Estimators in step regression models

Ursula U. Müller  Anton Schick  Wolfgang Wefelmeyer
Texas A&M University  Binghamton University  Universität zu Köln

Abstract

We consider nonparametric regression models in which the regression function is a step function, and construct a convolution estimator for the response density that has the same bias as the usual estimators based on the responses, but a smaller asymptotic variance.

1 Introduction

We consider the nonparametric regression model \( Y = r(X) + \varepsilon \), where \( \varepsilon \) has mean zero and \( X \) and \( \varepsilon \) are independent random variables with densities \( g \) and \( f \). We assume that the regression function is a step function with unknown jump points and jump heights. For simplicity we assume that \( m \), the number of jump points, is known. If \( f \) and \( g \) are positive and absolutely continuous, we can estimate the jump points with rate \( n^{-1} \) and the jump heights with rate \( n^{-1/2} \). This follows from results for more general stepwise linear and stepwise smooth and parametric regression functions. For deterministic covariates, Yao and Au (1989) estimate a step regression function for known and for bounded \( m \). Piecewise linear regression functions are studied by Quandt (1958, 1960), Hinkley (1969), Farley and Hinich (1970), Bai and Perron (1998), Koul and Qian (2002), and Koul, Qian and Surgailis (2003). For piecewise polynomial regression functions see Robison (1964), and for piecewise nonlinear regression functions see Feder (1975a, 1975b), Liu, Wu and Zidek (1997), Ciuperca (2004, 2009, 2011), Ciuperca and Dapzol (2008), and Launay, Philippe and Lamarche (2012). Our result extends to the case of an unknown but bounded number of jump points. An estimator for this number is obtained in Section 4 of Ciuperca (2011).

The estimators for the jump points and the jump heights determine an estimator \( \hat{r} \) for the regression function \( r \). We can use it to estimate the errors \( \varepsilon_i = Y_i - r(X_i) \) by residuals \( \hat{\varepsilon}_i = Y_i - \hat{r}(X_i) \), and to estimate the error density \( f \) by a residual-based kernel estimator \( \hat{f} \). We show in Lemma 1 that \( \hat{f} \) differs by a term of order \( n^{-1/2} \) from the kernel estimator based on the true errors \( \varepsilon_i \). There are similar results for the case that \( r \) is not a step function but smooth; see Schick and Wefelmeyer (2012) and (2013). Here the regression function has jumps, and the proof is different. In Theorem 1 we show that the residual-based kernel estimator \( \hat{f} \) is asymptotically normal with the same mean and variance as the error-based kernel estimator.
Assume for simplicity that the heights $b_j$ are pairwise different. Then the response density has a convolution representation

$$h(y) = \sum f(y - b_j)P(r(X) = b_j)$$

and can be estimated by a convolution estimator

$$\hat{h}(y) = \sum \hat{f}(y - \hat{b}_j)\hat{p}_j$$

with estimators $\hat{b}_j$ and $\hat{p}_j$ for $b_j$ and $P(r(X) = b_j)$. We show in Lemma 2 that it differs by a term of order $n^{-1/2}$ from the estimator with known covariate density and regression function. In Theorem 2 we show that the convolution estimator $\hat{h}$ has the same rate and asymptotic bias as the kernel estimator based only on the responses, and that it is asymptotically normal with the same mean, but with a considerably reduced variance.

This differs from results for the case that $r(X)$ is not discrete but has a smooth density. Then the corresponding convolution estimator can have the rate $n^{-1/2}$ of an empirical estimator; see again Schick and Wefelmeyer (2012) and (2013).

We show in Remark 1 that corresponding results hold when the covariate is discrete (and $r$ is arbitrary). Estimators for the regression function in this case are considered in particular by Bierens and Hartog (1988), Rahbar and Gardiner (1995) and Ouyang, Li and Racine (2009).

