

Cleveland Clinic Department of Quantitative Health Sciences
Technical Report Series *2006.01*

Varying Coefficients Model with Measurement Error

Liang Li

Department of Quantitative Health Sciences

Cleveland Clinic

9500 Euclid Ave, Wb4, Cleveland, OH 44195, U.S.A

E-mail: lil2@ccf.org

Tom Greene

Department of Quantitative Health Sciences

Cleveland Clinic

9500 Euclid Ave, Wb4, Cleveland, OH 44195, U.S.A

E-mail: greenet@ccf.org

August 31, 2006

Abstract

We propose a semi-parametric partially varying coefficient model to study the relationship between serum creatinine concentration and the glomerular filtration rate (GFR) among kidney donors and patients with chronic kidney disease. A regression model is used to relate serum creatinine to GFR and demographic factors in which coefficient of GFR is expressed as a function of age to allow its effect to be age-dependent. GFR measurements obtained from the clearance of a radioactively labeled isotope are assumed to be a surrogate for the true GFR, with the relationship between measured and true GFR expressed using an additive error model. We use local corrected score equations to estimate parameters and coefficient functions, and propose an expected generalized cross-validation (EGCV) method to select the kernel bandwidth. The performance of the proposed methods, which avoid distributional assumptions on the true GFR and residuals, is investigated by simulation. Application of the proposed model to several clinical data sets reveals the presence of a consistent relationship between serum creatinine and true GFR that appears to apply both to healthy kidney donors and to chronic renal disease patients.

Key Words: *varying coefficients model; measurement error; local polynomial; expected generalized cross-validation; local corrected score equations; Glomerular filtration rate*

1 Introduction

1.1 Relationship of Serum Creatinine with the Glomerular Filtration Rate

Accurate determination of renal function is essential in the treatment of kidney disease in order to identify patients with early renal impairment and to follow the course of established disease. The glomerular filtration rate (GFR), a measure of the clearance of waste products by the kidney, is regarded as the best overall index of kidney function. However, direct measurement of GFR is too expensive and logistically difficult for routine clinical practice. Accordingly, renal function is usually evaluated based on the concentration of serum creatinine (SCR), which, in "steady-state" conditions is related to GFR by the expression $SCR = (G - TS)/GFR$, where G and TS represents the rates of creatinine generation and tubular secretion of creatinine, respectively. In an attempt to account for variations in G - TS, a number of multiple regression models have been developed to relate GFR to serum creatinine in conjunction with demographic factors known to be correlated with creatinine generation. Presently, the most widely used is the so-called MDRD Study equation, which was obtained by regression of log GFR on log SCR, log age, gender, and race in a single study including 1628 patients with varying levels of chronic kidney disease (CKD) ([1] and [2]). In the past several years the external generalizability of the MDRD equation and other regression-based estimating equations for GFR have been evaluated in numerous studies (see, for example, [3],[4],[5],[6], and [7]). The results have been mixed, with particularly poor performance reported when the MDRD equation or other equations derived in CKD patients are applied in popula-

tions which differ from the study populations used to derive the equation. Generalizability appears to be especially problematic when the equations are applied to predominantly non-CKD populations.

The development of improved estimates of GFR is a major outstanding issue in the field of nephrology. To address this problem, the Chronic Kidney Disease - Epidemiology (CKD-EPI) consortium has been formed by the National Institutes of Health to jointly analyze approximately 20 research and clinical databases which include measurements of GFR and serum creatinine in an attempt to develop a more generalizable method for estimation of GFR. Borrowing concepts from analytic chemistry, one approach being used by the investigators involves the development of an "indirect regression" model to relate serum creatinine as the response variable to GFR and the demographic factors as predictor variables ([8] and [9]). This model is then inverted to provide estimates of GFR. The appeal of the indirect regression framework is that the designation of serum creatinine as the response variable accords with the direction of the underlying biological process in which creatinine is determined from GFR, rather than conversely. Inverse regression models with serum creatinine as the response may have greater generalizability between studies than direct regressions of GFR on serum creatinine, since in the indirect regression non-renal determinants of serum creatinine are properly incorporated in the residuals of the regression model, whereas in direct regressions with creatinine as a predictor variable these non-renal determinants act as uncontrolled confounding factors. Because assays for measurement of GFR are themselves subject to measurement error, the indirect regression models must be developed under an errors-in-variables framework.

In this paper, we address methodological issues associated with modeling the effect of age and interactions of age with the remaining predictor variables in the indirect regression model. Because creatinine generation, which depends closely on muscle mass, typically starts to decline with age after about 40 years, it is expected that age would modulate the model in a nonlinear fashion. However, the actual functional form for the age effect is not fully known. For this reason, we propose a semi-parametric varying coefficient model, which allows the coefficients of the covariates to depend on age in a general fashion. The varying coefficient model is developed under an errors-in-variables framework to account for measurement error in GFR.

1.2 Varying coefficient measurement error model

We formulate the indirect regression model as follows:

$$Y_i = Z_{1i}\beta_1(T_i) + \dots + Z_{pi}\beta_p(T_i) + \epsilon_i \quad (1)$$

$i = 1, 2, \dots, n$ is the index of study subjects; ϵ_i 's are independent residuals with mean zero. The model has p covariates including an intercept term for which $Z \equiv 1$. This model is distinguished from a standard linear model in that the regression coefficients are not parameters; they are smooth functions of a tuning variable T . In this application, Y is $\log(\text{Scr})$; T is age; $p = 4$; covariates include gender, race, $\log(\text{GFR})$ and the intercept corresponds to the main effect of age. Without loss of generality, we let Z_{pi} be the true value of $\log(\text{GFR})$. The measured $\log(\text{GFR})$ is denoted by X_{pi} . We use the classical additive error model, $X_{pi} = Z_{pi} + U_{pi}$, where U_{pi} is an independent noise with mean zero and variance σ_u^2 . For the purposes of this paper, we assume a constant measurement error variance on the log GFR scale of 0.015. This

value appears to be a reasonable estimate of the typical measurement error in GFR based on a variety of analyses, including assessments of longitudinal variation in log GFR over short time intervals in two of the CKD-EPI studies. Alternative values for σ_u^2 can be considered in sensitivity analyses.

The varying coefficient model (1) is useful for exploring complicated non-linear interactions between covariates while avoiding the curse of dimensionality. The entire model can be viewed as a function of age, which facilitates the study of the hypothesized confounding effect of age. No restrictions, other than smoothness, are placed on the coefficient functions in order to allow for enough flexibility. As a special case, some of the coefficient functions can be constant. This is sometimes called the partially linear varying coefficient model [10]. If a coefficient function is linear in T , then the model includes a main effect for the covariate, and its cross-product interaction with T as commonly used in linear regressions. Since we usually do not know, prior to the analysis, which covariates have constant or varying coefficients, we will start by allowing all the coefficients to be age-dependent. Coefficient functions that do not appear to depend on age will be treated as constant in a second step to obtain more efficient estimators.

1.3 Outline of the paper

Varying coefficient models without measurement error have been studied, for example, by [11] and [12]. The latter paper offers an approach to estimation of constant coefficients which is similar to those we present here, but did not discuss variance and confidence interval estimation. This is studied in this paper. For the case with measurement error, [13] studied a longitudinal

linear model with random intercepts and slopes which are allowed to depend on time. Regression splines are used to approximate the coefficient functions. The error correction algorithm is similar to regression calibration and some distributional assumptions were made regarding the true covariates. This approach is awkward to apply in situations such as the CKD-EPI project, which involves a joint analysis of multiple databases with widely varying and often irregular distributions of GFR, as it would be difficult to model the distribution of true GFR for each data set. This complication motivates the approach used in this paper in which varying coefficients are modeled using the local polynomial method. The advantage is that the error correction is easily handled by a modification of the kernel estimating equations, without requiring distributional assumptions on the true covariate. In addition, we propose an expected generalized cross-validation (EGCV) criterion for bandwidth selection in the presence of measurement error.

