

Nonlinear and Nonparametric Regression and Instrumental Variables

Raymond J. CARROLL, David RUPPERT, Ciprian M. CRAINICEANU,
Tor D. TOSTESON, and Margaret R. KARAGAS

We consider regression when the predictor is measured with error and an instrumental variable (IV) is available. The regression function can be modeled linearly, nonlinearly, or nonparametrically. Our major new result shows that the regression function and all parameters in the measurement error model are identified under relatively weak conditions, much weaker than previously known to imply identifiability. In addition, we exploit a characterization of the IV estimator as a classical “correction for attenuation” method based on a particular estimate of the variance of the measurement error. This estimate of the measurement error variance allows us to construct functional nonparametric regression estimators making no assumptions about the distribution of the unobserved predictor and structural estimators that use parametric assumptions about this distribution. The functional estimators uses simulation extrapolation or deconvolution kernels and the structural method uses Bayesian Markov chain Monte Carlo. The Bayesian estimator is found to significantly outperform the functional approach.

KEY WORDS: Bayesian methods; Identifiability; Measurement error; Simulation extrapolation; Splines; Structural modeling.

1. INTRODUCTION

1.1 Background and Problem Statement

Motivated by two problems in epidemiology (Sec. 5), we consider measurement errors in regression where the regression function could be modeled linearly, nonlinearly, or even nonparametrically. In the case where the measurement error variance is known or can be estimated by replication, functional (Carroll, Maca, and Ruppert 1999) and structural/Bayesian (Berry, Carroll, and Ruppert 2002) methods have been developed.

We use the notation of Carroll, Ruppert, and Stefanski (1995) so that Y is the response, X is an error-prone covariate whose measured value is W , Z is a vector of error-free covariates written as Z when univariate, and S is an instrument related to X .

In our first example, an epidemiologic study of skin cancer and arsenic exposure (Karagas et al. 2001), the error-prone predictor W is not replicated, so information on the measurement error is provided by a second measure of exposure S , which we use as an instrumental variable (IV).

Our second example uses data from the recent OPEN (observing protein and energy intake) study (Subar, Kipnis, and Triano 2003; Kipnis et al. 2003a), the first large study using biomarkers to investigate the properties of self-reporting instruments that measure nutrient intake. In this example, the response Y is the protein intake measured by a food frequency

questionnaire (FFQ); the error-prone covariate W is UN (urinary nitrogen), a biomarker for protein and an unbiased estimator of true protein intake X ; the instrument S is DLW (doubly labelled water), a biomarker for energy intake; and age is an additional covariate Z that is measured without error. It is suspected that the biases of FFQs are related to age.

In these examples, there is little prior knowledge about the form of the regression of the response and the instrument on the covariates. Therefore, we wish to model these functions using as few assumptions as possible. Our novel result is that all parameters are identified when Y follows a nonparametric model in (X, Z) and S follows a varying-coefficient model that is linear in X with coefficients smooth functions of Z . This level of generality is needed in, for example, the OPEN study where the slope of DLW on true protein intake depends on age, suggesting a varying-coefficient model. Not only do we identify the measurement error variance identified, but we also exhibit a root- n -consistent estimator. Because we can estimate the measurement error variance, we can apply methods from the measurement error literature to estimate the regression of Y on X .

Our identifiability result is related to a simple characterization of the IV estimator—specifically, that in simple linear regression with a scalar instrument, the usual IV estimator is a version of the classical “correction for attenuation” method based on a specific estimate of the measurement error variance.

There is some work on parametric but not necessarily linear regression with an IV (Fuller 1987; Hausman, Newey, Ichimura, and Powell 1991; Amemiya 1990; Carroll and Stefanski 1994; Stefanski and Buzas 1995; Buzas and Stefanski 1996; Cheng and Schneeweiss 1998; Wall and Amemiya 2000; Yalcin and Amemiya 2001). These methods are applicable either only for special parametric models or for general parametric models that rely on small-error approximations known to fail for some nonlinear and nonparametric models (Carroll et al. 1995). To the best of our knowledge, there are no techniques presently available for nonparametrically specified regression functions in the IV context.

Raymond J. Carroll is Distinguished Professor, Department of Statistics and Faculties of Nutrition and Toxicology, Texas A&M University, College Station, TX 77843-3143 (E-mail: carroll@stat.tamu.edu). David Ruppert is Andrew Schultz, Jr. Professor of Engineering, School of Operations Research & Industrial Engineering, Cornell University, Ithaca, NY 14853-3801 (E-mail: dr24@cornell.edu). Ciprian M. Crainiceanu is Assistant Professor, Department of Biostatistics, Johns Hopkins University, Baltimore, MD 21205 (E-mail: ccrainic@jhsph.edu). Tor D. Tosteson is Associate Professor, Community and Family Medicine, Dartmouth Medical School, Hanover, NH 03755 (E-mail: Tor.Devin.Tosteson@Dartmouth.edu). Margaret R. Karagas is Associate Professor, Community and Family Medicine, Dartmouth Medical School, Hanover, NH 03755 (E-mail: Margaret.R.Karagas@Dartmouth.edu). Carroll's research was supported by National Cancer Institute grant CA57030 and by the Texas A&M Center for Environmental and Rural Health via National Institute of Environmental Health Sciences grant P30-ES09106. Ruppert, Tosteson, and Karagas were supported by National Cancer Institute grant CA50597, and Tosteson and Karagas were supported by the National Cancer Institute grants CA50597 and CA57494 and National Institute of Environmental Health Sciences grant ES07373. The authors thank Tailen Hsing for showing us Durrent's (1996) counterexample used in the Appendix.

© 2004 American Statistical Association
Journal of the American Statistical Association
September 2004, Vol. 99, No. 467, Theory and Methods
DOI 10.1198/01621450400001088

1.2 Outline

In Section 2 we define our model and give characterizations of identifiability under increasingly weaker conditions. In Section 3 we outline the methods used, some making no assumptions about the distribution of the latent variable (functional case) and others assuming a specific form for this distribution (structural case). We present a small simulation study in Section 4. In Section 5 we illustrate the methods on the two examples. In response to those with concerns about the small-sample properties of the estimators, in Section 6 we describe some asymptotic calculations in the polynomial regression case, which illustrate how difficult the estimation problem of IV for nonlinear models can be. We present some concluding remarks in Section 7.

2. MODEL AND IDENTIFIABILITY

2.1 Introduction

Hausman et al. (1991) considered the most basic nonlinear model, polynomial regression. A polynomial is linear in the parameters but is nonlinear in the independent variable and therefore in the measurement error. Let Y be the response, let W be the unbiased measure of X , and let S be the instrument. They assume an iid sequence of vectors (Y, X, W, S) with only (Y, W, S) observed and also assume that the data satisfy the following assumption.

Assumption 1.

$$Y = m_{\text{poly}}(X, \boldsymbol{\beta}) + \epsilon, \tag{1}$$

$$W = X + U, \tag{2}$$

and

$$S = \alpha_0 + \alpha_1 X + \nu. \tag{3}$$

In (1), the function $m_{\text{poly}}(X, \boldsymbol{\beta})$ is a polynomial in X . In addition, ϵ , U , and ν have mean 0, and ν is independent of (X, ϵ, U) .

Model (2) is the classical measurement error model. Hausman et al. (1991) showed that the model (1)–(3) is identified when ϵ and U are correlated if there exists a known function $a(\cdot)$ such that $a(\alpha_0, \alpha_1) = 0$. We assume that ϵ and U are uncorrelated, which gives us a factor model, and then the existence of such a function is not necessary for identifiability. Fuller (1987, pp. 60–61) studied a linear factor model satisfying Assumption 1 with the additional assumption that $m_{\text{poly}}(X; \boldsymbol{\beta}) = \beta_0 + \beta_1 X$. He showed that all parameters are identified if $\text{var}(X) = \sigma_x^2 > 0$, $\beta_1 \neq 0$, and $\alpha_1 \neq 0$. Wall and Amemiya (2000) showed identifiability in polynomial structural models by exhibiting root- n -consistent estimators, but their methods rely on the polynomial structure. As for more general parametric nonlinear models, Yalcin and Amemiya (2001) considered the nonlinear factor model $\mathbf{Y}_i = \mathbf{g}(\mathbf{f}_i; \boldsymbol{\beta}) + \boldsymbol{\epsilon}_i$ and $\mathbf{X}_i = \mathbf{f}_i + \mathbf{u}_i$, where $\mathbf{g}(\mathbf{f}_i; \boldsymbol{\beta})$ is a nonlinear parametric model for a vector response \mathbf{Y} and \mathbf{f}_i is a vector of latent variables. The model (1)–(3) is a special case of their model. Yalcin and Amemiya mentioned some examples of their model that are identified but gave no general result. We are not aware of identifiability studies for nonparametric models. Our Theorem 1 generalizes these identifiability results to both parametric nonlinear and nonparametric regression.

2.2 Main Identifiability Results

Our proposed methods are based on the observation that the model (1)–(3) is identified without prior knowledge of α_1 even if the regression function is not a polynomial. Rather, α_1 can be determined from moments of the observable variables in (1)–(3). This means that $m(\cdot)$ can be estimated without any prior knowledge of parameters provided only that both the instrument S and the proxy W are observed.

Theorem 1. Suppose that (2) and (3) of Assumption 1 hold and that

$$(X, U, \epsilon, \nu) \text{ are mutually uncorrelated.} \tag{4}$$

Replace (1) by

$$Y = m(X) + \epsilon. \tag{5}$$

Then for any function $m(x)$ (not just polynomials), α_0 , α_1 , $\mu_x = E(X)$, $\sigma_x^2 = \text{var}(X)$, $\sigma_u^2 = \text{var}(U)$, and $\sigma_\nu^2 = \text{var}(\nu)$ are all identified if $\alpha_1 \neq 0$ and if

$$\text{cov}(Y, W) = \text{cov}\{X, m(X)\} \neq 0. \tag{6}$$

Proof. Note that $\alpha_1 = \text{cov}(Y, S) / \text{cov}(Y, W)$, $\mu_x = E(W)$, $\alpha_0 = E(S - \alpha_1 W)$, $\sigma_x^2 = \text{cov}(W, S) / \alpha_1$, $\sigma_u^2 = \text{var}(W) - \sigma_x^2$, and $\sigma_\nu^2 = \text{var}(S) - \alpha_1^2 \sigma_x^2$. Therefore, all parameters are functions of the moments of observable variables and so are identified.

One would generally expect Y and X to be related; so in linear regression, (6) should hold. However, (6) is less natural in nonlinear or nonparametric regression where (6) can fail, for example, if $m(\cdot)$ is an even function and X is symmetrically distributed about 0. If (6) were in fact necessary, then estimation might be unstable when the correlation between X and $m(X)$ was close to 0. Fortunately, if we strengthen (4) to the assumption that

$$\epsilon, U, \nu, \text{ and } X \text{ are mutually independent of one another,} \tag{7}$$

then (6) can be weakened to

$$\text{cov}[\{X - E(X)\}^k, m(X)] \text{ exists and is nonzero} \\ \text{for some positive integer } k. \tag{8}$$

Theorem 2. Assume (2), (3), (5), (7), and (8), that $\alpha_1 \neq 0$; and that $\sigma_x^2 > 0$. Then α_0 , α_1 , $\mu_x = E(X)$, $\sigma_x^2 = \text{var}(X)$, $\sigma_u^2 = \text{var}(U)$, and $\sigma_\nu^2 = \text{var}(\nu)$ are all identified; that is, they are determined by moments of observable variables.

The proof of the theorem is given in Appendix A.1. In Appendix A.5 we show that under weak assumptions, (8) will hold unless $m(\cdot)$ is constant. When $m(\cdot)$ is constant, $m(\cdot)$ is still identified, but α_1 appears to be not identified; see Section 3.5. Our theorem does not state explicitly whether m is identified, because this follows from Fan and Truong (1993), who stated conditions under which $m(\cdot)$ can be consistently estimated, and hence is identified if $\text{var}(U)$ is identified.

Some comments on the assumptions and implications of the characterization are in order. Assume now that (6) holds. Then the following conditions apply:

1. In the linear case, where $m(x) = \beta_0 + \beta_1 x$, the usual IV slope estimate is the sample version of $\beta_{1,iv} = \text{cov}(Y, S) / \text{cov}(W, S)$. If σ_u^2 is known, then the usual correction for attenuation estimate is the sample version of $\beta_{1,ca} = \text{cov}(Y, W) / (\text{var}(W) - \sigma_u^2)$. In our IV model the estimate of σ_u^2 is the sample version of $\sigma_u^2 = \text{var}(W) - \text{cov}(W, S) \text{cov}(Y, W) / \text{cov}(Y, S)$; that is, it replaces variances and covariances in this expression by sample versions. Substituting $\sigma_{u,iv}^2$ into the formula for $\beta_{1,ca}$ yields $\beta_{i,iv}$. Thus the usual IV estimator of the slope in linear regression is the correction for attenuation estimator when σ_u^2 is estimated via our proposal.
2. The connection between correction for attenuation and IV estimation offers the hope of more stable estimation. The attenuation is

$$\lambda = \frac{\sigma_x^2}{\sigma_x^2 + \sigma_u^2} = \frac{\sigma_w^2 - \sigma_u^2}{\sigma_w^2}. \quad (9)$$

The correction for an attenuation estimator is simply the least squares slope ignoring measurement error divided by an estimate of the attenuation. Because of this division, one can improve the IV estimator by bounding the attenuation away from 0 using prior knowledge when available.

3. Model (5) is more general than it looks, because the distribution of ϵ can depend on X ; for example, in our first application Y is binary, so that we can write the model as $\text{logit}\{\text{Pr}(Y = 1|X)\} = g(X)$, where $g(X) = \text{logit}\{m(X)\}$. Then ϵ is a Bernoulli variate minus its mean.
4. Because (5) is an unstructured regression model, the assumption of additivity in (2) and (3) is not as strong as it may seem. Instead, we are only assuming a common transformation of the original data to X , W , and S that satisfies these equations. If (5) holds for the original data, then it will also hold for the transformed data. For example, in our first application, we log-transform the data.
5. In practice, the methods are necessarily restricted to cases where Y and W are clearly related; otherwise α_1 will be poorly estimated. Indeed, if Y and X are independent, then the parameters in the (W, S) model are unidentifiable if (W, S) is jointly normal.

2.3 Estimation of σ_u^2

We now exhibit a root- n -consistent estimator $\hat{\sigma}_u^2$ of σ_u^2 , that is, one that satisfies $n^{1/2}(\hat{\sigma}_u^2 - \sigma_u^2) = O_p(1)$.

Assumption 2. Assume that for known K , the $2K$ th moments of (Y, W, S) are finite. In addition, for some unknown $1 \leq k \leq K$, $\rho_k = \text{cov}[m(X), \{X - E(X)\}^k] = \text{cov}[Y, \{W - E(W)\}^k] \neq 0$.

