*Biometrika* (2006), **93,** 1, *pp.* 75–84
© 2006 Biometrika Trust
*Printed in Great Britain*

# Efficient semiparametric estimator for heteroscedastic partially linear models

By YANYUAN MA

*Department of Statistics, Texas A&M University, College Station, Texas 77843, U.S.A.*

ma@stat.tamu.edu

JENG-MIN CHIOU

*Institute of Statistical Science, Academia Sinica, 128 Academia Road, Taipei 115, Taiwan*

jmchiou@stat.sinica.edu.tw

and NAISYIN WANG

*Department of Statistics, Texas A&M University, College Station, Texas 77843, U.S.A.*

nwang@stat.tamu.edu

## Summary

We study the heteroscedastic partially linear model with an unspecified partial baseline component and a nonparametric variance function. An interesting finding is that the performance of a naive weighted version of the existing estimator could deteriorate when the smooth baseline component is badly estimated. To avoid this, we propose a family of consistent estimators and investigate their asymptotic properties. We show that the optimal semiparametric efficiency bound can be reached by a semiparametric kernel estimator in this family. Building upon our theoretical findings and heuristic arguments about the equivalence between kernel and spline smoothing, we conjecture that a weighted partial-spline estimator could also be semiparametric efficient. Properties of the proposed estimators are presented through theoretical illustration and numerical simulations.

*Some key words*: Double robustness; Kernel estimation; Influence function; Partial spline; Semiparametric efficiency bound.

## 1. Introduction

We study a partially linear regression model,

$$Y_i = X_i^{\mathrm{T}}\beta + g(Z_i) + \varepsilon_i, \quad E(\varepsilon_i|X_i, Z_i) = 0 \quad (i = 1, \ldots, n), \tag{1}$$

where $X$ and $Z$ are random variables. In (1), the parametric component $X^{\mathrm{T}}\beta$ provides a simple summary of covariate effects which are of main scientific interest, while the smooth baseline component $g(.)$ is included to enrich model flexibility. Since its introduction by Engle et al. (1986), this model has been widely applied in diverse disciplines. Härdle et al. (2000, § 1.1) provides a comprehensive literature review. Even though an equal-variance assumption for $\varepsilon$ was often not made in various earlier papers, such as Liang et al. (1999), the proposed estimators often do not account for the heteroscedasticity of the model. Härdle et al. (2000, eqn 2.1.4) proposed a direct weighted extension of the estimator given in Liang et al. (1999) with the weights being inversely proportional to the variances. Nonetheless, our investigation indicates that this direct extension is still not efficient. Furthermore, when one cannot obtain a consistent estimator of $g$, the estimator of $\beta$ will be inconsistent. An intriguing finding here is that, when a constant weight function is adopted, the estimator of $\beta$ remains consistent even when $g$ is inconsistently estimated. Motivated by the robustness property of using a constant weight function, we derive a semiparametric efficient estimator which has the same robustness under nonconstant weights.

## 2. Weighted versus unweighted estimating schemes

Our goal for this section is to present an interesting finding about the relationship between weighted versus unweighted estimators and the consistency of $\hat{\beta}$ when the nuisance parameter $g$ cannot be consistently estimated. We suspect that this may be a rationale behind the use of unweighted estimators even when there exists heteroscedasticity. Nevertheless, to the best of our knowledge, this point has not been discussed in the literature. To ease the presentation of consistency considerations among different estimators, we present each estimator as a solution to an estimating equation. If the estimator is consistent, then the estimating equation converges to zero. Throughout, we denote the weight function by $w(X, Z)$. The commonly used inverse-to-variation weights have $w(X, Z) = \mathrm{var}(\varepsilon|X, Z)^{-1}$. We consider the following set of estimating equations:

$$0 = \Psi(Y, X, Z, w, \beta, \breve{g}) = n^{-1} \sum_{i=1}^{n} w_i \{Y_i - X_i^{\mathrm{T}} \beta - \breve{g}(Z_i, \beta)\} \{X_i - \hat{E}(X|Z_i)\}, \qquad (2)$$

where $\breve{g}(z, \beta)$ could be a given or estimated function of $z$ and may or may not be a function of $\beta$, $\hat{E}(.|z)$ denotes a consistent estimator of $E(.|Z = z)$, and $w_i = w(X_i, Z_i)$ denotes the weights. Almost all commonly seen $\hat{\beta}$ can be presented as solutions to one of these estimating equations. We focus on the case where $\breve{g}$ does not equal or converge to $g$ and evaluate, respectively, the limits of $\Psi$ for $w \equiv 1$ and for $w$ in general. To concentrate on the main concept, we temporarily assume here that $E(X|Z)$ is known.