Section 2 presents the elements of the proposed approach, including the point estimation, variance estimation, and bandwidth selection. Simulation results are presented in Section 3. We apply the proposed methods to two data sets from the joint CKD-EPI database in Section 4. Implications of this work are discussed in the final section of the paper.

2 Model Fitting

2.1 Estimating the coefficient functions

Let $\mathbf{Z}_i = (Z_{1i}, \dots, Z_{pi})^T$, and $\beta(t) = (\beta_1(t), \dots, \beta_p(t))^T$. Then model (1) becomes

$$Y_i = \mathbf{Z}_i^T \beta(T_i) + \epsilon_i. \quad (2)$$

We assume $\mathbf{X}_i = \mathbf{Z}_i + \mathbf{U}_i$. The error vector \mathbf{U} represents independent noise with mean zero and variance-covariance matrix $\boldsymbol{\Sigma}_u$. If some elements of \mathbf{Z} are measured without error, we define the corresponding elements of \mathbf{U} to be zero and set the related variance components in $\boldsymbol{\Sigma}_u$ to zero. This formulation is more general than (1), and all the statistical procedures below work in this general setting.

If \mathbf{Z} is known, i.e., there is no measurement error, we can fit model (2) using the local linear method [14]. Around any give point t_0 on the support of T , $\beta(T)$ can be approximated by a first-order Taylor series expansion in a neighborhood of t_0 : $\beta(T) = \beta(t_0) + \beta'(t_0)(T - t_0)$, where $\beta'(\cdot)$ is the first derivative of $\beta(\cdot)$. The estimation proceeds by minimizing the following weighted sum of squares with respect to $\theta = (\beta_0^T, \beta_1^T)^T$:

$$\sum_{i=1}^n K_h(T_i - t_0) \left\{ Y_i - \mathbf{Z}_i^T \beta_0 - \mathbf{Z}_i^T (T_i - t_0) \beta_1 \right\}^2 \quad (3)$$

$K_h(t) = K(t/h)/h$ is a kernel function. It is usually chosen to be unimodal and symmetric about zero so that it assigns greater weights to the observations in the neighborhood of t_0 , and lower weights to those away from t_0 . h is the bandwidth, which controls the size of the neighborhood. The linear approximation is good when h is sufficiently small. $\hat{\beta}_0$ and $\hat{\beta}_1$ are local linear estimators of $\beta(\cdot)$ and $\beta'(\cdot)$ at t_0 , respectively. The above weighted least squares estimation can be repeated at a grid of points on the support of T to establish the estimated coefficient functions. Usually the bandwidth h has a considerable impact on model-fitting, while the kernel function $K(\cdot)$ does not. Throughout this paper we use the Epanechnikov kernel $K(t) = 0.75(1 - t^2)\mathbb{I}_{\{-1 \leq t \leq 1\}}$.

When \mathbf{Z} is unknown and we observe \mathbf{X} instead, we use a *local correction* procedure as follows. Differentiating (3) with respect to θ leads to the following estimating equations:

$$\begin{aligned} 0 &= n^{-1} \sum_{i=1}^n \Phi_i(\theta; Y_i, \mathbf{Z}_i, T_i) \\ &= n^{-1} \sum_{i=1}^n K_h(T_i - t_0) (\eta_i \eta_i^T \theta - \eta_i Y_i) \end{aligned} \quad (4)$$

where $\eta_i = (\mathbf{Z}_i^T, \mathbf{Z}_i^T(T_i - t_0))^T$. Making use of the fact that

$$E(\mathbf{X}_i \mathbf{X}_i^T | \mathbf{Z}_i, Y_i) = \mathbf{Z}_i \mathbf{Z}_i^T + \Sigma_u \quad \text{and} \quad E(\mathbf{X}_i | \mathbf{Z}_i, Y_i) = \mathbf{Z}_i,$$

we have the following *locally corrected* estimating equations:

$$\begin{aligned} 0 &= n^{-1} \sum_{i=1}^n \Phi_i^*(\theta; Y_i, \mathbf{X}_i, T_i) \\ &= n^{-1} \sum_{i=1}^n K_h(T_i - t_0) \left\{ (\eta_i^* \eta_i^{*T} - \mathbf{C}_i) \theta - \eta_i^* Y_i \right\}. \end{aligned} \quad (5)$$

Here η_i^* equals η_i with \mathbf{Z}_i replaced by \mathbf{X}_i , and the matrix $\mathbf{C}_i = \{(1, T_i - t_0)^T(1, T_i - t_0)\} \otimes \Sigma_u$ corrects for bias resulting from the measurement error in X_i . This is an analog to Nakamura's corrected score equations [15] for models with finite dimensional parameters because

$$E\{\Phi_i^*(\theta; Y_i, \mathbf{X}_i, T_i) | Y_i, \mathbf{Z}_i, T_i\} = \Phi_i(\theta; Y_i, \mathbf{Z}_i, T_i). \quad (6)$$

In this sense, (5) is a "locally unbiased" estimating equation. The solution takes a closed form

$$\hat{\theta} = \left\{ \sum K_h(T_i - t_0) (\eta_i^* \eta_i^{*T} - \mathbf{C}_i) \right\}^{-1} \left(\sum K_h(T_i - t_0) \eta_i^* Y_i \right).$$

Its variance can be estimated by a sandwich estimator:

$$\text{var}(\hat{\theta}) = \left\{ \sum \frac{\partial}{\partial \theta^T} \Phi_i^* \right\}^{-1} \left\{ \sum \Phi_i^* \Phi_i^{*T} \right\} \left\{ \sum \frac{\partial}{\partial \theta^T} \Phi_i^* \right\}^{-1} \quad (7)$$

Condition (6) suggests that (4) and (5) and their associated solutions may be similar asymptotically, with or without measurement error. In the Appendix, we prove under usual regularity conditions that the asymptotic properties of the estimators are quite similar with or without the errors. The error corrected estimators for $\beta(t_0)$ are consistent and asymptotically normal with a regular nonparametric convergence rate $(nh)^{1/2}$. We also show that because the measurement error adds additional random noise to the data, the estimators have enlarged asymptotic variance. Equation (5) is a local estimating equation, for which bias and variance has been derived previously using conditional moments [16]. The difference here is that we establish the asymptotic distribution in our setting, which justifies the use of the normality-based point-wise confidence interval.

If some coefficient functions are known to be constant, that information can be used to derive more efficient estimators. Suppose we evaluate the curve $\beta_k(\cdot)$ at a finite (but reasonably large) number of pre-chosen grid points, denoted by t_j , $j = 1, 2, \dots, m$, that span the range of T . Let $\hat{\beta}_k(t_j)$ denote the estimated $\beta_k(\cdot)$ at t_j . When $\beta_k(\cdot) \equiv \beta_k$, a constant, each $\hat{\beta}_k(t_j)$ converges to β_k at a nonparametric rate $(nh)^{1/2}$. A more efficient estimator of β_k can be formed by simply taking the average:

$$\hat{\beta}_k = \frac{1}{m} \sum_{j=1}^m \hat{\beta}_k(t_j) \quad (8)$$

In the appendix, we show that $\hat{\beta}_k$ is approximately asymptotically normal with an asymptotic variance of order n^{-1} , similar to a typical parametric estimator. This is because we can only use a fraction of the data (roughly nh , if the range of T is standardized to be 1) to estimate varying coefficients, but

use the whole data set to estimate the constant coefficient. In practice, we use the bootstrap to estimate the variance of $\hat{\beta}_k$ empirically, and construct a normality-based Wald-type confidence interval. We can also compute the bootstrap confidence interval directly, but at the cost of additional computing resources. As with many nonparametric smoothers, the curve may be estimated with large variation at the two tails. This is sometimes called the tail effect. One may use a trimmed mean in (8) to reduce the tail effect.

2.2 Selecting the bandwidth

Despite a rich literature on bandwidth selection for local polynomial models, little has been done for the case with covariate measurement error. Here we propose the following expected generalized cross-validation (EGCV) criterion for bandwidth selection, which approximates the widely-used generalized cross-validation (GCV) criterion in the absence of measurement error.

