Published in final edited form as:

Biometrika. 2008 ; 95(3): . doi:10.1093/biomet/asn024.

Additive Partial Linear Models with Measurement Errors

HUA LIANG and **SALLY W. THURSTON**

Department of Biostatistics and Computational Biology, University of Rochester Medical Center, Rochester, New York 14642, U.S.A. hliang@bst.rochester.edu thurston@bst.rochester.edu

DAVID RUPPERT and **TATIYANA APANASOVICH**

College of Engineering, Cornell University, Ithaca, New York 14853, U.S.A.
dr24@cornell.edu tva2@cornell.edu

RUSS HAUSER

Department of Environmental Health, Harvard School of Public Health Boston, Massachusetts 02115, U.S.A. rhauser@hsph.harvard.edu

SUMMARY

We consider statistical inference for additive partial linear models when the linear covariate is measured with error. We propose attenuation-to-correction and SIMEX estimators of the parameter of interest. It is shown that the first resulting estimator is asymptotically normal and requires no undersmoothing. This is an advantage of our estimator over existing backfitting-based estimators for semiparametric additive models which require undersmoothing of the nonparametric component in order for the estimator of the parametric component be root- n consistent. This feature stems from a decrease of the bias of the resulting estimator which is appropriately derived using a profile procedure. A similar characteristic in semiparametric partially linear models was obtained by Wang et al. (2005). We also discuss the asymptotics of the proposed SIMEX approach. Finite-sample performance of the proposed estimators is assessed by simulation experiments. The proposed methods are applied to a dataset from a semen study.

Keywords

Backfitting; Correction-for-attenuation; Error-prone; Local linear regression; Semen quality study; Semiparametric estimation; SIMEX; Undersmoothing

1. INTRODUCTION

The standard additive model, a generalization of the multiple linear regression models that replaces linear components by one-dimensional nonparametric functions, has been used widely to explore the complicated relationship between a response to treatment and predictors of interest (Stone, 1985). Efforts have also been made to balance the interpretability of linear models and flexibility of additive models. Important results of those efforts are semiparametric additive partial linear models, in which some of the additive component functions are linear while the remaining ones are modelled nonparametrically (Opsomer & Ruppert, 1999). Additive partial linear models can be described as

$$Y = X^T \beta + \sum_{d=1}^D f_d(Z_d) + \varepsilon, \quad (1)$$

where $X = (x_1, \dots, x_p)^T$ and $Z = (Z_1, \dots, Z_D)^T$ are the linear and nonparametric components, $f_1, \dots, f_D$ are unknown smooth functions, $\beta = (\beta_1, \dots, \beta_p)$ is a vector of unknown parameters, and the model error ε has conditional mean zero given (X, Z) .

A special case of (1) with $D = 1$, the partial linear models, has been studied, for example by Engle et al. (1986), Speckman (1988), Robinson (1988), Severini & Staniswalis (1994), Zeger & Diggle (1994), Opsomer & Ruppert (1999) and Härdle et al. (2000). Liang et al. (1999) considered the partial linear models when the linear component X is measured with error. They proposed a correction-for-attenuation estimator of the parameter of interest, and established the asymptotic properties for the proposed estimators. The study of the partial linear models is slightly easier than that of the additive partial linear models, in part because there is only one nonparametric term. In addition, implementation for the partial linear models is simpler than for the additive partial linear models because the backfitting algorithm (Hastie & Tibshirani, 1990; Opsomer & Ruppert, 1999) is necessary for additive partial linear models, but not for partial linear models (Opsomer & Ruppert, 1999).

Our interest is in estimating β when the covariate X is measured with error. We assume an additive measurement error model relating X and its surrogate W :

$$W = X + U, \quad (2)$$

where the measurement error U is independent of (X, Z, Y) , and has mean zero and covariance matrix σ_U^2 . We term (1) and (2) the additive partial linear measurement error model.

Our motivation is a study of the relationship between environmental chemical exposures and semen quality, conducted by R. Hauser at Massachusetts General Hospital Fertility Center. We focus on the relationship between semen quality and phthalate exposure, a topic that was also considered in previous papers (Duty et al., 2003; Hauser et al., 2006). Phthalates are used in personal care products, to soften plastics, and are frequently used in many household items; see the website <http://www.atsdr.cdc.gov/toxprofiles/> of the Agency for Toxic Substances and Disease Registry. Phthalate exposure is measured by urinary metabolites which have half lives of hours in the body. The need to correct for measurement error in phthalate exposure arises because of temporal variability of phthalate exposure, which implies that a phthalate measurement at a single time-point only approximates total phthalate exposure within the 3-month time period needed for sperm development. A study of temporal variability in phthalate metabolites among a subset of the semen study found substantial variability, even within a 3-day window (Hauser et al., 2004). Nonlinear relationships between semen quality and several covariates are expected, especially for abstinence time (Carlsen et al., 2004) and age.

2. CORRECTION-FOR-ATTENUATION ESTIMATION

For notational simplicity, we first consider $D = 2$ in (1). To ensure identifiability of the non-parametric functions, we assume that $E\{f_1(Z_1)\} = E\{f_2(Z_2)\} = 0$. Without loss of generality, we also assume that X and Y are centred.