We rewrite the right-hand side of (2) as

$$n^{-1} \sum_{i} w(X_i, Z_i)\{\varepsilon_i + g(Z_i) - \breve{g}(Z_i)\}\{X_i - E(X|Z_i)\}. \qquad (3)$$

When $w \equiv 1$, the solution to (2) gives an unweighted estimator. Since the last term inside the curly brackets in (3) has mean zero given $Z$, (3) converges to zero. In contrast, for a weighted estimator, (3) converges to

$$E([E\{w(X, Z)X|Z\} - E\{w(X, Z)|Z\}E(X|Z)]\{g(Z) - \breve{g}(Z)\}),$$

which is zero if $w(X, Z)$ is independent of $X$ given $Z$, but may not be zero, otherwise; that is, the inclusion of weights in the commonly used estimating equations has a tendency to lead to inconsistent $\hat{\beta}$ when a wrong $g$ is in place.

## 3. Weighted estimating equations and efficient estimator

Given that weights are commonly included for various purposes, such as accounting for sampling strategies, and that $g$ could simply be a nuisance parameter, it seems desirable to construct an estimation scheme which does not mandate a well-estimated $g$ but still accommodates the inclusion of weights. Indeed, we can achieve this goal with a slight modification of (2). We consider a new family of weighted estimating equations,

$$0 = n^{-1} \sum_{i=1}^{n} w(X_i, Z_i)\{Y_i - X_i^{\mathrm{T}}\beta - \breve{g}(Z_i, \beta)\}\left[X_i - \frac{\hat{E}\{w(X, Z)X|Z_i\}}{\hat{E}\{w(X, Z)|Z_i\}}\right]. \tag{4}$$

If we use equivalent derivations to those in § 2 and note that

$$E[w\{X - E(wX|Z)/E(w|Z)\}|Z] = 0,$$

it is easily seen that the right-hand side of (4) converges to zero even when $\breve{g}$ is misspecified. This implies that even a wrong choice of $\breve{g} \equiv 0$ still leads to a consistent $\hat{\beta}$. Since (4) is a simple modification of (2), many theoretical or numerical tools developed under (2) can be easily adapted. For a given $\breve{g}$ and kernel/local-polynomial estimator $\hat{E}(.|z)$, asymptotic normality such as that given in Theorem 2.1.1, with $w_i \equiv 1$, or Theorems 2.1.2 and 2.1.3, the weighted versions, of Härdle et al. (2000) can be obtained and proved, following similar techniques. For illustrative purposes, we provide the asymptotic normality result when $\breve{g} = 0$ in the Appendix. The scenario of $\breve{g} = 0$ could be viewed as a 'worse-case' scenario when $g(z) \neq 0$, in that one completely ignores $Z$ as a predictor of $Y$.

The consistency and asymptotic normality properties described above apply to weights chosen for any reasons. When focusing on the purpose of increasing efficiency for a heteroscedastic partially linear model, we note that Chamberlain (1992) has given the semiparametric efficiency bound. However, none of the weighted estimators given in Härdle et al. (2000, § 2) has reached this bound. In what follows, we derive the semiparametric efficient score for $\beta$ and propose a semiparametric efficient estimator as the solution to a member of (4), which mimics the score. The semiparametric efficient score is defined as the projection of the score vector on to the orthogonal complement of the nuisance tangent space; for details, see Bickel et al. (1993, p. 70).

PROPOSITION 1. *Assume that the conditional probability density function of* $\varepsilon$ *given* $(X, Z)$, $p_\varepsilon(\varepsilon|X, Z)$, *is differentiable with respect to* $\varepsilon$ *and that* $0 < E(\varepsilon^2|X, Z) < \infty$ *almost everywhere. In estimating* $\beta$ *in model* (1), *write* $w = w(X, Z) = E(\varepsilon^2|X, Z)^{-1}$. *The semiparametric efficiency score is*

$$S_{\mathrm{eff}} = w\varepsilon\left\{X - \frac{E(wX|Z)}{E(w|Z)}\right\}. \tag{5}$$

The expression for $S_{\mathrm{eff}}$ suggests that, by carefully choosing $w$ and $\breve{g}$ in (4), we should obtain a semiparametric efficient $\hat{\beta}$ as the solution to (4). A sketch of derivations that lead to Proposition 1 is given in the Appendix.

It is well known that, in terms of numerical estimation, high-dimensional estimation suffers from the curse of dimensionality. To focus our presentation on the main concepts, we assume that there exists a variable $\xi = \xi(X, Z)$ such that $\mathrm{var}(\varepsilon|X, Z) = \mathrm{var}(\varepsilon|\xi)$. We will also concentrate on the case where the dimensions of $Z$ and $\xi$ are 1. It is worth noting that the structure holds in general. The model can be extended to include intermediate multivariate models as components, for example with additive structures, so that the univariate convergence rates remain achievable and that the structure is still very flexible.

If we use nonparametrically estimated $g$, $w$, $E(w|Z)$ and $E(wX|Z)$, the estimator can be written as the solution to

$$0 = \sum_{i=1}^{n} \{Y_i - X_i^{\mathrm{T}}\beta - \hat{g}(Z_i, \beta)\}\hat{w}(X_i, Z_i)\left[ X_i - \frac{\hat{E}\{\hat{w}(X, Z)X|Z_i\}}{\hat{E}\{\hat{w}(X, Z)|Z_i\}} \right]. \tag{6}$$

Up to this point, $Z$ can be continuous or discrete. Hereafter, we assume that $Z$ is a continuous random variable and estimate $E(.|Z)$ by local-linear kernel estimation on $Z$. Note that $Z$ can obviously be discrete but different assumptions are needed to ensure that certain desirable convergence properties of nonparametrically estimated $g$, $E(w|Z)$ and $E(wX|Z)$ still hold.

