Nonparametric Function Estimation for Clustered Data When the Predictor is Measured Without/With Error

Xihong Lin and Raymond J. Carroll [*]

October 4, 1999

Abstract

We consider local polynomial kernel regression with a single covariate for clustered data using estimating equations. We assume that at most $m < \infty$ observations are available on each cluster. In the case of random regressors, with no measurement error in the predictor, we show that it is generally the best strategy to ignore entirely the correlation structure within each cluster, and instead to pretend that all observations are independent. In the further special case of longitudinal data on individuals with fixed common observation times, we show that equivalent to the pooled data approach is the strategy of fitting separate nonparametric regressions at each observation time and constructing an optimal weighted average. We also consider what happens when the predictor is measured with error. Using the SIMEX approach to correct for measurement error, we construct an asymptotic theory for both the pooled and weighted average estimators. Surprisingly, for the same amount of smoothing, the weighted average estimators typically have smaller variances than the pooling strategy. We apply the proposed methods to the analysis of the AIDS Costs and Services Utilization Survey.

KEY WORDS: AIDS; Asymptotic bias and variance; Clustered data; Efficiency; Errors in variables; Estimating equations; Generalized linear models; Kernel regression; Longitudinal data; Measurement error; Nonparametric regression; Panel data; SIMEX.

Short title. Nonparametric Regression for Clustered Data

[*] Xihong Lin is Associate Professor, Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109-2029. Her research was supported by a grant from the National Cancer Institute (CA–76404). Raymond J. Carroll is University Distinguished Professor, Departments of Statistics and Biostatistics & Epidemiology, Texas A&M University, College Station TX 77843–3143. His research was supported by a grant from National Cancer Institute (CA–57030), and by the Texas A&M Center for Environmental and Rural Health via a grant from the National Institute of Environmental Health Sciences (P30-ESO9106).

Nonparametric Function Estimation for Clustered Data When the Predictor is Measured Without/With Error

Abstract

We consider local polynomial kernel regression with a single covariate for clustered data using estimating equations. We assume that at most $m < \infty$ observations are available on each cluster. In the case of random regressors, with no measurement error in the predictor, we show that it is generally the best strategy to ignore entirely the correlation structure within each cluster, and instead to pretend that all observations are independent. In the further special case of longitudinal data on individuals with fixed common observation times, we show that equivalent to the pooled data approach is the strategy of fitting separate nonparametric regressions at each observation time and constructing an optimal weighted average. We also consider what happens when the predictor is measured with error. Using the SIMEX approach to correct for measurement error, we construct an asymptotic theory for both the pooled and weighted average estimators. Surprisingly, for the same amount of smoothing, the weighted average estimators typically have smaller variances than the pooling strategy. We apply the proposed methods to the analysis of the AIDS Costs and Services Utilization Survey.

KEY WORDS: AIDS; Asymptotic bias and variance; Clustered data; Efficiency; Errors in variables; Estimating equations; Generalized linear models; Kernel regression; Longitudinal data; Measurement error; Nonparametric regression; Panel data; SIMEX.

Short title. Nonparametric Regression for Clustered Data

1 INTRODUCTION

There is a vast literature developed in the past decade on parametric regression for clustered data using estimating equations (Liang and Zeger, 1986), where generalized linear models are a special case. Such parametric assumptions may not always be desirable, since appropriate functional forms of the covariates may not be known in advance and the outcome may depend on the covariates in a complicated manner. There has been substantial interest recently in extending the existing parametric models to allow for nonparametric covariate effects (Zeger and Diggle, 1994; Severini and Staniswalis, 1994; Wild and Yee, 1996). Such nonparametric regression allows for more flexible functional dependence of the outcome variable on the covariates and can also be used to investigate whether an appropriate parametric function can be developed to describe the data well.

Another complication in the analysis of clustered data is the presence of covariate measurement error. For example, it has been well documented in the literature that covariates such as blood pressure (Carroll, Ruppert and Stefanski, 1995) and CD4 count (Tsiatis, Degruttola, Wulfsohn, 1995) are often subject to measurement error. We consider in this paper data from the AIDS Costs and Services Utilization Survey (ACSUS) (Berk, et al., 1993). The ACSUS sampled 2487 subjects in 10 randomly selected US cities with highest AIDS rates. A series of six interviews were conducted for each respondent every three months from 1991 to 1992. A main outcome of interest was whether an interviewee had had hospital admissions (yes/no) during the past three months. The collected covariates included demographic variables, HIV status, CD4 count, and treatments.