As before, we first consider the case without measurement error. Express both the outcome variable and the "design matrix" of (3) in matrix form as:

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \quad \mathbf{D}(t) = \begin{pmatrix} \mathbf{Z}_1^T & \mathbf{Z}_1^T(T_1 - t) \\ \vdots & \vdots \\ \mathbf{Z}_n^T & \mathbf{Z}_n^T(T_n - t) \end{pmatrix}.$$

Let $\theta(t) = (\beta_0(t)^T, \beta_1(t)^T)^T$ denote the parameter vector at $T = t$, and write $\mathbf{W}(t) = \text{diag}\{K_h(T_i - t)\}$. The local polynomial estimator for $\theta(t)$ is

$$\hat{\theta}(t) = \{\mathbf{D}(t)^T \mathbf{W}(t) \mathbf{D}(t)\}^{-1} \mathbf{D}(t)^T \mathbf{W}(t) \mathbf{Y}$$

This is a matrix presentation of the solution to (4). Since the fitted value of the outcome variable is $\hat{\mathbf{Y}}_i = (\mathbf{Z}_i^T, \mathbf{0})\hat{\theta}(T_i)$, the smoothing matrix \mathbf{S}_h , which

satisfies $\hat{\mathbf{Y}} = \mathbf{S}_h \mathbf{Y}$, takes the form:

$$\mathbf{S}_h = \begin{pmatrix} \vdots & & \\ (\mathbf{Z}_i^T, \mathbf{0}) \{ \mathbf{D}(T_i)^T \mathbf{W}(T_i) \mathbf{D}(T_i) \}^{-1} \mathbf{D}(T_i)^T \mathbf{W}(T_i) & & \\ \vdots & & \end{pmatrix}_{n \times n} \quad (9)$$

Bandwidth selection for varying coefficient models without measurement error has been studied by [17], who used the following GCV criterion:

$$GCV(h) = \frac{n^{-1} \|\mathbf{Y} - \mathbf{S}_h \mathbf{Y}\|^2}{\{1 - \text{tr}(\mathbf{S}_h)/n\}^2} \quad (10)$$

The numerator resembles the residual sums of squares term in linear regression; $\text{tr}(\mathbf{S}_h)$ in the denominator is related to the degrees of freedom used by fitting the model. The bandwidth h is taken to be the value that minimizes $GCV(h)$.

When there is measurement error, \mathbf{S}_h involves the unknown variable \mathbf{Z} . However, similar to the previous section, we can find a quantity of the observed variables that approximates (10) as n becomes large. To do that, we re-express (10) as:

$$GCV(h) = \frac{n^{-1} \{ \mathbf{Y}^T \mathbf{Y} - 2 \mathbf{Y}^T (\mathbf{S}_h \mathbf{Y}) + (\mathbf{S}_h \mathbf{Y})^T (\mathbf{S}_h \mathbf{Y}) \}}{\{1 - \text{tr}(\mathbf{S}_h)/n\}^2} \quad (11)$$

$$\text{tr}(\mathbf{S}_h) = \sum_{i=1}^n (\mathbf{Z}_i^T, \mathbf{0}) \{ \mathbf{D}^T(T_i) \mathbf{W}(T_i) \mathbf{D}(T_i) \}^{-1} (\mathbf{Z}_i^T, \mathbf{0})^T K_h(0)$$

Denote $\mathbf{D}^T(T_i) \mathbf{W}(T_i) \mathbf{D}(T_i)$ by $\mathbf{\Delta}(T_i)$. By similar arguments as in section 2.1, we can show that:

$$\mathbf{\Delta}(T_i) = \left\{ \mathbf{D}^{*T}(T_i) \mathbf{W}(T_i) \mathbf{D}^*(T_i) - \mathbf{C}(T_i) \right\} (1 + o_p(1)) \quad (12)$$

where $\mathbf{D}^*(t)$ is $\mathbf{D}(t)$ with \mathbf{Z} replaced by \mathbf{X} . The matrix $\mathbf{C}(t)$ performs the

error correction:

$$\mathbf{C}(T_i) = \left\{ \begin{pmatrix} 1 & & & \\ & \ddots & & \\ & & 1 & \\ & T_1 - T_i & \cdots & T_n - T_i \end{pmatrix} \mathbf{W}(T_i) \begin{pmatrix} 1 & T_1 - T_i \\ \vdots & \vdots \\ 1 & T_n - T_i \end{pmatrix} \right\} \otimes \boldsymbol{\Sigma}_u$$

(12) suggests that $\boldsymbol{\Delta}(T_i)$ can be replaced by $\boldsymbol{\Delta}^*(T_i) = \mathbf{D}^{*T}(T_i)\mathbf{W}^*(T_i)\mathbf{D}^*(T_i) - \mathbf{C}(T_i)$ in the subsequent derivation. In this sense, we call $\boldsymbol{\Delta}(T_i)$ *correctable*.

Let $\boldsymbol{\Delta}_{-11}^*(T_i)$ denote the upper left $p \times p$ matrix of $\boldsymbol{\Delta}^*(T_i)^{-1}$. We can show that:

$$\begin{aligned} \text{tr}(\mathbf{S}_h) &= K_h(0) \sum_{i=1}^n \mathbf{Z}_i^T \boldsymbol{\Delta}_{-11}^*(T_i) \mathbf{Z}_i (1 + o_p(1)) \\ &= K_h(0) \sum_{i=1}^n \left[\mathbf{X}_i^T \boldsymbol{\Delta}_{-11}^*(T_i) \mathbf{X}_i - \text{tr}\{\boldsymbol{\Delta}_{-11}^*(T_i) \boldsymbol{\Sigma}_u\} \right] (1 + o_p(1)) \end{aligned}$$

Therefore, $\text{tr}(\mathbf{S}_h)$ is correctable. Next, we will show both $\mathbf{Y}^T(\mathbf{S}_h \mathbf{Y})$ and $(\mathbf{S}_h \mathbf{Y})^T(\mathbf{S}_h \mathbf{Y})$ are also correctable.

$$\mathbf{Y}^T(\mathbf{S}_h \mathbf{Y}) = \sum_{i=1}^n \left[Y_i(\mathbf{Z}_i^T, 0) \boldsymbol{\Delta}(T_i)^{-1} \{ \mathbf{D}^T(T_i) \mathbf{W}(T_i) \mathbf{Y} \} \right]$$

$\boldsymbol{\Delta}(T_i)^{-1}$ is correctable. $\boldsymbol{\Psi}(T_i) = \mathbf{D}^T(T_i) \mathbf{W}(T_i) \mathbf{Y}$ can be corrected by $\boldsymbol{\Psi}^*(T_i) = \mathbf{D}^{*T}(T_i) \mathbf{W}(T_i) \mathbf{Y}$. Hence, $\mathbf{Y}^T(\mathbf{S}_h \mathbf{Y})$ can be corrected as follows:

$$\mathbf{Y}^T(\mathbf{S}_h \mathbf{Y}) = \sum_{i=1}^n \left[Y_i(\mathbf{X}_i^T, 0) \boldsymbol{\Delta}^*(T_i)^{-1} \boldsymbol{\Psi}^*(T_i) \right] (1 + o_p(1))$$

$$\begin{aligned} (\mathbf{S}_h \mathbf{Y})^T(\mathbf{S}_h \mathbf{Y}) &= \sum_{i=1}^n (\mathbf{Z}_i^T, 0) \boldsymbol{\Delta}(T_i)^{-1} \boldsymbol{\Psi}(T_i) \boldsymbol{\Psi}(T_i)^T \boldsymbol{\Delta}(T_i)^{-T} (\mathbf{Z}_i^T, 0)^T \\ &\approx \sum_{i=1}^n (\mathbf{Z}_i^T, 0) \boldsymbol{\Delta}^*(T_i)^{-1} \boldsymbol{\Psi}^*(T_i) \boldsymbol{\Psi}^*(T_i)^T \boldsymbol{\Delta}^*(T_i)^{-T} (\mathbf{Z}_i^T, 0)^T \end{aligned}$$