Under Assumption 2,

$$\alpha_1 = \text{sign}\{\text{cov}(W, S)\} |\text{cov}[Y, \{S - E(S)\}^k] / \rho_k|^{1/k}. \quad (10)$$

Thus the idea is to first test the hypothesis that $\rho_k = 0$ for $k = 1, \dots, K$ and for the selected value of k , form $\hat{\alpha}_1$ from the sample moments of (10). A valid (in the sense of type I error) test statistic for the hypothesis is defined as follows. Let $Y_{ic} = Y_i - \bar{Y}$, $W_{ic} = W_i - \bar{W}$, and $S_{ic} = S_i - \bar{S}$. Then a test statistic is $T_k = n^{-1/2} \sum_{i=1}^n Y_{ic} W_{ic}^k / (n^{-1} \sum_{i=1}^n Y_{ic}^2 W_{ic}^{2k})^{1/2}$. This statistic is asymptotically $N(0, 1)$ when $\rho_k = 0$. Our algorithm

is to test each hypothesis at level $c_n \rightarrow 0$, separately for $k = 1, \dots, K$; c_n tends to 0 slower than $1 - \Phi(n^{1/2-a})$ for some $a > 0$. The estimate of k is either the first k for which the hypothesis is rejected or, if none are rejected, the k corresponding to the smallest p -value is selected. This is clearly a consistent model selector for the first k such that $\rho_k \neq 0$, and hence $n^{1/2}(\hat{\alpha}_1 - \alpha_1) = O_p(1)$.

We next turn to estimation of σ_u^2 . Let $\hat{\sigma}_{ws}$ be the sample covariance between W and S , and let $\hat{\sigma}_w^2$ be the sample variance of W . Then $\hat{\sigma}_x^2 = \hat{\sigma}_{ws} / \hat{\alpha}_1$ is a root- n -consistent estimator of σ_x^2 , and hence $\hat{\sigma}_u^2 = \hat{\sigma}_w^2 - \hat{\sigma}_x^2$ necessarily satisfies $n^{1/2}(\hat{\sigma}_u^2 - \sigma_u^2) = O_p(1)$. Finally, $\hat{\lambda}_n = \hat{\sigma}_x^2 / \hat{\sigma}_w^2$ is a root- n -consistent estimate of the attenuation λ .

The estimator σ_u^2 is not necessarily positive, so we suggest the following modification. Set a user-specified lower bound on the attenuation (9), say λ_L . Let $\hat{\lambda}$ be the estimate of λ obtained by replacing σ_w^2 in (9) by the sample variance $\hat{\sigma}_w^2$ of the W 's and by replacing σ_u^2 by its estimate. If $\lambda_L \leq \hat{\lambda}$, then use this estimate of $\hat{\sigma}_u^2$. If $\hat{\lambda} < \lambda_L$, then form a new estimator by solving $\lambda_L = (\hat{\sigma}_w^2 - \sigma_u^2) / \hat{\sigma}_w^2$ so that $\hat{\sigma}_u^2 = \hat{\sigma}_w^2(1 - \lambda_L)$.

2.4 More General Models

The model (2), (3), and (5) can be generalized to the following.

Assumption 3. Assume that

$$Y = \mathcal{G}\{\mathbf{Z}, X, \epsilon\}, \quad (11)$$

$$W = X + U, \quad (12)$$

and

$$S = \alpha_0(\mathbf{Z}) + \alpha_1(\mathbf{Z})X + \nu, \quad (13)$$

where $\mathcal{G}(\cdot, \cdot, \cdot)$ is a smooth function, \mathbf{Z} is a vector of covariates measured without error, (ϵ, U, ν) have mean 0 and variances $(\sigma_\epsilon^2, \sigma_u^2, \sigma_\nu^2)$ and are mutually independent of one another and of (X, \mathbf{Z}) , and $\alpha_0(\cdot)$ and $\alpha_1(\cdot)$ are unknown smooth functions. We observe (Y, W, S, \mathbf{Z}) .

Model (11) includes generalized linear models; the partially linear model

$$Y = \boldsymbol{\gamma}^T \mathbf{Z} + m(X) + \epsilon, \quad (14)$$

where m is an unknown smooth function; and the additive model where, with g also an unknown smooth function,

$$Z \text{ is univariate and } Y = g(Z) + m(X) + \epsilon. \quad (15)$$

In this section we not only show that σ_u^2 is identified, but also display a root- n -consistent estimate of it. Our next assumption is similar to one made previously, but includes the assumption that α_1 is constant, which will be removed in Section 2.5.

Assumption 4. Assume that Z is univariate with support $[0, 1]$ and density $f_Z(\cdot)$ bounded away from 0 on $[0, 1]$, and there exists a known bound $L \geq 1$ such that for some ℓ , $1 \leq \ell \leq L$, $\text{cov}[Y, \{W - E(W|Z)\}^k | Z] = 0$ for $k < \ell$, and $\text{cov}[Y, \{W - E(W|Z)\}^\ell | Z]$ is bounded away from zero for $0 \leq Z \leq 1$. Also, $\alpha_1(\cdot)$ is constant.

Because the following is easy to verify, we omit the proof.

Lemma 1. Under Assumptions 3 and 4, for all ℓ , $\alpha_1^\ell \text{cov}[Y, \{W - E(W|Z)\}^\ell | Z] = \text{cov}[Y, \{S - E(S|Z)\}^\ell | Z]$.

With Assumptions 3 and 4 and Lemma 1, the method becomes easy to describe. Make the following definitions: $\eta_{\ell w}(z) = \text{cov}[Y, \{W - E(W|Z=z)\}^\ell | Z=z]$, $\eta_{\ell s}(z) = \text{cov}[Y, \{S - E(S|Z=z)\}^\ell | Z=z]$, $\gamma_w(z) = \text{var}(W|Z=z) = \text{var}(X|Z=z) + \sigma_u^2$, $\gamma_s(z) = \text{var}(S|Z=z) = \alpha_1^2 \text{var}(X|Z=z) + \sigma_v^2$, and $\xi_{ws}(z) = \text{cov}(W, S|Z=z) = \alpha_1 \text{var}(X|Z=z)$.

Assume that ℓ and L satisfy Assumption 1, and note that $\alpha_1^\ell = \eta_{\ell s}(z)/\eta_{\ell w}(z)$ for $0 \leq z \leq 1$. Let $0 < a < b < 1$, and define $\hat{p}_{ab} = n^{-1} \sum_{i=1}^n I(a \leq Z_i \leq b)$ and $p_{ab} = \Pr(a \leq Z \leq b)$. Shortly we give estimators of $\eta_{\ell w}(\cdot)$, $\eta_{\ell s}(\cdot)$, $\gamma_w(z)$, $\gamma_s(z)$, and $\xi_{ws}(z)$. Given these estimators, our estimate of α_1^ℓ becomes $\hat{\alpha}_1^\ell = (n\hat{p}_{ab})^{-1} \sum_{i=1}^n I(a \leq Z_i \leq b) \hat{\eta}_{\ell s}(Z_i) / \hat{\eta}_{\ell w}(Z_i)$. Given $\hat{\xi}_{ws}(\cdot)$, because $\text{sign}(\alpha_1) = \text{sign}\{\xi_{ws}(z)\}$, our estimate of $\text{sign}(\alpha_1)$ is $\text{sign}\{n^{-1} \sum_{i=1}^n I(a \leq Z_i \leq b) \hat{\xi}_{ws}(Z_i)\}$, and our estimate of σ_u^2 is $\hat{\sigma}_u^2 = (n\hat{p}_{ab})^{-1} \sum_{i=1}^n I(a \leq Z_i \leq b) \{\hat{\gamma}_w(Z_i) - \hat{\xi}_{ws}(Z_i) / \hat{\alpha}_1\}$.

To construct the various estimators, we use kernel regression, primarily for ease of theoretical explication; we expect that most other nonparametric estimators will also work. Let $K(\cdot)$ be a symmetric density function with bounded support, let h be a bandwidth, and define $K_h(x) = h^{-1}K(x/h)$. In what follows, we use an undersmoothed kernel estimator, with $C_1 n^{-1/3} \leq h \leq C_2 n^{-1/3}$ for all n , for some $0 < C_1 \leq C_2 < \infty$.

Assumption 5. Suppose that Assumptions 3 and 4 hold. Define $m_{jkp}(z) = E(Y^j W^k S^p | Z=z)$. Let $\hat{m}_{jkp}(\cdot)$ be a kernel regression estimator of the regression of $Y^j W^k S^p$ on Z with the property that for $j = 0, 1$, $0 \leq k, p \leq L$, uniformly for $0 < a \leq z \leq b < 1$,

$$\hat{m}_{jkp}(z) - m_{jkp}(z) = \left[\{nf_Z(z)\}^{-1} \sum_{i=1}^n K_h(Z_i - z) \epsilon_{jkpi} \right] + O_P\{n^{-2/3} \log(n)\}, \quad (16)$$

for any $d > 0$, where $E\{\epsilon_{jkpi} | Z_i\} = 0$ and $\text{var}\{\epsilon_{jkpi} | Z_i\} \leq A < \infty$ for some A and all Z . The term in square brackets in (16) is $O_P(n^{-1/3} \log(n)^{1/2})$.

Under standard regularity conditions (e.g., those in Mack and Silverman 1982), kernel regression estimators satisfy (16); see Appendix A.2.

Each of $\eta_{\ell w}(z)$, $\eta_{\ell s}(z)$, $\gamma_w(z)$, $\gamma_s(z)$, and $\xi_{ws}(z)$ is a function $\{m_{jkp}(z) : j = 0, 1, 0 \leq k, p \leq L\}$. Define $\hat{\eta}_{\ell w}(z)$, $\hat{\eta}_{\ell s}(z)$, $\hat{\gamma}_w(z)$, $\hat{\gamma}_s(z)$, and $\hat{\xi}_{ws}(z)$ to be the analogous functions of $\{\hat{m}_{jkp}(z) : j = 0, 1, 0 \leq k, p \leq L\}$. The proof of the following is given in Appendix A.3.

Theorem 3. Suppose that Assumption 5 holds. Then $n^{1/2} \times (\hat{\alpha}_1 - \alpha_1) = O_P(1)$. In addition, there are random variables ϵ_{ui} such that $E(\epsilon_{ui} | Z_i) = 0$ and $n^{1/2}(\hat{\sigma}_u^2 - \sigma_u^2) = n^{-1/2} \sum_{i=1}^n \epsilon_{ui} + o_P(1) = O_P(1)$.

Finally, we discuss the estimation of ℓ satisfying Assumption 4. Let $0 < a_* < a < b < b_* < 1$, where (a, b) are used in constructing the estimators. Having chosen (a, b) , we merely need to identify an ℓ such that $Q_\ell(z)$ is bounded away from 0

on $[a_*, b_*]$, where $Q_\ell(z) = \text{cov}[Y, \{W - E(W)\}^\ell | Z=z] = \text{cov}[m(X), \{X - E(X)\}^\ell | Z=z]$. Because of Assumption 5, we can find an estimator $\hat{Q}_\ell(z)$ that converges to $Q_\ell(z)$ uniformly on compact interior subsets of $[0, 1]$ at the rate $O_P\{n^{-1/3} \log(n)^{1/2}\}$. Let $c_n \rightarrow 0$ be a fixed sequence converging to 0 more slowly than $n^{-1/3} \log(n)^{1/2}$. Then the smallest ℓ such that $\hat{Q}_\ell(z) > c_n$ for $a_* \leq z \leq b_*$ will do.

In practice, we make the following modifications to this algorithm:

- Because $\gamma_w(z) - \gamma_s(z)/\alpha_1 > 0$ and $\gamma_s(z)/\alpha_1 > 0$, in the definition of $\hat{\sigma}_u^2$ we take an average over those $a \leq Z \leq b$ such that $\hat{\gamma}_w(Z) - \hat{\xi}_{ws}(Z)/\hat{\alpha}_1 > 0$ and $\hat{\xi}_{ws}(Z)/\hat{\alpha}_1$.
- Let $\hat{\sigma}_w^2$ be the sample variance of the W_i . For a lower bound $\lambda_{Ln} > 0$ on the attenuation, if $\hat{\lambda}_n = (\hat{\sigma}_w^2 - \hat{\sigma}_u^2) / \hat{\sigma}_w^2 < \lambda_{Ln}$, then we set $\hat{\sigma}_u^2 = \hat{\sigma}_w^2(1 - \lambda_{Ln})$. If $\lambda_{Ln} = o(n^{-1/2})$, then the modifications do not affect the rate of convergence for $\hat{\sigma}_u^2$.

2.5 Varying-Coefficient Instruments

Consider the model (11)–(13), with $\alpha_1(\cdot)$ being a smooth function. Minor changes in the algorithm and the results of Section 2.4 are needed. We first note that we have the estimate $\hat{\alpha}_1^\ell(z) = \hat{\eta}_{\ell s}(z) / \hat{\eta}_{\ell w}(z)$. This suggests the estimator $\hat{\sigma}_{u,vc}^2 = (n\hat{p}_{ab})^{-1} \sum_{i=1}^n I(a \leq Z_i \leq b) \{\hat{\gamma}_w(Z_i) - \hat{\xi}_{ws}(Z_i) / \hat{\alpha}_1(Z_i)\}$. The proof of the next theorem is given in Appendix A.4.

Theorem 4. Suppose that ℓ satisfying Assumption 4 is known, that Assumption 5 holds, and that $\alpha_1(z)$ is bounded away from 0. Then $\text{sign}\{\alpha_1(z)\}$ is estimated consistently and $n^{1/2}(\hat{\sigma}_{u,vc}^2 - \sigma_u^2) = O_P(1)$. In addition, there are random variables $\epsilon_{ui,vc}$ with the property that $E(\epsilon_{ui,vc} | Z_i) = 0$ such that $n^{1/2}(\hat{\sigma}_{u,vc}^2 - \sigma_u^2) = n^{-1/2} \sum_{i=1}^n \epsilon_{ui,vc} + o_P(1)$.

2.6 Additional Considerations

Having constructed a root- n -consistent estimate of the measurement error variance σ_u^2 , along with an asymptotic linear expansion for it, allows us to apply any estimator for the measurement error problem with σ_u^2 known to the IV problem. Various special cases are of interest.

Suppose that (14) holds, $\alpha_1(\mathbf{z}) \equiv \alpha_1$, and $\alpha_0(\mathbf{z}) = \alpha_0 + \alpha_{0,1}^T \mathbf{z}$. Then if we treat $S = Y^*$ as a response and $\epsilon^* = v$ as the regression errors, we have a linear model, $Y^* = \alpha_{0,1} + \alpha_{0,1}^T \mathbf{z} + \alpha_1 X + \epsilon^*$, and the parameters can be estimated using the method of moments. In addition, the model of primary interest would be $Y = \beta^T \mathbf{Z} + m(X) + \epsilon$, a partially linear measurement error model with the error in X . Because we now have a root- n -consistent estimate of σ_u^2 , the methods of Liang (2000) can be used.

Suppose that $g(\mathbf{z})$ and $\alpha_0(\mathbf{z})$ are left unspecified and the regression of Y on (Z, X) is an additive model. Because the root- n rate for estimating σ_u^2 is faster than the possible rate of convergence for estimating either $g(\mathbf{z})$ or $m(x)$, we can treat this problem asymptotically as if σ_u^2 were known. Hence, once the problem of additive models with known measurement error is solved (and to the best of our knowledge it has not been), the same problem is solved for the IV problem.

