how assumption NK implies assumption 2.

With regard to Table 1, our choice of \( h = 1 \) is larger than the values .571 and .706 of Tables 2 and 3, with the bias evident from Table 1 larger than those given in the latter tables. There are some arguments that "asymptotically optimal" bandwidths such as we have computed are too small in finite samples (c.f. Marron and Wand (1992) for several examples in density estimation); we just note that such arguments would indicate larger bias values than those in Tables 2 and 3.

With a unimodal design, we also see how the flattening effect is affected by the tail structure of \( \phi_h \). Fatter tails than a normal distribution (e.g. exponential) imply that the score \( \lambda_h \) increases less quickly than a linear function, with a smaller "flattening" impact. Thinner tails than a normal distribution implies the opposite, with a greater flattening impact.

Stoker (1992b) demonstrates several connections between the true density \( f \) and the convolution \( \phi_h \). For instance, for a small enough bandwidth, it is easy to show that the modal structure of \( \phi_h \) will be analogous to that of \( f \).
The average bias in derivatives from (2.17) is $\mathbb{E}[\gamma_h'(x)] = \beta^T(I - h^2 \mathbb{E}(-\delta^2 \ln \phi_h/\delta x \delta^T)}].$
The expectation in the latter expression is $\mathbb{E}(-\delta^2 \ln \phi_h/\delta x \delta^T} = \int -\delta^2 \ln \phi_h/\delta x \delta^T f(x) \, dx,$
which is not the information matrix of $\phi_h$, but nevertheless will be positive if weighting by $f$ does not differ much from weighting by $\phi_h$.

As before, Proposition 5 is consistent with standard approximation results; namely as $h \to 0$, we have $C_g(z) \to 0$ and $\gamma_h(z) \to g(z)$.

I thank an anonymous referee for pointing out this example.

In very rare drawings with $x < 4$, we set $g(x) = \ln(.001) = -6.908$ in the computations.

In the case of kernel density estimation, Stoker(1992b) points out how the "curvature adjustment" itself implies attenuation bias in density derivative estimates. This occurs because smooth densities must go to zero in their tails, but this structure is not available in the regression case here.

A previous version of the paper suggested using the leading term of (1.6), namely $N^{-1} \sum \hat{x}'(x_i)^T l_i,$ as a diagnostic statistic and correction for derivative bias. This term estimates $A_h$ and performs well as an IV correction for bias in Table 1. While of merit, we have not highlighted this idea here because of the issues raised with nonlinearity and shifting of the evaluation point. For instance, this statistic would signal greater derivative bias for the log model of Figure 2.5 than actually exists.

These can be computed directly, although we have found it more convenient to compute the overall aggregates and apply the results of Stoker(1986a); for instance, we have $E(g') = \delta P/\delta E(x)$, where $P = E(g)$ is the overall probability that $y = 1$ when the mean of $x$ is $E(x)$.

Standard nonparametric theory is surveyed in Prakasa-Rao (1983) and Härdle (1991), and Bierens (1987) discusses kernel regression in an econometric context. Robinson (1992) gives several new results available when the true model has a normal additive disturbance, results that characterize the joint density of regression estimates at different evaluation points, and refined rates of convergence. These references also provide a good guide to recent contributions in the statistical literature on kernel regression.
Our choice of standard deviations fixes the location of negative and positive side lobes, for example locating them farther from the center than if \( \psi_2, \psi_3, \psi_4 \) had standard deviations of 1.1, 1.2, 1.3, for instance. My experience is that the positioning of the side lobes is similar to the effect of bandwidth - namely using a kernel with "closer" side lobes has the same impact as using a smaller bandwidth. While this fits with intuition, it is only based on my casual observation, and the construction of the kernel definitely merits further study.

This conclusion is similar to that of Marron and Wand (1992) for density estimation with higher order kernels.
References