Let $\boldsymbol{\Lambda}^*(T_i) = \boldsymbol{\Delta}^*(T_i)^{-1} \boldsymbol{\Psi}^*(T_i) \boldsymbol{\Psi}^*(T_i)^T \boldsymbol{\Delta}^*(T_i)^{-T}$, and use $\boldsymbol{\Lambda}_{11}^*(T_i)$ to denote the upper left $p \times p$ matrix of $\boldsymbol{\Lambda}^*(T_i)$, $(\mathbf{S}_h \mathbf{Y})^T(\mathbf{S}_h \mathbf{Y})$ can be corrected as:

$$(\mathbf{S}_h \mathbf{Y})^T(\mathbf{S}_h \mathbf{Y}) = \sum_{i=1}^n \left[\mathbf{X}_i^T \boldsymbol{\Lambda}_{11}^*(T_i) \mathbf{X}_i - \text{tr}\{\boldsymbol{\Lambda}_{11}^*(T_i) \boldsymbol{\Sigma}_u\} \right] (1 + o_p(1))$$

We propose replacing $\text{tr}(\mathbf{S}_h)$, $\mathbf{Y}^T(\mathbf{S}_h\mathbf{Y})$ and $(\mathbf{S}_h\mathbf{Y})^T(\mathbf{S}_h\mathbf{Y})$ in the GCV formula by their corresponding error corrected versions. We refer to the resulting function of h as the EGCV function (expected GCV). The bandwidth is defined to be the minimizer of the EGCV function.

3 Simulation

We conducted a simulation to study the finite sample performance of the proposed estimation and bandwidth procedures. The data were generated from a simplified version of model (1):

$$Y_i = \beta_1(T_i)Z_{1i} + \beta_2(T_i)Z_{2i} + \epsilon_i \quad \text{and} \quad X_{2i} = Z_{2i} + U_{2i} \quad (13)$$

T_i has a uniform distribution between 0 and 1. Z_{1i} and Z_{2i} each have bivariate normal distribution with mean zero and variance c ($c = 1/3$, or 1); their correlation coefficient is 0.5. ϵ_i is normal with mean zero and variance 0.6, so that the signal-to-noise variance ratios vary between 0.5 : 1 to 3.5 : 1, depending on the setting. The measurement error variance is 0.2. $\beta_1(t) = \cos(1.5\pi t)$ or 1; $\beta_2(t) = \sin(1.3\pi t)$. The sample size $n = 250$ or 1000.

We consider four scenarios:

- (A) $N = 250$; $h = 0.22$; $c = 1$; $\beta_1(\cdot) \equiv 1$.
- (B) $N = 250$; $h = 0.22$; $c = 1$; $\beta_1(\cdot) = \cos(1.5\pi t)$.
- (C) $N = 1000$; $h = 0.12$; $c = 1$; $\beta_1(\cdot) = \cos(1.5\pi t)$.
- (D) $N = 1000$; $h = 0.12$; $c = 1/3$; $\beta_1(\cdot) = \cos(1.5\pi t)$.

The bandwidth was selected in some preliminary runs using our bandwidth selection procedure (see below). The coefficient functions were evaluated on $m = 101$ equally spaced grid points from 0 to 1. For each scenario, 500 simulations were conducted. Figure 1 shows the median estimated coefficient functions at these grid points. Table 1 reports the median integrated mean squared error (IMSE) of the varying coefficients estimators, and the bias and mean squared error (MSE) of the constant coefficients. $IMSE(\hat{\beta}_k) = \int_0^1 \{\hat{\beta}_k(t) - \beta_k(t)\}^2 d t \approx \delta \sum_{j=1}^m \{\hat{\beta}_k(t_j) - \beta_k(t_j)\}^2$. $\delta = 0.01$ is the squared distance between adjacent grid points. In practice, we deleted 10% of the grid points on each end to reduce the tail effect of the estimated curves on IMSE.

Figure 1 and Table 1 reveal that a common phenomenon with parametric measurement error models is also seen in the varying coefficients model: the naive method results in attenuation of the coefficient function of the error-prone covariate Z_2 toward zero, and the bias can be carried over to the estimated coefficient Z_1 , due to the correlation between them. The proposed method corrects most of the bias. The efficiency improves as the sample size gets larger. While the measurement error variance was fixed, we considered a smaller variance for the true covariates in simulation D. As shown in Figure 1, it enlarges the bias caused by measurement error, which resembles the situation seen in the application we provide below.

[Table 1 about here.]

[Figure 1 about here.]

We also conducted an additional simulation under setting (A) to assess

whether the estimation of constant coefficient is sensitive to the choice of the bandwidth. We set the bandwidth to be a series of values around the chosen bandwidth $h = 0.22$: 0.10, 0.15, 0.19, 0.22, 0.25, 0.30. The bias and standard deviation of the estimated constant coefficients, averaged across 500 simulation runs, are: -0.0114 (0.0757), -0.00102 (0.0732), 0.000843 (0.0751), -0.00899 (0.0753), -0.00778 (0.0693), -0.00828 (0.0743). This result suggests that the estimated constant coefficient is not sensitive to the choice of bandwidth.

We simulated additional data to study the proposed bandwidth selection criterion. Two hundred data sets were generated according to the model in simulation B (N=250) or C (N=1000). For each data set, the GCV function was calculated using the simulated true covariate, and the EGCV function was calculated using the simulated error contaminated covariate. Figure 2 shows that the two functions are close to each other and have similar bowl shapes. The minimizers of the curves are at similar locations. The EGCV function has a small negative bias of less than 2%. The bias becomes smaller when the sample size increases. Further analysis reveals that the bias generally comes from the approximation to the quadratic function in the numerator of (10), but the corrected $tr(\mathbf{S}_h)$ is quite accurate. The bias does not appear to influence the choice of the bandwidth.

[Figure 2 about here.]

4 Application

In this section, we apply the proposed methods to characterize the influence of age on the regression coefficients relating log transformed serum creatinine

(SCR) to log transformed GFR, gender, and race that was described in Section 1. We first consider a pooled cross-sectional data set constructed from three of the studies participating in the CKD-EPI project in which patients were known to have chronic kidney disease. The pooled data set includes the baseline data from enrollees into two clinical trials, the MDRD study and the AASK (African American Study of Kidney Disease) study, and a registry, the Cleveland Clinic research laboratory database. As described in Section 1, in this report we assume a measurement error variance in log GFR of $\sigma_u^2 = 0.015$.

After excluding of a small group of patients with diabetes or with age older than 70, where data are sparse, the pooled data set contains 2510 patients, of whom approximately 54% are black, and 62% are male. The median age is 53. Since we do not know which coefficients are constant or varying in age, we initially assume that all of the coefficients are varying coefficients, and apply model (2) with correction for the measurement error in GFR. The covariate vector Z includes binary indicator variables for black race and male gender, the varying intercept term of age, and $\log(\text{GFR})$, with age serving as the tuning variable T . The EGCV chooses $h = 7$ (years) as the bandwidth. Figure 3 shows the estimated varying coefficients together with pointwise 95% confidence intervals, evaluated on the grid of integers between ages 18 and 70 years. The coefficient for gender appears to be constant. The coefficient of black race appears to be linear with a small suggestion of a negative slope, indicating a possible cross-product interaction between black race and age corresponding to an attenuation of the effect of race at higher ages. The main effect of age and the interaction of age with $\log(\text{GFR})$ appear

to be nonlinear, with the slopes of the relationships apparently increasing in magnitude after age 40. We can see from the above that the varying coefficients model is a useful tool to quickly identify complicated interactions between a confounding factor and other covariates.

[Figure 3 about here.]