Let $(X_1, Z_{11}, Z_{12}, W_1, Y_1), \dots, (X_n, Z_{n1}, Z_{n2}, W_n, Y_n)$ be a random sample of size n from model (1) and (2). Let $X = (X_1^T, \dots, X_n^T)^T$, $W = (W_1^T, \dots, W_n^T)^T$, $Y = (Y_1, \dots, Y_n)^T$ and similarly for Z_1 and Z_2 . We write the vectors of additive functions at the observations as $F_1 = \{f_1(Z_{11}), \dots, f_1(Z_{n1})\}^T$ and $F_2 = \{f_1(Z_{12}), \dots, f_1(Z_{n2})\}^T$. In matrix notation, (1) can be expressed as

$$Y = X\beta + F_1 + F_2 + \varepsilon. \quad (3)$$

Before deriving our correction-for-attenuation estimator, we briefly review the backfitting algorithm for model (3) without measurement errors, and refer readers to Hastie &

Tibshirani (1990, pp. 118-21) and Opsomer & Ruppert (1999) for further details. We only consider local linear regression as the smoothing method in what follows. Extensions to higher-order local polynomial estimation are straightforward. In what follows, we write $A^2 = AA^T$ and $A^{*2} = A^T A$. Also, let $\tilde{X}_i = X_i - E(X_i|Z_{i1}) - E(X_i|Z_{i2})$ and $\tilde{W}_i = W_i - E(W_i|Z_{i1}) - E(W_i|Z_{i2})$, and denote $\text{cov}(\tilde{X}_1)$ by X/Z .

Let s_{1,z_1}^T, s_{2,z_2}^T be the equivalent kernels for the local linear regression at z_1 and z_2 :

$$s_{1,z_1}^T = e_1^T (Z_1^T \Omega_1 Z_1)^{-1} Z_1^T \Omega_1, \quad s_{2,z_2}^T = e_1^T (Z_2^T \Omega_2 Z_2)^{-1} Z_2^T \Omega_2,$$

where $e_1 = (1, 0)^T$,

$$\Omega_1 = \text{diag} \left\{ \frac{1}{h_1} K \left(\frac{Z_{11} - z_1}{h_1} \right), \dots, \frac{1}{h_1} K \left(\frac{Z_{n1} - z_1}{h_1} \right) \right\},$$

$$\Omega_2 = \text{diag} \left\{ \frac{1}{h_2} K \left(\frac{Z_{12} - z_2}{h_2} \right), \dots, \frac{1}{h_2} K \left(\frac{Z_{n2} - z_2}{h_2} \right) \right\},$$

for a kernel function $K(\cdot)$ and bandwidths h_1 and h_2 , and

$$Z_1 = \begin{bmatrix} 1 & Z_{11} - z_1 \\ \vdots & \vdots \\ 1 & Z_{n1} - z_1 \end{bmatrix}, \quad Z_2 = \begin{bmatrix} 1 & Z_{12} - z_2 \\ \vdots & \vdots \\ 1 & Z_{n2} - z_2 \end{bmatrix}$$

are $n \times 2$ design matrices. Let S_1 and S_2 be the smoother matrices whose rows are the equivalent kernels at the observations $(Z_{11}, \dots, Z_{n1})^T$ and $(Z_{12}, \dots, Z_{n2})^T$; i.e.,

$$S_1 = \begin{bmatrix} s_{1Z_{11}} \\ \vdots \\ s_{1Z_{n1}} \end{bmatrix}, \quad S_2 = \begin{bmatrix} s_{2Z_{12}} \\ \vdots \\ s_{2Z_{n2}} \end{bmatrix}.$$

Then $S_1^c = (I - 11^T/n) S_1$ is the centred smoother matrix corresponding to S_1 , and similarly for S_2^c , where 1 is the one-vector of length n . A backfitting algorithm for the additive partial linear models was defined as the solution to the following equations (Hastie & Tibshirani, 1990, pp. 118-21; Opsomer & Ruppert, 1999):

$$\begin{aligned} \hat{\beta} &= (X^T X)^{-1} X^T \left(Y - \sum_{d=1}^2 \hat{F}_d \right), \\ F_1^{(m)} &= S_1^c \left(Y - X \hat{\beta} - F_2^{(m-1)} \right), \\ F_2^{(m)} &= S_2^c \left(Y - X \hat{\beta} - F_1^{(m-1)} \right), \end{aligned} \quad (4)$$

where $F_1^{(m)}$ and $F_2^{(m)}$ denote the estimators at the m stage of the backfitting algorithm. As a consequence, noniterative estimators of are of the form

$$\hat{\beta}_{\text{or}} = \left\{ X^T (I - S_{12}) X \right\}^{-1} X^T (I - S_{12}) Y, \quad (5)$$

where $S_{12} = \left\{ I - (I - S_1^c S_2^c)^{-1} (I - S_1^c) \right\} + \left\{ I - (I - S_2^c S_1^c)^{-1} (I - S_2^c) \right\}$.

To ensure that $\hat{\beta}_{or}$ is root- n consistent, undersmoothing is necessary. We are going to remove this restriction by using a profile likelihood procedure (Severini & Staniswalis, 1994). The basic idea can be described as follows. For each z_1, z_2 and β , let $\hat{f}_1(\beta, z_1)$ and $\hat{f}_2(\beta, z_2)$ be the backfitting estimators of $f_1(z_1)$ and $f_2(z_2)$, which are given by the second formula of (4) except for replacing $\hat{\beta}$ by β . Correspondingly, $\hat{F}_1$ and $\hat{F}_2$ can be expressed as

$$\begin{aligned}\hat{F}_1(\beta) &= \left\{ I - (I - S_1^c S_2^c)^{-1} (I - S_1^c) \right\} \left(Y - X^T \beta \right), \\ \hat{F}_2(\beta) &= \left\{ I - (I - S_2^c S_1^c)^{-1} (I - S_2^c) \right\} \left(Y - X^T \beta \right).\end{aligned}\quad (6)$$

Substituting $\hat{F}_1(\beta)$ and $\hat{F}_2(\beta)$ in (3) and using the least-squares principle, we obtain the profile-based estimators of β as follows:

$$\hat{\beta}_{pb} = \left\{ X^T (I - S_{12})^{*2} X \right\}^{-1} X^T (I - S_{12})^{*2} Y. \quad (7)$$

As discussed in Hastie & Tibshirani (1990, pp.118-21) and Opsomer & Ruppert (1997), the centring of S_1^c and S_2^c is necessary to ensure that the algorithm converges and the estimators given in (7) are well defined by assuming that $\sum_{i=1}^n f_1(Z_{i1}) = \sum_{i=1}^n f_2(Z_{i2}) = 0$. The difference between $\hat{\beta}_{pb}$ and $\hat{\beta}_{or}$ is that two more factors of $I - S_{12}$ appears in $\hat{\beta}_{pb}$. This term avoids an undersmoothing restriction for additive partial linear models. Since the bias of $\hat{\beta}_{pb}$ is determined by $X^T (I - S_{12})^{*2} (F_1 + F_2)$, each element of which is of order $h_1^4 + h_2^4 + o(h_1^4) + o(h_2^4)$, the usual 'optimal' bandwidth $n^{-1/5}$ implies that $\hat{\beta}_{pb}$ is $\sqrt{n}$ -consistent.