One option for $\hat{g}(Z_i, \beta)$ is

$$\hat{E}(wY|Z_i)/\hat{E}(w|Z_i) - \{\hat{E}(wX|Z_i)/\hat{E}(w|Z_i)\}\beta, \tag{7}$$

even though other consistently estimated $g$ could reach equivalent asymptotic properties. Müller & Stadtmüller (1987) and Chiou & Müller (1999) give thorough discussions about nonparametric variance estimation. Chiou & Müller (1999) also use inverse-to-variation weights in semiparametric estimation. In this respect, our estimated $w$ or $\mathrm{var}(\varepsilon|\xi)$ is simply a direct application of their work; see the original papers for details.

Proposition 2. *Assume that $\hat{\beta}$ solves* (6). *Then, under the regularity conditions given in the Appendix, when $n \to \infty$,*

$$\sqrt{n}(\hat{\beta} - \beta) \to N(0, V)$$

*in distribution, where*

$$V = \{E(S_{\mathrm{eff}} S_{\mathrm{eff}}^{\mathrm{T}})\}^{-1} = \left[ E\left\{ wXX^{\mathrm{T}} - \frac{E(wX|Z)E(wX|Z)^{\mathrm{T}}}{E(w|Z)} \right\} \right]^{-1}. \tag{8}$$

A sketch of the proof of Proposition 2 is given in the Appendix. Note that Chamberlain (1992) gave the same efficiency bound as in (8), although an estimator to achieve this bound has not been derived. The above results suggest that various approaches to nonparametric estimation do not cause any efficiency loss. In other words, $\hat{\beta}$ is asymptotically equivalent to the solution to (6) with known $g$, $w$, $E(w|Z)$ and $E(wX|Z)$.

There are several slightly different ways of calculating the estimator. We outline our way of implementation below.

*Step* 1. Estimate $E(Y|Z)$ and $E(X|Z)$ nonparametrically and obtain the initial estimates $\hat{\beta} = [\{X - \hat{E}(X|Z)\}^{\mathrm{T}}\{X - \hat{E}(X|Z)\}]^{-1}[\{X - \hat{E}(X|Z)\}^{\mathrm{T}}\{Y - \hat{E}(Y|Z)\}]$, and

$$\hat{g}(Z_i) = \hat{E}(Y|Z_i) - \hat{E}(X|Z_i)^{\mathrm{T}}\hat{\beta}.$$

*Step* 2. Calculate $\hat{\varepsilon}_i = Y_i - X_i^{\mathrm{T}}\hat{\beta} - \hat{g}(Z_i)$.

*Step* 3. Obtain nonparametrically estimated $\hat{w}_i = \hat{E}(\hat{\varepsilon}^2|X_i, Z_i)^{-1}$, for $i = 1, \ldots, n$.

*Step* 4. For the set of $\{\hat{w}_i\}$, obtain $\hat{E}(\hat{w}_i|Z_i)$, $\hat{E}(\hat{w}X|Z_i)$ and $\hat{E}(\hat{w}Y|Z_i)$, for $i = 1, \ldots, n$, in the form of local-linear estimators.

*Step* 5. Let $\tilde{X}_i = X_i - \hat{E}(\hat{w}X|Z_i)/\hat{E}(\hat{w}|Z_i)$ and $\tilde{Y}_i = Y_i - \hat{E}(\hat{w}Y|Z_i)/\hat{E}(\hat{w}|Z_i)$, and let $W$ be a diagonal matrix with $\hat{w}_i$ being the $i$th diagonal element. Calculate

$$\hat{\beta} = (\tilde{X}^{\mathrm{T}} W \tilde{X})^{-1} \tilde{X}^{\mathrm{T}} W \tilde{Y}, \tag{9}$$

and $\hat{g}$ as in (7) with $\beta = \hat{\beta}$.

*Step* 6. Iterate Steps 2 to 5 until convergence and use the sandwich covariance estimate based on (9) to estimate the asymptotic variance matrix of $\hat{\beta}$, with the estimated variance matrix of $\tilde{Y}$ being a diagonal matrix with $i$th diagonal element equal to $\hat{E}(\hat{\varepsilon}^2|X_i, Z_i)$.

