LOCAL POLYNOMIAL DERIVATIVE ESTIMATION:
ANALYTIC OR TAYLOR?

JEFFREY S. RACINE

ABSTRACT. Local polynomial regression is extremely popular in applied settings. Recent developments in shape constrained nonparametric regression allow practitioners to impose constraints on local polynomial estimators thereby ensuring that the resulting estimates are consistent with underlying theory. However, it turns out that local polynomial derivative estimates may fail to coincide with the analytic derivative of the local polynomial regression estimate which can be problematic, particularly in the context of shape constrained estimation. In such cases practitioners might prefer to instead use analytic derivatives along the lines of those proposed in the local constant setting by Rilstone & Ullah (1989). Demonstrations and applications are considered.

1. Introduction

Local polynomial regression has emerged as a dominant method for nonparametric estimation and inference. The local linear variant was proposed by Stone (1977) and Cleveland (1979); see Fan (1992) and Fan & Gijbels (1996) for an extensive treatment of the local polynomial estimator. Local estimators are nowadays being used in the context of shape constrained estimation and inference; see Hall & Huang (2001) and Du, Parmeter & Racine (2013) by way of illustration. However, one feature of local polynomial estimators that may not be widely appreciated is that the local polynomial derivative estimator does not, in general, coincide with the analytic derivative of the local polynomial regression estimator in

Date: March 30, 2015.
Key words and phrases. Nonparametric, Smoothing, Constrained Estimation.
I would like to express my deepest gratitude to Aman Ullah for his encouragement and guidance over the course of my education and career. I would also like to acknowledge Chris Parmeter for his encouragement and feedback, and with whom previous work and discussions provided the groundwork for this paper, and to thank Carlos Ordas Criado for the illustrative example used herein. Finally, I would like to gratefully acknowledge support from the Natural Sciences and Engineering Research Council of Canada (NSERC:www.nserc.ca), the Social Sciences and Humanities Research Council of Canada (SSHRC:www.sshrc.ca), and the Shared Hierarchical Academic Research Computing Network (SHARCNET:www.sharcnet.ca).
finite-sample settings. This can cause problems, particularly in the context of shape con-
strained estimation. The problem arises when the object of interest is the regression function
itself and constraints are to be imposed on derivatives of the regression function, however the
regression estimate and derivative estimate are not internally consistent, i.e., the derivative
of the local polynomial regression estimate does not coincide with the local polynomial deriv-
ative estimate. In such cases the ‘analytic’ derivative of the regression estimate is required,
otherwise the shape constrained estimate may fail to satisfy the constraints imposed, which
would be troubling for the practitioner. As well, the divergence between the local polynomial
derivative estimate and the analytic derivative of the local polynomial regression estimate
may be cause for concern in settings other than shape constrained estimation.

In this paper we derive the relationship between the two derivative estimators, demonstrate
conditions under which they are equal, consider a measure of divergence between the two,
and alert practitioners to the pitfalls associated with using local polynomial smoothers in a
shape constrained framework and elsewhere.

2. Locally Weighted Regression

Consider a model of the form

$$y_i = g(x_i) + \epsilon_i, \quad i = 1, \ldots, n,$$

where $g(x) = E(Y|X = x)$ is an unknown but smooth conditional mean function evaluated
at the value $X = x$, $\epsilon_i$ is a stochastic error term, and $n$ the number of sample realizations.
To fix ideas, we restrict attention to the single predictor case in what follows (extensions to
the multiple predictor case will hopefully be obvious to the reader).

A fairly broad class of nonparametric kernel regression estimators of $g(x)$ can be written
as a linear combination of locally weighted $y_i$, i.e.,

$$\hat{g}(x) = \sum_{i=1}^{n} W\left(\frac{x_i - x}{h}\right) y_i,$$
where \( W((x_i - x)/h) \) is a local weight function, examples of which will be given shortly. Local weights will be used below, with the defining feature being that observations lying close to the point \( x \) will receive higher weight than those lying far from \( x \). Having selected a kernel function, the bandwidth \( h \) controls the amount of weight given the each sample realization. For notational simplicity we shall write \( W_i(x) = W((x_i - x)/h) \) in what follows, hence we can succinctly express this class of locally weighted estimators as \( \hat{g}(x) = \sum_{i=1}^{n} W_i(x)y_i \).