2 Results

Let $Y = r(X) + \varepsilon$, where $X$ and $\varepsilon$ are independent random variables with positive and absolutely continuous densities $g$ and $f$, and $E\varepsilon = 0$ and $E\varepsilon^2 < \infty$. We assume that the regression function is a step function

$$r = b_11_{(-\infty, a_1]} + \sum_{j=2}^{m} b_j 1_{[a_{j-1}, a_j]} + b_{m+1}1_{[a_m, \infty)}$$

with unknown jump points $a_1 < \cdots < a_m$ and unknown heights $b_1, \ldots, b_{m+1}$, and known $m$. We observe independent copies $(X_1, Y_1), \ldots, (X_n, Y_n)$ of $(X, Y)$. By Ciuperca (2009), there are estimators $\hat{a}_j$ for $a_j$ with $n(\hat{a}_j - a_j)$ bounded in probability, and estimators $\hat{b}_j$ for $b_j$ with $n^{1/2}(b_j - \hat{b}_j)$ bounded in probability. They are obtained by minimizing, for an appropriate convex function $\varrho$, the process $\sum_{i=1}^{n} \varrho(Y_i - r(X_i))$ in the parameters $a_1, \ldots, a_m$ and $b_1, \ldots, b_{m+1}$ of $r$. This is an $M$-estimator for $r$. For the choice $\varrho(y) = y^2$ it is a least squares estimator. The minimizing values $\hat{a}_j$ and $\hat{b}_j$ determine an estimator for the regression function,

$$\hat{r} = \hat{b}_11_{(-\infty, \hat{a}_1]} + \sum_{j=2}^{m} \hat{b}_j 1_{[\hat{a}_{j-1}, \hat{a}_j]} + \hat{b}_{m+1}1_{[\hat{a}_m, \infty)}.$$

2
It can be used to estimate the errors \( \varepsilon_i \) by residuals \( \hat{\varepsilon}_i = Y_i - \hat{r}(X_i) \), and the error density \( f \) by a residual-based kernel estimator
\[
\hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} K_b(x - \hat{\varepsilon}_i)
\]
with \( K_b(x) = K(x/b)/b \), where \( K \) is a kernel and \( b \) a bandwidth.

In order to show that \( \hat{f}(x) \) is asymptotically normal, we compare it first with the kernel estimator based on the true errors,
\[
\bar{f}(x) = \frac{1}{n} \sum_{i=1}^{n} K_b(x - \varepsilon_i).
\]
Lemma 1 shows that \( \hat{f}(x) \) differs from \( \bar{f}(x) \) by a term of order \( O_p(n^{-1/2}) \). The proof is in Section 3.

**Lemma 1.** Let \( g \) be positive and absolutely continuous, and let \( f \) be twice continuously differentiable at \( x \). Choose a kernel \( K \) with bounded support that is twice differentiable with second derivative fulfilling a Lipschitz condition, and a bandwidth \( b \) with \( b \to 0 \) and \( n^{1/4}b \to \infty \). Then
\[
\hat{f}(x) = \bar{f}(x) + f'(x) \sum_{j=1}^{m+1} p_j(\hat{b}_j - b_j) + o_p(n^{-1/2}).
\]

Denote by \( \mathcal{K}_r \) the bounded functions \( K \) on the real line that vanish outside a compact set, and that are (signed) kernels of order \( r \), i.e., \( \int K(t) dt = 1 \), \( \int t^j K(t) dt = 0 \) for \( j = 1, \ldots, r-1 \), and \( \int t^r K(t) dt \neq 0 \).