It has been shown in section 2 that coefficients can be estimated with increased efficiency if they can be assumed to be constant. Therefore, we re-fit the model above with three constant coefficient terms: gender, black race, and the age by black race interaction. The coefficient (bootstrap standard error) estimates are 0.247(0.00904) for male gender, 0.249(0.0737) for black race, and $-0.00201(0.00158)$ for the interaction. The two main effects are statistically significant at 0.05 level, while the interaction is not ($p=0.2$). In this analysis, the curve estimates for age and $\log(\text{GFR})$ are essentially unchanged. Note that the estimation procedure based on (3), as discussed above, requires a small modification when the cross-product term of Z and T is included in the model. In this situation, we do not use the first order Taylor expansion to approximate the constant coefficients, because if we do, there will be exact collinearity between the cross-product term and a term in Z 's Taylor expansion. Instead, we minimize the following weighted sum of squares:

$$\sum_{i=1}^n K_h(T_i - t_0) \left\{ Y_i - \mathbf{ZC}_i^T \alpha - \mathbf{Z}_i^T \beta_0 - \mathbf{Z}_i^T (T_i - t_0) \beta_1 \right\}^2$$

where \mathbf{ZC} and \mathbf{Z} include the covariates with constant and varying coefficients, respectively. The remainder of the estimation procedure proceeds as

described in section 2. Again, we must average across grid points to obtain a single point estimator for α .

Next, we apply the model to a pooled data set (N=700) constructed from two databases of individuals without prior evidence of kidney disease who were being screened as possible donors for kidney transplantation. As above, we assume a measurement error variance of \log GFR of 0.015. Figure 4 (right) shows the estimated coefficient of $\log(\text{GFR})$ as a function of age. Compared to the result from CKD data (left), the most obvious distinction is that the naive analysis produces quite different curves, with the coefficient of $\log(\text{GFR})$ approximating -0.8 for the CKD patients and -0.4 for the donors. After error correction, the coefficient functions become closer throughout most of the age range. This result suggests that measurement error in GFR explains much of the apparent discrepancy in the regression functions between the two populations.

[Figure 4 about here.]

In simulation D, we saw that the same amount of measurement error causes more attenuation as the variance of the true covariate becomes smaller compared to the variance of the measurement error. This explains what we see in this example. The variance of $\log(\text{GFR})$ is 0.368 for CKD, but only 0.029 for Donors. Figure 5 shows the histograms of GFR for CKD and donors. The normal range of GFR is usually around 90-110 ml/min/1.73m², while people with CKD may have a wide spectrum of reduced GFRs depending on the stage of disease.

[Figure 5 about here.]

Figure 4 suggests that the coefficient of $\log(\text{GFR})$ is independent of age throughout most of the age range, but appears to decrease sharply after age 55. This may really be the case, or it could be a result of the relatively large measurement error that blurs the underlying structure. The curve drops sharply after age 55. This is likely a tail effect. Eighty percent of the kidney donors are healthy people aged between 30 and 55, and hence the sample size for the local regression near the boundary is very small. This can be particularly problematic when there is measurement error. The proposed moment-based correction can have substantial bias in sparse local neighborhoods when the measurement error is large.

5 Discussion

The local correction procedure we propose can be viewed as a generalization of the corrected score equations approach, first proposed by [15] for parametric models, to the case of semi-parametric varying coefficient models. Our work also extends the work of [18] by employing a local polynomial method, which can be viewed as a generalization of the local constant method used in that paper.

In this paper, we assume the error variance is a known constant. In fact, there is uncertainty regarding the true measurement error of GFR in the studies including in the datasets we have analyzed, in part because it is possible that the precision of the GFR assay varies between studies. Therefore, the analyses presented here may best be viewed as a sensitivity analysis for the effect of measurement error. In this setting it is not necessary to account for sampling error in estimating σ_u^2 . If the sampling error must be considered,

the estimators presented here will remain consistent so long as the estimator of σ_u^2 is consistent, but the finite sample variances of the estimators would be expected to be larger than calculated in Section 2.

The work presented in this paper has clinical implications and highlight the importance of measurement error analysis in this clinical study. As we see in Section 4, without error adjustment, the regression models for CKD patients and healthy donors look quite different, which appears to suggest different underlying relationships between serum creatinine and GFR in the two populations. As a result, we might want to develop different GFR prediction equations for different populations, which could be difficult to use in clinical practice. However, the measurement error analysis suggests that most of the observed discrepancy between populations is due to the measurement error, which results in different signal to noise ratios in different populations, and hence the attenuated regression coefficients. Therefore, a unified model for the relationship between serum creatinine and GFR may still be used, with proper adjustment for the measurement error. The analysis also suggests that accurate prediction of GFR in healthy population could be difficult, due to the low signal to noise ratio.

6 Appendix

In this section, we prove that the error corrected estimators for model (2) are consistent and asymptotically normal. The following regularity conditions are needed:

- (1) The kernel function $K(\cdot)$ is a symmetric density function centered around zero with compact support.

- (2) The variable T has bounded support; its density function $f(T)$ is Lipschitz continuous and $f(T) > 0$, *a.s.*
- (3) $\beta(\cdot)$ has continuous second derivative in T .
- (4) The matrix of the second moment of \mathbf{Z} , conditional on T , is non-singular for any T in its support, and both the matrix and its inverse are Lipschitz continuous.
- (5) $h \rightarrow 0$, $nh^2 \rightarrow \infty$ and h satisfies the conditions in Lemma 1.

The following lemma, which was used in [12], is useful in deriving the asymptotic bias for kernel estimators. The conditions of the lemma are easily satisfied in practical situations such as ours since all variables considered are essentially bounded.

Lemma 1. Let $\{(X_i, Y_i), i = 1, \dots, n\}$ be independent and identically distributed random vectors, where the Y_i 's are scalar random variables. Assume that $E|Y|^s < \infty$ and $\sup_x \int |Y|^s f(X, Y) dY < \infty$, where f denotes the joint density of (X, Y) . K is a bounded positive function with a bounded support, satisfying a Lipschitz condition. Then

$$\sup_x \left| \frac{1}{n} \sum_{i=1}^n [K_h(X_i - x)Y_i - E\{K_h(X_i - x)Y_i\}] \right| = O_p\left(\left\{\frac{\log(1/h)}{nh}\right\}^{1/2}\right)$$

provided that $n^{2\epsilon-1}h \rightarrow \infty$ for some $\epsilon < 1 - s^{-1}$.

We use the following notation throughout the proof: $\mu_j = \int t^j K(t) dt$, $\nu_j = \int K^j(t) dt$, and $e_n = \left\{\frac{\log(1/h)}{nh}\right\}^{1/2}$.

Lemma 2. Suppose $\{(Y_i, T_i), i = 1, \dots, n\}$ satisfy the conditions in Lemma 1; T_i 's satisfy the regularity conditions (1), (2), and (5); $E(Y|T)$ is Lipschitz

continuous. Then as $n \rightarrow \infty$:

$$\begin{aligned}
\frac{1}{n} \sum_{i=1}^n K_h(T_i - t_0) Y_i &= E(Y|t_0) f(t_0) + O_p(h^2) + O_p(e_n) \\
\frac{1}{n} \sum_{i=1}^n K_h(T_i - t_0) (T_i - t_0) Y_i &= O_p(h^2) + O_p(h^4) + O_p(e_n) \\
\frac{1}{n} \sum_{i=1}^n K_h(T_i - t_0) (T_i - t_0)^2 Y_i &= h^2 E(Y|t_0) f(t_0) \mu_2 + O_p(h^4) + O_p(e_n).
\end{aligned} \tag{14}$$

The proof of this result uses Lemma 1 and a Taylor series expansion of $E(Y|T)f(T)$ around t_0 .