Since X is measured with error, $\hat{\beta}_{pb}$ is not available, while a direct substitution of X by W results in an inconsistent estimator. The correction-for-attenuation approach proposed by Liang et al. (1999) remains applicable with a straightforward modification. We first assume that Σ_{uu} is known, and will consider the unknown case later. In an analogous way to Liang et al. (1999), the parameter of interest, β , in the setting of additive partial linear measurement error models, is estimated by

$$\hat{\beta}_{n,ca} = \left\{ W^T (I - S_{12})^{*2} W - n \Sigma_{uu} \right\}^{-1} W^T (I - S_{12})^{*2} Y. \quad (8)$$

This corrects bias because, as we will verify, the terms $n^{-1} W^T (I - S_{12})^{*2} W$ and $n^{-1} W^T (I - S_{12})^{*2} Y$ converge to $X/Z + \Sigma_{uu}$ and X/Z , respectively. Thus $\hat{\beta}_{n,ca}$ is consistent. We present the conditions that are sufficient for this consistency as well as the asymptotic normality of $\hat{\beta}_{n,ca}$.

Assumption 1

- (a) The matrix X/Z is a positive-definite, $E(X/Z) = 0$ and $E(|X/Z|^3) < \infty$.
- (b) The bandwidths h_d are of order $n^{-1/5}$, for $d = 1, 2$.
- (c) The function $K(\cdot)$ is a bounded symmetric density function with compact support and satisfies $\int K(u) du = 1$, $\int K(u) u du = 0$ and $\int u^2 K(u) du = 1$.

(d) The density functions of Z_1 and Z_2 are bounded away from zero and have bounded continuous second partial derivatives.

(e) We assume that $E(U^3) < \infty$.

Remark 1

Assumption 1 (b), in conjunction with Theorem 1 below, indicates that the band-width convergence rate that is optimal for the nonparametric components is also appropriate for estimating β . This finding about convergence rates is attractive in theory but, of course, does not give explicit guidance for bandwidth selection in practice. However, h is commonly the parameter of interest, whereas the nonparametric functions are regarded as nuisance parameters, and the exact value of the bandwidth is much less important when estimating compared to when estimating f_1 and f_2 . In fact, Theorem 1 shows that, if $h_1 = h_2 = Cn^{-1/5}$ for $C > 0$, then the asymptotic distribution of $\hat{\beta}$ does not depend on C . Also, our limited finite-sample experience indicates that $\hat{\beta}$ is insensitive to small fluctuations in bandwidth. In our simulation study, we let $C = 1$ to reduce computation time, while, in the semen study, we use Ruppert et al.'s (1995) bandwidth selector, which is known to be of order $n^{-1/5}$.

THEOREM 1

Suppose that Assumption 1 holds. Then $\hat{\beta}_{n,ca}$ is asymptotically normal:

$$n^{1/2} (\hat{\beta}_{n,ca} - \beta) \rightarrow N(0, \Gamma_{X|Z}^{-1} \Sigma_{ca} \Gamma_{X|Z}^{-1}),$$

in distribution, where

$$\Sigma_{ca} = E \left[\left\{ (\varepsilon - U^T \beta) \tilde{X} \right\}^{\otimes 2} \right] + E \left\{ (UU^T - \Sigma_{uu}) \beta \right\}^{\otimes 2} + E(UU^T \varepsilon^2).$$

If ε is homoscedastic and independent of (X, Z_1, Z_2) with variance σ^2 , Σ_{ca} can be simplified to $E(\varepsilon - U^T \beta)^2 \text{cov}(\tilde{X}) + E \left\{ (UU^T - \Sigma_{uu}) \beta \right\}^{\otimes 2} + \Sigma_{uu} \sigma^2$.

The asymptotic variance matrix of $\hat{\beta}_{n,ca}$, which we denote by $\hat{\Sigma}_{\beta,ca}$, can be estimated by a sandwich method as follows. Let $\widehat{W}_i$ be the i th row of $(I - S_{12})W$, and let $\widehat{\varepsilon}_i$ be the i th element of $(I - S_{12})(Y - E\hat{\beta}_{n,ca})$. Write

$$\begin{aligned} \widehat{\Gamma}_{X|Z} &= W^T(I - S_{12})^* W / n - \Sigma_{uu}, \\ \widehat{\Sigma}_{ca} &= \frac{1}{n} \sum_{i=1}^n \left(\widehat{W}_i \widehat{\varepsilon}_i + \Sigma_{uu} \hat{\beta}_{n,ca} \right)^{\otimes 2} \end{aligned}$$

and $\widehat{\Sigma}_{\beta,ca} = \widehat{\Gamma}_{X|Z}^{-1} \widehat{\Sigma}_{ca} \widehat{\Gamma}_{X|Z}^{-1}$. Then it is easily shown that $\widehat{\Sigma}_{\beta,ca}$ is a consistent estimator of $\Sigma_{\beta,ca}$.