3. ESTIMATING THE REGRESSION FUNCTION

Once an estimator $\hat{\sigma}_u^2$ is available, several estimators of m are available from the measurement error literature. To simplify the discussion and to save space, we assume that (5) holds, although we indicate when more general assumptions can be accommodated.

In Section 3.1 we describe a consistent estimator of m due to Fan and Truong (1993). Although consistent, this method was much less efficient than competitors in a finite-sample study of a measurement error problem where W was replicated but there was no instrument S (Carroll et al. 1999). Because of the inefficiency of the Fan and Truong estimator, we describe alternative estimators even though these have not been proven to be consistent. In Section 3.2 we describe penalized splines. In Section 3.3 we propose the functional simulation extrapolation method (Cook and Stefanski 1994), which makes no assumptions about the distributions of the random variables (X, U, ϵ, ν) . In the nonparametric regression problem with σ_u^2 known or estimated by replication of W , Berry et al. (2002) showed that a Bayesian approach using splines could achieve significant gains in efficiency when compared to the SIMEX method. In Section 3.4 we extend the Bayesian method to the IV problem and to problems such as binary regression.

3.1 Deconvolution Kernels

Assume that σ_u is known and that without loss of generality equals 1. Let $K(\cdot)$ be a density function with Fourier transform $(1 - t^2)_+^3 = (1 - t^2)^3 I(t^2 \leq 1)$. The deconvoluting kernel (Fan and Truong 1993, sec. 5.1), $K_n(x, h) = (\pi h)^{-1} \int_0^1 \cos(tx/h) (1 - t^2)^3 \exp\{\sigma_u^2 t^2 / (2h^2)\} dt$, has the property $E\{K_n(W - x_0, h) | X\} = h^{-1} K\{(X - x_0)/h\}$ that controls the bias. The deconvoluting kernel density estimator is $\hat{f}_X(x_0) = n^{-1} \sum_{i=1}^n K_n(W_i - x_0, h)$, and the deconvoluting kernel regression estimator is $\hat{m}(x_0) = \{\hat{f}_X(x_0)\}^{-1} n^{-1} \times \sum_{i=1}^n K_n(W_i - x_0, h) Y_i$. Then we have that, for some constant $\delta > 0$ and all x ,

$$\hat{f}_X(x) - f_X(x) = O_p\{h^2 + (nh^3)^{-1} \exp(\delta/h^2)\} \quad (17)$$

and

$$\hat{m}(x) - m(x) = O_p\{h^2 + (nh^3)^{-1} \exp(\delta/h^2)\}. \quad (18)$$

The former was given by Carroll and Hall (1988); the latter by Fan and Truong (1993).

If σ_u is unknown, it can be replaced in the root- n -consistent estimator in Section 2.3. Because $\hat{\sigma}_u^2$ is root- n consistent, (17) and (18) continue to hold.

3.2 Penalized Splines

A general approach to spline fitting is to use *penalized splines*, or *P-splines* (Eilers and Marx 1996). We introduce the idea in this section (see Ruppert, Wand, and Carroll 2003 for further discussion of P-splines).

Let $\mathbf{C}(x) = \{B_1(x), \dots, B_N(x)\}^T$, $N \leq n$, be a basis of linearly independent piecewise polynomial functions. The P-spline model specifies that $m(x) = m(x, \boldsymbol{\beta}) = \mathbf{C}(x)^T \boldsymbol{\beta}$ for some $\boldsymbol{\beta}$. Popular bases are B-splines (Eilers and Marx 1996) and the truncated power series basis (Ruppert 2002). B-splines are more stable numerically than the truncated power basis,

but the roughness penalty adds numerical stability, making the truncated power basis computationally feasible (Ruppert 2002). We use a p th-degree polynomial spline with k knots, t_1, \dots, t_k . We choose the knots at the quantiles of the W 's to accommodate possible regions of sparse data. Equally spaced knots could be used if the W_i 's were roughly uniformly distributed. A convenient basis for these splines is the set of monomials plus the truncated power functions so that $\mathbf{C}(x) = (1, x, x^2, \dots, x^p, (x - t_1)_+^p, \dots, (x - t_k)_+^p)^T$, where $a_+^p = \{\max(0, a)\}^p$. Then $N = 1 + p + k$, $\beta_1, \dots, \beta_{p+1}$ are the monomial coefficients, and $\beta_{2+p}, \dots, \beta_N$ are the sizes of the jumps in the p th derivative of $g(x) = \mathbf{C}(x)^T \boldsymbol{\beta}$ at the knots.

If measurement error is ignored, then it is typical to fit the function $m(x, \boldsymbol{\beta})$ by penalized maximum likelihood. Consider the truncated power series basis. Let \mathbf{D}_* be the $N \times N$ diagonal matrix with $p + 1$ 0's followed by k 1's along the diagonal. Let γ be a smoothing parameter. The penalized estimator $\hat{\boldsymbol{\beta}}(\gamma)$ ignoring measurement error minimizes the log-likelihood in (Y, W) minus $\gamma \boldsymbol{\beta}^T \mathbf{D}_* \boldsymbol{\beta}$. More formally, suppose that the log-likelihood in (Y, X) is $\mathcal{L}(Y, X, \boldsymbol{\beta})$. Then the penalized regression spline ignoring measurement error is the solution to

$$\max_{\boldsymbol{\beta}} \left[\left\{ \sum_{i=1}^n \mathcal{L}(Y_i, W_i, \boldsymbol{\beta}) \right\} - \gamma \boldsymbol{\beta}^T \mathbf{D}_* \boldsymbol{\beta} \right]. \quad (19)$$

One can use cross-validation (CV) or generalized cross-validation (GCV) to choose γ . Also, a P-spline can be viewed as a mixed model, and then γ can be estimated as a ratio of variance components using either restricted maximum likelihood or a Bayes estimator (Ruppert et al. 2003). Other penalties such as on the integral of the squared second derivative can be imposed by other choices of \mathbf{D}_* .

3.2.1 Selecting the Knots. Readers may wish for clear guidelines on choosing the number of knots k and may find our recommendation to place knots at sample quantiles somewhat arbitrary. Unfortunately, it is difficult to give definite recommendations, simply because the choices of number of knots and their locations are not critical, as we now discuss. The choice of k was discussed by Ruppert (2002), who found that for P-splines the exact value of k has relatively little effect on the estimator, provided that k is at least a certain minimum value; the reason for this is that smoothness is controlled by the penalty parameter. Generally, $k = 20$ more than suffices for the types of regression functions that are found in practice and that can be recouped when there is measurement error, although of course there can be exceptions. Berry et al. (2002) found that P-splines and smoothing splines, which have a knot at every value of the covariate, generally give very similar answers, so one can use a large number of knots even if it is not necessary to do so.

Powell's (1981) results on the approximation properties of splines help explain why relatively few knots are needed and why their exact locations are not crucial. Let $a < b$ be two real numbers and consider spline approximation of the regression function m on $[a, b]$. Powell's theorem 20.3 gives the error of the best approximation of a $C^l[a, b]$ (l times continuously differentiable) function m by a k th-degree spline. For convenience, we now restate the theorem.

Theorem 5 (Powell 1981). Let \mathcal{S} be the set of k th-degree splines on $[a, b]$ with knots $a < \kappa_1 < \dots < \kappa_k < b$. Define $\kappa_0 = a$ and $\kappa_{k+1} = b$. Let $h = \max\{\kappa_j - \kappa_{j-1} : j = 1, \dots, k + 1\}$. Suppose that m is $C^l[a, b]$. Then

$$\inf_{s \in \mathcal{S}} \|m - s\|_\infty \leq \frac{(k + 1)!}{(k + 1 - j)!} (h/2)^j \|m^{(j)}\|_\infty \quad (20)$$

for every $j \in \{1, 2, \dots, \min(l, k + 1)\}$, where $\|\cdot\|_\infty$ is the L^∞ norm on $[a, b]$.

This bound is independent of the specific knot locations, instead depending only on the maximum distance between any two consecutive knots. For example, if one uses quadratic splines and $m \in C^3[a, b]$, then (20) holds for $j = 3$. If the covariate has a density on $[a, b]$ bounded away from 0, then h will be proportional to $1/k$ for knots at sample quantiles with equal probability spacing. Thus the bias due to approximation of m by a spline will be $O(k^{-3})$. If $k \propto n^{1/6}$, then the squared bias due to spline approximation will be $O(1/n)$, the *parametric* rate for the variance. Thus having the k grow proportional to $n^{1/6}$ would be more than adequate. Typically, k could grow much more slowly. For example, the best possible rate of convergence for nonparametric regression with normal measurement errors is a power of $\log(n)$ (Fan and Truong 1993), so in the presence of measurement error, if k grows faster than any power of $\log(n)$ (e.g., n^δ for any $\delta > 0$), then the bias due to spline approximation will be smaller than the best possible rate of convergence.

3.3 SIMEX

SIMEX needs a *base estimator* that one would use if there were no measurement error. Carroll et al. (1999) described two base estimators: local linear kernel regression of Y on W with bandwidths estimated either by GCV or by empirical bias bandwidth selection (EBBS), and P-splines of Y on W as described previously, with smoothing parameter estimated by GCV.

Our SIMEX-IV estimator applies the SIMEX method of Cook and Stefanski (1994), using $\hat{\sigma}_u^2$ as the estimate of error variance. For any fixed $\zeta > 0$, one repeatedly adds to W , via simulation, additional error with mean 0 and variance $\hat{\sigma}_u^2 \zeta$, forming *pseudovalues*. One generates sets of pseudovalues repeatedly, computes the base estimator on each set, and averages the estimators, calling the average $g(\zeta)$. Generally, 50–200 sets of pseudovalues will suffice, and one uses $\lambda = 0, .5, 1.0, 1.5$, and 2.0. The idea is to plot $g(\zeta)$ against $\zeta \geq 0$, fit a model to this plot, and then extrapolate back to $\zeta = -1$. In our calculations, we used a quadratic function to model the plot of $g(\zeta)$ against ζ .

3.3.1 Asymptotic Theory for SIMEX. Asymptotic theory for the SIMEX method is easy if one uses kernel methods. Because $\hat{\sigma}_u^2$ converges at the rate $O_p(n^{-1/2})$, and the kernel estimator converges at a slower rate, the asymptotics are the same as if σ_u^2 were known. This means that the SIMEX-kernel IV estimator has the same asymptotic distribution and expansion as described by Carroll et al. (1999). In the interest of space, we do not rewrite the details.

There are no known limiting results for P-splines and an estimated smoothing parameter. If the number of knots is fixed and the smoothing parameter γ in (19) is held fixed or converges

to 0, then the solution to (19) is the solution to an estimating equation. The limiting distribution of SIMEX for estimating equations is already known (see Stefanski and Cook 1995; Carroll, Küchenhoff, Lombard, and Stefanski 1996).

3.4 Bayesian P-Splines

Our Bayesian methods are similar to those described by Berry et al. (2002). For simplicity, we describe Bayesian P-splines only for the model (2), (3), and (5) with Y either Gaussian or binary, although the extension to the more general models in Section 2.4 is not difficult. Partition $\mathbf{C}(x) = \{\mathbf{C}_1^T(x), \mathbf{C}_2^T(x)\}^T$, where $\mathbf{C}_1^T(x) = (1, x, x^2, \dots, x^p)$. Partition $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T)^T$ similarly. As is common with P-splines, we will assume that $\boldsymbol{\beta}_2 = \mathbf{N}(\mathbf{0}, \sigma^2 \mathbf{I})$, where \mathbf{I} is the identity matrix. Other formulations are possible. The parameters then become $\alpha_0, \alpha_1, \mu_x, \sigma_x^2, \sigma_u^2, \sigma_v^2, \boldsymbol{\beta}$, and σ^2 .

The formulas to implement the Gibbs sample are detailed. In Sections A.6 and A.7 we give these formulas for Gaussian and binary Y . Section 4.3 describes an implementation in WinBUGS and our experience with it.

3.4.1 Asymptotic Theory for the Bayesian Estimator. With the number of knots fixed, the model is parametric and the Bayesian estimator is asymptotic efficient at the model (e.g., by Lehmann 1983, thm. 7.2; or Bernardo and Smith 1994, p. 291). Thus for smooth m , a fixed-knot spline model can approximate m arbitrarily closely by Theorem 5, and the Bayesian estimator is asymptotically efficient for any m that is a spline with these knots. This result quite possibly explains the Bayesian estimator’s good performance in the simulation studies of Section 4 and in the study of Carroll, Maca, and Ruppert (1999).

If the number of knots increases to infinity, then we conjecture that P-splines are consistent by results of Barron, Schervish, and Wasserman (1999), but we have been unable to verify that their assumptions hold.

3.5 The Case Where $\text{cov}[m(X), \{X - E(X)\}^k] = 0$ for All k

It would appear that when $\text{cov}[m(X), \{X - E(X)\}^k] = 0$ for all k , the function $m(\cdot)$ may not be identifiable. However, in the important subcase where $m(\cdot)$ is constant, say equal to c , $m(\cdot)$ is identifiable, at least when a lower bound on the attenuation is specified.

We sketch the proof using the SIMEX estimator. Assume that the extrapolant function is parametric and includes the constant function as a special case. Recall that if the regression function is constant, then $E(Y|X) \equiv E(Y|W) \equiv c$. Thus the naive estimator that ignores measurement error consistently estimates $m(\cdot)$.

Consider what happens in the SIMEX algorithm. If the attenuation is bounded below by λ_L , then for sufficiently large samples, we have that $(1/2)\lambda_L \leq (\hat{\sigma}_w^2 - \sigma_u^2)/\hat{\sigma}_w^2$. This means that for sufficiently large samples we can find an interval $[a, b]$ for σ_u^2 , and the attenuation on this interval of values always exceeds 0. Fix any σ_{u*}^2 in this interval. Consider the construction of pseudovalues $W(\text{pseudo}, \zeta, \sigma_{u*}^2) = W + \zeta^{1/2} \sigma_{u*} Z$, where Z is $\mathbf{N}(0, 1)$. The pseudovalues also satisfy $E\{Y|W(\text{pseudo}, \zeta, \sigma_{u*}^2)\} \equiv c$. Thus the naive estimator applied to the pseudovalues consistently estimates $m(\cdot) \equiv c$.

Because the extrapolant function includes the constant function as a special case, for any $\sigma_{u^*}^2$ in $[a, b]$, applying SIMEX leads to a consistent estimate of $m(\cdot)$. To complete the proof, one must show that this argument holds uniformly in $\sigma_{u^*}^2 \in [a, b]$, and hence that SIMEX is consistent if one bounds the attenuation away from 0. SIMEX is used only to prove identifiability and other methods could be used to estimate m .

4. A SIMULATION STUDY

We performed a modest simulation experiment using the model (2), (3), and (5) with Y Gaussian and $n = 100$, a small sample size given the difficulty of nonparametric regression with measurement error. We took $\sigma_x^2 = 1$, $\sigma_u^2 = .33$, $\sigma_v^2 = 1$, $\alpha_0 = 0$, $\alpha_1 = 1$, and $\sigma_\epsilon^2 = .09$. The true attenuation was $\lambda = .75$. The attenuation estimator $\hat{\lambda}$ was constrained to lie in $[\.60, 1.00]$. We calculated mean squared biases and mean squared errors (MSEs) for $x \in [-2.0, 2.0]$.