Applying these results to the corrected estimating equation (5), we have:

$$\begin{aligned}
&\frac{1}{n} \sum_{i=1}^n K_h(T_i - t_0) [\eta_i^* \eta_i^{*T} - \mathbf{C}_i] \\
&= f(t_0) \begin{pmatrix} \mathbf{A}_{11} + O_p(h^2 + e_n) & O_p(h^2 + h^4 + e_n) \\ O_p(h^2 + h^4 + e_n) & \mathbf{A}_{22} + O_p(h^4 + e_n) \end{pmatrix}
\end{aligned} \tag{15}$$

where $\mathbf{A}_{11} = E(\mathbf{Z}_i \mathbf{Z}_i^T | t_0)$, $\mathbf{A}_{22} = h^2 \mu_2 \mathbf{A}_{11}$. We are interested in β_0 , the first p elements of θ , whose corresponding estimating equations are

$$f(t_0) \mathbf{A}_{11} \hat{\beta}_0 = \frac{1}{n} \sum_{i=1}^n K_h(T_i - t_0) \mathbf{X}_i Y_i + O_p(h^2 + e_n)$$

Denote the first term on the right hand side as \mathbf{M}_n . Then,

$$\hat{\beta}_0 = f(t_0)^{-1} \mathbf{A}_{11}^{-1} \mathbf{M}_n + O_p(h^2 + e_n) \tag{16}$$

Since $\mathbf{M}_n = f(t_0) \mathbf{B}_1 \{1 + O_p(h^2 + e_n)\}$ by Lemma 2, where $\mathbf{B}_1 = \mathbf{A}_{11} \beta(t_0)$,

$$\hat{\beta}_0 = \beta(t_0) + O_p(h^2 + e_n).$$

By condition (5), the estimators are consistent. The asymptotic bias is of the same order as if there were no measurement error. This is because the error is an independent mean zero noise and its effect is averaged out in the

corrected estimating equation. However, the error adds noise into the model and increase the asymptotic variance of $\hat{\beta}_0$, which will be shown below.

The asymptotic normality of $\hat{\beta}_0$ follows that of \mathbf{M}_n , which is the average of independent and identically distributed (for given n) random vectors. By the Crámer-Wold device, it suffices to show that for any unit vector \mathbf{c} of the same dimension as $\hat{\beta}_0$, $n^{-1} \sum_{i=1}^n K_h(T_i - t_0) \mathbf{c}^T \mathbf{X}_i Y_i$ is asymptotically normal. Let $\mathbf{X}_i Y_i = \xi_i$, and $\varphi_i = \mathbf{c}^T \xi_i$. Note that φ_i 's are scalar random variables. In order to apply the Linderberg's central limit theorem, we need to check the Liapunov condition. Similar to Lemma 2, we can verify the following useful formula:

$$E\{K_h(T_i - t_0)^d \varphi_i^d\} = h^{1-d} E(\varphi_i^d | t_0) f(t_0) \nu_d (1 + O_p(h^2)), \quad d = 1, 2, \dots$$

Liapunov condition requires that $n^{-1} \text{Var}(K_h \varphi_i)^{-2} E\{K_h \varphi_i - E(K_h \varphi_i)\}^4 \rightarrow 0$, which is implied by $n^{-1} \text{Var}(K_h \varphi_i)^{-2} E(K_h \varphi_i)^4 \rightarrow 0$. The latter can be shown to be a result of $nh \rightarrow \infty$ using the formula above. Therefore, the Liapunov condition holds, and the estimator $\hat{\beta}_0$ is asymptotically normal with asymptotic variance being approximately $(nh)^{-1} \mathbf{A}_{11}^{-1} E(\xi_i \xi_i^T | t_0) \mathbf{A}_{11}^{-1} f(t_0)^{-1} \nu_2$. Since

$$\begin{aligned} E(\xi_i \xi_i^T | t_0) &= E((\mathbf{Z}_i \mathbf{Z}_i^T + \mathbf{U}_i \mathbf{Z}_i^T + \mathbf{Z}_i \mathbf{U}_i^T + \mathbf{U}_i \mathbf{U}_i^T) Y_i^2 | t_0) \\ &= E\{(\mathbf{Z}_i \mathbf{Z}_i^T) Y_i^2 | t_0\} + E\{(\mathbf{U}_i \mathbf{U}_i^T) Y_i^2 | t_0\} \end{aligned}$$

The first term equals $E(\xi_i \xi_i^T | t_0)$ when $U_i \equiv 0$ (*i.e.*, no error). The added noise is shown in the second term, which is strictly positive.

Next, we study the asymptotic properties of the estimator in (8). By (16), we have

$$\begin{aligned}
\hat{\beta}_k &= \frac{1}{m} \sum_{j=1}^m \hat{\beta}_k(t_j) = \frac{1}{m} \sum_{j=1}^m \mathbf{s}_k^T \hat{\beta}(t_j) \\
&= \frac{1}{m} \sum_{j=1}^m \mathbf{s}_k^T f(t_j)^{-1} \mathbf{A}_{11}(t_j)^{-1} \left\{ \frac{1}{n} \sum_{i=1}^n K_h(T_i - t_j) \mathbf{X}_i Y_i \right\} \\
&= \mathbf{s}_k^T \frac{1}{n} \sum_{i=1}^n \left\{ \frac{1}{m} \sum_{j=1}^m K_h(T_i - t_j) f(t_j)^{-1} \mathbf{A}_{11}(t_j)^{-1} \right\} \mathbf{X}_i Y_i
\end{aligned}$$

\mathbf{s}_k is a vector of zeros except that the k th element is 1. Without loss of generality, we assume the support of T is $[0, 1]$ and $\{t_j\}_{j=1}^m$ are sampled with probability density $g(t)$. If m is large, the quantity inside the bracket is approximately

$$\int_0^1 f(x)^{-1} \mathbf{A}_{11}(x)^{-1} K_h(T_i - x) g(x) dx \approx f(T_i)^{-1} \mathbf{A}_{11}(T_i)^{-1} g(T_i).$$

Therefore,

$$\begin{aligned}
\hat{\beta}_k &\approx \mathbf{s}_k^T \frac{1}{n} \sum_{i=1}^n f(T_i)^{-1} \mathbf{A}_{11}(T_i)^{-1} g(T_i) \mathbf{X}_i Y_i \\
E(\hat{\beta}_k) &\approx E \left\{ E \left\{ \mathbf{s}_k^T f(T_i)^{-1} \mathbf{A}_{11}(T_i)^{-1} g(T_i) \mathbf{X}_i Y_i \mid T_i \right\} \right\} \\
&= E \left\{ \mathbf{s}_k^T f(T_i)^{-1} \mathbf{A}_{11}(T_i)^{-1} g(T_i) \mathbf{A}_{11}(T_i) \beta(T_i) \right\} \\
&= E \left\{ f(T_i)^{-1} g(T_i) \mathbf{s}_k^T \beta(T_i) \right\} \\
&= E \left\{ f(T_i)^{-1} g(T_i) \beta_k \right\} \\
&= \beta_k
\end{aligned}$$

It follows from the central limit theorem that $\hat{\beta}_k$ is asymptotically normal with variance of order n^{-1} .

7 Acknowledgement

This work was partially supported by grant 5U01DK053869-06 from the National Institute of Diabetes and Digestive and Kidney Diseases.

References

- [1] A.S. Levey, J.P. Bosch, J.B. Lewis, T. Greene, and *et al.* A more accurate method to estimate glomerular filtration rate from serum creatinine: a new prediction equation. *Annals of Internal Medicine*, 130:461–470, 1999.
- [2] A.S. Levey, T. Greene, J. Kusek, and G. Beck. A simplified equation to predict glomerular filtration rate from serum creatinine (abstract). *Journal of American Society of Nephrology*, 11:155A, 2000.
- [3] J. Lewis, L. Agodoa, D. Cheek, T. Greene, J. Middleton, D. O’Connor, A. Ojo, R. Phillips, M. Sika, and J. Jr Wright. Comparison of cross-sectional renal function measurements in african-americans with hypertensive nephrosclerosis and of primary formulas to estimate glomerular filtration rate. *American Journal of Kidney Disease*, 38:744–753, 2001.
- [4] A. Bostom, F. Kronenberg, and E. Ritz. Predictive performance of renal function equations for patients with chronic kidney disease and normal serum creatinine levels. *Journal of American Society of Nephrology*, 13:2140–2144, 2002.
- [5] A. Rule, T. Larson, E. Bergstralh, J. Slezak, S. Jacobsen, and F. Cosio. Using serum creatinine to estimate glomerular filtration rate: Accuracy

- in good health and in chronic kidney disease. *Ann Intern Med*, 141:929–937, 2004.
- [6] E. Poggio, X. Wang, T. Greene, F. Van Lente, and P. Hall. Performance of the mdrd and cockcroft-gault equations in the estimation of glomerular filtration rate in health and in chronic kidney disease. *Journal of American Society of Nephrology*, 16:459–466, 2005.
- [7] H. Ibrahim, M. Mondress, A. Tello, Y. Fan, J. Koopmeiners, and W. Thomas. An alternative formula to the cockcroft-gault and the modification of diet in renal diseases formulas in predicting gfr in individuals with type 1 diabetes. *Journal of American Society of Nephrology*, 16:1051–1060, 2005.
- [8] Krutchkoff. Classical and inverse regression methods of calibration. *Technometrics*, 9:425–439, 1967.
- [9] R. Sunberg. Multivariate calibration - direct and indirect regression methodology. *Scandinavian Journal of Statistics*, 26:161–207, 1999.
- [10] Ibrahim Ahmad, Sittisak Leelahanon, and Qi Li. Efficient estimation of a semiparametric partially linear varying coefficient model. *Annals of Statistics*, 33 (1):258–283, 2005.
- [11] Jianqing Fan and Tao Huang. Profile likelihood inferences on semiparametric varying-coefficient partially linear models. *Bernulli*, 2005.
- [12] Wenyang Zhang, Sik-Yum Lee, and Xinyuan Song. Local polynomial fitting in semivarying coefficient model. *Journal of Multivariate Analysis*, 82:166–188, 2002.