Remark 2

For the general additive partial linear models given in (1) with a D -dimensional error-prone X , we can construct a profile-based estimator in a similar way to (8) except that we need more complex notation for S_{12} . The corresponding estimator is also asymptotically normal under suitable assumptions with covariance matrix given by the same expression as for $D = 2$ except that $\tilde{X}_1 = X_1 - \sum_{d=1}^D E(X_1 | Z_{1d})$.

In the previous discussion, we assumed that the measurement error variance matrix Σ_{uu} is known, which is unlikely in practice. Prior knowledge of this matrix may not exist. Instead, Σ_{uu} is sometimes estimated using an external dataset, or internal replicates are used, so that we observe $W_{ij} = X_i + U_{ij}$ for $i = 1, \dots, n$ and $j = 1, \dots, m_i - 1$, where the measurement errors, U_{ij} , are independent, zero-mean, normally distributed random variables with variance matrix Σ_{uu} . We assume that $m_j \geq 2$ for notational simplicity. An estimator for Σ_{uu} is given in Carroll et al. (2006) or Liang et al. (1999); that is, $\hat{\Sigma}_{uu}$ for either external or internal data, is estimated at ordinary parametric rates $op(n^{-1/2})$, and thus,

$$\hat{\Sigma}_{uu} = \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^{m_i-1} (W_{ij} - \bar{W}_i) (W_{ij} - \bar{W}_i)^T,$$

where $\bar{W}_i$ is the mean of the replicate measurements W_{ij} . This estimator is consistent and unbiased with a root- n rate of convergence. The corresponding estimator of $\beta_{n,ca}^*$ is

$$\hat{\beta}_{n,ca}^* = \left\{ \bar{W}^T (I - S_{12})^{*2} \bar{W} - \frac{n}{2} \hat{\Sigma}_{uu} \right\}^{-1} \bar{W}^T (I - S_{12})^{*2} Y. \quad (9)$$

where $\bar{W} = \begin{pmatrix} \bar{W}_1^T, \dots, \bar{W}_n^T \end{pmatrix}^T$.

THEOREM 2

Under the conditions of Theorem 1, the estimator $\hat{\beta}_{n,ca}^*$ given in (9) is consistent and

asymptotically normal with covariance matrix $\Gamma_{X|Z}^{-1} \Sigma_{ca}^* \Gamma_{X|Z}^{-1}$, where

$$\Sigma_{ca}^* = E \left[\left\{ \left(\varepsilon - \bar{U}^T \beta \right) \tilde{X} \right\}^{\otimes 2} \right] + E \left[\left\{ \left(\bar{U} \bar{U}^T - \Sigma_{uu}/2 \right) \beta \right\}^{\otimes 2} \right] + E \left(\bar{U} \bar{U}^T \varepsilon^2 \right), \text{ and } \bar{U} \text{ refers to the mean of two } U\text{'s.}$$

By a straightforward but burdensome derivation, Theorem 2 can be proved in a manner similar to Theorem 1. We omit the details.

The standard error estimators can also be derived. A consistent estimator of $\Gamma_{X|Z}$ in this case is given by $\bar{W}^T (I - S_{12})^{*2} \bar{W} / n - \hat{\Sigma}_{uu}/2$, and Σ_{ca}^* is estimated as follows. Let

$$R_i^* = \widehat{W}_i^* \widehat{\varepsilon}_i^* + \frac{1}{2} \left\{ (W_{i1} - W_{i2})^{\otimes 2} - \hat{\Sigma}_{uu} \right\} \hat{\beta}_{n,ca}^*,$$

where $\widehat{W}_i^*$ and $\widehat{\varepsilon}_i^*$ are similar to $\widehat{W}_i$ and $\widehat{\varepsilon}_i$ except that $\bar{W}$ and $\hat{\beta}_{n,ca}^*$ are replaced by $\widehat{W}$ and $\hat{\beta}_{n,ca}^*$. Then a consistent estimator of Σ_{ca}^* is the sample covariance matrix of the R_i^* 's; see Liang et al. (1999) for a similar but detailed discussion.

3. SIMEX ESTIMATION

In this section, we combine the SIMEX technique and the backfitting principle to handle the additive partial linear measurement errors models. The SIMEX method, introduced by Cook & Stefanski (1994) and Stefanski & Cook (1996), is a functional one in that it can be applied

without any assumptions about the distribution of the unobservable X . It can be applied to general estimation methods such as least squares and nonparametric estimation. Carroll et al. (1996) developed its large sample properties and modified the theory.

The SIMEX algorithm consists of simulation and extrapolation steps. In the simulation step, known or estimated measurement error with a mean 0 and a covariance matrix Σ_{uu} is added to the independent variable. Here λ takes on values such as (0.0, 0.1, ..., 1.0) and controls how much error is added to the independent variable. At each level of λ , the parameter of interest is estimated. This process is usually repeated a number of times, B , at each λ value, and the average of those B estimates is calculated. With these estimates, one can obtain the SIMEX estimate by extrapolating the averages back to $\lambda = -1$, corresponding to no measurement error. Implementation for additive partial linear measurement errors models is as follows.

Let $W_{\lambda,b,i} = W_i + \lambda^{1/2} \Sigma_{uu}^{1/2} U_{b,i}$, for $b = 1, \dots, B$ and $i = 1, \dots, n$, where the $U_{b,i}$ are mutually independent, independent of the observed data, and distributed as $N(0, I)$. Here $\Sigma_{uu}^{1/2}$ denotes a square root of the matrix Σ_{uu} . Let $U_{\lambda,b,i} = U_i + \lambda^{1/2} \Sigma_{uu}^{1/2} U_{b,i}$.

With the simulated data for a given λ , we estimate f_1 and f_2 as functions of X by using the backfitting algorithm, and obtain $\hat{f}_{1,\lambda}(\beta, z_1)$ and $\hat{f}_{2,\lambda}(\beta, z_2)$. We then estimate the parametric components. The procedure is similar to the arguments for (6) and (7), and can be described by the following algorithm.