Estimators of the form (1) include the local constant estimator (Nadaraya (1965), Watson (1964)), the Priestley-Chao estimator (Priestley & Chao (1972)), the Gasser-Müller estimator (Gasser & Müller (1979)), and the local polynomial estimator (Stone (1977), Cleveland (1979), Fan (1992)), among others. The most popular estimators among practitioners are the local constant estimator (Nadaraya (1965), Watson (1964)) and the local linear variant of Fan’s (1992) local polynomial estimator, the latter being dominant in applied settings.

2.1. Shape Constrained Locally Weighted Regression. Shape constrained nonparametric regression is sometimes required. When imposing shape constraints on local estimators of the form (1) derivative estimates may be necessary (i.e., derivatives of \( g(x) \) with respect to \( x \)). For the local constant estimator, the analytic derivative of \( \hat{g}(x) \) has been studied by Rilstone & Ullah (1989). For the local polynomial estimator, derivatives are delivered by the estimation procedure itself, i.e., the local linear estimator delivers estimators of \( g(x) \) and \( g'(x) \), the local quadratic estimators of \( g(x), g'(x), \) and \( g''(x) \) and so forth. To impose shape constraints one can introduce \( n \) weights \( p_i, i = 1, \ldots n \), in the class of estimators given in (1) as follows:

\[
\hat{g}(x|p) = \sum_{i=1}^{n} W_i(x)y_ip_i.
\]

Shape constraints of the form

\[
l(x) \leq \frac{d^j}{dx^j}\hat{g}(x|p) \leq u(x),
\]

3
for arbitrary $l(\cdot)$, $u(\cdot)$, and $j$, are obtained by solving a nonlinear program to obtain the $p_i$ (hence $\hat{g}(x|p)$) satisfying the constraint (3); see Hall & Huang (2001) and Du et al. (2013) for further details. When $j = 0$ the constraint is on the conditional mean $g(x)$, when $j = 1$ on its first derivative, etc. For constraints such as monotonicity and concavity, we require the $j$th derivative of (2) which involves the $j$th derivative of $W_i(x)$ with respect to $x$, i.e.,

$$
\frac{d^j}{dx^j} \hat{g}(x|p) = \sum_{i=1}^{n} W_i^{(j)}(x)y_i p_i
$$

where $W_i^{(j)}(x) = \frac{d^j}{dx^j} W_i(x)$. From (4) it is evident that imposing constraints on derivatives necessitates computation of analytic derivatives, i.e., ones that coincide with the derivative of the estimator $\hat{g}(x|p)$. We now turn to the local constant and local linear derivative estimators as they play a central role in what follows.

2.2. **Local Constant Derivative Estimation.** We first consider the local constant derivative estimator studied by Rilstone & Ullah (1989) as it nicely illustrates the role played by the analytic derivative. The local constant (Nadaraya (1965), Watson (1964)) estimator of $g(x)$ is obtained as arg min$_a S$ where $S$ is given by the locally weighted sum of squares,

$$
S = \sum_{i=1}^{n} (y_i - a)^2 K(z_i),
$$

where $K(z_i) \equiv K((x_i - x)/h)$ is a kernel function such as the Gaussian or Epanechnikov weight function, which yields the local constant estimator

$$
\hat{g}(x) = \frac{\sum_{i=1}^{n} K(z_i)y_i}{\sum_{i=1}^{n} K(z_i)} = \sum_{i=1}^{n} W_i(x)y_i.
$$

where the local weights for this estimator are of the form $W_i(x) = K(z_i)/\sum_{i=1}^{n} K(z_i)$. The local constant estimator of the first derivative of $g(x)$ is the analytic derivative of $\hat{g}(x)$ with respect to $x$, i.e.,

$$
\hat{\beta}(x) = \frac{d}{dx} \hat{g}(x) = \sum_{i=1}^{n} \frac{d}{dx} W_i(x)y_i = \sum_{i=1}^{n} W_i'(x)y_i.
$$
and was studied by Rilstone & Ullah (1989). The local constant estimator, however, is known to suffer from so-called ‘boundary bias’. In applied settings local polynomial estimation is often preferred as it possesses the same bias in the interior of the support as it does at the boundaries, unlike its local constant peer. The local linear variant is dominant in applied settings, so we consider the local linear estimator by way of illustration.