Let \( f \) be \( r \) times continuously differentiable at \( x \) and \( K \in \mathcal{K}_r \). The following results are well known, also under mixing conditions and for linear processes. See Parzen (1962), Chanda (1983), Bradley (1983), Tran (1992), Hallin and Tran (1996) and Lu (2001). A convenient reference is Müller and Wefelmeyer (2014), Lemma 1 and Proposition 1. Set
\[
\mu_r = \frac{(-1)^r}{r!} \int t^r K(t) dt, \quad \sigma^2 = \int K^2(t) dt.
\]
If \( b \to 0 \), then
\[
b^{-r} E(\hat{f}(x) - f(x)) \to f^{(r)}(x)\mu_r.
\]
If \( nb \to \infty \), then
\[
nb \text{Var } \hat{f}(x) \to f(x)\sigma^2.
\]
The optimal rate is achieved by a bandwidth of the form \( b = cn^{-1/(2r+1)} \) for some constant \( c \). We absorb the factor \( c \) as a scale factor into \( K \) and work with \( b = n^{-1/(2r+1)} \). With this bandwidth,
\[
n^{r/(2r+1)}(\hat{f}(x) - f(x)) \Rightarrow N(f^{(r)}(x)\mu_r, f(x)\sigma^2).
\]
Since \( \hat{b}_j - b_j = O_p(n^{-1/2}) \) is asymptotically negligible, \( n^{r/(2r+1)}(\hat{f}(x) - f(x)) \) has the same asymptotic distribution as \( n^{r/(2r+1)}(\hat{f}(x) - f(x)) \). Together with Lemma 1 we obtain the following result.
Theorem 1. Let \( g \) be positive and absolutely continuous, and let \( f \) be positive and \( r \) times continuously differentiable at \( x \) for an \( r \geq 2 \). Let \( K \in K_r \) with \( K'' \) Lipschitz. Set \( b = n^{-1/(2r+1)} \). Then \( n^{r/(2r+1)}(\hat{f}(x) - f(x)) \) is asymptotically normal with mean \( f^{(r)}(x)\mu_r \) and variance \( f(x)\sigma^2 \).

We turn now to estimation of the response density \( h \). A simple estimator is the kernel estimator based on the responses,

\[
\hat{h}(y) = \frac{1}{n} \sum_{i=1}^{n} K_b(y - Y_i).
\]

If \( h \) is \( r \) times continuously differentiable at \( y \), then we obtain as above, for \( K \in K_r \) and \( b = n^{-1/(2r+1)} \),

\[
n^{r/(2r+1)}(\tilde{h}(y) - h(y)) \Rightarrow N(h^{(r)}(y)\mu_r, h(y)\sigma^2).
\]

A better estimator than \( \tilde{h} \) can be based on the convolution representation

\[
h(y) = \int f(y - r(x))g(x) \, dx = \sum_{j=1}^{m+1} f(y - b_j)p_j
\]

with \( p_j = P(r(X) = b_j) \). Here we have assumed for notational simplicity that the heights \( b_1, \ldots, b_{m+1} \) are pairwise different. Then \( r(X) \) is supported by \( b_1, \ldots, b_{m+1} \), with probabilities

\[
p_1 = P(r(X) = b_1) = \int_{-\infty}^{a_1} g(x) \, dx,
p_{m+1} = P(r(X) = b_{m+1}) = \int_{a_{m}}^{\infty} g(x) \, dx,
p_j = P(r(X) = b_j) = \int_{a_{j-1}}^{a_j} g(x) \, dx, \quad j = 2, \ldots, m.
\]

We estimate the \( p_j \) empirically,

\[
\hat{p}_1 = \# \{ i : -\infty < X_i < \hat{a}_1 \} / n,
\hat{p}_{m+1} = \# \{ i : \hat{a}_m \leq X_i < \infty \} / n,
\hat{p}_j = \# \{ i : \hat{a}_{j-1} \leq X_i < \hat{a}_j \} / n, \quad j = 2, \ldots, m.
\]

From \( \hat{a}_j - a_j = O_p(n^{-1}) \) it follows that \( \hat{p}_j - p_j = O_p(n^{-1/2}) \). The convolution representation for the response density \( h \) now suggests the convolution estimator

\[
\hat{h}(x) = \sum_{j=1}^{m+1} \hat{f}(y - \hat{b}_j)\hat{p}_j.
\]

Similarly as for \( \hat{f} \), we compare \( \hat{h} \) first with the convolution estimator based on the true jump points \( a_j \) and heights \( b_j \) and on the true probabilities \( p_j \),

\[
\hat{h}(y) = \sum_{j=1}^{m+1} \hat{f}(y - b_j)p_j.
\]
As in Lemma 1 we now obtain that \( \hat{h}(y) \) differs from \( h(y) \) by a term of order \( O_p(n^{-1/2}) \).