Step 1

Let $\hat{\beta}_{n,\lambda,b}$ be the estimate of β from the backfitting algorithm given by (7) with X replaced by $W_{\lambda,b} = (W_{\lambda,b,1}^T, \dots, W_{\lambda,b,n}^T)^T$.

Step 2

For each $\lambda = (\lambda_1, \dots, \lambda_m)$, the estimates $\hat{\beta}_{n,\lambda}$ are obtained by averaging $\hat{\beta}_{n,\lambda,b}$ over $b = 1, \dots, B$. Here λ is the grid of λ values for extrapolation and typically ranges from 0 to 2.

Step 3

Either the quadratic function $\mathcal{G}_Q(\lambda, \Gamma) = \gamma_1 + \gamma_2 \lambda + \gamma_3 \lambda^2$ or the rational linear function $\mathcal{G}_R(\lambda, \Gamma) = \gamma_1 + \gamma_2 / (\gamma_3 + \lambda)$ with $\Gamma = (\gamma_1, \gamma_2, \gamma_3)^T$ is used to establish a regression relationship between $\hat{\beta}$ and λ by fitting the estimated values $\{\hat{\beta}_{n,\lambda}, \lambda \in \Lambda\}$ on $\{\lambda\}$. We therefore obtain an estimate $\hat{\Gamma}$ of Γ . We will denote the extrapolant function by $\mathcal{G}$, regardless of whether the quadratic or rational linear extrapolant function is used. The SIMEX estimator of β is defined as $\hat{\beta}_{\text{sim}} = \mathcal{G}(-1, \hat{\Gamma})$.

Note that the estimator $\hat{\beta}_{n,\lambda,b}$ from Step 1 can be expressed as

$$\begin{aligned} \hat{\beta}_{n,\lambda,b} &= \left\{ W_{\lambda,B}^T (I - S_{12})^{*2} W_{\lambda,B} \right\}^{-1} \left\{ W_{\lambda,B}^T (I - S_{12})^{*2} Y \right\} \\ &\doteq \left\{ W^T (I - S_{12})^{*2} W + n \lambda \Sigma_{uu} \right\}^{-1} \left\{ W^T (I - S_{12})^{*2} Y \right\}. \end{aligned} \quad (10)$$

The approximation originates from the definition of $W_{\lambda,B}$ and an argument similar to the proof of (A.4). Here $\hat{\beta}_{n,\lambda,b}$ is a function of λ of the form $\mathcal{G}_R(\lambda, \Gamma)$ for suitably defined Γ .

$(\beta_1, \beta_2, \beta_3)^T$. We can show, using the standard SIMEX derivation (Carroll et al., 2006), that, under the conditions of Theorem 1, if the extrapolant function is theoretically exact, then the estimator $\hat{\beta}_{\text{sim}}$ is asymptotically normal. The asymptotic covariance matrix of the estimator $\hat{\beta}_{\text{sim}}$ can be estimated by using a sandwich version. We omit the details, which can be found in an earlier version of this article, available from the first author.

Note that $\hat{\beta}_{n,\lambda,b}$ is similar to the method-of-moments estimator. Recalling the approximation in (10), we may find that $\hat{\beta}_{n,\lambda,b}$ is a function of $\mathcal{G}_R$ approximately when W is univariate. This indicates that the rational linear extrapolant is asymptotically exact for estimation of β . Our numerical experience indicates that the quadratic extrapolation is more easily implemented and has smaller variance although this extrapolation is slightly biased.

4. SIMULATION STUDY

We generated $n = 200$ and $n = 400$ observations from models (1) and (2) with $D = 2$, $f_1(z_1) = \exp(2z_1) - 3.75$ and $f_2(z_2) = 0.2z_2^{11}\{10(1 - z_2)\}^6 + 10^4 z_2^3(1 - z_2)^{10} - 1.4$. We used the measurement error model $W = X + U$, where $U \sim N(0, 0.2^2)$. In our simulations, $\lambda = 4$, and

X, Z_1 and Z_2 are independently generated from $\text{Un}(0, 1)$. Let $\varepsilon \sim N\{0, \sigma_\varepsilon^2(x, z)\}$. We considered two cases: (1) $(x, z) = (0.1, 0.25, 0.5, 1)$; and (2)

$\sigma_\varepsilon^2(x, z) = \{\sigma_0 \sin(2\pi x^3) + 0.5z_1 + 0.5z_2 + 0.3\}^2$, using the same 4 values for σ_0^2 . The latter case is used for examining the effect of heteroscedastic errors on the estimators.

In our nonparametric estimation procedure, we used the kernel function $K(u) = 0.75(1 - u^2)I^{|u| \leq 1}$ and bandwidths $h_1 = h_2 = n^{-1/5}$. We generated 200 datasets in each of the cases. To estimate the variance of U , we generated duplicate samples of W . We computed the naive, attenuated-correction and SIMEX estimates of the parametric components. The estimated values and associated standard errors are presented in Table 1, in which we also present the coverage probability of the estimators based on the two proposed methods, but omit the results for the naive estimator because its coverage probabilities are very poor and close to zero. Rational linear and quadratic extrapolating functions are used for the SIMEX estimator. Not surprisingly, the naive estimated values are substantially biased, while the estimated values based on both the attenuated-correction and SIMEX methods with rational linear extrapolant are unbiased. However, the estimated values based on the SIMEX method with quadratic extrapolant are slightly biased. These results are in accord with the theoretical assertions. On the other hand, the standard errors of the corrected estimators are larger than those of the naive estimator, and the SIMEX method with the rational linear extrapolant has the largest standard error. The coverage probabilities of the proposed methods are quite close to the nominal level. Overall, the proposed methods outperform the naive method in terms of bias-reduction and coverage probability, and show promise for use in real applications.