Constraining the attenuation estimator with a lower bound can improve the estimation of the regression function, because it avoids gross overcorrection when the attenuation estimator is too low. This improvement can be quite significant; this was noticed in simulation results (not reported here) when $\hat{\lambda}$ was unconstrained. We recommend using such bounds when subject matter knowledge provides them. Unfortunately, it is often true that no such lower bound is known. An example is the OPEN study described in Section 5.2, for which we did not constrain the attenuation estimator. When the sample size is small, there is another reason to constrain $\hat{\lambda}$; the effect of overcorrection (i.e., using an attenuation smaller than λ) is worse than the effect of undercorrection. This can be appreciated by considering the example of linear regression, $m(x) = \beta_0 + \beta_1 x$. Often $\hat{\lambda}$ is independent of the naive least squares estimator $\hat{\beta}_{LS,1}$. For simplicity, we will assume this condition for $\hat{\lambda}$. We also assume that $E(\hat{\beta}_{LS,1}) = \lambda\beta_1$, although this is only an asymptotic approximation. Then

$$\begin{aligned} \text{MSE}(\hat{\lambda}^{-1}\hat{\beta}_{LS,1}|\hat{\lambda}) &\approx \left(\frac{\lambda}{\hat{\lambda}}\beta_1 - \beta_1\right)^2 + \frac{\text{var}(\hat{\beta}_{LS,1})}{\hat{\lambda}^2} \\ &= \hat{\lambda}^{-2}\{(\lambda - \hat{\lambda})^2\beta_1^2 + \text{var}(\hat{\beta}_{LS,1})\}. \end{aligned}$$

Thus for a given value of $\lambda - \hat{\lambda}$, the MSE is smaller if $\hat{\lambda} > \lambda$ (undercorrection) than if $\hat{\lambda} < \lambda$ (overcorrection). Figure 1 shows this approximation to the MSE as a function $\hat{\lambda}$ when $\lambda = .5$, $\beta_1 = 1$, and $\text{var}(\hat{\beta}_{LS,1}) = .1$. Interestingly, the MSE is higher when using $\hat{\lambda} = .5$ than when using $.5 < \hat{\lambda} \leq 1$. Thus, any truncation of $\hat{\lambda}$ should improve the MSE in this example. Also, underestimation of λ can greatly increase the MSE, which is the motivation for using truncation. Even a modest amount of truncation, say to the interval $[\.3, 1]$, would be helpful. Of course, if more data were available, so that $\text{var}(\hat{\beta}_{LS,1})$ was sufficiently smaller than .1 but β_1 was still 1, then truncation of $\hat{\lambda}$ to a value well above λ would increase the MSE. Truncation should not be used blindly, but rather should be done with knowledge of the subject matter and the relative size of the bias and variance of the naive estimator.

Although we assumed that the X 's were normally distributed, to test robustness we considered three distributions for the X 's: $N(0, 1)$, uniform on $[-2, 2]$, and skewed normal with

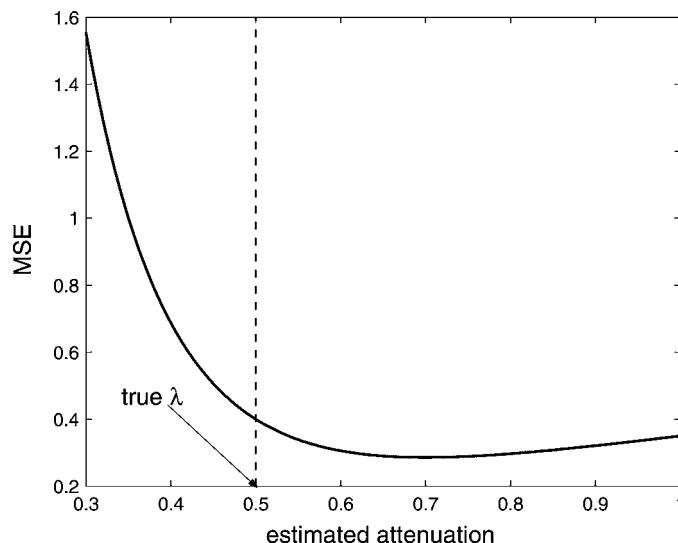


Figure 1. MSE of the Corrected Slope Estimator as a Function of the Estimated Attenuation. The true attenuation is .5, the true β is 1, and the variance of the LS estimate is .1.

index $\alpha = 5$. The skewed normal distribution has density proportional to $f(x|\alpha) = 2\phi(x)\Phi(\alpha x)$, where ϕ and Φ represent the standard normal density and distribution (Azzalini 1985). This density is reasonably skewed for any value of $\alpha \geq 5$.

We considered three regression models: $1/\{1 + \exp(4x)\}$ (case 1), $\sin(\pi x/2)/(1 + [2x^2\{1 + \sin(\pi x/2)\}])$ (case 2), and $\sin(\pi x/2)/(1 + [2x^2\{1 + \text{sign}(x)\}])$ (case 3).

4.1 Bias–Variance Trade-Offs in P-Spline Estimation

Carroll et al. (1999, sec. 4.4) described theoretical calculations in the classical measurement error problem showing that if one uses regression splines and maximum likelihood estimation, then the variance of the fits “blows up” as the smoothing parameter converges to 0.

What does this mean, and why is it important? The essential point is that with a sample of size 100, our methods must penalize the spline to make it reasonably stable. There is a cost for such smoothing bias. Specifically, for such sample sizes, it is impossible to estimate the details of difficult functions with, for example, deep valleys, such as cases 2 and 3.

4.2 Results

The results are given in Table 1 for a 25-knot quadratic regression spline; similar results were obtained for the linear spline. In this table, and MSE mean squared bias are averages over 101 grid points on the interval $[-2, 2]$ and over all Monte Carlo samples.

We see that the Bayes estimator clearly dominates the SIMEX estimators and the naive estimator that ignores measurement error, in terms of both bias and MSE. The SIMEX estimator with a quadratic extrapolant is far less biased than the naive estimator, but it has large variance.

Figure 2 corresponds to Table 1, normal X , case 3. Figures 2(a), 2(b), and 2(c) show three simulated datasets; Figure 2(d) shows the mean over all simulated datasets. This is a problem for which the naive estimator is only somewhat worse than the Bayes estimator (from Table 1: naive squared bias,

Table 1. $100 \times$ Mean Squared Bias and $100 \times$ MSE for the Simulation for the Spline Gaussian Error Model

Distribution	Method	Case 1		Case 2		Case 3	
		Mean squared bias	MSE	Mean squared bias	MSE	Mean squared bias	MSE
Normal	Naive	1.40	1.98	7.27	8.43	2.99	3.72
	SIMEX(L)	.82	1.61	6.56	8.19	2.72	3.77
	SIMEX(Q)	.52	3.31	4.60	11.25	1.92	5.90
	Bayes	.21	1.02	2.51	4.40	1.29	2.97
Uniform	Naive	.91	1.64	5.94	7.09	2.61	3.34
	SIMEX(L)	.57	1.40	5.32	6.59	2.31	3.14
	SIMEX(Q)	.43	3.33	2.86	7.34	1.29	4.20
	Bayes	.19	.78	2.61	3.80	1.62	2.44
Skewed normal	Naive	1.38	2.11	9.64	10.91	3.28	4.12
	SIMEX(L)	.84	1.68	9.87	11.26	3.36	4.26
	SIMEX(Q)	.58	3.57	8.36	13.17	2.59	5.34
	Bayes	.29	1.21	4.71	6.76	1.44	3.28

NOTE: In case 1 the regression function is $1/(1 + \exp(4x))$. In case 2 the regression function is $\sin(\pi x/2)/(1 + [2x^2(1 + \sin(\pi x/2))])$. In case 3 the regression function is $\sin(\pi x/2)/(1 + [2x^2 \times \{1 + \text{sign}(x)\}])$. The sample size is $n = 100$ throughout.

2.99; Bayes squared bias, 1.29; naive MSE, 3.72; Bayes MSE, 2.97). Careful inspection of the plot shows that the naive estimate often misses or just barely finds the inflection points. The SIMEX estimator has excessive variability, as shown in Table 1. This means that when the naive estimator is not too bad relative to the Bayes estimator, the differences between the SIMEX and Bayes estimator's are real but subtle. The perception from Figure 2 may be that none of the estimators is performing well. The problem is that the sample size is small considering the size of σ_u and σ_ϵ . Figure 3 shows another dataset, but now with $n = 500$. The (W, Y) pairs are also plotted to show the sizable scatter. With $n = 500$, the Bayesian estimator is rather close to the true curve.

Figure 4 corresponds to Table 1, case 2, where the Bayes estimator is a large improvement over the SIMEX estimator. This can be seen in Figure 4(a), which shows a dataset where the naive estimate is poor and the SIMEX is even worse. Table 1

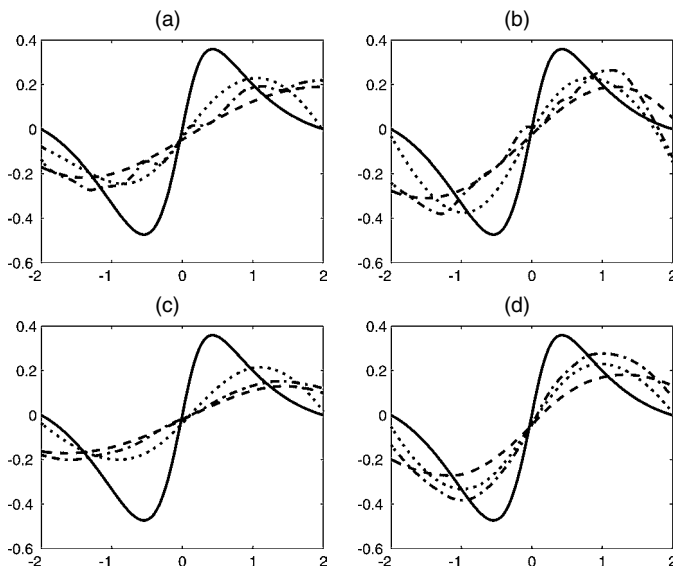


Figure 2. Results From the Simulations Corresponding to Table 1, Case 3. (a), (b), and (c) Simulated datasets; (d) the mean over all simulated datasets (— true; - - naive; - · - · SIMEX; ··· Bayes).

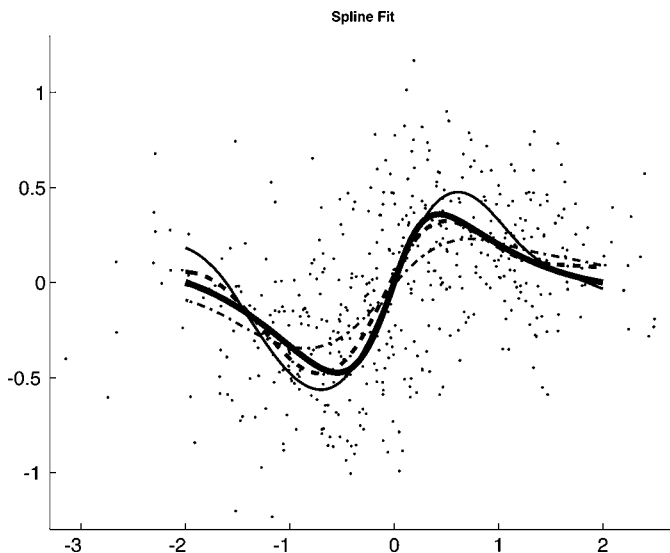


Figure 3. Results for Case 3 but With $n = 500$ (— true; - - Bayes; ··· naive; — SIMEX). The dots are the pairs (W, Y) .

shows the same thing, real dominance by the Bayesian estimator. Notice that in the bottom right the mean of SIMEX is close to that of the Bayes estimator, so that these two estimators have similar bias. This implies that the substantial MSE improvement of the Bayes estimator over SIMEX seen in Table 1 is due to the lower variability of the Bayes estimator.

4.3 Implementation and Comparison With WinBUGS

We have implemented the methods in MATLAB for the model (2), (3), and (5) with Y either Gaussian or binary. The programs are available at the website stat.tamu.edu/~carroll/matlab_programs/software.php. In addition, we have constructed 20 simulated datasets for each of the cases in the simulation, along with case 4, $m(x) = x^2$. This case is interesting because (6) is violated. Theorem 2 shows that the parameters

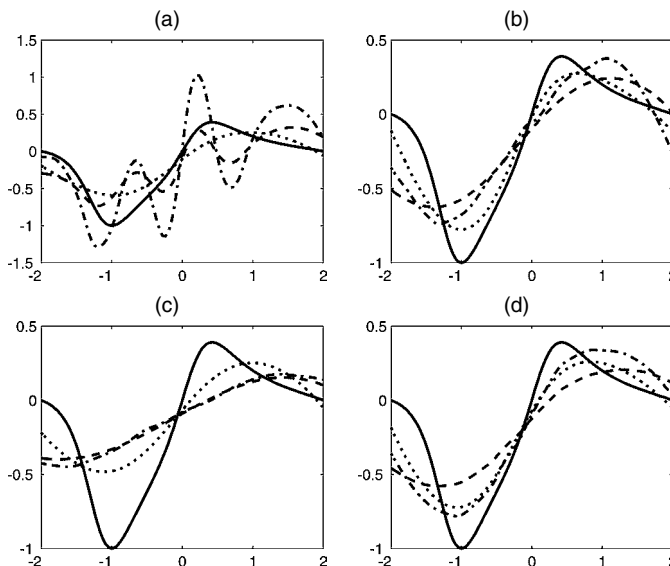


Figure 4. Results From the Simulations Corresponding to Table 1, Case 2. (a), (b), and (c) Simulated datasets; (d) the mean over all simulated datasets (— true; - - naive; - · - · SIMEX; ··· Bayes).

are still identified, but one might expect some instability because they are identified by higher moments. We have provided the naive and Bayes estimates of the regression functions.

It is also possible to implement the Bayesian method using software designed for Markov chain Monte Carlo (MCMC) simulations, such as WinBUGS. On the website we provide our WinBUGS implementation of the Gaussian model.

The MATLAB and the WinBUGS implementations use the same set of priors but different proposal distributions. The MATLAB program takes advantage of the specific features of the model for which all but two complete conditionals are explicit. Carefully tailored Metropolis–Hastings steps are used for these two complete conditionals. These features of the MATLAB program improve simulation speed (simulations per second) and, more important, MCMC mixing. For example, for a dataset with $n = 100$ for case 2, 1,000 MCMC simulations were obtained in 14 seconds with MATLAB and in 48 seconds with WinBUGS (on a 2.66-GHz CPU, 1 GB RAM). For the MATLAB program, 30,000 simulations, including 10,000 for burn-in, proved to be sufficient to achieve convergence. Due to differences in mixing quality, the WinBUGS program required 1,000,000 simulations, including 500,000 for burn-in, to achieve the same results. Despite the computational efficiency of the MATLAB program, one might prefer WinBUGS because it is much quicker and easier to program. WinBUGS would be a valuable tool in the initial phase of research when many models are considered, and also to validate other programs. For example, to implement the models in Section 2.4 that are more complex than the model (2), (3), and (5), we might prefer WinBUGS over MATLAB, to save programming time.