A question of interest in this study is how CD4 count affects the risk of hospitalization. The analysis of this data set has two major complications. The first complication is that even though it is believed that a lower CD4 count is associated with a higher risk of hospitalization, the functional form of this relationship is not known. We are interested in whether the relationship is simply linear, or whether there is a change point, or whether the relationship has a complex form. The second complication is that CD4 count was measured with error. One source of error came from its substantial variability, e.g., the coefficient of variation could be as large as 50% (Tsiatis, et al. 1995). The other source of error came from the fact that CD4 count was not measured at the time of each interview but the most recent CD4 count was abstracted from each respondent's medical record using his/her usual source of care. In view of these complications, we are interested in modeling the effect of CD4 count nonparametrically and accounting for its measurement error. Our nonparametric approach allows us to model the relationship between hospitalization and CD4 count using a flexible function without restricting any particular functional form and to investigate whether we can identify a simple parametric function to capture this relationship. Another advantage is that nonparametric regression can often help recover unexpected patterns of the relationship.

We consider in this paper nonparametric regression estimation for clustered data with a single covariate using estimating equations when the covariate is measured accurately or with error. We estimate the nonparametric function using the local polynomial kernel methods and extend these methods to the measurement error case using the SIMEX method (Cook and Stefanski, 1994). We study the asymptotic biases and variances of the proposed estimators.

We develop two main striking results in this paper:

When the Covariate is Measured Accurately

Several authors have tried to account for within-correlation when constructing an estimator for the nonparametric function (Severini and Staniswalis, 1994; Wild and Yee, 1996; Verbyla, et al., 1999). We however show that it is generally the best strategy to ignore the correlation structure entirely, and pretend as if the data within a cluster were independent (i.e., the working independence model in GEE terminology). Furthermore, correctly specifying the correlation structure in estimating the nonparametric function in fact has adverse effects, i.e., it results in an asymptotically less efficient estimator. This result is dramatically different from the parametric regression situation for clustered data, where correctly specifying the correlation structure gives the most efficient estimators of regression coefficients (Liang and Zeger, 1986). While the result was a surprise to us, it may result from the local property of local polynomial estimation. As the bandwidth becomes smaller, the chance that correlated observations from the same cluster fall in the same bandwidth vanishes and the observations essentially behave independently.

"Panel Data" with Measurement Error

In "panel data", observations for different subjects are obtained at a series of common time points during a longitudinal followup. We show that it is preferable to fit separate functions to each time period and then combine the methods via weighted averaging, rather than try to perform a single measurement error analysis by pooling all the data from different panels. This result is also dramatically different from parametric measurement error regression, where pooled analysis gives an asymptotically more efficient estimator.

The paper is organized as follows. In Section 2, we introduce the model. In Section 3, we consider local polynomial methods for nonparametric regression in clustered data when the predictor is observed exactly. We study the asymptotic biases and variances of the local polynomial kernel estimators. Ruckstuhl, Welsh and Carroll (1999) have investigated this issue in the Gaussian model when the covariance structure of observations within a cluster is that of the usual one–way random effects analysis of variance model. One part of this paper consists of extending their work to generalized linear models, allowing for an arbitrary correlation structure and working correlation models. The results of the generalization are surprising to us and much in line with those of Ruckstuhl, et al. Specifically, we show that the asymptotically most efficient estimator of the nonparametric function is obtained by entirely ignoring the correlation within each cluster. This result has by the way been conjectured in the Gaussian case by Hoover, et al. (1998) and Wu, Chiang and Hoover (1998) and used as the basis for their methods.

Two methods emerge from our analysis. The first simply pools the data and runs a standard nonparametric regression analysis, possibly with weighting for variability. The second method applies to the "panel data" problem, in which case it makes sense to compute regression estimates separately for each time point, and form a weighted average of the resulting estimates. We show in Section 3 that the methods of pooling and weighted averaging yield asymptotically equivalent estimates.

In Section 4 we take up the issue of measurement error. We consider the behavior of the SIMEX methodology (Cook and Stefanski, 1994) for correcting measurement error, obtaining asymptotic theory for the pooling method and for the weighted average method. Surprisingly, the two methods are no longer asymptotically equivalent in the "panel data" context, where the weighted average method can have a smaller variance. In Section 5, we apply the proposed methods to the analysis of the ACSUS data, followed by discussion in Section 6.