**Lemma 2.** Let \( f \) be positive and twice continuously differentiable at \( y - b_1, \ldots, y - b_{m+1} \). Take \( g, K \) and \( b \) as in Lemma 1. Then

\[
\hat{h}(y) = h(y) - \sum_{j=1}^{m+1} (f'(y - b_j) - h'(y))p_j(\hat{b}_j - b_j) + \sum_{j=1}^{m+1} f(y - b_j)(\hat{p}_j - p_j) + o_p(n^{-1/2}).
\]

Since \( \hat{b}_j - b_j \) and \( \hat{p}_j - p_j \) are of order \( O_p(n^{-1/2}) \) and therefore asymptotically negligible, \( n^{r/(2r+1)}(\hat{h}(y) - h(y)) \) has the same asymptotic distribution as \( n^{r/(2r+1)}(\hat{h}(y) - h(y)) \). For large enough \( n \), the supports of \( K_b(\cdot) \) are disjoint, and hence \( \hat{f}(y - b_j) \) use disjoint subsets of \( \varepsilon_1, \ldots, \varepsilon_n \) for different \( j \). Together with the representation \( \hat{h}(y) = \sum_{j=1}^{m+1} \hat{f}(y - b_j)p_j \) this implies

\[
\text{Var} \hat{h}(y) = \sum_{j=1}^{m+1} p_j^2 \text{Var} \hat{f}(y - b_j).
\]

Lemma 2 and Theorem 1 therefore give the following.

**Theorem 2.** Let \( f \) be positive and \( r \) times continuously differentiable at \( y - b_1, \ldots, y - b_{m+1} \). Take \( g, K \) and \( b \) as in Theorem 1. Then \( n^{r/(2r+1)}(\hat{h}(y) - h(y)) \) is asymptotically normal with mean \( h^{(r)}(y)\mu_r \) and variance \( \sum_{j=1}^{m+1} f(y - b_j)p_j^2\sigma^2 \).

From the convolution representation of \( h \) it follows that \( h \) and \( f \) are smooth of the same order. The corresponding mean for the kernel estimator \( \hat{h}(y) = \frac{1}{n} \sum_{i=1}^{n} K_b(y - Y_i) \) based on the responses only is again \( h^{(r)}(y)\mu_r \), but the variance is

\[
h(y)\sigma^2 = \sum_{j=1}^{m+1} f(y - b_j)p_j\sigma^2,
\]

while the convolution estimator \( \hat{h}(y) \) has \( p_j^2 \) in place of \( p_j \). This is a variance reduction. It is noticeable if no weight is close to one, and it is considerable if there are many small weights. In particular, if \( r(X) \) is uniformly distributed, so that \( p_j = 1/(m+1) \), the variance is reduced by the factor \( 1/(m+1) \).

**Remark 1.** Our approach also works when the covariate \( X \) is discrete, say with values \( a_1, \ldots, a_m \) in an arbitrary space. Then the regression function \( r \) may be arbitrary, because it enters the model only through the values \( b_j = r(a_j) \). Again we assume that \( m \) is known, and that the \( b_j \) are pairwise different. Then \( r(X) \) is discrete with values \( b_j \) having probabilities

\[
P(r(X) = b_j) = P(X = a_j) = p_j.
\]

The values \( a_1, \ldots, a_m \) are eventually observed and need not be estimated. We estimate \( p_j \) empirically, by

\[
\hat{p}_j = N_j/n \quad \text{with} \quad N_j = \#\{i : X_i = a_j\}.
\]

An estimator for \( b_j \) is

\[
\hat{b}_j = \frac{1}{N_j} \sum_{i : X_i = a_j} Y_i.
\]
The response density has the representation

\[ h(y) = \sum_{j=1}^{m} f(y - b_j)p_j. \]

The error \( \varepsilon_i = Y_i - r(X_i) \) is estimated by the residual \( \hat{\varepsilon}_i = Y_i - \hat{b}_j \) if \( X_i = a_j \). Let \( \hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} K_b(x - \hat{\varepsilon}_i) \) denote the residual-based kernel estimator. The convolution estimator for \( h(y) \) is

\[ \hat{h}(y) = \sum_{j=1}^{m} \hat{f}(y - \hat{b}_j)p_j. \]

Under the same assumptions on \( f, K \) and \( b \) as before, the above results continue to hold, with \( m \) in place of \( m + 1 \).