5. EXAMPLES

5.1 The Arsenic Example

Arsenic exposure has been clearly linked with skin, bladder, and lung cancer occurrence in highly exposed populations either occupationally, medicinally, or through contaminated drinking water (National Research Council 1999; International Agency for Research on Cancer 1987). An ongoing population-based study in New Hampshire (Karagas et al. 1998, 2001) is examining the effects of arsenic on the incidence of skin and bladder cancer in response to low to moderate exposures, due primarily to natural sources of arsenic contamination in well water. Because of intense regulatory interest in the effects of abatement strategies, the shape of the exposure–response relationship at lower exposures is important, and strategies for nonlinear modeling are currently being explored (Karagas and Tosteson 2002).

Exposure assessment is accomplished through the measurement of arsenic concentrations in both tap water from home water supplies and toenail samples for individuals newly diagnosed with skin or bladder cancer (cases) and individuals belonging to an age- and gender-matched sample of other state residents (controls). For our example, we consider data for 215 controls and 233 basal cell skin cancer cases with both water and toenail samples. Because we are interested in characterizing changes in cancer incidence due to changes in arsenic water contamination, we specify the water measurement as the unbiased exposure, taking X to be $\log(.005 + \text{arsenic level in}$

tap water sample) and W to be the measured value of this quantity. The toenail arsenic measurements are interpreted as the IV, so that S is specified as $\log(.005 + \text{arsenic level in the toenail sample})$. Log transformations were chosen to make W and S both reasonably close to normally distributed, although some skewness remains. We used the model (2), (3), and (5) with Y the binary indicator of a case.

Preliminary analysis ignoring measurement error showed a positive but not statistically significant linear trend between arsenic in tap water and basal cell cancer incidence. For the purposes of this analysis, the results were not adjusted for possible confounding factors such as age and gender. The results for the regression spline analysis are given in Figure 5. The naive fit ignoring measurement error shows a modest increase in the logit of basal cell cancer incidence over the range of observed tap water arsenic levels, with some indication of nonlinearity. The Bayes fit adjusting for measurement error shows a somewhat more-uniform increase, with the impression of less nonlinearity. The confidence bands indicate that the overall increase is not statistically significant.

The posterior means were $\alpha_0 = -2.0$, $\alpha_1 = .20$, $\sigma_x^2 = 2.61$, $\sigma_u^2 = .54$, $\sigma_v^2 = .28$, $\mu_x = -1.11$, and $\lambda = .83$, the latter indicating that the amount of attenuation is not as great as might be supposed.

In our example, the designation of tap water as the unbiased exposure measure reflects a certain interpretation of the fitted regression curve, that the curve is the probability of skin cancer given a level of exposure in drinking water. However, in practice, total arsenic exposure includes not only the amount consumed, but also exposure from other sources, such as food. Another formulation would focus on the dose response for a biologically active arsenic exposure, for which toenail concentrations could be taken as an unbiased measure. Conceptually, this would introduce an additional latent variable to represent the biologically active exposure, D , which would depend on true tap water concentrations in a linear fashion. Retaining the designation of W for transformed value of measured tap water arsenic and S for toenail, we could rewrite our model as

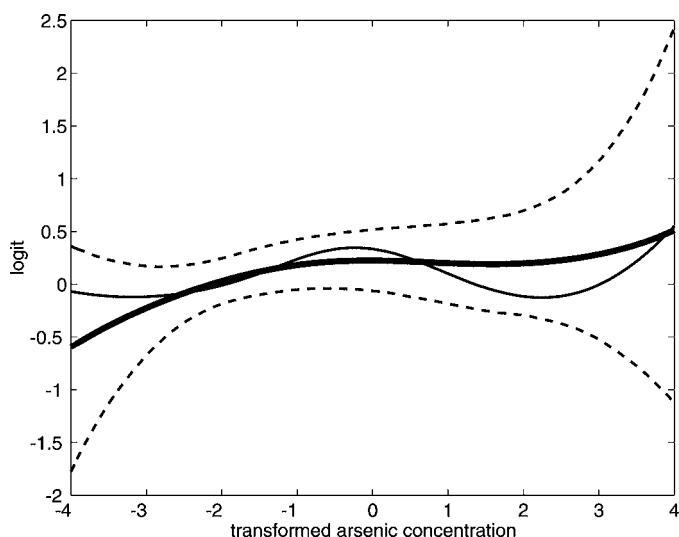


Figure 5. Logit of the Probability of Basal Cell Cancer as a Function of X , the Transformed Value of the Arsenic Concentration in Drinking Water (— uniform Bayes fit; — Bayes fit; — naive).

$$Y = m_{\text{poly}}(D, \beta) + \epsilon, \quad W = X + U, \quad D = \alpha_0 + \alpha_1 X + \nu, \quad \text{and} \\ S = D + \xi.$$

5.2 The OPEN Study

The OPEN (observing protein and energy intake) Study (Subar et al. 2003; Kipnis et al. 2003a) is the first large study using biomarkers to investigate the properties of self-reporting instruments for measuring nutrient intakes. We first briefly describe the study and its rationale; full details are given in the two references.

Much of the recent literature on the relationship between diet and cancer has been based on analytic epidemiological studies using food frequency questionnaires (FFQs). A number of large prospective studies of this kind have failed to find a consistent relationship between dietary components (such as fat, fiber, fruits, and vegetables) and cancers of the breast, colon, or rectum. This may be explained by a true lack of diet–cancer associations or, alternatively, by serious methodologic limitations of the studies themselves, especially due to FFQ measurement error.

Because the reported nutrient intake values from FFQs are subject to substantial measurement error, both systematic and random, there is considerable interest in the properties of this instrument. In the OPEN study, this was addressed by measuring two biomarkers for nutrient intake, doubly labelled water (DLW) for energy (caloric) intake, and urinary nitrogen (UN) for protein intake. The biomarkers and a FFQ were observed in 484 healthy volunteers. We use a subset of these data. The first FFQ was used to measure self-reported intake of protein and energy. We take the problem of interest to be relating the true protein intake (X) and the reported protein intake via the FFQ (Y). The instrument that we use is $S = \text{DLW}$ as a measure of energy intake, which is closely related to protein intake. The measure W of protein intake is the protein biomarker. In all cases, nutrient values were in the logarithmic scale.

In the OPEN study, the protein biomarker was replicated, but very closely in time, namely twice in a 2-week period. Kipnis et al. (2003a) then fit a linear measurement error model to the data, and used the replicates to obtain the estimate $\hat{\sigma}_u^2 \approx .03$. One can argue that Kipnis et al. failed to correctly estimate within-person variability because the short time period between repeats might lead to correlation in the measurement errors, and hence possibly a large underestimate of the measurement error variance in the protein biomarker. Such an argument was made by Willett (2003) in a slightly different context.

Our methods can be used to partially answer this issue. We take the covariate measured without error Z to be patient age, transformed linearly to $[-1, 1]$. We use the model given by (12), (13), and (15). After some data analysis, it seemed reasonable to take $m(\cdot)$ in (15) to be linear in X . Spline fits indicated that especially $g(\cdot)$ in (15) and also perhaps $\alpha_0(\cdot)$ in (13) were nonlinear in Z (Fig. 6). Thus the model is $W = X + U$ and $S = \alpha_0(Z) + \alpha_1(Z)X + \nu$. A specific form for $E(Y|X, Z)$ is not needed to estimate $\hat{\sigma}_u^2$, but for other purposes, $Y = g(Z) + X\beta + \epsilon$ could be used.

We used an Epanechnikov kernel. For every kernel regression, we used the EBBS method (Ruppert 1997) to estimate the bandwidth locally. We then averaged these estimated bandwidths, and then achieved the undersmoothing necessary in the

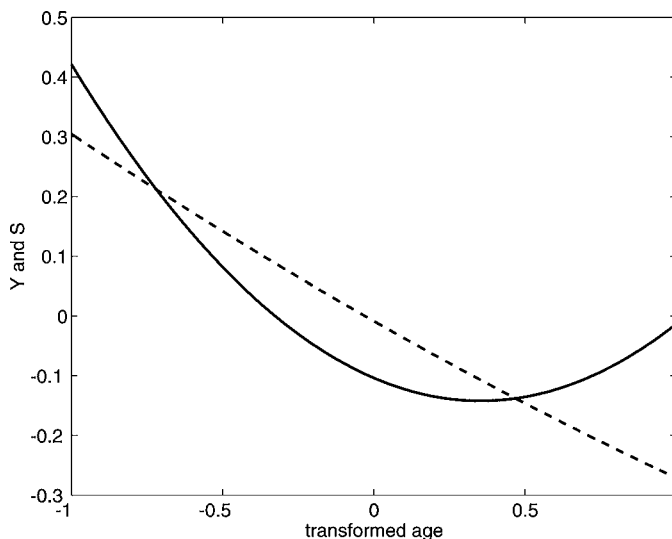


Figure 6. OPEN Study Data: P-Spline Fits of the Protein FFQ (Y) (—) and the Energy Biomarker (S) (---), Both Standardized to Have Mean 0 and Variance 1, versus Age Linearly Transformed to $[-1, 1]$ (Z).

theory by multiplying the result by $n^{-2/15}$. The resulting estimate of $\alpha_1(\cdot)$ is displayed in Figure 7, which contrasts the theoretically important undersmoothed estimator with a regression spline estimate of $\alpha_1(\cdot)$ that does not undersmooth. There appears to be some evidence showing that $\alpha_1(\cdot)$ is not constant.

Figure 8 shows the results from bootstrapping (500 resamples) the estimates of $\hat{\sigma}_u^2$ using three methods:

- The method that ignores $Z = \text{age}$ entirely, so that $g(z) = \beta_0$, $\alpha_0(z) = \alpha_0$, and $\alpha_1(z) = \alpha_1$. For this method, $\hat{\sigma}_{u, \text{no } z}^2 = .019$, with bootstrap mean .022 and bootstrap standard error .005. This estimate is much smaller than that from the OPEN study using replicates ($\hat{\sigma}_{u, \text{OPEN}}^2 = .030$).
- The method that uses $Z = \text{age}$, but assumes that $\alpha_1(\cdot)$ is constant. Now $\hat{\sigma}_{u, \text{constant } \alpha_1}^2 = .023$, with bootstrap mean

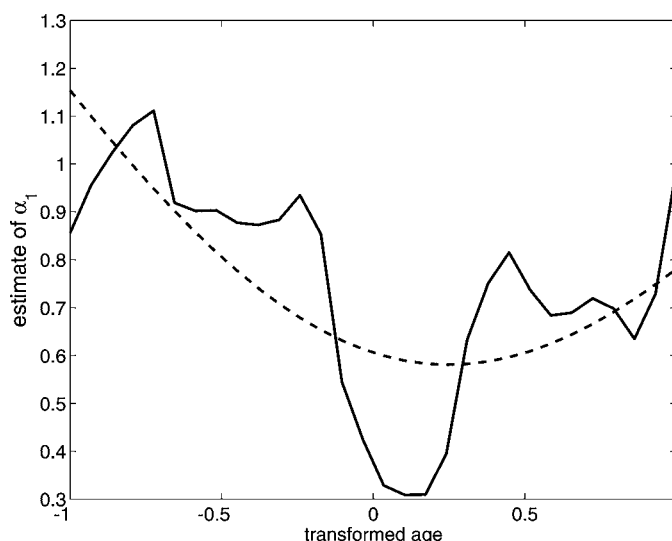


Figure 7. OPEN Study Data: The Undersmoothed Kernel Estimate of $\alpha_1(\cdot)$ (—) and the Same Estimate Using a P-Spline Without Undersmoothing (---). On the X-axis is age linearly transformed to $[-1, 1]$.

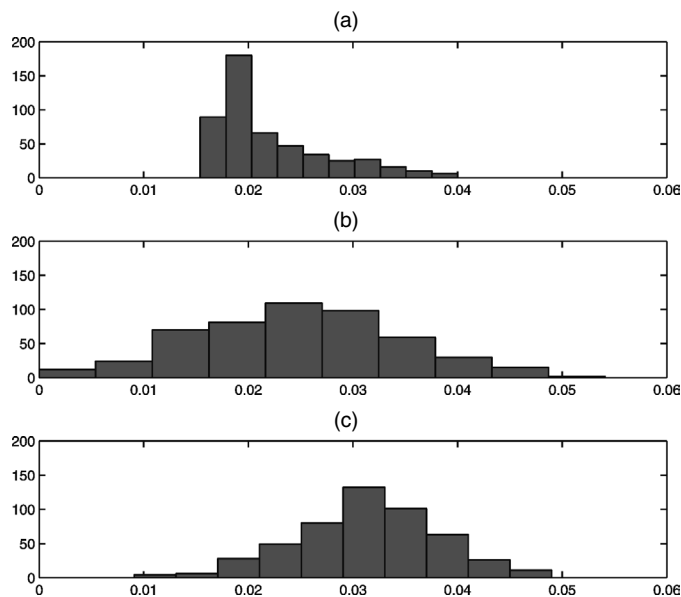


Figure 8. OPEN Study Data: Histogram of 500 Bootstrap Estimates of σ_u^2 . (a) Age not considered. (b) Age considered, but α_1 does not vary with age. (c) α_1 varies with age.

.025 and bootstrap standard error .010. This estimate is also much smaller than that from the OPEN study.

- The method that allows $\alpha_1(\cdot)$ to vary with $Z = \text{age}$, as suggested by Figure 7. With this method, $\hat{\sigma}_{u, \text{varying } \alpha_1}^2 = .029$, with bootstrap mean .031 and bootstrap standard error .007. This method almost exactly reproduces the result ($\hat{\sigma}_{u, \text{OPEN}}^2 = .030$) from the OPEN study using the replicates.

Of course, a short interval between repeats *might* downwardly bias an estimate of σ_u^2 , but the IV analysis suggests that in this instance the bias was not appreciable. Also, if we had not used the varying-coefficient model for S , we would have gotten the counterintuitive result that the OPEN estimate of σ_u^2 is *upwardly* biased, implying a *negative* correlation between the replicates.

6. SOME ASYMPTOTICS FOR POLYNOMIAL REGRESSION

The trade-off between bias and variance is familiar in non-parametric regression but is less well known in measurement error modeling, where the effect is even more profound. Measurement error leads to bias, often as attenuation (shrinkage) toward 0. To remove bias, one “unshrinks” the estimator, increasing variability. Thus the naive estimator is biased but less variable than estimators correcting bias.