2 THE MODEL

Suppose that the data consist of n clusters with the ith ($i = 1, \cdots, n$) cluster having m_i observations. Let Y_{ij} and (X_{ij}, W_{ij}) be the response variable, the true unobserved covariate and the observed X-related error-prone covariate of the jth ($j = 1, \cdots, m_i$) observation in the ith cluster, respectively. The observations within the same cluster might be correlated. Given the true covariate X_{ij}, the mean and variance of Y_{ij} are $E(Y_{ij}|X_{ij}) = \mu_{ij}$ and $\text{var}(Y_{ij}|X_{ij}) = \phi_j w_{ij}^{-1} V(\mu_{ij})$, where ϕ_j is a scale parameter, w_{ij} is a weight and $V(\cdot)$ is a variance function. The marginal mean μ_{ij} depends on X_{ij} through a known monotonic link function $\mu(\cdot)$:

$$\mu_{ij} = \mu\{\theta(X_{ij})\}, \tag{1}$$

where $\theta(\cdot)$ is an unknown smooth function and the link function $\mu(\cdot)$ is differentiable. Note that so far we have not specified a within-cluster correlation structure for the observations Y_{ij}.

The model is completed by assuming the unobserved covariate X_{ij} is related to the observed covariate W_{ij} by an additive measurement error model

$$W_{ij} = X_{ij} + U_{ij}, \tag{2}$$

where U_{ij} is a measurement error and $\mathbf{U}_i = (U_{i1}, \cdots, U_{im_i})^T$ follows Normal$(0, \boldsymbol{\Sigma}_{i,uu})$. Note that we have not assumed a distribution for the X_{ij} and they may be correlated within the same cluster.

In some examples, the index j takes on a special meaning. For example, there could be $j = 1, ..., m$ sampling times at which an individual is measured, e.g., in a panel study, or j could refer to a family member, e.g., mother, daughter, etc. With some abuse of terminology, we call such situations "panel data" problems. In this special case, it makes sense to distinguish among the values of j, e.g., allowing different scale parameters, different density functions for the X's, or even different measurement error variances. Outside of this special case, with no meaning attached to j, it makes more sense to let the scale parameters, densities, etc. be independent of j. In what follows, we do our calculations as if there was special meaning attached to j, but all calculations cover the general case.

3 ESTIMATION WHEN THERE IS NO MEASUREMENT ERROR

3.1 Local Polynomial Kernel Estimators

For independent data, local polynomial kernel smoothing has been widely used in nonparametric regression. We now extend local polynomial kernel smoothing to model (1) for clustered data. To motivate the estimating equations for the kernel estimators of the nonparametric function $\theta(\cdot)$, we first consider estimating equations when $\theta(\cdot)$ a parametric pth polynomial function $\theta(\cdot) = \boldsymbol{G}_p(\cdot)^T\boldsymbol{\beta}$, where $\boldsymbol{G}_p(z) = (1, z, \cdots, z^p)^T$ and $\boldsymbol{\beta} = (\beta_0, \cdots, \beta_p)^T$. Let $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{im_i})^T$ and $\boldsymbol{G}_{ip} = \{\boldsymbol{G}_p(X_{i1}), \cdots, \boldsymbol{G}_p(X_{im_i})\}^T$. The regression coefficients $\boldsymbol{\beta}$ can be estimated using the conventional generalized estimating equations (GEEs) (Liang and Zeger, 1986)

$$\sum_{i=1}^{n} \mathbf{G}_{ip}^T \boldsymbol{\Delta}_i \mathbf{V}_i^{-1}(\mathbf{Y}_i - \boldsymbol{\mu}_i) = 0, \tag{3}$$

where $\boldsymbol{\mu}_i = E(\mathbf{Y}_i)$ with its jth component $\mu_{ij} = \mu\{\boldsymbol{G}_p^T(X_{ij})\boldsymbol{\beta}\}$, $\boldsymbol{\Delta}_i = \text{diag}\left[\mu^{(1)}\{\boldsymbol{G}_p^T(X_{ij})\boldsymbol{\beta}\}\right]$, $\mu^{(1)}(\cdot)$ is the first derivative of $\mu(\cdot)$, $\mathbf{V}_i = \mathbf{S}_i^{1/2}\mathbf{R}_i(\boldsymbol{\delta})\mathbf{S}_i^{1/2}$, $\mathbf{S}_i = \text{diag}[\phi_j w_{ij}^{-1} V\{\mu_{ij}\}]$ and $\mathbf{R}_i$ is an invertible working correlation matrix, possibly depending a parameter vector $\boldsymbol{\delta}$, which can be estimated using the method of moments. Liang and Zeger (1986) showed that the GEE estimator $\widehat{\boldsymbol{\beta}}$ is asymptotically consistent if the mean function μ_{ij} is correctly specified even when the working correlation matrix $\mathbf{R}_i$ is misspecified. The most efficient estimator of $\boldsymbol{\beta}$ is obtained by correctly specifying $\mathbf{R}_i$.