5. SEMEN STUDY

Our analysis is based on a larger and more recent version of the semen study database than that described in Duty et al. (2003) and Hauser et al. (2006), with complete data on 455 subjects. Our main objective is to quantify the bias-corrected relationship between sperm concentration and monobutyl phthalate, MBP, the metabolite of di-n-butyl phthalate. Since urinary metabolite concentrations, in nanograms/millilitre, are calculated by dividing the measured mass of the metabolite by the urine volume, it is necessary to account for urine dilution. To adjust for urinary dilution we used specific gravity, consistent with Duty et al.

(2003) and Hauser et al. (2006). Both sperm concentration, in millions of sperm per ml., and MBP, in ng monoester per ml. of urine, were log-transformed to satisfy regression assumptions better.

Our models adjusted for four covariates, namely age, abstinence time, race, coded as 1 for whites and 0 otherwise, and body mass index, BMI. The first three covariates were important predictors in models reported by Duty et al. (2003) and Hauser et al. (2006), and BMI has also been associated with semen parameters (Kort et al., 2006). Although Duty et al. (2003) and Hauser et al. (2006) dichotomized sperm concentration and categorized model covariates, we did not categorize the continuous variables. A preliminary study indicated that log (sperm concentration) was linearly related to log(MBP) and abstinence time but nonlinearly related to age and, perhaps, to BMI; see Fig. 1. As seen in Fig. 1, the effect of BMI is estimated with considerable error, and it is impossible to say whether or not the effect is reasonably close to being linear. We err on the side of caution and take the effect to be nonlinear. As noted earlier, log(MBP) is subject to measurement error. In the present study we have 78 replicate observations of MBP, with mean on the logarithmic scale of 3.14 and 3.52 for the first and second samples respectively. Their sample correlation on the logarithmic scale is 0.101. This implies that the measurement error accounts for most of the variability of the observed exposure. Thus, the need to correct for measurement error is even stronger than in most studies.

A possible concern with using the log transformation is that, if the observed MBP is unbiased for the true MBP, then the observed log(MBP) will be biased for the true log(MBP). Fortunately, a constant bias in a covariate will cause bias in the estimated intercept, a nuisance parameter in this study, but not in the estimated slopes. The bias in log(MBP) will be constant, if the measurement errors are additive and homoscedastic on the log scale. A plot of $|W_{j1} - W_{j2}|$ against $(W_{j1} + W_{j2})/2$, which is not included here to save space, shows no association. This suggests additive errors on the log scale so that the slope estimators are unbiased; see Carroll et al. (2006, §1.7) for methods for checking assumptions about the measurement error model.

We now apply the model and estimation method described in §§2 and 3 to explore this dataset and address the concerns mentioned in §1. We calculate the estimated values using the naive, attenuated-correction and SIMEX methods. The standard errors of the SIMEX-based estimators are calculated by the method presented at the end of the Appendix. The results are presented in Table 2. It turned out that MBP was negatively associated with sperm concentration under all three methods, consistent with results from Duty et al. (2003) and Hauser et al. (2006). Comparing the magnitudes of these estimates under the three scenarios, we see that the naive method and the two measurement error methods give remarkably different estimates. The estimated MBP effect is attenuated toward zero when measurement error is not considered. The bias-corrected MBP effect is estimated with a larger standard error than when measurement error is not considered. These findings are similar to those in standard linear or nonlinear regression models with measurement error (Carroll et al., 2006). The estimated MBP effect based on the SIMEX method with quadratic extrapolant reflects a negative relationship between sperm concentration and MBP that is of similar magnitude to that based on the attenuated-correction method, while the estimated relationship based on the SIMEX method with rational linear extrapolant is more strongly negative. The latter estimation procedure also has the largest standard error among the three adjusted estimators. For an illustration of the SIMEX quadratic extrapolant, in Fig. 2 we depict the parameter estimates reflecting the MBP effect for different values and for the extrapolant value at -1 .

The estimated effects of the linear covariates, abstinence time and race, are very similar for all three methods. As expected (Carlsen et al., 2004), a greater abstinence time is associated with a greater sperm concentration. Although not significant, white men on average had a higher sperm concentration than non-white men, consistent with other studies (Swan et al., 2003).

Acknowledgments

The authors thank a referee, the associate editor and the editor for their insightful comments and suggestions which greatly improved this paper. Liang's research was partially supported by two grants from the U.S. National Institute of Allergy and Infectious Diseases. Ruppert's research was supported by grants from the U.S. National Science Foundation and National Institutes of Health. Hauser's research was supported by two grants from the U.S. National Institute of Environmental Health Sciences.

APPENDIX: Technical details

Proof of Theorem 1

The proof follows arguments similar to those used by Liang et al. (1999) to prove their Theorem 3.1, and Opsomer & Ruppert (1997) to prove their Theorem 4.1. Only the key steps are outlined below.

$$\begin{aligned} \hat{\beta}_{n,ca} - \beta &= \left\{ W^T(I - S_{12})^{*2}W - n\Sigma_{uu} \right\}^{-1} W^T(I - S_{12})^{*2} (F_1 + F_2) \\ &\quad + \left\{ W^T(I - S_{12})^{*2}W - n\Sigma_{uu} \right\}^{-1} \left[W^T(I - S_{12})^{*2}\varepsilon + \left\{ n\Sigma_{uu}\beta - W^T(I - S_{12})^{*2}U\beta \right\} \right]. \end{aligned}$$