2.3. Local Linear Taylor-Based Derivative Estimation. Assuming that the second derivative of \( E[Y|x] \) exists, then in a small neighbourhood of a point \( x \), by Taylor approximation we can write \( E[Y|x_0] \approx E[Y|x] + (\partial E[Y|x]/\partial x)(x_0 - x) = a + b(x_0 - x) \). The problem of estimating \( g(x) \equiv E[Y|x] \) is equivalent to the local linear regression problem of estimating the intercept \( a \). The problem of estimating the derivative \( \beta(x) \equiv \partial E[Y|x]/\partial x \) is equivalent to the local linear regression problem of estimating the slope \( b \). The local linear estimators of \( a \) and \( b \), denoted by \( \hat{a} \) and \( \hat{b} \), are obtained as \( \arg \min_{a,b} S \) where \( S \) is given by the locally weighted sum of squares

\[
S = \sum_{i=1}^{n} (y_i - a - b(x_i - x))^2 K(z_i).
\]

In this univariate predictor setting, the local linear regression estimator \( \hat{g}(x) = \hat{a} \) can be shown to be

\[
\hat{g}(x) = \frac{\sum_{i=1}^{n} (S_2(x) - S_1(x)(x_i - x)) K(z_i)y_i}{S_2(x)S_0(x) - S_1^2(x)} = \frac{\sum_{i=1}^{n} C_i(x)y_i}{D(x)} = \sum_{i=1}^{n} W_i(x)y_i,
\]

where

\[
S_j(x) = \sum_{i=1}^{n} (x_i - x)^j K(z_i),
\]

while the local linear (i.e., Taylor-based) derivative estimator, \( \hat{\beta}(x) = \hat{b} \), can be shown to be

\[
\hat{\beta}(x) = \frac{\sum_{i=1}^{n} (S_0(x)(x_i - x) - S_1(x)) K(z_i)y_i}{D(x)} = \frac{\sum_{i=1}^{n} E_i(x)y_i}{D(x)} = \sum_{i=1}^{n} B_i(x)y_i.
\]

See Fan & Gijbels (1996) for further details.
2.4. **Local Linear Analytic Derivative Estimation.** An alternative derivative estimator, the ‘analytic’ derivative of \( \hat{g}(x) \) which we denote by \( \hat{\beta}^a(x) \), is obtained by taking the derivative of \( \hat{g}(x) \) defined in (5), and is given by

\[
\hat{\beta}^a(x) = \sum_{i=1}^{n} W'_i(x) y_i,
\]

where \( W'_i(x) \) is the derivative of \( W_i(x) \) from (5) with respect to \( x \) and is given by

\[
W'_i(x) = \frac{D(x)C'_i(x) - C_i(x)D'(x)}{D(x)^2}
= \frac{D(x)R_{1,i}(x)K(z_i) + D(x)R_{2,i}(x)K'(z_i) - C_i(x)R_3(x)}{D(x)^2}
= \frac{R_{1,i}(x)K(z_i) + R_{2,i}(x)K'(z_i)}{D(x)} - W_i(x) \frac{R_3(x)}{D(x)},
\]

where

\[
R_{1,i}(x) = T_2(x) - T_1(x)(x_i - x) + S_1(x),
\]
\[
R_{2,i}(x) = S_2(x) - S_1(x)(x_i - x),
\]
\[
R_3(x) = T_2(x)S_0(x) + S_2(x)T_0(x) - 2S_1(x)T_1(x),
\]

and where \( T_j(x) \) is given below.