While we focus on kernel methods in demonstrating the efficiency of our estimator, in practice, estimating $g$ by a spline estimator is as common as estimating $g$ by the kernel method. Heckman (1986), Rice (1986) and others adopted smoothing spline techniques in the estimation of $\beta$ and $g$ in model (1). Ruppert et al. (2003, Ch. 7, 9) provide various applications using partially linear penalised splines. A weighted version of spline estimation is given in (4.4) of Green & Silverman (1994):

$$\arg\min_{\beta, g} \sum_{i=1}^{n} w_i \{Y_i - X_i^{\mathrm{T}}\beta - g(Z_i)\}^2 + \lambda \int g''(z)^2 dz.$$

Although the pointwise asymptotic equivalence between kernel- and spline-estimated $g$ has been established for both uncorrelated data (Silverman, 1984; Nychka, 1995) and correlated data (Lin et al., 2004), without uniform equivalence results, it is nontrivial to establish rigorously equivalence between the ways of estimating $\beta$. The arguments given below nonetheless suggest that such equivalence could be established provided that 'uniform' equivalence between regular spline and kernel estimators existed.

Consider a model, $Y^* = g(Z_i) + \varepsilon$, let

$$\hat{g}_{\mathrm{S}}(z) = n^{-1} \sum_i \gamma_{\mathrm{S}}(z, Z_i) Y_i^*, \quad \hat{g}_{\mathrm{K}}(z) = n^{-1} \sum_i \gamma_{\mathrm{K}}(z, Z_i) Y_i^*,$$

and let $G_{\mathrm{S}}$ and $G_{\mathrm{K}}$ denote the hat matrices with their $(i, j)$th elements being $\gamma_{\mathrm{S}}(Z_i, Z_j)$ and $\gamma_{\mathrm{K}}(Z_i, Z_j)$, with the subscripts 'S' and 'K' standing for spline and kernel, respectively. It is straightforward to show that $\hat{\beta}_{\mathrm{K}}$ satisfies

$$X^{\mathrm{T}}(I - G_{\mathrm{K}})^{\mathrm{T}} W (I - G_{\mathrm{K}}) X \hat{\beta}_{\mathrm{K}} = X^{\mathrm{T}}(I - G_{\mathrm{K}})^{\mathrm{T}} W (I - G_{\mathrm{K}}) Y,$$

while, by Green & Silverman (1994, § 4.2), $\hat{\beta}_{\mathrm{S}}$ satisfies $X^{\mathrm{T}} W (I - G_{\mathrm{S}}) X \hat{\beta}_{\mathrm{S}} = X^{\mathrm{T}} W (I - G_{\mathrm{S}}) Y$. As noted in Hu et al. (2004, § 3) and Opsomer & Ruppert (1999) for independent errors and with proper undersmoothing, $\hat{\beta}_{\mathrm{K}}$ is asymptotically equivalent to $\hat{\beta}_{\mathrm{K}}^*$ which solves $X^{\mathrm{T}} W (I - G_{\mathrm{K}}) X \hat{\beta}_{\mathrm{K}}^* = X^{\mathrm{T}} W (I - G_{\mathrm{K}}) Y$. If the uniform equivalence between $G_{\mathrm{S}}$ and $G_{\mathrm{K}}$ were established, then one should be able to identify a proper $\lambda$ corresponding to the undersmoothed version of the kernel estimator, $\hat{\beta}_{\mathrm{K}}^*$, and show that $\hat{\beta}_{\mathrm{S}}$ and $\hat{\beta}_{\mathrm{K}}^*$ share the same asymptotic distribution. A rigorous proof to show the uniform equivalence is nonetheless beyond the scope of this paper.

## 4. SIMULATIONS

We use a small simulation study to examine the finite-sample performance of seven different estimators, all with the estimating equation structure of

$$n^{-1}\sum_i w_i r_i H(X_i, Z_i) = 0,$$

where $r$ is the residual from fitting $Y$:

(a) $n^{-1}\sum_{i=1}^{n} \hat{w}_i\{Y_i - X_i^{\mathrm{T}}\beta - \hat{g}(Z_i,\beta)\}\left\{X_i - \dfrac{\hat{E}(\hat{w}X|Z_i)}{\hat{E}(\hat{w}|Z_i)}\right\} = 0,$

(b) $n^{-1}\sum_{i=1}^{n} \hat{w}_i\{Y_i - X_i^{\mathrm{T}}\beta - \hat{g}(Z_i,\beta)\}\{X_i - \hat{E}(X|Z_i)\} = 0,$

(c) $n^{-1}\sum_{i=1}^{n} \{Y_i - X_i^{\mathrm{T}}\beta - \hat{g}(Z_i,\beta)\}\{X_i - \hat{E}(X|Z_i)\} = 0,$

(d) $n^{-1}\sum_{i=1}^{n} \hat{w}_i(Y_i - X_i^{\mathrm{T}}\beta)\left\{X_i - \dfrac{\hat{E}(\hat{w}X|Z_i)}{\hat{E}(\hat{w}|Z_i)}\right\} = 0,$

(e) $n^{-1}\sum_{i=1}^{n} \hat{w}_i\{Y_i - X_i^{\mathrm{T}}\beta - p_2(Z_i)\}\left\{X_i - \dfrac{\hat{E}(\hat{w}X|Z_i)}{\hat{E}(\hat{w}|Z_i)}\right\} = 0,$

(f) $n^{-1}\sum_{i=1}^{n} \hat{w}_i(Y_i - X_i^{\mathrm{T}}\beta)\{X_i - \hat{E}(X|Z_i)\} = 0,$