We now consider how to extend the parametric GEE (3) to model (1) when $\theta(\cdot)$ is a nonparametric function using the kernel method. In what follows, the order of the local polynomial is p, the bandwidth is h, and the symmetric kernel density function is $K(\cdot)$, normalized without loss of generality to have unit variance. Let $K_h(v) = h^{-1}K(v/h)$. The idea is to approximate $\theta(\cdot)$ at any given x using a local polynomial satisfying $\theta(X) = \{\boldsymbol{G}_p(X - x)\}^T\boldsymbol{\beta}$, where $\boldsymbol{G}_p(\cdot)$ and $\boldsymbol{\beta}$ were defined above. Having estimated $\boldsymbol{\beta}$ at x, the estimated $\theta(x)$ satisfies $\widehat{\theta}(x) = \widehat{\beta}_0$.

Let $\boldsymbol{G}_{ip}(x) = \{\boldsymbol{G}_p(X_{i1} - x), \cdots, \boldsymbol{G}_p(X_{im_i} - x)\}^T$. Kernel estimation of the nonparametric function $\theta(\cdot)$ at any given x requires incorporating the kernel weight function $K_h(\cdot)$ in GEE (3). Two ways are possible and they give two sets of kernel estimating equations for $\theta(x)$

$$\sum_{i=1}^{n} \boldsymbol{G}_{ip}(x)^T \boldsymbol{\Delta}_i(x)\mathbf{V}_i(x)^{-1}\mathbf{K}_{ih}(x)\{\mathbf{Y}_i - \boldsymbol{\mu}_i(x)\} = 0, \tag{4}$$

or

$$\sum_{i=1}^{n} \boldsymbol{G}_{ip}(x)^T \boldsymbol{\Delta}_i(x)\mathbf{K}_{ih}^{1/2}(x)\mathbf{V}_i^{-1}(x)\mathbf{K}_{ih}^{1/2}(x)\{\mathbf{Y}_i - \boldsymbol{\mu}_i(x)\} = 0, \tag{5}$$

where $\mathbf{K}_{ih}(x) = \mathrm{diag}\{K_h(X_{ij} - x)\}$, and $\{\boldsymbol{\mu}_i(x), \boldsymbol{\Delta}_i(x), \mathbf{V}_i(x), \mathbf{S}_i(x)\}$ are the same as those in (3) except that they are evaluated at $\mu_{ij}(x) = \mu\{\boldsymbol{G}_p^T(X_{ij} - x)\boldsymbol{\beta}\}$. The working correlation matrix $\mathbf{R}_i$ in $\mathbf{V}_i(x)$ may depend a parameter vector $\boldsymbol{\delta}$, which again can be estimated using the method of moments.

One can easily see that the two estimating equations (4) and (5) are often different except when $\mathbf{V}_i(x)$ is a diagonal matrix (assuming independence). Equation (5) weights the residuals $\{\mathbf{Y}_i - \boldsymbol{\mu}_i(x)\}$ symmetrically, while equation (4) does not. They hence often give different estimators of $\theta(x)$. We denote by $\widehat{\theta}_p(x;h)$ the local pth order kernel estimator using (4) and by $\widehat{\theta}_p^*(x;h)$ the local pth order kernel estimator using (5). These two estimators are identical when $\mathbf{V}_i(x)$ is a diagonal matrix. We will show in Sections 3.2-3.3 that the symmetric property of (4) and (5) results in different asymptotic properties of $\widehat{\theta}_p(x;h)$ and $\widehat{\theta}_p^*(x;h)$.

We have allowed the scale parameters ϕ_j to depend on j. In many problems it is reasonable to suppose that they do not depend on j, then we can set $\mathbf{S}_i(x) = \mathrm{diag}[w_{ij}^{-1}V\{\mu_{ij}(x)\}]$. If the ϕ_j do depend on j, then they will have to be estimated, again by the method of moments.