To obtain this result we note that

\[
C'_i(x) = \frac{d}{dx} \left[ (S_2(x) - S_1(x)(x_i - x)) K(z_i) \right]
= (T_2(x) - T_1(x)(x_i - x) + S_1(x)) K(z_i) + (S_2(x) - S_1(x)(x_i - x)) K'(z_i).
\]
Note that when \( j \geq 1 \), we have

\[
T_j(x) = S'_j(x) = \sum_{i=1}^{n} \frac{d}{dx} (x_i - x)^j K(z_i)
\]

\[
= \sum_{i=1}^{n} \left( (x_i - x)^j K'(z_i) - j(x_i - x)^{j-1} K(z_i) \right),
\]

while, when \( j = 0 \),

\[
S_0(x) = \sum_{i=1}^{n} K(z_i),
\]

hence

\[
T_0(x) = S'_0(x) = \sum_{i=1}^{n} K'(z_i).
\]

Finally,

\[
D'(x) = \frac{d}{dx} \left[ S_2(x)S_0(x) - S_1^2(x) \right]
\]

\[
= T_2(x)S_0(x) + S_2(x)T_0(x) - 2S_1(x)T_1(x).
\]

Therefore, \( D(x)C'_i(x) - C_i(x)D'(x) \) is seen to be

\[
D(x) \{ (T_2(x) - T_1(x)(x_i - x) + S_1(x)) K(z_i) + (S_2(x) - S_1(x)(x_i - x)) K'(z_i) \}
\]

\[
- C_i(x) \{ T_2(x)S_0(x) + S_2(x)T_0(x) - 2S_1(x)T_1(x) \}.
\]

As it turns out, the Taylor-based and analytic local linear derivative estimators given in (6) and (7), respectively, may not coincide. We now turn to an assessment of the degree to which they may differ and the conditions under which they are equal.

From the definition of the local linear Taylor-based and analytic derivatives, we see that

\[
\hat{\beta}(x) = \sum_{i=1}^{n} B_i(x)y_i + \left( \hat{\beta}^a(x) - \sum_{i=1}^{n} W'_i(x)y_i \right)
\]

\[
= \hat{\beta}^a(x) - \sum_{i=1}^{n} (B_i(x) - W'_i(x)) y_i.
\]

Observe that when \( h \to \infty \), \( K(z_i) = K(0) = c_0 \), \( T_0(x) = 0 \), \( T_1(x) = -nc_0 \), \( T_2(x) = -2c_0 \sum_{i=1}^{n}(x_i - x) \), \( S_0(x) = nc_0 \), \( S_1(x) = c_0 \sum_{i=1}^{n}(x_i - x) \), and \( S_2(x) = c_0 \sum_{i=1}^{n}(x_i - x)^2 \).

Then

\[
R_{1,i}(x) = T_2(x) - T_1(x)(x_i - x) + S_1(x),
\]

\[
= -c_0 \sum_{i=1}^{n}(x_i - x) + nc_0(x_i - x),
\]

\[
R_{2,i}(x) = S_2(x) - S_1(x)(x_i - x),
\]

\[
= c_0 \sum_{i=1}^{n}(x_i - x)^2 - c_0(x_i - x) \sum_{i=1}^{n}(x_i - x),
\]

\[
R_3(x) = T_2(x)S_0(x) + S_2(x)T_0(x) - 2S_1(x)T_1(x)
\]

\[
= -2nc_0^2 \sum_{i=1}^{n}(x_i - x) + 2nc_0^2 \sum_{i=1}^{n}(x_i - x) = 0,
\]

hence

\[
W'_i(x) = \frac{R_{1,i}(x)K(z_i)}{D(x)}
\]

\[
= -\frac{c_0^2 \sum_{i=1}^{n}(x_i - x) + nc_0^2(x_i - x)}{D(x)}
\]

\[
= \frac{E_i(x)}{D(x)}
\]

\[
= B_i(x),
\]

8
hence

$$\lim_{h \to \infty} \hat{\beta}(x) = \hat{\beta}^a(x),$$

that is, as $h$ increases the analytic and local linear derivatives will converge. Otherwise, they may differ, sometimes substantially so. Furthermore,

$$\|\hat{\beta}(x) - \hat{\beta}^a(x)\|/\|\hat{\beta}(x)\|$$

could serve as a useful measure of the degree to which these derivative estimates diverge, where $\|\cdot\|$ denotes Euclidean norm. The issue at hand is whether, in applied settings, any such divergence between the Taylor-based and analytic derivatives arises and, if so, what impact it might have. We now consider a simple illustration that will serve to highlight this divergence.