(g) $n^{-1}\sum_{i=1}^{n} \hat{w}_i\{Y_i - X_i^{\mathrm{T}}\beta - p_2(Z_i)\}\{X_i - \hat{E}(X|Z_i)\} = 0,$

where $p_2(Z)$ in (e) and (g) represents a quadratic polynomial function of $Z$. Estimator (a) is the semiparametric efficient kernel estimator we propose, which is calculated as described in § 3. Estimators (b) and (c) correspond to two existing weighted and unweighted partially linear kernel estimators described in Härdle et al. (2000, § 2), respectively. Estimators (d) and (f) correspond to the scenario in which $\breve{g} = 0$; that is, one does not consider $Z$ as a covariate in the main component of the asymptotically unbiased estimating equation but only includes $Z$ through $H(X, Z)$. Estimators (e) and (g) correspond to the scenario in which $\breve{g}$ is approximated by a quadratic polynomial. Since the choices of both not including $Z$ in $r_i$ and specifying $\breve{g}$ as quadratic are incorrect, the residuals may not be centred at zero. We thus use the estimated weights obtained in (a) here. The theoretical outcomes suggest that estimators (a)–(e) are asymptotically unbiased. We expect estimators (f) and (g) to be biased, as explained in § 2. In practice, estimator (d) or estimator (e) could provide a quick assessment of $\beta$. One can simply let $w(X_i, Z_i) \equiv 1$ or use historical weights, if appropriate, to obtain a consistent solution.

We generated $Z_i$ from a Un(2, 4) distribution, $X_i$ from a $N(2Z_i, 4Z_i^2)$ distribution, and $Y_i$ from a normal distribution with mean $X_i + 10\sin(2Z_i)$ and variance $\mathrm{var}(\varepsilon_i|X_i, Z_i) = w_i^{-1} = X_i^2 + 1$. Thus, the true regression parameter $\beta$ is 1 and the true $g(Z)$ is $g(Z_i) = \tau\sin(2Z_i)$ with $\tau = 10$. Simulations with sample size $n = 250$ were repeated 1000 times. The simulation results are presented in Table 1, where the bias, standard error and mean squared error are sample statistics calculated from the estimates of 1000 replicates.

Table 1: *Simulation results for estimators* (a)–(g). *The results are based on* 1000 *replicates of data with sample size* 250 *for* $g(Z) = \tau \sin(2z)$, *and* $\tau = 10$, 5 *and* 1, *respectively.*

| | $\tau = 10$ | | | $\tau = 5$ | | | $\tau = 1$ | | |
|---|---|---|---|---|---|---|---|---|---|
| Est | Bias | SE | MSE | Bias | SE | MSE | Bias | SE | MSE |
| (a) | 0·0001 | 0·1007 | 0·0101 | −0·0010 | 0·0984 | 0·0097 | −0·0017 | 0·0970 | 0·0094 |
| (b) | 0·0161 | 0·1111 | 0·0126 | 0·0085 | 0·1105 | 0·0123 | 0·0028 | 0·1111 | 0·0123 |
| (c) | −0·0016 | 0·1371 | 0·0188 | −0·0016 | 0·1371 | 0·0188 | −0·0015 | 0·1370 | 0·0188 |
| (d) | −0·0093 | 0·1164 | 0·0136 | −0·0060 | 0·1096 | 0·0121 | −0·0023 | 0·1122 | 0·0126 |
| (e) | −0·0045 | 0·1065 | 0·0114 | −0·0047 | 0·1139 | 0·0130 | −0·0045 | 0·1217 | 0·0148 |
| (f) | 0·3443 | 7·8879 | 62·3383 | −0·1267 | 12·7871 | 163·5264 | 0·0513 | 0·4473 | 0·2027 |
| (g) | 0·0017 | 2·5419 | 6·4613 | −0·1880 | 4·0902 | 16·7652 | −0·2949 | 13·8599 | 192·1851 |

Est, estimator; Bias, Monte Carlo average of the estimates minus the true value; SE, Monte Carlo standard error; MSE, mean squared error of the estimates.

We repeated the simulation with $\tau = 5$ and $\tau = 1$, that is $g(Z) = 5\sin(2Z)$ and $g(Z) = \sin(2Z)$. Unlike in the previous set-up, the signals from $g$ are smaller in scale and thus it is harder to identify them correctly with the existing noise level. The results from all three sets of simulations are reported in Table 1.

The Monte Carlo biases for the semiparametric estimators (a)–(e) are relatively small, whereas the biases for estimators (f) and (g) are large compared to those of the estimators (a)–(e). In addition, estimator (a) has the smallest Monte Carlo standard errors and mean squared errors. When $\tau = 10$, 5 and 1, the Monte Carlo standard errors are 0·1007, 0·0984 and 0·097, while the averages of the estimated standard errors of estimator (a) are 0·0963, 0·0946 and 0·0932, respectively. The latter set of values are somewhat smaller than the former. The corresponding coverage probabilities for 95% confidence intervals are 92·7%, 92·9% and 92·6%, respectively. Apparently, even though there should be no effect asymptotically resulting from the estimation of $w$, it is not exactly the case when the sample size is 250. When we increase the sample size to 500, all three coverage probabilities are between 94% and 95%.