Application of the Fisher-scoring algorithm to equation (4) shows that the estimator $\widehat{\boldsymbol{\beta}}$ can be updated using iteratively reweighted least squares:

$$\left[\sum_{i=1}^{n} \boldsymbol{G}_{ip}(x)^T \boldsymbol{C}_i(x) \boldsymbol{G}_{ip}(x)\right] \widehat{\boldsymbol{\beta}} = \sum_{i=1}^{n} \boldsymbol{G}_{ip}(x)^T \boldsymbol{C}_i(x) \boldsymbol{y}_i, \tag{6}$$

where $\boldsymbol{C}_i(x) = \boldsymbol{\Delta}_i(x)\mathbf{V}_i(x)^{-1}\mathbf{K}_{ih}(x)\boldsymbol{\Delta}_i(x)$ is a working weight matrix and $\boldsymbol{y}_i = \boldsymbol{G}_{ip}(x)^T\boldsymbol{\beta} + \boldsymbol{\Delta}_i^{-1}(x)\{\mathbf{Y}_i - \boldsymbol{\mu}_i(x)\}$ is a working vector. The variance of $\widehat{\theta}_p(x;h)$ is equal to $\mathrm{var}\{\widehat{\beta}_0(x)\}$ and can be estimated using a sandwich estimator, which takes the form $\mathrm{cov}\{\widehat{\theta}_p(x;h)\} = \boldsymbol{e}^T\boldsymbol{\Omega}_1^{-1}\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^{-1}\boldsymbol{e}$, where $\boldsymbol{e} = (1, 0, \cdots, 0)^T$, and

$$\begin{aligned}
\boldsymbol{\Omega}_1 &= \sum_{i=1}^{n} \boldsymbol{G}_{ip}(x)^T \boldsymbol{\Delta}_i(x)\mathbf{V}_i^{-1}(x)\mathbf{K}_{ih}(x)\boldsymbol{\Delta}_i(x)\boldsymbol{G}_{ip}(x)\\
\boldsymbol{\Omega}_2 &= \sum_{i=1}^{n} \boldsymbol{G}_{ip}(x)^T \boldsymbol{\Delta}_i(x)\mathbf{V}_i^{-1}(x)\mathbf{K}_{ih}(x)\{\mathbf{Y}_i - \boldsymbol{\mu}_i(x)\}\{\mathbf{Y}_i - \boldsymbol{\mu}_i(x)\}^T\mathbf{K}_{ih}(x)\mathbf{V}_i^{-1}(x)\boldsymbol{\Delta}_i(x)\boldsymbol{G}_{ip}(x).
\end{aligned}$$

A similar Fisher scoring algorithm can be constructed to solve (5) for $\widehat{\theta}_p^*(x;h)$ and to calculate its variance. Specifically, one simply replaces $\mathbf{V}_i(x)\mathbf{K}_{ih}(x)$ by $\mathbf{K}_{ih}^{1/2}(x)\mathbf{V}_i(x)\mathbf{K}_{ih}^{1/2}(x)$ in $(\boldsymbol{C}_i, \boldsymbol{\Omega}_1, \boldsymbol{\Omega}_2)$.

Some versions of (4) have been proposed previously. There are three obvious choices: (a) let $\mathbf{R}_i$ be an estimator of the actual correlation matrix; (b) let $\mathbf{V}_i^{-1}$ be the diagonal values of the inverse of the covariance matrix of $\mathbf{Y}_i$; and (c) let $\mathbf{R}_i$ be the identity matrix, thus effectively ignoring the correlation structure within clusters. We call method (c) the *weighted pooled estimator*. Method (a) was proposed by Severini and Staniswalis (1994) in their equation (18) for average kernel $p = 0$. Method (b) is a generalization, from the Gaussian case to generalized linear models, of the modified quasilikelihood proposal by Ruckstuhl, et al. (1999). Method (c) is a generalization and modification, from the Gaussian case to generalized linear models, of the "pooled" method of Ruckstuhl, et al. (1999), allowing for different values of ϕ depending on the value of j. We will consider in Section 3.5 another estimator called "the weighted average estimator."

Ruckstuhl, et al. (1999) considered (4) for Gaussian data, i.e., $w_{ij} \equiv 1$ and $V(\mu_{ij}) \equiv 1$, They showed that under the simple variance component model $\mathbf{V}_i = \phi\mathbf{I} + \delta\boldsymbol{J}$, where $\mathbf{I}$ is an identity matrix and $\boldsymbol{J}$ is a matrix of ones, when $m_i \equiv m$, $p = 1$ so that local linear regression is employed, and the X_{ij}'s are independent and identically distributed, methods (b) and (c) are asymptotically equivalent and have uniformly smaller asymptotic mean squared error than method (a). Methods (b) and (c) can also be shown to have faster rates of convergence for local quadratics, $p = 2$.