3. Demonstration: The Prestige Data

We consider the ‘Prestige’ data taken from the R package ‘car’ (Fox & Weisberg (2011)) and use variables ‘income’ (average income of incumbents, in thousands of dollars, in 1971) and the logarithm of ‘prestige’ (the Pineo-Porter prestige score for occupation, from a social survey conducted in the mid-1960s, collected by Census Canada). This dataset contains $n = 102$ observations. We conduct local linear estimation with least-squares cross-validation for bandwidth selection using the R package np (Hayfield & Racine (2008)). The cross-validated bandwidth using a second order Gaussian kernel is $h = 3.48$ thousand dollars. The local linear estimate $\hat{g}(x)$ is presented in Figure 1.

The local linear Taylor-based derivative $\hat{\beta}(x)$ is presented in Figure 2 along with the analytic derivative $\hat{\beta}^a(x)$ and that obtained by numerical differentiation of $\hat{g}(x)$ itself, denoted $\Delta \hat{g}(x)/\Delta x$. The inclusion of the numerical derivative is to verify that the analytic derivative coincides with the derivative of $\hat{g}(x)$. For a fairly large range of income, Figure 2 reveals the
substantial discrepancy between the Taylor-based and analytic derivative estimates. The value of the divergence metric given in (8) is \( \| \hat{\beta}(x) - \hat{\beta}^a(x) \| / \| \hat{\beta}(x) \| = 0.102 \).

Close inspection of figures 1 and 2 reveal that the regression function \( \hat{g}(x) \) is non-monotonic for large values of income (i.e., \( \hat{g}(x) \) falls slightly, while a close examination of Figure 2 reveals that for high levels of income, the analytic/numeric derivatives assume negative values). In addition, the Taylor-based estimate is substantially higher than the analytic derivative over much of the support, i.e., it overstates the actual derivative of \( \hat{g}(x) \).

Suppose, by way of illustration, that we required a monotonic estimate \( \hat{g}(x) \), and imposed monotonicity on the resulting regression function, i.e., we imposed \( \hat{g}'(x|p) \geq 0 \). The resulting shape constrained estimator is presented in Figure 3, and the unconstrained and constrained derivative estimates are presented in Figure 4.
Close inspection of Figure 3 reveals a rather curious artifact of using the Taylor-based derivative to impose monotonicity, namely that the Taylor-based shape constrained estimate is non-monotone in income. This is most unwelcome as we imposed monotonicity on the resulting estimate. The analytic-based shape constrained estimate is, of course, monotonic. This simple illustration demonstrates the pitfalls of using local polynomial derivatives in settings where one requires the derivative of $\hat{g}(x)$, not just a consistent estimator of $\beta(x) = d\hat{g}(x)/dx$.

3.1. Implications for Inference. The use of Taylor-based derivatives for shape constrained estimation is not the only illustration of the potential pitfalls associated with local polynomial estimation, and the use of analytic derivatives might also be germane for inference and prediction. For instance, nonparametric tests of significance may involve imposing the
null $H_0: \beta(x) = 0$ almost everywhere, with rejection of the null implying, for instance, predictability in the conditional mean in a time-series setting. But given that the Taylor-based derivative estimates may fail to correspond to the conditional mean estimate, the outcome of the test may be misleading when it comes to testing for predictability in the conditional mean of a series.