Referring to the proof of Theorem 4.1 in Opsomer & Ruppert (1997), we obtain that

$$\begin{aligned} (I - S_1^c S_2^c)^{-1} &= I + o_p(1/n), \\ (I - S_d^c)X &= \{X - E(X|Z_d)\} \{1 + o_p(1)\}, \end{aligned}$$

for $d = 1, 2$. It follows that

$$\begin{aligned} \frac{1}{n} X^T(I - S_{12})^{*2}X &= \frac{1}{n} X^T \left\{ (I - S_1^c S_2^c)^{-1} (I - S_1^c) - I + (I - S_2^c S_1^c)^{-1} (I - S_2^c) \right\}^{*2} X \quad (\text{A.1}) \\ &= E\{X - E(X|Z_1) - E(X|Z_2)\}^{\otimes 2} \{1 + o_p(1/n)\}. \end{aligned}$$

Similarly

$$\begin{aligned} \frac{1}{n} X^T(I - S_{12})^{*2}U &= o_p(1), \\ \frac{1}{n} U^T(I - S_{12})^{*2}X &= o_p(1), \\ \frac{1}{n} U^T(I - S_{12})^{*2}U &= \Sigma_{uu} \{1 + o_p(1/n)\}. \end{aligned}$$

We therefore have that, in probability,

$$\frac{1}{n} \left\{ W^T(I - S_{12})^{*2}W - n\Sigma_{uu} \right\} \rightarrow \Gamma_{X|Z}. \quad (\text{A.2})$$

In addition, direct calculations yield that

$$\begin{aligned} \frac{1}{n} W^T (I - S_{12})^{*2} (F_1 + F_2) &= \frac{1}{n} \sum_{i=1}^n \left[W_i - \sum_{d=1}^2 \{ E(X_i | Z_{id}) + o(h_d^2) \} \right] \\ &\quad \times \left\{ h_1^2 f_1^{(2)}(z_1) + h_2^2 f_2^{(2)}(z_2) + o(h_1^2) + o(h_2^2) \right\} \\ &= o_p \left(n^{-1/2} h_1^2 + n^{-1} h_2^2 \right) + o_p \left(h_1^4 + h_2^4 \right), \end{aligned}$$

$$\frac{1}{n} W^T (I - S_{12})^{*2} \varepsilon = \frac{1}{n} \sum_{i=1}^n \left(\tilde{X}_i + U_i \right) \varepsilon_i + O_p \left(n^{-1/2} h_1^2 + n^{-1/2} h_2^2 \right), \quad (\text{A.3})$$

$$\frac{1}{n} W^T (I - S_{12})^{*2} U \beta = \frac{1}{n} \sum_{i=1}^n \left(\tilde{X}_i + U_i \right) U_i^T \beta + O_p \left(n^{-1/2} h_1^2 + n^{-1/2} h_2^2 \right). \quad (\text{A.4})$$

Note that $n^{1/2} (h_1^4 + h_2^4) \rightarrow 0$ and $h_1, h_2 \rightarrow 0$ as $n \rightarrow \infty$. A combination of (A.3) and (A.4) yields that

$$\begin{aligned} \frac{1}{\sqrt{n}} \left[W^T (I - S_{12})^{*2} \varepsilon + \left\{ n \Sigma_{uu} \beta - W^T (I - S_{12})^{*2} U \beta \right\} \right] \\ = \frac{1}{\sqrt{n}} \sum_{i=1}^n \left\{ \tilde{X}_i \left(\varepsilon_i - U_i^T \beta \right) + U_i \varepsilon_i - \left(U_i U_i^T - \Sigma_{uu} \right) \beta \right\} + o_p(1). \end{aligned}$$

The mean and the variance matrix of $\tilde{X}_i \left(\varepsilon_i - U_i^T \beta \right) + U_i \varepsilon_i - \left(U_i U_i^T - \Sigma_{uu} \right) \beta$ are equal to zero and Σ_{ε} , respectively. The proof of the theorem follows from a central limit theorem.

REFERENCES

- Carlsen E, Petersen JH, Andersson A-M, Skakkebaek NE. Effects of ejaculatory frequency and season on variations in semen quality. *Fertil. Steril.* 2004; 82:358–66. [PubMed: 15302284]
- Carroll RJ, Küchenhoff H, Lombard F, Stefanski LA. Asymptotics for the SIMEX estimator in structural measurement error models. *J. Am. Statist. Assoc.* 1996; 91:242–50.
- Carroll R, Maca JD, Ruppert D. Nonparametric estimation in the presence of measurement error. *Biometrika.* 1999; 86:541–54.
- Carroll, RJ.; Ruppert, D.; Stefanski, LA.; Crainiceanu, CM. *Measurement Error in Nonlinear Models.* 2nd ed. Chapman and Hall; New York: 2006.
- Cook JR, Stefanski LA. Simulation-extrapolation estimation in parametric measurement error models. *J. Am. Statist. Assoc.* 1994; 89:1314–28.
- Duty SM, Silva MJ, Barr DB, Brock JW, Ryan L, Chen ZY, Herrick R, Christiani DC, Hauser R. Phthalate exposure and human semen parameters. *Epidemiology.* 2003; 14:269–77. [PubMed: 12859026]
- Engle RF, Granger CWJ, Rice J, Weiss A. Semiparametric estimates of the relation between weather and electricity sales. *J. Am. Statist. Assoc.* 1986; 81:310–20.
- Hauser R, Meeker JD, Park S, Silva MJ, Calafat AM. Temporal variability of urinary phthalate metabolite levels in men of reproductive age. *Envir. Health Perspect.* 2004; 112:1734–9.
- Hauser R, Meeker JD, Duty S, Silva MJ, Calafat AM. Altered semen quality in relation to urinary concentrations of phthalate monoester and oxidative metabolites. *Epidemiology.* 2006; 17:682–91. [PubMed: 17003688]
- Härdle, W.; Liang, H.; Gao, J. *Partially Linear Models.* Springer Physica-Verlag; Heidelberg: 2000.
- Hastie, T.; Tibshirani, R. *Generalized Additive Models.* Chapman and Hall; London: 1990.