To understand the asymptotic behavior of the estimates, we performed some exact calculations. For each of the three cases in Table 1 and with case 4, $m(x) = x^2$, we fit the polynomial that best captures the function on the interval $\mu_x \pm 3(\sigma_x^2 + \sigma_u^2)^{1/2}$; that is, we set up a grid on this interval and fit polynomials to the function $y = m(x)$ on the grid. The degrees of the polynomials chosen were 7, 7, 7, and 5 for cases 1–4. The polynomial functions are shown in Figure 9 on the interval $\mu_x \pm 2\sigma_x$. Although not perfect representations of cases 1–4,

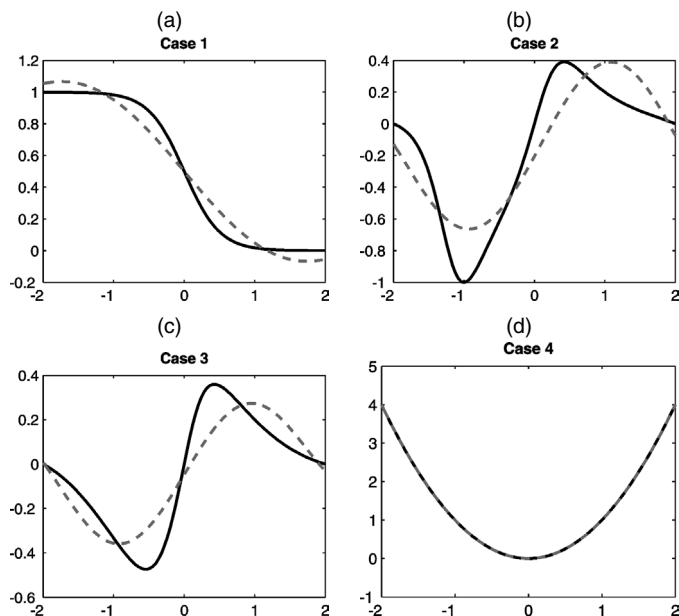


Figure 9. The “Actual” Functions (—) From Cases (a) 1, (b) 2, (c) 3, and (d) 4 for the Gaussian Simulation and Their “True” Polynomial Approximations (---) Used in Computing Theoretical Asymptotic Distributions.

they are sufficiently close to yield some insight. In what follows, these polynomials are treated as the true regression functions.

With X and U normally distributed, define $\sigma_{x|w}^2 = \text{var}(X|W) = \lambda\sigma_u^2$. Recall that if Z has a standard normal distribution, then $E(Z^{2r}) = (2r)!/(2^r r!)$. Then, if the true polynomial is $m(x, \beta) = \sum_{k=1}^d \beta_k x^k$, the observed data have the regression function $E(Y|W) = \sum_{j=0}^d \beta_{j, \text{naive}} W^j$, where $\beta_{j, \text{naive}} = \sum_{k=j}^d \beta_k k! \lambda^j \sigma_{x|w}^{k-j} E(Z^{k-j}) / \{j!(k-j)!\}$. If the true regression function is $m(x, \beta)$, then the naive estimator of β converges to β_{naive} , the minimizer of $E\{[m(X, \beta) - m(W, \beta_{\text{naive}})]^2\}$. Once the expectation is computed as a function of β_{naive} , it can be minimized by any standard minimizer. Using the notation $h(z) = \mu_x + \sqrt{2}\sigma_x z$, the expectation is given as

$$\begin{aligned} & (\sigma_x^2 \sigma_u^2)^{-1/2} \int \{m(x, \beta) - m(x + u, \beta_{\text{naive}})\}^2 \\ & \quad \times \phi\{(x - \mu_x)/\sigma_x\} \phi(u/\sigma_u) dx du \\ & = (\pi \sigma_u^2)^{-1/2} \int [m\{h(z), \beta\} - m\{h(z) + u, \beta_{\text{naive}}\}]^2 \\ & \quad \times \exp(-z^2) \phi(u/\sigma_u) dz du \\ & \quad \times \pi^{-1/2} E \left(\int [m\{h(z), \beta\} - m\{h(z) + U, \beta_{\text{naive}}\}]^2 \right. \\ & \quad \quad \left. \times \exp(-z^2) dz \right), \end{aligned}$$

where the expectation is over the distribution of U . The integral can be computed via Gaussian quadrature and the expectation via simulation.

Let $m^{(1)}(\cdot)$ be the derivative of $m(\cdot)$ with respect to β . Because $m(\cdot)$ is linear in β , this derivative does not involve β . In a sample of size n , the naive estimator is the solution to the

equation $\sum_{i=1}^n m^{(1)}(W_i)\{Y_i - m(W_i, \beta^*)\}$. Asymptotically, its variance is $n^{-1} \mathbf{B}_{\text{naive}}^{-1} \mathbf{A}_{\text{naive}} \mathbf{B}_{\text{naive}}^{-1}$, where

$$\mathbf{B}_{\text{naive}} = E[m^{(1)}(W)\{m^{(1)}(W)\}^T]$$

and

$$\mathbf{A}_{\text{naive}} = E\left[\left(\sigma_\epsilon^2 + \{m(X, \beta) - m(W, \beta_{\text{naive}})\}^2\right) \times m^{(1)}(W)\{m^{(1)}(W)\}^T\right].$$

Both $\mathbf{A}_{\text{naive}}$ and $\mathbf{B}_{\text{naive}}$ can be computed either directly by simulation or by a combination of simulation and Gaussian quadrature.

Under a parametric model, the Bayes estimator will be asymptotically equivalent to the maximum likelihood estimator, and hence it will be asymptotically consistent and its variance is $n^{-1} \mathcal{I}^{-1}$, where \mathcal{I} is the information matrix for β and the measurement error model parameters are $\mu_x, \alpha_0, \alpha_1, \sigma_x^2, \sigma_u^2, \sigma_\epsilon^2$, and σ_v^2 . Again, a combination of Gaussian quadrature and simulation is used. By simple calculations, again using the notation $h(z) = \mu_x + \sqrt{2}\sigma_x z$, the likelihood is

$$(\sigma_\epsilon^2 \sigma_u^2 \sigma_v^2 \pi)^{-1/2} \int \phi\left\{\frac{Y - m\{h(z), \beta\}}{\sigma_\epsilon}\right\} \phi\left(\frac{W - h(z)}{\sigma_u}\right) \times \phi\left\{\frac{S - \alpha_0 - \alpha_1 h(z)}{\sigma_v}\right\} \exp(-z^2) dz.$$

We computed the score $\mathcal{L}(Y, W, S, \cdot)$, the derivative of the log-likelihood, via numerical differentiation. The information, $\mathcal{I} = E\{\mathcal{L}(Y, W, S, \cdot)\mathcal{L}^T(Y, W, S, \cdot)\}$ can be computed by simulation.

On a grid of values x_i for $i = 1, \dots, n_{\text{grid}}$, the mean squared bias of the naive estimator is $n_{\text{grid}}^{-1} \sum_{i=1}^{n_{\text{grid}}} \{m(x_i, \beta) - m_{\text{naive}}(x_i, \beta_{\text{naive}})\}^2$. Because $m(\cdot)$ and $m_{\text{naive}}(\cdot)$ are linear in β , the average variance based on a sample of size n is $n^{-1} \text{trace}(\mathbf{B}_{\text{naive}}^{-1} \mathbf{A}_{\text{naive}} \mathbf{B}_{\text{naive}}^{-1} \mathbf{C})$, where $\mathbf{C} = n_{\text{grid}}^{-1} \times \sum_{i=1}^{n_{\text{grid}}} m^{(1)}(x_i)\{m^{(1)}(x_i)\}^T$.

The sample $n = 100$ was substituted into these asymptotic formulas, thus ignoring any small-sample bias in the maximum likelihood estimator; the results are given in Table 2. The message from this table is the same that we have noted previously; for such small sample sizes, the excess variance of the (asymptotically) best parametric method for correcting bias due to measurement error makes the naive approach at least comparable in terms of MSE. This (asymptotic) fact of life shows up somewhat in our simulations, although we have noted that

Table 2. Asymptotic Calculations for Polynomial Approximations to Four Functions in the Gaussian Case

Case	Naive 100 × RASB	Naive 100 × RMSE	MLE 100 × RMSE	Variance ratio: MLE to naive
1	9.82	12.49	13.43	3.03
2	14.09	16.43	14.70	3.03
3	9.33	11.92	14.74	3.95
4	63.42	68.65	35.44	1.82

NOTE: Here RASB means 100 times the square root of the average squared bias, whereas RMSE is 100 times the square root of the MSE. In these calculations it was assumed that the sample size was $n = 100$, and that the MLE had no small-sample bias. In case 1 the target regression function is $1/(1 + \exp(4x))$. In case 2 the target regression function is $\sin(\pi x/2)/(1 + [2x^2\{1 + \sin(\pi x/2)\}])$. In case 3 the target regression function is $\sin(\pi x/2)/(1 + [2x^2\{1 + \text{sign}(x)\}])$. In case 4 the regression function is x^2 .

the Bayesian methods actually perform quite a bit better than our asymptotics would suggest. Note that the values in Tables 1 and 2 are not identical, because the latter uses asymptotics, different functions, and different estimation methods. The qualitative message is, however, the same; for a relatively small sample size (e.g., $n = 100$), one must accept either large bias or large variance.

As n increases, variances decrease but the bias of the naive estimator is unchanged, and eventually bias dominates. Estimators that remove bias efficiently (e.g., the Bayesian estimator) are rather accurate for reasonably large sample sizes (see Fig. 3).

The topic of “weak instruments” in the econometric literature may be of relevance to the use of IV methods for measurement error corrections. Stock, Wright, and Yogo (2002) have reviewed recent developments in this field. For the most part, the interest has been focused on improved asymptotic properties for IV estimation in the linear model setting, but some of the methods have been extended to nonlinear models using generalized methods of moments (Hansen and Singleton 1984). Further work in this area may usefully extend these results to the nonparametric spline models used in our application.

7. SUMMARY

Our main theoretical contribution is to show that all parameters are identified for a rather general class of models relating the response to the covariates when the linear regression of the instrument on the error-prone covariate has coefficients that are smooth functions of the error-free covariate. We exhibit a root- n -consistent estimate of the measurement error variance. These results extend the applicability of IV estimation to many interesting examples.

APPENDIX: PROOFS AND TECHNICAL DETAILS

A.1 Proof of Theorem 1

Let M be the smallest positive integer k such that $\text{cov}[m(X), \{X - E(X)\}^k]$ is not 0. By (7), ϵ, X , and v are mutually independent, so

$$\begin{aligned} & \text{cov}[Y, \{S - E(S)\}^M] \\ &= \text{cov}[m(X), \{\alpha_1(X - E(X)) + v\}^M] \\ &= \sum_{j=0}^M \binom{M}{j} \alpha_1^j \text{cov}[m(X), \{X - E(X)\}^j v^{M-j}] \\ &= \alpha_1^M \text{cov}[m(x), \{X - E(X)\}^M], \end{aligned} \tag{A.1}$$

because $\text{cov}[m(X), \{X - E(X)\}^j v^{M-j}] = \text{cov}[m(X), \{X - E(X)\}^j] E(v^{M-j})$, and by the definition of M , we have $\text{cov}[m(X), \{X - E(X)\}^j] = 0$ for $1 \leq j < M$.

Similarly, ϵ, X , and U are independent, so that

$$\text{cov}[Y, \{W - E(W)\}^M] = \text{cov}[m(x), \{X - E(X)\}^M]. \tag{A.2}$$

Then, by (A.1) and (A.2),

$$\alpha_1^M = \frac{\text{cov}[Y, \{S - E(S)\}^M]}{\text{cov}[Y, \{W - E(W)\}^M]}. \tag{A.3}$$

If M is odd, then (A.3) determines α_1 from moments of observables. If M is even, then α_1 is only determined up to its sign by (A.3), but then its sign can be determined by the relation $\text{cov}(W, S) = \alpha_1 \sigma_x^2$ and the assumption that $\sigma_x^2 > 0$.

A.2 Kernel Estimators Satisfying Assumption 5

Let (Z_i, V_i) be iid with joint density $f(z, v)$. Suppose that the support of Z_i is $[0, 1]$, $f(z)$ is the marginal density of \mathbf{Z} , and $\inf_{z \in [0,1]} f(z) > 0$. Define $g(z) = E(V_i|Z_i = z)$ and $r(z) = \int v f(z, v) dv = f(z)g(z)$. As before, $K_h(z) = h^{-1}K(h^{-1}z)$ and $C_1 n^{-1/3} \leq h \leq C_2 n^{-1/3}$. Let $\hat{f}(z) = n^{-1} \sum_{i=1}^n K_h(Z_i - z)$, $\hat{r}(z) = n^{-1} \sum_{i=1}^n K_h(Z_i - z)V_i$, and $\hat{g}(z) = \hat{r}(z)/\hat{f}(z)$. Define $\epsilon_i(z) = \{V_i - E(V_i|Z_i = z)\}$ and $D_{i,h}(z) = \{n f_{\mathbf{Z}}(z)\}^{-1} K_h(Z_i - z)$. Then, following Marron and Härdle (1986, p. 99),

$$\begin{aligned} \hat{g}(z) - g(z) &= \frac{\hat{r}(z) - \hat{f}(z)g(z)}{\hat{f}(z)} \\ &= \left\{ 1 + \frac{f(z) - \hat{f}(z)}{f(z)} \right\} \sum_{i=1}^n D_{i,h}(z) \epsilon_i(z). \end{aligned}$$

Under standard regularity conditions (see, e.g., Mack and Silverman 1982, thm. B), uniformly on compact interior subsets of $[0, 1]$, we have $\{f(z) - \hat{f}(z)\}/f(z) = O_P\{n^{-1/3} \log(n)^{1/2}\}$ and $\sum_{i=1}^n D_{i,h}(z) \times \epsilon_i(z) = O_P\{n^{-1/3} \log(n)^{1/2}\}$, so that $\hat{g}(z) - g(z) = \sum_{i=1}^n D_{i,h}(z) \times \epsilon_i(z) + O_P\{n^{-2/3} \log(n)\}$. Let $\epsilon_i = \epsilon_i(Z_i)$, so that $\epsilon_i(z) - \epsilon_i = g(z) - g(Z_i)$. By standard calculations, $\sum_{i=1}^n D_{i,h}(z) \{\epsilon_i(z) - \epsilon_i\} = O_P(n^{-2/3})$, so that $\hat{g}(z) - g(z) = \sum_{i=1}^n D_{i,h}(z) \epsilon_i + O_P\{n^{-2/3} \times \log(n)\}$. Also, $E(\epsilon_i|Z_i) = E\{V_i - E(V_i|Z_i)|Z_i\} = 0$.