4. Finite-Sample Comparison of Analytic and Taylor-Based Derivatives

The question naturally arises, “why not simply always use the analytic derivatives?” This is certainly a reasonable query. One would expect that the analytic derivatives are less biased in finite-sample than the Taylor-based ones given the results presented above. The issue is whether, from a square error loss perspective, the use of analytic derivatives over its Taylor-based peer can be justified in general.
To address this issue we consider a Monte-Carlo simulation in which the data is generated from

\( y = \sin(2\pi x) + \epsilon, \)

where \( x \sim U[0, 1] \) and \( \epsilon \sim N(0, \sigma^2) \) with \( \sigma = 0.25, 0.50 \). We consider \( M = 1,000 \) Monte Carlo draws from this data generating process. Here \( g'(x) = 2\pi \cos(2\pi x) \), and we compare the analytic and Taylor-based estimates. We present boxplots of the bias and MSE of the estimator where each is averaged pointwise over the empirical support of the data. Results are presented in Figure 5 for samples of size \( n = 100, 200, 400, \) and 800 and \( \sigma = 0.25 \). Median bias and MSE are summarized in Table 1 for the above sample sizes and for \( \sigma = 0.25, 0.50 \).
Table 1. Median Bias and MSE.

<table>
<thead>
<tr>
<th>$n$</th>
<th>Analytic</th>
<th>Taylor</th>
<th>Analytic</th>
<th>Taylor</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>0.035</td>
<td>0.068</td>
<td>1.493</td>
<td>1.248</td>
</tr>
<tr>
<td>200</td>
<td>0.015</td>
<td>0.036</td>
<td>1.022</td>
<td>0.862</td>
</tr>
<tr>
<td>400</td>
<td>0.001</td>
<td>0.016</td>
<td>0.761</td>
<td>0.645</td>
</tr>
<tr>
<td>800</td>
<td>-0.001</td>
<td>0.009</td>
<td>0.566</td>
<td>0.482</td>
</tr>
</tbody>
</table>

For $\sigma = 0.25$

<table>
<thead>
<tr>
<th>$n$</th>
<th>Analytic</th>
<th>Taylor</th>
<th>Analytic</th>
<th>Taylor</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>0.072</td>
<td>0.145</td>
<td>2.802</td>
<td>2.438</td>
</tr>
<tr>
<td>200</td>
<td>0.036</td>
<td>0.072</td>
<td>2.024</td>
<td>1.716</td>
</tr>
<tr>
<td>400</td>
<td>0.040</td>
<td>0.069</td>
<td>1.377</td>
<td>1.210</td>
</tr>
<tr>
<td>800</td>
<td>0.020</td>
<td>0.036</td>
<td>1.042</td>
<td>0.882</td>
</tr>
</tbody>
</table>

For $\sigma = 0.50$

Figure 5 reveals that the analytic derivatives display less bias than the Taylor-based ones, while the median bias tends toward zero for both approaches as $n$ increases. However, the analytic derivatives display higher MSE for all $n$, so from an MSE perspective using analytic derivatives may not be warranted in general.

5. Summary

We consider the impact of using local polynomial Taylor-based derivative estimates for shape constrained estimation, and caution practitioners that these derivatives may be internally inconsistent with the estimated regression function. However, in cases where internal consistency is required, an option is to instead use analytic derivatives. Such derivatives possess lower bias than their Taylor-based counterparts, however they also are less efficient from the mean square error perspective so their use as a general replacement for Taylor-based derivatives cannot be justified, at least not from a square error perspective. For shape constrained estimation and other uses that require internal consistency between the regression function and derivative estimates, analytic derivatives cannot be avoided.
Figure 5. Boxplots of bias and mean square error, $\sigma = 0.25$, $n = 100, 200, 400, 800$. 
References


URL: [http://socserv.socsci.mcmaster.ca/jfox/Books/Companion](http://socserv.socsci.mcmaster.ca/jfox/Books/Companion)


URL: [http://www.jstatsoft.org/v27/i05/](http://www.jstatsoft.org/v27/i05/)


Department of Economics, Department of Mathematics and Statistics (Graduate Program in Statistics), McMaster University, Hamilton, Ontario L8S 4M4, Canada

E-mail address: racinej@mcmaster.ca