When the magnitude of $g(Z)$ decreases with $\tau$, the mean function is increasingly dominated by the linear part, $X\beta$, and the performance of estimator (d) improves while that of estimator (g) deteriorates badly. In all the scenarios we studied, the semiparametric efficient estimator (a) consistently performed well.

## 5. Conclusion

Our final note is to point out that this estimator has the so-called double robustness feature; that is, its consistency is ensured if either $g$ or $E(wX|Z)/E(w|Z)$ can be consistently estimated. Estimating both quantities nonparametrically seems also to provide an estimator with a small finite-sample bias, as observed in our simulation study.

Equivalent arguments and conclusions hold when model (1) is generalised to the scenario in which the partially linear function $X_i^{\mathrm{T}}\beta + g(Z_i)$ is replaced by an arbitrary semiparametric function $m(X_i, Z_i, \beta, g)$. The estimator will have the form

$$\sum_{i=1}^{n} \{Y_i - \hat{m}_i(\beta)\}\hat{w}_i \left[ \frac{\partial \hat{m}_i(\beta)}{\partial \beta} - \frac{\hat{E}\{\hat{w}_i \partial \hat{m}_i(\beta)/\partial\beta | Z_i\}}{\hat{E}(\hat{w}_i|Z_i)} \right] = 0,$$

where $\hat{m}_i(\beta) = m(X_i, Z_i, \beta, \hat{g})$. The estimator remains consistent and efficient provided that $E(\hat{w}\partial\hat{m}/\partial\beta|Z)$ and $E(\hat{w}|Z)$ are properly estimated nonparametrically.

Finally, when the dimensions of $Z$ and $\xi$ are high, the partially additive model represents one strategy to overcome the curse of dimensionality. To be specific, one lets $g(Z) = \sum_j g_j(Z_j)$ and $v(\xi) = \sum_j v(\xi_j)$, where $Z_j$ and $\xi_j$ are the $j$th components of $Z$ and $\xi$. The method and procedure can be extended to obtain efficient estimators for such partially additive linear models. Recent progress in this area has been nicely summarised in Fan & Li (2003).

## Acknowledgement

This work was supported by grants from the U.S. National Institutes of Health. Chiou's research is partially supported by the National Science Council, Taiwan.

## Appendix

### *Technical details*

We briefly outline the essential technical materials here. The detailed conditions and derivations can be obtained from Y. Ma.

*Asymptotic normality of $\hat{\beta}$ with $\breve{g} = 0$.* Let

$$A = E\left[ w\left\{X - \frac{E(wX|Z)}{E(w|Z)}\right\} X^{\mathrm{T}} \right],$$

$$B = E\left[ w^2(Y - X^{\mathrm{T}}\beta)^2 \left\{X - \frac{E(wX|Z)}{E(w|Z)}\right\}\left\{X - \frac{E(wX|Z)}{E(w|Z)}\right\}^{\mathrm{T}} \right],$$

and note that the influence function is proportional to $w(Y - X^{\mathrm{T}}\beta)\{X - E(wX|Z)/E(w|Z)\}$. When $A^{-1}$ exists, the estimator is consistent. Furthermore, as $n \to \infty$, $\sqrt{n}(\hat{\beta} - \beta_0) \to N(0, A^{-1}BA^{-\mathrm{T}})$ in distribution.

*Proof of Proposition* 1. Denote the score function for $\beta$ by $S_\beta$. We obtain that

$$S_\beta = -Xp_\varepsilon'(\varepsilon|X, Z)/p_\varepsilon(\varepsilon|X, Z),$$

where $p_\varepsilon'(.|X, Z)$ is the derivative with respect to $\varepsilon$. It can be verified that the nuisance tangent space $\Lambda$, for the three nuisance parameters $p(X, Z)$, $p_\varepsilon(\varepsilon|X, Z)$ and $g(Z)$, is

$$\Lambda = \{f(X, Y, Z) : E(f) = 0, E(\varepsilon f|X, Z) \text{ being a function of } Z \text{ only}\},$$

and its orthogonal complement $\Lambda^\perp = \{\varepsilon f(X, Z) : E(f|Z) = 0\}$. The function $S_{\mathrm{eff}}$, as defined in (5), certainly satisfies $S_{\mathrm{eff}} \in \Lambda^\perp$. For an arbitrary element $\varepsilon f(X, Z) \in \Lambda^\perp$, $E(f|Z) = 0$, and, dropping the arguments $(X, Z)$ from $f(X, Z)$, we have that $E\{(S_\beta - S_{\mathrm{eff}})^{\mathrm{T}}\varepsilon f\}$ equals

$$E\left[ -X^{\mathrm{T}} f E\left(\frac{\varepsilon p_\varepsilon'}{p_\varepsilon}|X, Z\right) - E(\varepsilon^2|X, Z) w \left\{X^{\mathrm{T}} f - \frac{E(wX^{\mathrm{T}}|Z)}{E(w|Z)} f\right\} \right].$$