3 Proofs

Proof of Lemma 1. In a first step we show that asymptotically it makes no difference if we replace the \( \hat{a}_j \) by \( a_j \). We simplify the notation by writing

\[ A_1 = (-\infty, a_1), \quad A_{m+1} = [a_m, \infty), \quad A_j = [a_{j-1}, a_j) \quad \text{for} \quad j = 2, \ldots, m. \]

We write \( \hat{A}_j \) if the \( a_j \) are replaced by \( \hat{a}_j \). We write \( \hat{r} \) for the estimator obtained from \( \hat{r} \) by replacing \( \hat{a}_j \) by \( a_j \), i.e., \( \hat{r} = \sum_{j=1}^{m+1} 1_{A_j}(\hat{b}_j) \). We define the residuals associated with \( \hat{r} \) by

\[ \hat{\varepsilon}_j = Y_j - \hat{r}(X_j) \]

and set

\[ \hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} K_b(x - \hat{\varepsilon}_i) = \sum_{j=1}^{m+1} \frac{1}{n} \sum_{i=1}^{n} 1_{A_j}(X_i)K_b(x - \varepsilon_i + \hat{b}_j - b_j). \]

With \( A \) the complement of the union of the intervals \( \hat{A}_j \cap A_j, j = 1, \ldots, m + 1 \), we can express

\[ \hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} K_b(x - \hat{\varepsilon}_i) = \frac{1}{n} \sum_{i=1}^{n} 1_A(X_i)K_b(x - \varepsilon_i) + \sum_{j=1}^{m+1} \frac{1}{n} \sum_{i=1}^{n} 1_{A_j \cap \hat{A}_j}(X_i)K_b(x - \varepsilon_i + \hat{b}_j - b_j). \]

Note that \( N = \sum_{i=1}^{n} 1_A(X_i) \) is bounded in probability. Indeed, for positive constants \( B \) and \( C \) and \( D_n = n \max_{1 \leq j \leq m} |\hat{a}_j - a_j| \), we have

\[
P(N > B) \leq P(D_n > C) + P\left( \sum_{j=1}^{m} \sum_{i=1}^{n} 1_{[a_j - C/n, a_j + C/n]}(X_i) > B \right)
\]

\[
\leq P(D_n > C) + \sum_{j=1}^{m} nP(a_j - C/n \leq X \leq a_j + C/n)/B
\]

\[
\leq P(D_n > C) + \sup_{y} g(y)2mC/B.
\]
It is now easy to see that

\begin{equation}
(3.1) \quad \sup_x |\hat{f}(x) - \bar{f}(x)| \leq 2 \sup_y |K_b(y)|N/n = O_p((nb)^{-1}) = o_p(n^{-1/2}).
\end{equation}

In a second step, we replace \( \hat{b}_j \) by \( b_j \). For \( X_i \in A_j \), a Taylor expansion yields

\[\begin{align*}
|K_b(x - \bar{\varepsilon}_i) - K_b(x - \varepsilon_i) - (\hat{b}_j - b_j)K'_b(x - \varepsilon_i) - \frac{1}{2}(\hat{b}_j - b_j)^2K''_b(x - \varepsilon_i)| &\leq \frac{L|\hat{b}_j - b_j|^3}{66^4},
\end{align*}\]