We study in the next two sections the asymptotic biases and variances of the general kernel estimators $\widehat{\theta}_p(x;h)$ and $\widehat{\theta}_p^*(x;h)$ under the kernel GEEs (4) and (5). This investigation will allow us to compare the asymptotic performance of methods (a)-(c) and to identify an *optimal* working correlation matrix $\mathbf{R}_i$. Our main conclusions from the asymptotic analyses are:

(1) The two kernel estimators $\widehat{\theta}_p(x;h)$ and $\widehat{\theta}^*_p(x;h)$ often have different asymptotic properties and the asymptotic properties of $\widehat{\theta}_p(x;h)$ is much harder to study;

(2) Unlike the parametric GEE estimator in (3), if $\theta(x)$ is a nonparametric function, the asymptotically most efficient estimators of both $\widehat{\theta}_p(x;h)$ and $\widehat{\theta}^*_p(x;h)$ are obtained when ignoring the within-cluster correlation entirely, i.e., assuming working independence $\mathbf{R}_i = \mathbf{I}_i$. Correctly specifying the correlation matrix in fact results in an asymptotically <u>**less**</u> efficient estimator of $\theta(x)$.

3.2 Asymptotic Theory for the Kernel Estimator $\widehat{\theta}_p(x;h)$ from (4)

Asymptotic bias and variance analysis of $\widehat{\theta}_p(x;h)$ under (4) is often difficult for general local pth polynomial estimation, a general working correlation matrix $\mathbf{R}_i$ and non–Gaussian data. Hence, for general working correlation matrix $\mathbf{R}_i$, we first focus on average kernel estimation ($p = 0$) for both Gaussian and non–Gaussian data (Theorem 1), and then study local linear kernel estimation ($p = 1$) for Gaussian data (Theorem 2). If working independence $\mathbf{R}_i = \mathbf{I}$ is assumed, asymptotic bias and variance analysis of general local pth polynomial estimation for both Gaussian and non–Gaussian data is simple and is presented in Theorem 3.

In what follows, let $m_i = m < \infty$. We allow $\mathbf{X}_i = (X_{i1}, \cdots, X_{im})^T$ to be correlated unless stated otherwise, and denote by $f_j(.)$ the marginal density of X_{ij}. We further assume that the $(\mathbf{Y}_i, \mathbf{X}_i)$ $(i = 1, \cdots, n)$ are independent and identically distributed pairs, and $\mathbf{V}_i(\boldsymbol{\mu}_i, \boldsymbol{\delta}) = \mathbf{V}(\boldsymbol{\mu}_i, \boldsymbol{\delta})$. Denote by $g^{(r)}(.)$ the rth derivative of $g(.)$, and by v^{jk} the (j,k)th element of $\mathbf{V}^{-1}(.)$. Let $c_K(r) = \int z^r K(z)dz$, with $c_K(0) = c_K(2) = 1$, $\gamma_K(r) = \int z^r K^2(z)dz$, $\mathbf{E}_c(L) = \{c_K(L), c_K(L+1), ..., c_K(L+p)\}^T$, $\mathbf{E}_p(c)$ and $\mathbf{E}_p(\gamma)$ the $(p+1) \times (p+1)$ matrices with (j,k) element $c_K(j+k-2)$ and $\gamma_K(j+k-2)$, respectively. We further assume $nh \to \infty$ as $n \to \infty$ and $h \to 0$.

Theorem 1 *Let $\widehat{\theta}_0(x;h)$ be the solution of (4) for $p = 0$ and for any given weight matrix $\mathbf{V}_i$.*