Note that $w = E(\varepsilon^2|X, Z)^{-1}$ and $E(\varepsilon p_\varepsilon'/p_\varepsilon|X, Z) = -1$, and consequently we obtain

$$E\{(S_\beta - S_{\text{eff}})^{\mathrm{T}} \varepsilon f(X, Z)\} = E\left\{X^{\mathrm{T}} f - X^{\mathrm{T}} f + \frac{E(wX^{\mathrm{T}}|Z)}{E(w|Z)} f\right\}$$

$$= E\left\{\frac{E(wX^{\mathrm{T}}|Z)}{E(w|Z)} E(f|Z)\right\} = 0.$$

Thus, $S_{\text{eff}}$ is the projection of $S_\beta$ on to $\Lambda^\perp$. □

*Regularity conditions and Proof of Proposition* 2. The regularity conditions are analogous to (M1–M7) and (K1–K4) of Chiou & Müller (1999).

*Condition* 1. The errors $\varepsilon_i$ $(i = 1, \ldots, n)$ are independent and $0 < E(\varepsilon_i^2) < \infty$ and $\text{var}(\varepsilon_i^2) > 0$.

*Condition* 2. There exist a variance function, $v(.)$, $\xi_i = \xi(X_i, Z_i)$ and a positive constant $\gamma$, such that $E(\varepsilon^2|x_i, z_i) = v(\xi_i)$, with $v(.) > \gamma > 0$.

*Condition* 3. The functions $g(z)$, $E(X|z)$, $E(Y|z)$, $v(\xi)$, $E(w|z)$ and $E(wX|z)$ are twice continuously differentiable with finite derivatives. As a function of $(x, z)$, $\xi$ is three times differentiable with finite derivatives.

*Condition* 4. There exists a function $\mu_4(.)$ such that $E(\varepsilon_i^4) = \mu_4(\xi_i)$. The function $\mu_4(.)$ is continuous; furthermore, there exists an $s > 2$ such that $\max_{1 \leqslant i \leqslant n} E\varepsilon_i^{2s} < c < \infty$ for some $c > 0$. This condition is necessary for obtaining uniform consistency of the estimator for the variance function $v(.)$.

*Condition* 5. Assume that the random variables $X_i$ have a density, $f_X$, and that the support of $f_X$ is a compact interval. This condition ensures that $X_i^{\mathrm{T}}\beta$ is bounded.

*Condition* 6. Assume that the random variables $\xi_i$ and $Z_i$ have densities, $f_\xi$ and $f_Z$, respectively, that the supports of $f_\xi$ and $f_Z$ are compact intervals and that $f_\xi$ and $f_Z$ are twice continuously differentiable, satisfying $0 < \inf f_\xi(.) \leqslant \sup f_\xi(.) < \infty$ and $0 < \inf f_Z(.) \leqslant \sup f_Z(.) < \infty$. This condition enables us to simplify asymptotic expressions of certain sums of functions of variables. This condition also excludes pathological cases where the number of observations in a window defined by the bandwidth may not increase to infinity when $n \to \infty$.

*Condition* 7. The kernel function $K$ is symmetric and continuously differentiable with compact support $[-1, 1]$.

*Condition* 8. The bandwidth $h$ used in the kernel estimators satisfies $h \to 0$, $nh^3 \to \infty$ and $nh^8 \to 0$ as $n \to \infty$, and

$$\liminf_{n \to \infty} (nh/\log n)^{1/2} n^{-2/r} > 0$$

for a constant $r$ $(0 < r < s)$ where $s$ is given in Condition 4.

*Condition* 9. The estimators of the variance function $v(.)$ are such that $\hat{v}_n(.)$ are truncated below by a sequence $\zeta_n > 0$, where $\zeta_n \to 0$. This sequence satisfies $h/\zeta_n \to 0$, $nh^2\zeta_n^2 \to \infty$ and $nh\zeta_n^2/\log n \to \infty$.

The sketch of the proof is as follows. Following the equivalent steps that prove Theorems 4.1 and 4.2 in Chiou & Müller (1999), we can show that $\widehat{\text{var}}(\varepsilon|\xi)$ converges uniformly to $\text{var}(\varepsilon|\xi)$ and that $\hat{\beta}$, based on the estimated $w$, shares the same asymptotic distribution as that of $\hat{\beta}$ based on the known $w$. Therefore, we only outline the procedures under the scenario of known $w$.