- Kort HI, Massey JB, Elsner CW, Mitchell-Leef D, Shapiro DB, Witt MA, Roudebush WE. Impact of body mass index values on sperm quantity and quality. *J. Androl.* 2006; 27:450–2. [PubMed: 16339454]
- Liang H, Härdle W, Carroll RJ. Estimation in a semiparametric partially linear errors-in-variables model. *Ann. Statist.* 1999; 27:1519–35.
- Opsomer JD, Ruppert D. Fitting a bivariate additive model by local polynomial regression. *Ann. Statist.* 1997; 25:186–211.
- Opsomer JD, Ruppert D. A root- n consistent backfitting estimator for semiparametric additive modeling. *J. Comp. Graph. Statist.* 1999; 8:715–32.
- Robinson PM. Root- n -consistent semiparametric regression. *Econometrica.* 1988; 56:931–54.
- Ruppert D, Sheather SJ, Wand MP. An effective bandwidth selector for local least squares regression. *J. Am. Statist. Assoc.* 1995; 90:1257–70.
- Severini TA, Staniswalis JG. Quasilikelihood estimation in semiparametric models. *J. Am. Statist. Assoc.* 1994; 89:501–11.
- Speckman P. Kernel smoothing in partial linear models. *J. R. Statist. Soc. B.* 1988; 50:413–36.
- Stefanski LA, Cook JR. Simulation-extrapolation: the measurement error jackknife. *J. Am. Statist. Assoc.* 1995; 90:1247–56.
- Stone CJ. Additive regression and other nonparametric models. *Ann. Statist.* 1985; 13:689–705.
- Swan SH, Brazil C, Drobnis EZ, Liu F, Kruse R, Hatch M, Redmon JB, Wang C, Overstreet JW, Study For Future Families Research Group. Geographic differences in semen quality of fertile U.S. males. *Envir. Health Perspect.* 2003; 111:414–20.
- Wang NS, Carroll R, Lin XH. Efficient semiparametric marginal estimation for longitudinal/clustering data. *J. Am. Statist. Assoc.* 2005; 100:147–57.
- Zeger SL, Diggle PJ. Semiparametric models for longitudinal data with application to CD4 cell numbers in HIV seroconverters. *Biometrics.* 1994; 50:689–99. [PubMed: 7981395]

Figure 1. Semen study. The patterns of (a) age and (b) BMI with dotted curves representing estimates $\pm$ one standard error, obtained by using the R gam function.

Figure 2.

Semen study. Coefficient extrapolant performances. The simulated estimates $\{\hat{\beta}_\lambda, \lambda\}$ are plotted (filled circles), and the fitted quadratic extrapolant (solid line) is extrapolated to $\lambda = -1$ (dashed line). This extrapolation results in the SIMEX estimate (open circle).

Table 1

Simulation study. Point estimates for β_1 , true value 4, and the standard errors based on the naive (naive), correction-for-attenuation (AC), and SIMEX (SIMEX.qua: quadratic extrapolation; SIMEX.ra: rational linear extrapolation) methods and the coverage probabilities (COV) of the adjusted estimators for the simulated data.

n	2 0	Naive			AC			SIMEX.qua			SIMEX.ra		
		Est	s.e.	Est	s.e.	COV	Est	s.e.	COV	Est	s.e.	COV	
Constant error													
200	0.1	2.72	0.12	4.04	0.13	94.8	3.90	0.22	93.5	3.99	0.31	96.3	
	0.25	2.71	0.12	4.00	0.15	95.9	3.81	0.24	96.5	3.99	0.32	97.5	
	0.5	2.71	0.12	4.00	0.17	95.8	3.90	0.29	97.5	3.99	0.36	97.1	
400	1	2.72	0.18	4.02	0.21	93.8	3.91	0.39	93.4	4.05	0.47	95.1	
	0.1	2.71	0.11	4.01	0.12	97.0	3.90	0.19	96.5	4.03	0.25	97.2	
	0.25	2.69	0.11	4.00	0.13	96.9	3.86	0.18	98.1	3.98	0.23	97.3	
	0.5	2.73	0.12	4.03	0.13	96.0	3.93	0.20	95.8	4.06	0.26	96.1	
	1	2.72	0.15	4.00	0.14	96.0	3.93	0.26	95.4	4.02	0.31	96.8	
	Heteroscedastic error												
200	0.1	2.71	0.17	4.03	0.16	96.0	3.85	0.31	95.6	4.02	0.38	93.7	
	0.25	2.72	0.18	4.04	0.20	96.0	3.86	0.34	94.0	4.05	0.42	95.9	
	0.5	2.73	0.19	4.02	0.21	93.0	3.90	0.33	93.9	4.05	0.40	93.4	
400	1	2.72	0.20	4.01	0.23	98.0	3.88	0.36	98.8	4.07	0.42	96.9	
	0.1	2.70	0.12	4.00	0.19	96.9	3.89	0.22	97.8	4.00	0.25	96.6	
	0.25	2.71	0.13	4.01	0.14	96.1	3.86	0.25	93.5	4.02	0.28	93.1	
	0.5	2.71	0.14	4.02	0.15	93.0	3.92	0.26	92.8	4.00	0.30	94.4	
	1	2.72	0.15	4.02	0.16	93.9	3.90	0.27	93.6	4.06	0.32	96.3	

Est, estimate; s.e., standard error

Table 2

Semen study. Parameter estimates based on three methods, with estimated standard errors in parentheses

	Naive	AC	SIMEX.qua	SIMEX.ra
Abstinence time	0.077(0.034)	0.076 (0.032)	0.076 (0.033)	0.075 (0.033)
Race (white)	0.278(0.194)	0.255 (0.184)	0.255 (0.195)	0.210 (0.208)
Log(MBP)	−0.09(0.066)	−0.132 (0.101)	−0.132(0.104)	−0.162 (0.124)

AC, correction-for-attenuation; SIMEX.qua, quadratic extrapolation; SIMEX.ra, rational linear extrapolation.