(1) The asymptotic bias and variance of $\widehat{\theta}_0(x;h)$ are given by

$$\begin{aligned} bias\{\widehat{\theta}_0(x;h)\} &\approx h^2 \left\{ \theta^{(1)}(x) \frac{\sum_{j=1}^m v^{j\cdot}(x) f_j^{(1)}(x)}{\sum_{j=1}^m v^{j\cdot}(x) f_j(x)} + \frac{\theta^{(2)}(x)}{2} \right\}; \\ var\{\widehat{\theta}_0(x;h)\} &\approx \frac{\gamma_K(0)}{nh} \frac{\sum_{j=1}^m \{v^{j\cdot}(x)\}^2 \sigma_{jj}(x) f_j(x)}{[\mu^{(1)}\{\theta(x)\}]^2 \left\{\sum_{j=1}^m v^{j\cdot}(x) f_j(x)\right\}^2}, \end{aligned}$$

respectively, where $\sigma_{jj}(x) = var(Y_{ij}|X_{ij} = x) = w_j^{-1}\phi_j V[\mu\{\theta(x)\}]$, and $v^{j\cdot}(x) = \sum_{l=1}^m v^{jl}(x)$. If $f_j(.) = f(.)$, the bias of $\widehat{\theta}_0(x;h)$ is free of $\mathbf{V}$.

(2) The asymptotic variance of $\widehat{\theta}_0(x;h)$ is minimized when one assumes the working correlation matrix $\mathbf{R} = \mathbf{I}$ (independence), and is equal to

$$min_{\mathbf{V}}\left[\{var\{\widehat{\theta}_0(x;h)\}\right] \approx \{\gamma_K(0)/nh\}\left(\left[\mu^{(1)}\{\theta(x)\}\right]^2\sum_{j=1}^{m}\{f_j(x)/\sigma_{jj}(x)\}\right)^{-1}.$$

The proof of Theorem 1 is given in Appendix A.1. We discuss the implication of Theorem 1 after presenting Theorem 2. For linear kernel estimation ($p = 1$), it is difficult to study asymptotic properties for general $\mathbf{V}$ and non–Gaussian data. This is because for any given weight matrix $\mathbf{V}$, asymptotic bias and variance analysis depends on the forms of $\mu(.)$ and $V(.)$. We hence concentrate on the Gaussian case and study in Theorem 2 its asymptotic bias and variance. The proof of Theorem 2 is given in Appendix A.2.

Theorem 2 *Let $\widehat{\theta}_{1,G}(x;h)$ be the solution of (4) for Gaussian data with $V(\cdot) = 1$, $w_{ij} = 1$, and $p = 1$ and any given weight matrix $\mathbf{V}$.*

(1) The asymptotic bias and variance of $\widehat{\theta}_{1,G}(x;h)$ are $h^2\theta^{(2)}(x)/2$ and $c/(nh)$, respectively, where the expression of c is complicated and is given in Appendix A.2 and Ruckstuhl, et al. (1999). Note that the asymptotic bias of $\widehat{\theta}_{1,G}(x;h)$ is free of the distribution of the X_{ij} and $\mathbf{V}$.

(2) If the $\mathbf{X}_{ij}$ are independent and identically with common density $f(.)$, the asymptotic variance of $\widehat{\theta}_{1,G}(x;h)$ is minimized when one assumes the working correlation matrix $\mathbf{R} = \mathbf{I}$ (independence), and is equal to

$$min_{\mathbf{V}}\left[var\{\widehat{\theta}_{1,G}(x)\}\right] \approx \{\gamma_K(0)/nh\}\left[f(x)\sum_{j=1}^{m}\{1/\sigma_{jj}\}\right]^{-1}.$$

where $\sigma_{jj} = var(Y_{ij}|X_{ij} = x) = \phi_j$.

Parts (2) in Theorems 1 and 2 are the most important results. They suggest that at least for average kernel estimation $p = 0$ (Gaussian and non–Gaussian data) and for local linear kernel estimation $p = 1$ (Gaussian), it is optimal to simply assume independence for kernel regression using (4) for clustered data and method (c) dominates methods (a) and (b). In other words, the asymptotically most efficient estimators $\hat{\theta}_0(x;h)$ and $\hat{\theta}_{1,G}(x)$ are obtained by completely ignoring the within-cluster correlation and correctly specifying the correlation results in less efficient estimators.

Study of the asymptotic properties of general local pth polynomial estimation under a general working correlation matrix $\mathbf{R}$ in (4) is difficult, even for Gaussian data. However, such calculations are possible when assuming independence $\mathbf{R} = \mathbf{I}$, i.e., for the weighted pooled estimator (method c). These results are stated in Theorem 3, whose proof is given in Appendix A.3.