Under the true $\beta$, write $g(Z) = g(Z, \beta) = E(wY|Z)/E(w|Z) - E(wX|Z)/E(w|Z)\beta$, and use local-linear estimation to replace $E(.|Z)$ by $\hat{E}(.|Z)$ in $\hat{g}(Z, \beta)$. Note that this equality in $g(Z, \beta)$ holds for general choices of weights. Let

$$D = E\left[w\left\{X - \frac{E(wX|Z)}{E(w|Z)}\right\}\left\{X - \frac{E(wX|Z)}{E(w|Z)}\right\}^{\mathrm{T}}\right].$$

Some derivations lead to

$$\sqrt{n}D(\hat{\beta}-\beta)+o_p(1)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varepsilon_i w_i\left\{X_i-\frac{E(wX|Z_i)}{E(w|Z_i)}\right\}$$

$$-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varepsilon_i w_i\left\{\frac{\hat{E}(wX|Z_i)}{\hat{E}(w|Z_i)}-\frac{E(wX|Z_i)}{E(w|Z_i)}\right\} \quad \text{(A1)}$$

$$-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{\hat{g}(Z_i,\beta)-g(Z_i)\}w_i\left\{X_i-\frac{E(wX|Z_i)}{E(w|Z_i)}\right\} \quad \text{(A2)}$$

$$+\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{\hat{g}(Z_i,\tilde{\beta})-g(Z_i)\}w_i\left\{\frac{\hat{E}(wX|Z_i)}{\hat{E}(w|Z_i)}-\frac{E(wX|Z_i)}{E(w|Z_i)}\right\}. \quad \text{(A3)}$$

If we use the properties of local-linear regression and note that $E(\varepsilon)=0$, it is easy to show that (A1) is $o_p(1)$. Similarly, using the fact that $E[w\{X-E(wX|Z)/E(w|Z)\}|Z]=0$ and $\hat{g}(Z,\beta)=g(Z)+o_p(1)$, one can prove (A2) to be $o_p(1)$. Finally, with $n^{1/2}h^4\to 0$ and $(nh)^{-1}\to 0$, we have that (A3) is $o_p(1)$. Hence $\sqrt{n}(\hat{\beta}-\beta)\to N(0,\Sigma_\beta)$ in distribution with $\Sigma_\beta=D^{-1}\,\mathrm{var}[w\varepsilon\{X-E(wX|Z)/E(w|Z)\}]D^{-1}$. Note that $S_{\mathrm{eff}}(Y_i,X_i,Z_i,\beta)=w_i\varepsilon_i\{X_i-E(wX|Z_i)/E(w|Z_i)\}$. Recall that, in Proposition 2, $w=w(\xi)=\mathrm{var}^{-1}(\varepsilon|\xi)$. It is easily seen that $\Sigma_\beta=D^{-1}=V$, where $V$ is defined in (8). □

## References


Bickel, P., Klaassen, C. A. J., Ritov, Y. & Wellner, J. A. (1993). *Efficient and Adaptive Inference in Semiparametric Models*. Baltimore: Johns Hopkins University Press.

Chamberlain, G. (1992). Efficiency bounds for semiparametric regression. *Econometrica* **60**, 567–96.

Chiou, J. M. & Müller, H.-G. (1999). Nonparametric quasi-likelihood. *Ann. Statist.* **27**, 36–64.

Engle, R. F., Granger, C. W. J., Rice, J. & Weiss, A. (1986). Semiparametric estimates of the relation between weather and electricity sales. *J. Am. Statist. Assoc.* **81**, 310–20.

Fan, Y. & Li, Q. (2003). A kernel-based method for estimating additive partially linear models. *Statist. Sinica* **13**, 739–62.

Green, P. J. & Silverman, B. W. (1994). *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*. London: Chapman and Hall.

Härdle, W., Liang, H. & Gao, J. (2000). *Partially Linear Models*. Heidelberg: Physica-Verlag.

Heckman, N. E. (1986). Spline smoothing in a partly linear model. *J. R. Statist. Soc.* B **48**, 244–8.

Hu, Z., Wang, N. & Carroll, R. J. (2004). Profile-kernel versus backfitting in the partially linear models for longitudinal/clustered data. *Biometrika* **91**, 251–62.

Liang, H., Härdle, W. & Carroll, R. J. (1999). Estimation in a semiparametric partially linear errors-in-variables model. *Ann. Statist.* **5**, 1519–35.

Lin, X., Wang, N., Welsh, A. H. & Carroll, R. J. (2004). Equivalent kernels of smoothing splines in nonparametric regression for clustered data. *Biometrika* **91**, 177–93.

Müller, H.-G. & Stadtmüller, U. (1987). Estimation of heteroscedasticity in regression analysis, *Ann. Statist.* **15**, 610–25.

Nychka, D. (1995). Splines as local smoothers. *Ann. Statist.* **23**, 1175–97.

Opsomer, J. D. & Ruppert, D. (1999). A root-n consistent backfitting estimator for semiparametric additive modeling. *J. Comp. Graph. Statist.* **8**, 715–32.

Rice, J. (1986). Convergence rates for partially splined models. *Statist. Prob. Lett.* **4**, 203–8.

Ruppert, D., Wang, M. P. & Carroll, R. J. (2003). *Semiparametric Regression*. New York: Cambridge University Press.

Silverman, B. (1984). Spline smoothing: the equivalent variable kernel method. *Ann. Statist.* **12**, 898–916.


[*Received December* 2004. *Revised October* 2005]