A.3 Proof of Theorem 3

The functions $\eta_{\ell w}(\cdot)$, $\eta_{\ell s}(\cdot)$, $\gamma_w(\cdot)$, $\gamma_s(\cdot)$, and $\xi_{ws}(\cdot)$ can all be expressed smoothly in terms of the functions $m_{jkp}(\cdot)$. Define $D_{i,h}(z) = \{n f_{\mathbf{Z}}(z)\}^{-1} K_h(Z_i - z)$. In what follows, we use the term ‘‘uniformly’’ to mean uniformly on compact interior subsets of $[0, 1]$. In addition, any random variables with ϵ ’s in the arguments are conditionally mean 0 with uniformly bounded variances given Z_i . By the delta method using Assumption 5, uniformly to order $o_P(n^{-1/2})$, we have that $\hat{\eta}_{\ell w}(z) = \eta_{\ell w}(z) + \sum_{i=1}^n D_{i,h}(z) \epsilon_{\eta_{\ell w} i}$, $\hat{\eta}_{\ell s}(z) = \eta_{\ell s}(z) + \sum_{i=1}^n D_{i,h}(z) \epsilon_{\eta_{\ell s} i} + o_P(n^{-1/2})$, $\hat{\gamma}_w(z) = \gamma_w(z) + \sum_{i=1}^n D_{i,h}(z) \epsilon_{\gamma_w i}$, $\hat{\gamma}_s(z) = \gamma_s(z) + \sum_{i=1}^n D_{i,h}(z) \epsilon_{\gamma_s i}$, and $\hat{\xi}_{ws}(z) = \xi_{ws}(z) + \sum_{i=1}^n D_{i,h}(z) \epsilon_{\xi_{ws} i}$. Our estimate of $\text{sign}(\alpha_1)$ is therefore root- n consistent.

We see that uniformly to order $o_P(n^{-1/2})$, $\hat{\eta}_{\ell s}(z) \hat{\eta}_{\ell w}(z) = \alpha_1^\ell + \sum_{i=1}^n D_{i,h}(z) \epsilon_{\eta_{\ell s} w i}$ and

$$\begin{aligned} n^{1/2} \hat{p}_{ab}(\hat{\alpha}_1^\ell - \alpha_1^\ell) &= n^{-1/2} \sum_{i=1}^n I(a \leq Z_i \leq b) \sum_{j=1}^n D_{i,h}(Z_j) \epsilon_{\eta_{\ell s} w j} \\ &= n^{-1/2} \sum_{i=1}^n \epsilon_{\eta_{\ell s} w i} n^{-1} \sum_{j=1}^n I(a \leq Z_j \leq b) D_{i,h}(Z_j). \end{aligned} \quad (A.4)$$

Define $G_n(z) = n^{-1} \sum_{j=1}^n I(a \leq z \leq b) D_{i,h}(z)$. Conditional on the Z ’s, the variance of the last expression in (A.4) is uniformly bounded by a constant times $n^{-1} \sum_{i=1}^n G_n^2(Z_i)$. Because $K(\cdot)$ has compact support, for large enough n there exists $0 < a_* < a < b < b_* < 1$ such that $G_n(z) = 0$ if $z \leq a_*$ or if $z \geq b_*$. In addition, $G_n(z)$ is uniformly converging on all compact subsets of $[0, 1]$, and hence in particular on $[a_*, b_*]$. Hence, $G_n(\cdot)$ is uniformly bounded with probability approaching 1, and thus we have shown that $n^{1/2}(\hat{\alpha}_1^\ell - \alpha_1^\ell) = O_P(1)$.

We note in passing that because $G_n(z)$ converges pointwise to $I(a \leq z \leq b)$ for $a < z < b$ to terms of order $o_P(n^{-1/2})$, $n^{1/2} \hat{p}_{ab}(\hat{\alpha}_1^\ell - \alpha_1^\ell) = n^{-1/2} \sum_{i=1}^n \epsilon_{\eta_{\ell s} w i} I(a \leq Z_i \leq b) + o_P(1)$, and

hence if $p_{ab} = \Pr(a \leq Z \leq b)$, then

$$n^{1/2}(\hat{\alpha}_1 - \alpha_1) = \frac{\alpha_1^{1-\ell}}{\ell p_{ab}} n^{-1/2} \sum_{i=1}^n \epsilon_{\eta_{\ell s} w i} I(a \leq Z_i \leq b) + o_P(1). \quad (A.5)$$

The same argument shows that $n^{1/2}(\hat{\sigma}_u^2 - \sigma_u^2) = O_P(1)$. Specifically, to order $o_P(n^{-1/2})$, we have

$$\begin{aligned} \hat{\gamma}_w(z) - \frac{\hat{\xi}_{ws}(z)}{\hat{\alpha}_1} &= \hat{\gamma}_w(z) - \frac{\hat{\xi}_{ws}(z)}{\alpha_1} + \xi_{ws}(z) \alpha_1^{-2} (\hat{\alpha}_1 - \alpha_1) \\ &= \sigma_u^2 + \sum_{i=1}^n D_{i,h}(z) \left(\epsilon_{\gamma_w i} - \frac{\epsilon_{\xi_{ws} i}}{\alpha_1} \right) + \frac{\xi_{ws}(z)}{\alpha_1^2} (\hat{\alpha}_1 - \alpha_1). \end{aligned} \quad (A.6)$$

Substituting (A.6) into the definition of $\hat{\sigma}_u^2$, we have

$$\begin{aligned} n^{1/2}(\hat{\sigma}_u^2 - \sigma_u^2) &= (n^{1/2} p_{ab})^{-1} \sum_{i,j=1}^n I(a \leq Z_i \leq b) D_{i,h}(Z_j) (\epsilon_{\gamma_w j} - \epsilon_{\xi_{ws} j} / \alpha_1) \\ &\quad + (n p_{ab} \alpha_1^2)^{-1} \sum_{i=1}^n \xi_{ws}(Z_i) I(a \leq Z_i \leq b) n^{1/2} (\hat{\alpha}_1 - \alpha_1) \\ &\quad + o_P(n^{-1/2}). \end{aligned}$$

Interchanging the summation in the first line, defining $d_{\gamma ab} = E\{I(a \leq Z \leq b) \gamma_s(Z)\}$, and substituting (A.5), we find that

$$\begin{aligned} n^{1/2}(\hat{\sigma}_u^2 - \sigma_u^2) &= (n^{1/2} p_{ab})^{-1} \sum_{i=1}^n I(a \leq Z_i \leq b) \left\{ \epsilon_{\gamma_w i} - \frac{\epsilon_{\xi_{ws} i}}{\alpha_1} + \frac{d_{\gamma ab} \epsilon_{\eta_{\ell s} w i}}{\alpha_1^{1+\ell}} \right\} \\ &\quad + o_P(1). \end{aligned}$$

A.4 Proof of Theorem 4

The proof is a minor modification of what we did before. Uniformly to order $o_P(n^{-1/2})$, we have $\hat{\alpha}_1^\ell(z) - \alpha_1^\ell(z) = \hat{\eta}_{\ell s}(z)/\hat{\eta}_{\ell w}(z) - \alpha_1^\ell(z) = \sum_{i=1}^n D_{i,h}(z) \epsilon_{\eta_{\ell s} w i}$, so that $\hat{\alpha}_1(z) - \alpha_1(z) = \{\alpha_1^{1-\ell}(z)/\ell\} \times \sum_{i=1}^n D_{i,h}(z) \epsilon_{\eta_{\ell s} w i} + o_P(n^{-1/2})$ and, by a Taylor series,

$$\begin{aligned} \hat{\gamma}_w(z) - \frac{\hat{\xi}_{ws}(z)}{\hat{\alpha}_1(z)} - \sigma_u^2 &= \hat{\gamma}_w(z) - \gamma_w(z) - \frac{\hat{\xi}_{ws}(z) - \xi_{ws}(z)}{\alpha_1(z)} + \frac{\xi_{ws}(z)}{\alpha_1^2(z)} \{\hat{\alpha}_1(z) - \alpha_1(z)\} \\ &= \sum_{i=1}^n D_{i,h}(z) \left\{ \epsilon_{\gamma_w i} - \frac{\epsilon_{\xi_{ws} i}}{\alpha_1(z)} + \frac{\xi_{ws}(z) \epsilon_{\eta_{\ell s} w i}}{\alpha_1^{1+\ell}(z) \ell} \right\}, \end{aligned}$$

so that

$$\begin{aligned} n^{1/2}(\hat{\sigma}_u^2 - \sigma_u^2) &= (n^{1/2} p_{ab})^{-1} \\ &\quad \times \sum_{i=1}^n I(a \leq Z_i \leq b) \\ &\quad \times \left[\epsilon_{\gamma_w i} - \frac{\epsilon_{\xi_{ws} i}}{\alpha_1(Z_i)} + \epsilon_{\eta_{\ell s} w i} \left\{ \xi_{ws}(Z_i) \alpha_1^{-(1+\ell)}(Z_i) / \ell \right\} \right] \\ &\quad + o_P(1). \end{aligned}$$

A.5 On Condition (8)

We now prove the following result showing that if X is compactly supported and m is continuous, then (8) holds unless $m(\cdot)$ is constant. The condition that X is compactly supported cannot be removed. A counterexample can be constructed using counterexample 1 of Durrett (1996, p. 107). That counterexample demonstrates that there are densities distinct from the lognormal density but with the same moments as the lognormal. If f_X is the density of X and if $m \cdot f_X$ is the difference between two distinct densities with the same moments, then clearly $E[m(X)\{X - E(X)\}^k] = 0$ for all k .

Theorem A.1. Suppose that the support of X is contained in a compact interval $[a, b]$ and that $m(\cdot)$ is continuous on $[a, b]$. If

$$\text{cov}[m(X), \{X - E(X)\}^k] = 0 \quad \text{for all } k, \quad (\text{A.7})$$

then $\text{var}\{m(X)\} = 0$, so that $P[m(X) = E\{m(X)\}] = 1$.

Proof. By the Weierstrass approximation theorem, for all $\delta > 0$ there exists a polynomial $m_{\text{poly}}(\cdot)$ such that $|m(x) - E\{m(X)\} - m_{\text{poly}}(x)| < \delta$ for all $x \in [a, b]$. By (A.7), $m(X)$ and $m_{\text{poly}}(X)$ have covariance 0, so that $\text{var}\{m(X)\} = E[m(X) - E\{m(X)\} - m_{\text{poly}}(X)]^2 - E[m_{\text{poly}}(X)]^2 \leq \delta^2(b - a) - E[m_{\text{poly}}(X)]^2 \leq \delta^2(b - a)$. Because $\delta > 0$ is arbitrary, the result follows.

A.6 Markov Chain Monte Carlo Calculations in the Gaussian Case

In the Gaussian case, $m(x) = \mathbf{C}_1(x)\boldsymbol{\beta}_1 + \mathbf{C}_2(x)\boldsymbol{\beta}_2$, where $\boldsymbol{\beta}_2 \sim \mathbf{N}(\mathbf{0}, \sigma^2 \mathbf{D})$ and \mathbf{D} is a $k \times k$ matrix. For the regression spline, \mathbf{D} was chosen as the identity matrix. Priors for $\alpha_0, \alpha_1, \mu_x$ and $\boldsymbol{\beta}_1$ were independent normals with mean 0 and (large) variances $\sigma_\alpha^2, \sigma_\alpha^2, \sigma_\mu^2$, and $\sigma_\beta^2 \mathbf{I}$. The prior for the attenuation λ was beta on the interval $[\lambda_L, \lambda_H]$; $[\lambda] \propto (\lambda - \lambda_L)^{\Delta_1 - 1} (\lambda_H - \lambda)^{\Delta_2 - 1}$. Of course, by simple algebra, $\sigma_x^2 = \lambda \sigma_u^2 / (1 - \lambda)$. Priors for $\sigma_\epsilon^2, \sigma_u^2, \sigma_v^2$, and σ^2 were inverse gamma with parameters $(a_\epsilon, b_\epsilon), (a_u, b_u), (a_v, b_v)$, and (a_σ, b_σ) , where the $\text{IG}(A, B)$ density is given by $\{\Gamma(A)B^A x^{A-1}\}^{-1} \exp\{-1/(Bx)\}$. Let $\mathbf{C}(x) = \{\mathbf{C}_1^T(x), \mathbf{C}_2^T(x)\}^T$, $\mathbf{H} = \sum_{i=1}^n \mathbf{C}(X_i)Y_i/\sigma_\epsilon^2$, $\mathbf{Q} = \{\sum_{i=1}^n \mathbf{C}(X_i)\mathbf{C}^T(X_i)/\sigma_\epsilon^2 + \text{diag}(\mathbf{I}/\sigma_\beta^2, \mathbf{I}_k/\sigma^2)\}^{-1}$, $\mathcal{D} = \{\sum_{i=1}^n (1, X_i)^T(1, X_i)/\sigma_v^2 + \mathbf{I}_2/\sigma_\alpha^2\}^{-1}$, and $\mathcal{A} = \sum_{i=1}^n (1, X_i)^T S_i/\sigma_v^2$. The joint density of the data and the parameters (i.e., the unnormalized posterior density) is proportional to

$$\begin{aligned} & \exp\left[-\frac{\sum_{i=1}^n \{Y_i - \mathbf{C}_1(X_i)\boldsymbol{\beta}_1 - \mathbf{C}_2(X_i)\boldsymbol{\beta}_2\}^2}{2\sigma_\epsilon^2} - \frac{\sum_{i=1}^n (W_i - X_i)^2}{2\sigma_u^2} - \frac{\sum_{i=1}^n (S_i - \alpha_0 - \alpha_1 X_i)^2}{2\sigma_v^2} \right. \\ & - \frac{\sum_{i=1}^n (1 - \lambda)(X_i - \mu_x)^2}{2\lambda\sigma_u^2} - \frac{\mu_x^2}{2\sigma_\mu^2} - \frac{\alpha_0^2}{2\sigma_\alpha^2} - \frac{\alpha_1^2}{2\sigma_\alpha^2} \\ & \left. - \frac{\boldsymbol{\beta}_1^T \boldsymbol{\beta}_1}{2\sigma_\beta^2} - \frac{\boldsymbol{\beta}_2^T \mathbf{D}^{-1} \boldsymbol{\beta}_2}{2\sigma^2} - \frac{1}{b_\epsilon \sigma_\epsilon^2} - \frac{1}{b_u \sigma_u^2} - \frac{1}{b_v \sigma_v^2} - \frac{1}{b_\sigma^2} \right] \\ & \times (\sigma_\epsilon^2)^{-(a_\epsilon + n/2 + 1)} (\sigma_u^2)^{-(a_u + n + 1)} \\ & \times (\sigma_v^2)^{-(a_v + n/2 + 1)} (\sigma^2)^{-(a_\sigma + k/2 + 1)} \\ & \times (\lambda^{-1} - 1)^{n/2} (\lambda - \lambda_L)^{\Delta_1 - 1} (\lambda_H - \lambda)^{\Delta_2 - 1}. \end{aligned}$$

Let \bar{X} be the arithmetic mean of the X_i 's and let $[W]$ denote the density of W for any random variable W . The complete conditionals are $\mu_x \stackrel{d}{=} \mathbf{N}(\bar{X}\{n(1 - \lambda)\sigma_\mu^2\}/\{n(1 - \lambda)\sigma_\mu^2 + \lambda\sigma_u^2\}, \{\lambda\sigma_u^2\sigma_\mu^2\}/\{\lambda\sigma_u^2 + n(1 - \lambda)\sigma_\mu^2\})$, $\sigma_u^2 \stackrel{d}{=} \text{IG}(a_u + n, [1/b_u + (1/2)\{(1 - \lambda)/\lambda\} \sum_{i=1}^n (X_i - \mu_x)^2 + (1/2)\sum_{i=1}^n (W_i - X_i)^2]^{-1})$, $\sigma_\epsilon^2 \stackrel{d}{=} \text{IG}(a_\epsilon + n/2, [1/b_\epsilon +$