with \( L \) the Lipschitz constant of \( K'' \). Hence

\[\begin{align*}
\hat{f}(x) &= \bar{f}(x) + \sum_{j=1}^{m+1}(\hat{b}_j - b_j)\frac{1}{n} \sum_{i=1}^{n}1_{A_j}(X_i)K'_b(x - \varepsilon_i) \\
&\quad + \frac{1}{2} \sum_{j=1}^{m+1}(\hat{b}_j - b_j)^2\frac{1}{n} \sum_{i=1}^{n}1_{A_j}(X_i)K''_b(x - \varepsilon_i) + o_p(n^{-1/2})
\end{align*}\]

holds in view of \( n^{-3/2}b^{-4} = o_p(n^{-1/2}) \). For the second term in the expansion of \( \hat{f}(x) \) we use

\[\begin{align*}
E1_{A_j}(X)K'_b(x - \varepsilon) &= P(X \in A_j) \int K'_b(x - y)f(y) \, dy \\
&= p_j \int K(t)f'(x - bt) \, dt \to p_jf'(x)
\end{align*}\]

and

\[\begin{align*}
\text{Var} 1_{A_j}(X)K'_b(x - \varepsilon) &\leq E(K'_b(x - \varepsilon)^2) = b^{-3} \int f(x - bt)(K'(t))^2 \, dt.
\end{align*}\]

Since \( n^{-1}b^{-3} \to 0 \), we obtain

\[\begin{align*}
\frac{1}{n} \sum_{i=1}^{n}1_{A_j}(X_i)K'_b(x - \varepsilon_i) &= p_jf'(x) + o_p(1).
\end{align*}\]

Similarly,

\[\begin{align*}
\frac{1}{n} \sum_{i=1}^{n}1_{A_j}(X_i)K''_b(x - \varepsilon_i) &= p_jf''(x) + o_p(1) + O_p(n^{-1/2}b^{-5/2}).
\end{align*}\]

The assertion follows.

**Proof of Lemma 2.** The proof follows along the lines of the proof of Lemma 1. We continue using the notation introduced there. Write

\[\begin{align*}
\hat{f}(y - \hat{b}_j) &= \frac{1}{n} \sum_{i=1}^{n}K_b(y - \bar{\varepsilon}_i - \hat{b}_j) \\
&= \frac{1}{n} \sum_{i=1}^{n}K_b(y - \varepsilon_i - b_j + \bar{r}(X_i) - r(X_i) - (\hat{b}_j - b_j)) \\
&= \sum_{k=1}^{m+1} \frac{1}{n} \sum_{i=1}^{n}1_{A_k}(X_i)K_b(y - \varepsilon_i + \hat{b}_k - b_k - (\hat{b}_j - b_j)).
\end{align*}\]
By the same Taylor expansion as in the proof of Lemma 1,
\[
\tilde{f}(y - \hat{b}_j) = \bar{f}(y - \hat{b}_j) + \sum_{k=1}^{m+1} (\hat{b}_k - b_k)p_kf'(y - b_j) - (\hat{b}_j - b_j)f'(y - b_j) + o_p(n^{-1/2}).
\]

Using (3.1), we obtain
\[
\hat{h}(y) = \sum_{j=1}^{m+1} \hat{f}(y - \hat{b}_j)\hat{p}_j = \sum_{j=1}^{m+1} \tilde{f}(y - \hat{b}_j)\hat{p}_j + o_p(n^{-1/2})
\]
\[
= \sum_{j=1}^{m+1} \tilde{f}(y - b_j)p_j + \sum_{j=1}^{m+1} \tilde{f}(y - \hat{b}_j)(\hat{p}_j - p_j) + o_p(n^{-1/2})
\]
\[
= \sum_{j=1}^{m+1} \tilde{f}(y - b_j)p_j + \sum_{j=1}^{m+1} f'(y - b_j)p_j \sum_{k=1}^{m+1} p_k(\hat{b}_k - b_k) - \sum_{j=1}^{m+1} f'(y - b_j)p_j(\hat{b}_j - b_j)
\]
\[
+ \sum_{j=1}^{m+1} f(y - b_j)(\hat{p}_j - p_j) + o_p(n^{-1/2}).
\]

The assertion now follows with
\[
\sum_{j=1}^{m+1} f'(y - b_j)p_j = h'(y).
\]

References