- [13] Hua Liang, Hulin Wu, and Raymond J. Carroll. The relationship between virologic and immunologic responses in aids clinical research using mixed-effects varying-coefficient models with measurement error. *Biostatistics*, 4):297–312, 2004.
- [14] Jianqing Fan and Wenyang Zhang. Statistical estimation in varying coefficient models. *Annals of Statistics*, 27 (5):1491–1518, 1999.
- [15] Tsuyoshi Nakamura. Corrected score function for errors-in-variables models: methodology and applicaiton to generalized linear models. *Biometrika*, 77:127–137, 1990.
- [16] Raymond J. Carroll, David Ruppert, and Alan H. Welsh. Local estimating equations. *Journal of the American Statistical Association*, 93(441):214–227, 1998.
- [17] Chunming Zhang. Calibrating the degrees of freedom for automatic data smoothing and effective curve checking. *Journal of the American Statistical Association*, 98(463):609–628, 2003.
- [18] Hua Liang, Wolfgang Härdle, and Raymond J. Carroll. Estimation in a semiparametric partially linear errors-in-variables model. *Annals of Statistics*, 27(5):1519–1535, 1999.

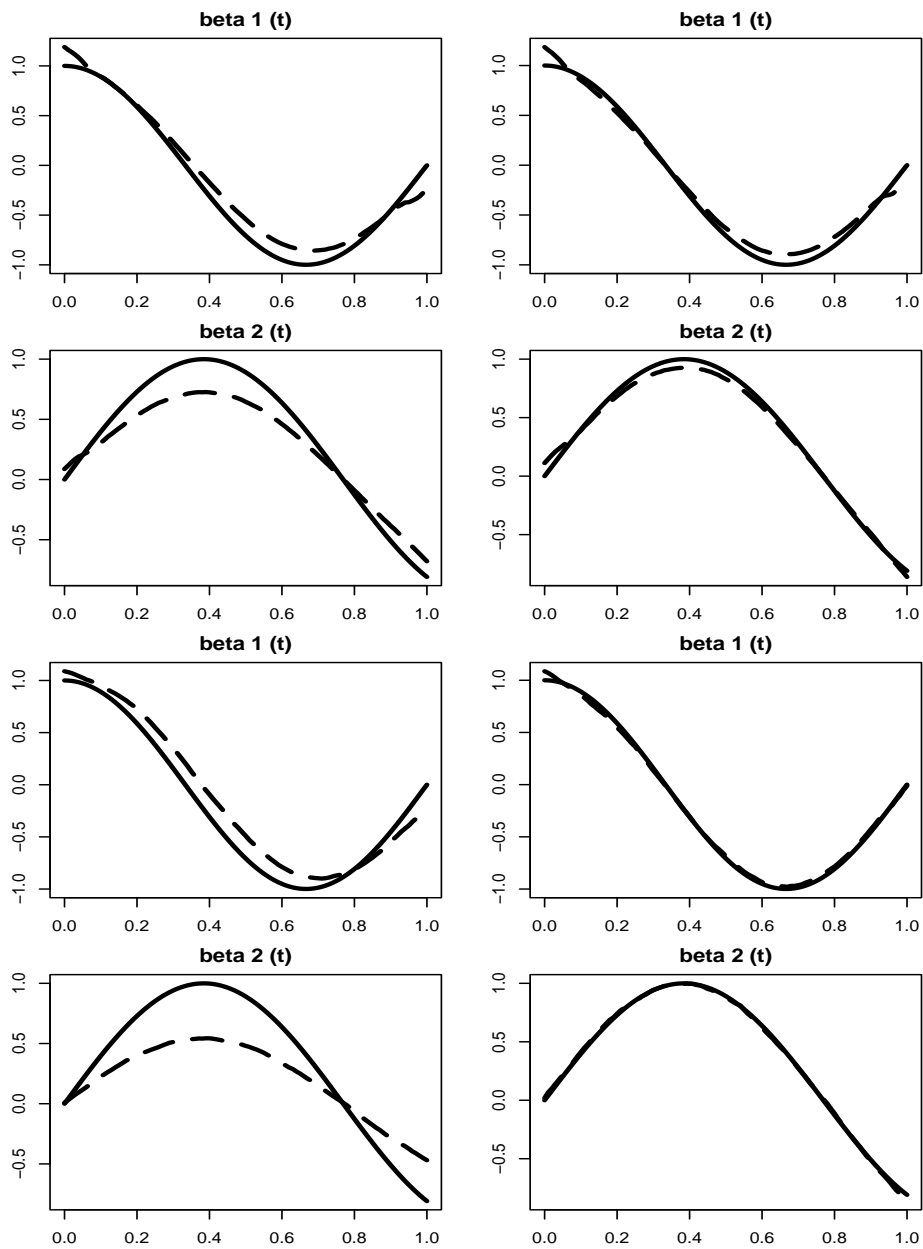


Figure 1: The median estimated curves. The four plots on the top are from simulation B; the other four are from simulation D. Naive estimators are on the left, the corresponding error corrected ones are on the right. Solid curves are true coefficients; dashed curves are estimated coefficients. Simulations A and C produce similar results and were omitted for simplicity.

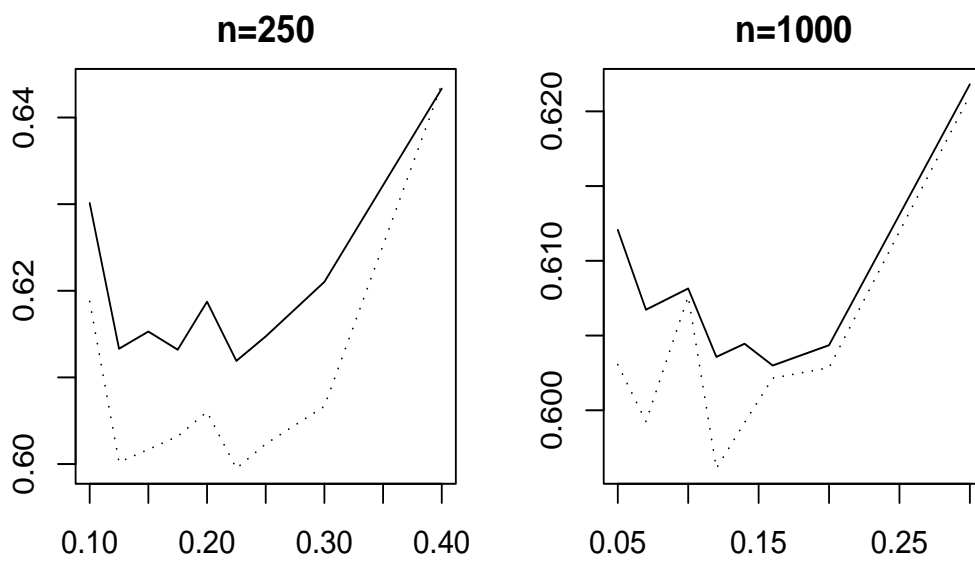


Figure 2: Simulation on bandwidth selection. The curves are median over 200 simulation runs evaluated at selected h . The solid lines are GCV functions calculated using the true covariate; the dotted lines are the proposed EGCV functions incorporating error correction.