$(1/2)\sum_{i=1}^n \{Y_i - \mathbf{C}_1(X_i)\boldsymbol{\beta}_1 - \mathbf{C}_2(X_i)\boldsymbol{\beta}_2\}^2]^{-1})$, $\sigma_v^2 \stackrel{d}{=} \text{IG}(a_v + n/2, [1/b_v + (1/2)\sum_{i=1}^n (S_i - \alpha_0 - \alpha_1 X_i)^2]^{-1})$, $\sigma^2 \stackrel{d}{=} \text{IG}(a_\sigma + k/2, [1/b + (1/2)\boldsymbol{\beta}_2^T \mathbf{D}^{-1} \boldsymbol{\beta}_2]^{-1})$, $(\alpha_0, \alpha_1) \stackrel{d}{=} \mathbf{N}(\mathcal{D}\mathcal{A}, \mathcal{D})$, $(\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T)^T \stackrel{d}{=} \mathbf{N}(\mathbf{QH}, \mathbf{Q})$, and $[X_i] \propto \exp\{-\{Y_i - \mathbf{C}_1(X_i)\boldsymbol{\beta}_1 - \mathbf{C}_2(X_i)\boldsymbol{\beta}_2\}^2/2\sigma_\epsilon^2 - (W_i - X_i)^2/2\sigma_u^2 - (S_i - \alpha_0 - \alpha_1 X_i)^2/2\sigma_v^2 - (1 - \lambda)(X_i - \mu_x)^2/2\lambda\sigma_u^2\}$. In addition,

$$\begin{aligned} & [\lambda] \propto I(\lambda_L \leq \lambda \leq \lambda_H) \\ & \times \left\{ \frac{1 - \lambda}{\lambda} \right\}^{n/2} (\lambda - \lambda_L)^{\Delta_1 - 1} (\lambda_H - \lambda)^{\Delta_2 - 1} \\ & \times \exp\left\{-\frac{\sum_{i=1}^n (1 - \lambda)(X_i - \mu_x)^2}{2\lambda\sigma_u^2}\right\}. \quad (\text{A.8}) \end{aligned}$$

All of the complete conditionals except for λ and the X 's are easily generated. For λ , in our simulations we discretized the set $\lambda \in [\lambda_L, \lambda_H]$ into 41 different values, computed (A.8) for these values, turned the result into probabilities, and sampled λ according to these probabilities. This gridded Gibbs estimator is not strictly correct, of course, but it is convenient and provides good mixing. We also implemented a full Metropolis–Hastings step: mixing was not quite as good, thus requiring somewhat more MCMC samples, but in selective test cases we found that the final fits to the regression function were virtually identical to our gridded method.

For the X 's, the complete conditional is not explicit. We used Metropolis–Hastings steps where the candidate density was normal with the current value of X as the mean and the variance being 1/2 times the conditional variance for X given (W, S) , with the latter variance evaluated at the current parameter values.

In our simulations, the prior distributions were $\sigma_x^2 \stackrel{d}{=} \text{IG}(1, 1)$, $\sigma_v^2 \stackrel{d}{=} \text{IG}(1, 1)$, $\sigma_\epsilon^2 \stackrel{d}{=} \text{IG}(1, 1)$, $\lambda \stackrel{d}{=} \text{U}[.60, 1.00]$, $\mu_x \stackrel{d}{=} \mathbf{N}(0, 100)$, $\alpha_0 \stackrel{d}{=} \mathbf{N}(0, 100)$, $\alpha_1 \stackrel{d}{=} \mathbf{N}(0, 100)$, and $\sigma^2 \stackrel{d}{=} \text{IG}(1, 1000)$. We also used $\sigma^2 \stackrel{d}{=} \text{IG}(.01, 100)$ without appreciable differences in some test cases.

The model can be extended to incorporate possible prior information regarding the means and covariances of the parameters $\mu_x, \alpha_1, \alpha_2, \boldsymbol{\beta}_1$, and $\boldsymbol{\beta}_2$. Because we have no such prior information, we did not do this.

A.7 Markov Chain Monte Carlo Calculations in the Probit Model

We fit a probit regression model, turning it into a logistic fit by the usual device: If the probability is p , then the logit function is $\log\{p/(1 - p)\}$. Note that we are *not* approximating the logit model by a probit model. Rather, our method is exact, because if the logit of $P(Y = 1|X)$ is a smooth function of X , then the probit of $P(Y = 1|X)$ is another smooth function of X .

For the probit model, we modified the method of Albert and Chib (1993). Specifically, one defines latent variable Z_i that is normally distributed with mean $\mathbf{C}_1(X_i)\boldsymbol{\beta}_1 + \mathbf{C}_2(X_i)\boldsymbol{\beta}_2$ and variance 1.0, so that $Y_i = I(Z_i > 0)$. Given the values of Z_i , the MCMC steps of Section A.6 apply without change, with two exceptions: (a) $\sigma_\epsilon^2 = 1$ is known a priori, and (b) Z_i replaces Y_i in that section. This means that the only thing necessary in the MCMC steps is to generate values of the Z_i from their complete conditional distribution. Write $\mu_i = \mathbf{C}_1(X_i)\boldsymbol{\beta}_1 + \mathbf{C}_2(X_i)\boldsymbol{\beta}_2$. The density of Z_i given the rest is $f(Z_i|\text{rest}) \propto \{Y_i I(Z_i > 0) + (1 - Y_i) I(Z_i \leq 0)\} \exp\{-(1/2)(Z_i - \mu_i)^2\}$. This means that if $Y_i = 1$, then Z_i is a truncated normal random variable, that is, a normal random variable with mean μ_i , variance 1.0, with left truncation point 0. Also, if $Y_i = 0$, then Z_i is a normal random variable with mean μ_i , and variance 1.0, with right truncation point 0. Define $R_i = 1 - 2I(Y_i = 1)$, and $\text{TN}(a, b)$ to be a normal random variable with mean a and variance 1.0 with left truncation point b . Then it follows that complete conditional of Z_i is $Z_i \sim \mu_i - R_i \text{TN}_L(0, R_i \mu_i)$.

To generate these truncated normals, we used the accept–reject algorithm of Robert (1995), with the following modification. If we want to generate a normal random variable truncated from the left (right) at 0 and with a positive (negative) mean, we did not use Robert's algorithm but instead generated normals at random until one was positive (negative).

Although the candidate density for X discussed in the Gaussian case (Sec. A.6) worked well enough in that case, we found that it was not nearly so efficient in the probit model. The following gave better mixing and faster convergence of the sampler. Suppose that the current X_i is $X_{\text{curr},i}$, and the latent variable is \mathbf{Z} . Let β_{lin} be the simple linear regression estimate of $\{\mathbf{Z}\}_{i=1}^n$ on $\{X_{\text{curr}}\}_{i=1}^n$. Our candidate density was the density of X given (\mathbf{Z}, W, S) assuming a linear model for \mathbf{Z} and X with coefficients β_{lin} , and assuming the current values of μ_x , σ_x^2 , α_0 , α_1 , σ_u^2 , and σ_v^2 . In all cases investigated, the percentage of mixing for X was over 95%.

[Received March 2002. Revised March 2004.]

REFERENCES

- Albert, J. H., and Chib, S. (1993), "Bayesian Analysis of Binary and Polychotomous Response Data," *Journal of the American Statistical Association*, 88, 669–679.
- Amemiya, Y. (1990), "Two-Stage Instrumental Variable Estimators for the Nonlinear Errors-in-Variables Model," *Journal of Econometrics*, 44, 311–332.
- Azzalini, A. (1985), "A Class of Distributions Which Includes the Normal Ones," *Scandinavian Journal of Statistics*, 12, 171–178.
- Barron, A., Schervish, M. J., and Wasserman, L. (1999), "The Consistency of Posterior Distributions in Nonparametric Problems," *The Annals of Statistics*, 27, 536–561.
- Bernardo, J. M., and Smith, A. F. M. (1994), *Bayesian Theory*, New York: Wiley.
- Berry, S. A., Carroll, R. J., and Ruppert, D. (2002), "Bayesian Smoothing and Regression Splines for Measurement Error Problems," *Journal of the American Statistical Association*, 97, 160–169.
- Buzas, J. S., and Stefanski, L. A. (1996), "Instrumental Variable Estimation in Generalized Linear Measurement Error Models," *Journal of the American Statistical Association*, 91, 999–1006.
- Carroll, R. J., and Hall, P. (1988), "Optimal Rates Convergence for Deconvolving a Density," *Journal of the American Statistical Association*, 83, 1184–1186.
- Carroll, R. J., Küchenhoff, H., Lombard, F., and Stefanski, L. A. (1996), "Asymptotics for the SIMEX Estimator in Structural Measurement Error Models," *Journal of the American Statistical Association*, 91, 242–250.
- Carroll, R. J., Maca, J. D., and Ruppert, D. (1999), "Nonparametric Regression With Errors in Covariates," *Biometrika*, 86, 541–554.
- Carroll, R. J., Ruppert, D., and Stefanski, L. A. (1995), *Measurement Error in Nonlinear Models*, New York: Chapman & Hall.
- Carroll, R. J., and Stefanski, L. A. (1994), "Meta-Analysis, Measurement Error and Corrections for Attenuation," *Statistics in Medicine*, 13, 1265–1282.
- Cheng, C. L., and Schneeweiss, H. (1998), "Polynomial Regression With Errors in the Variables," *Journal of the Royal Statistical Society. Ser. B*, 60, 189–199.
- Cook, J. R., and Stefanski, L. A. (1994), "Simulation-Extrapolation Estimation in Parametric Measurement Error Models," *Journal of the American Statistical Association*, 89, 1314–1328.
- Durrett, R. (1996), *Probability: Theory and Examples* (2nd ed.), Belmont, CA: Duxbury.
- Eilers, P. H. C., and Marx, B. D. (1996), "Flexible Smoothing With B-Splines and Penalties" (with discussion), *Statistical Science*, 11, 89–102.
- Fan, J., and Truong, Y. (1993), "Nonparametric Regression With Errors in Variables," *The Annals of Statistics*, 21, 1900–1925.
- Fuller, W. A. (1987), *Measurement Error Models*, New York: Wiley.
- Hansen, L. P., and Singleton, K. J. (1984), "Generalized Instrumental Variables Estimation of Nonlinear Rational Expectations Models," *Econometrica*, 52, 267–268.
- Hastie, T., and Tibshirani, R. (1990), *Generalized Additive Models*, New York: Chapman & Hall.
- Hausman, J. A., Newey, W. K., Ichimura, H., and Powell, J. L. (1991), "Identification and Estimation of Polynomial Errors-in-Variables Models," *Journal of Econometrics*, 50, 273–295.
- Karagas, M. R., Stukel, T. A., Morris, J. S., Tosteson, T. D., Weiss, J. A., Spencer, S. K., and Greenberg, E. R. (2001), "Skin Cancer Risk in Relation to Toenail Arsenic Concentrations in a U.S. Population-Based Case-Control Study," *American Journal of Epidemiology*, 153, 559–565.
- Karagas, M. R., and Tosteson, T. D. (2002), "Assessment of Cancer Risk and Environmental Levels of Arsenic in New Hampshire," *International Journal of Hygiene and Environmental Health*, 21, 4894–4899.
- Karagas, M. R., Tosteson, T. D., Blum, J., Morris, J. S., Baron, J. A., and Klaue, B. (1998), "Design of an Epidemiologic Study of Drinking Water Arsenic Exposure and Skin and Bladder Cancer Risk in a U.S. Population," *Environmental Health Perspectives*, 106, 1047–1050.
- Kipnis, V., Subar, A. F., Midthune, D., Freedman, L. S., Ballard-Barbash, R., Troiano, R., Bingham, S., Schoeller, D. A., Schatzkin, A., and Carroll, R. J. (2003a), "The Structure of Dietary Measurement Error: Results of the OPEN Biomarker Study," *American Journal of Epidemiology*, 158, 14–21.
- Kipnis, V., Subar, A. F., Schatzkin, A., Midthune, D., Troiano, R., Schoeller, D. A., Bingham, S., and Freedman, L. S. (2003b), "Response to Willett's OPEN Questions," *American Journal of Epidemiology*, 158, 25–26.
- LARC (1987), "Arsenic and Arsenic Compounds (Group 1)," in *Monographs on the Evaluation of Carcinogenic Risk of Chemicals to Humans, Supplement 7*, Lyon: International Agency for Research on Cancer, pp. 100–106.
- Lehmann, E. L. (1983), *Theory of Point Estimation*, New York: Wiley.
- Liang, H. (2000), "Asymptotic Normality of Parametric Part in Partially Linear Models With Measurement Error in the Nonparametric Part," *Journal of Statistical Planning and Inference*, 86, 51–62.
- Mack, Y. P., and Silverman, B. W. (1982), "Weak and Strong Uniform Consistency of Kernel Regression Estimates," *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 61, 405–415.
- Marron, J. S., and Härdle, W. (1986), "Random Approximations to Some Measures of Accuracy in Nonparametric Curve Estimation," *Journal of Multivariate Analysis*, 20, 91–113.
- National Research Council (1999), *Arsenic in Drinking Water*, Washington, DC: National Academy Press.
- Powell, M. J. D. (1981), *Approximation Theory and Methods*, Cambridge, U.K.: Cambridge University Press.
- Robert, C. P. (1995), "Simulation of Truncated Normal Variables," *Statistics and Computing*, 5, 121–125.
- Ruppert, D. (1997), "Empirical-Bias Bandwidths for Local Polynomial Nonparametric Regression and Density Estimation," *Journal of the American Statistical Association*, 92, 1049–1062.
- (2002), "Selecting the Number of Knots for Penalized Splines," *Journal of Computational and Graphical Statistics*, 11, 735–757.
- Ruppert, D., Wand, M. P., and Carroll, R. J. (2003), *Semiparametric Regression*, Cambridge, U.K.: Cambridge University Press.
- Stefanski, L. A., and Buzas, J. S. (1995), "Instrumental Variable Estimation in Binary Regression Measurement Error Variables," *Journal of the American Statistical Association*, 90, 541–550.
- Stefanski, L. A., and Cook, J. R. (1995), "Simulation-Extrapolation: The Measurement Error Jackknife," *Journal of the American Statistical Association*, 90, 1247–1256.
- Stock, J. H., Wright, J. H., and Yogo, M. (2002), "A Survey of Weak Instruments and Weak Identification in Generalized Method of Moments," *Journal of Business & Economic Statistics*, 20, 518–529.
- Subar, A. F., Kipnis, V., and Triano, R. P. (2003), "Using Intake Biomarkers to Evaluate the Extent of Dietary Misreporting in a Large Sample of Adults," *American Journal of Epidemiology*, 158, 1–13.
- Wall, M. M., and Amemiya, Y. (2000), "Estimation in Polynomial Structural Equation Models," *Journal of the American Statistical Association*, 95, 929–940.
- Willett, W. (2003), "Invited Commentary: OPEN Questions," *American Journal of Epidemiology*, 158, 22–24.
- Yalcin, I., and Amemiya, Y. (2001), "Nonlinear Factor Analysis as a Statistical Method," *Statistical Science*, 16, 275–295.