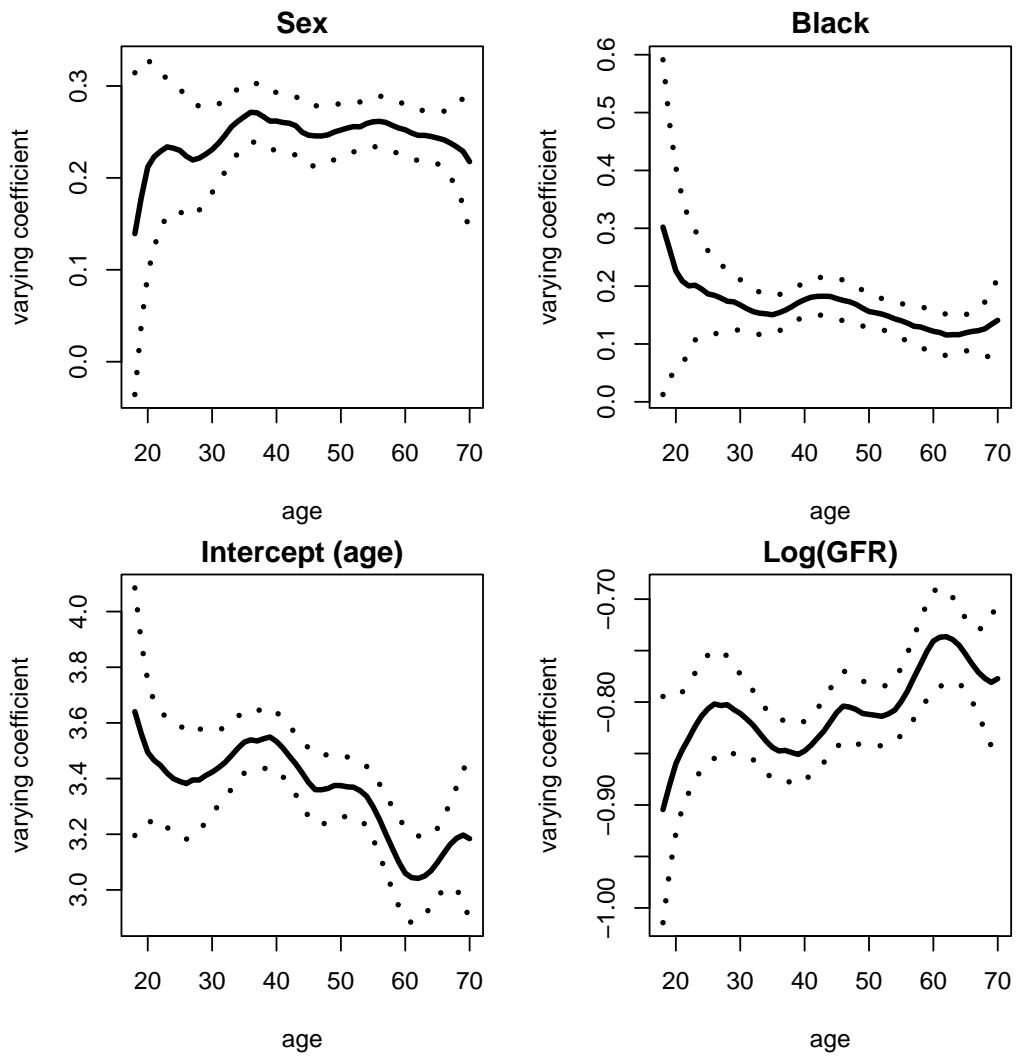


Figure 3: Estimated varying coefficients from the CKD data, with error correction. The solid lines are estimates; the dotted lines are pointwise 95% confidence intervals.

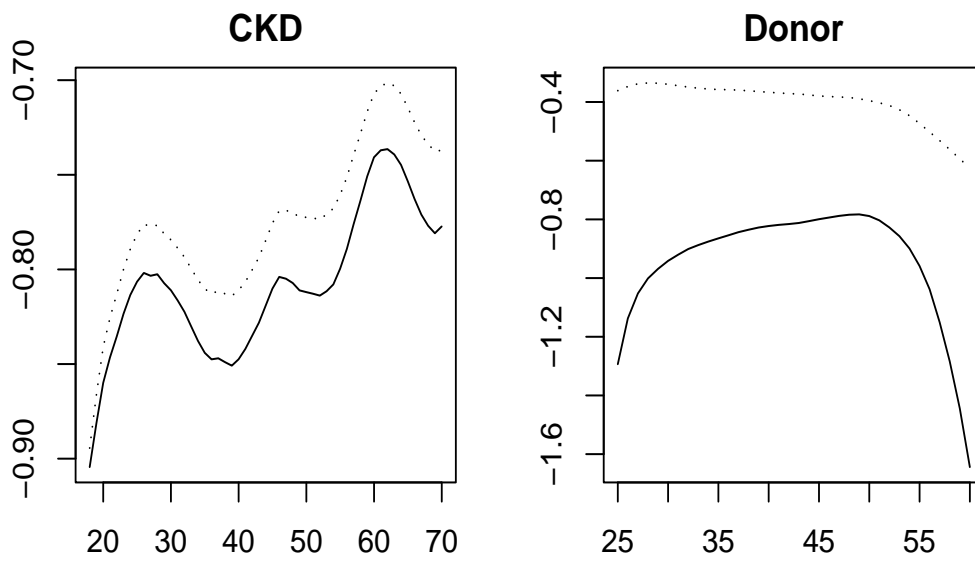


Figure 4: Naive (dotted line) and error corrected (solid line) coefficient functions of $\log(\text{GFR})$ for CKD and Donor data sets.

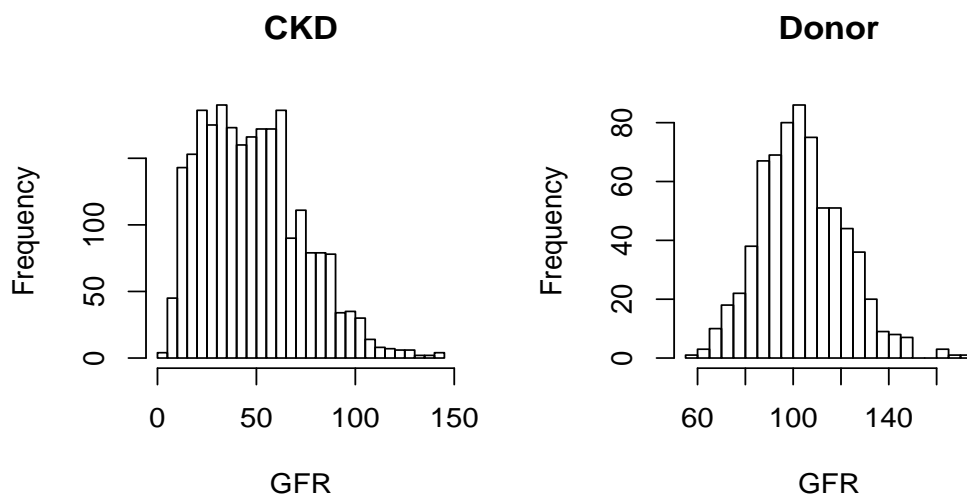


Figure 5: Distribution of GFR (ml/min/1.73m²) in CKD and Donor data sets.

Table 1: Results from simulations A-D. Median IMSEs under each scenario are presented. For the constant coefficient in simulation A: b bias; m square root of MSE

	method	$\hat{\beta}_1(t)$	$\hat{\beta}_2(t)$
A	naive	0.0498^b 0.0811^m	0.0323
	corrected	-0.0100^b 0.0713^m	0.0125
B	naive	0.0158	0.0335
	corrected	0.0117	0.0113
C	naive	0.00773	0.0228
	corrected	0.00402	0.00438
D	naive	0.0265	0.0868
	corrected	0.0135	0.0215