Theorem 3 *Let $\widehat{\theta}_{p,wpe}(x;h)$ be the weighted pooled estimator, i.e., the solution of (4) for any given p and $\mathbf{R} = \mathbf{I}$ (working independence).*

(1) The asymptotic bias of $\widehat{\theta}_{p,wpe}(x;h)$ is,

(i)if $p = 0$, $\quad bias\{\widehat{\theta}_{0,wpe}(x;h)\} \approx h^2 \left\{ \theta^{(1)}(x) \frac{\sum_{j=1}^m f_j^{(1)}(x)/\sigma_{jj}(x)}{\sum_{j=1}^m f_j(x)/\sigma_{jj}(x)} + \frac{\theta^{(2)}(x)}{2} \right\};$

(ii)if $p = odd$, $\quad bias\{\widehat{\theta}_{p,wpe}(x;h)\} = h^{p+1} \frac{\theta^{(p+1)}(x)}{(p+1)!} \boldsymbol{e}^T \mathbf{E}_p^{-1}(c) \mathbf{E}_c(p+1);$

(iii) if p*=even and* $p > 0$,

$$bias\{\widehat{\theta}_{p,wpe}(x;h)\} \approx h^{p+2} \left\{ \frac{\theta^{(p+1)}(x)}{(p+1)!} \frac{\sum_{j=1}^m \partial\{L_j(x) f_j(x)\}/\partial x}{\sum_{j=1}^m L_j(x) f_j(x)} + \frac{\theta^{(p+2)}(x)}{(p+2)!} \right\} \boldsymbol{e}^T \mathbf{E}_p^{-1}(c) \mathbf{E}_c(p+2);$$

where $L_j(x) = \left[\mu^{(1)}\{\theta(x)\}\right]^2 / \sigma_{jj}(x)$ *and* $\sigma_{jj}(x) = var(Y_{ij}|X_{ij} = x) = w_j^{-1} \phi_j V[\mu\{\theta(x)\}]$.

(2) The asymptotic variance of $\widehat{\theta}_{p,wpe}(x;h)$ *is*

$$var\{\widehat{\theta}_{p,wpe}(x;h)\} \approx \frac{\gamma_K(0)}{nh} \left(\left[\mu^{(1)}\{\theta(x)\}\right]^2 \sum_{j=1}^m f_j(x)/\sigma_{jj}(x) \right)^{-1} \boldsymbol{e}^T \mathbf{E}_p^{-1}(c) \mathbf{E}_p(\gamma) \mathbf{E}_p^{-1}(c) \boldsymbol{e}. \quad (7)$$

Using the results in Theorem 3, one can easily show that, for example, the asymptotic bias and variance of the weighted pooled local linear kernel estimator $\widehat{\theta}_{1,wpe}(x;h)$ are

$$\begin{aligned} \text{bias}\{\widehat{\theta}_{1,wpe}(x;h)\} &\approx h^2 \theta^{(2)}(x)/2 \\ \text{var}\{\widehat{\theta}_{1,wpe}(x)\} &\approx (\gamma_K(0)/nh) \left(\left[\mu^{(1)}\{\theta(x)\}\right]^2 \sum_{j=1}^m \{f_j(x)/\sigma_{jj}(x)\} \right)^{-1}. \end{aligned}$$

3.3 Asymptotic Theory of the Kernel Estimator $\widehat{\theta}_p^*(x;h)$ Using (5)

We study in Theorem 4 the asymptotic bias and variance of $\widehat{\theta}_p^*(x;h)$, the solution of the estimating equation (5), for a general local pth order polynomial and a general weight matrix $\mathbf{V}_i$ for both Gaussian and non–Gaussian cases. Unlike $\widehat{\theta}_p(x;h)$ whose bias and variance analysis under this general condition is difficult, such a general analysis is feasible for $\widehat{\theta}_p^*(x;h)$ and the results are much simpler and are different from those of $\widehat{\theta}_h(x;h)$. This is due to the symmetric nature of the estimating equation (5). These results allow us to easily study the *optimal* choice of the working correlation matrix $\mathbf{R}_i$.

The key result in Theorem 4 below is given in part (3), i.e., the asymptotically most efficient estimator $\hat{\theta}_p^*(x;h)$ is obtained by entirely ignoring the within-cluster correlation and assuming the data were independent.

Note that under working independence, the two kernel estimators $\hat{\theta}_p(x;h)$ and $\hat{\theta}_p^*(x;h)$ are identical and have the same asymptotic properties. The proof of Theorem 4 is given in Appendix A.4.

Theorem 4 *Suppose* $\int s^r K^{1/2}(s)ds < \infty$ *for integers* $r \leq p$. *Let* $\widehat{\theta}_p^*(x;h)$ *be the solution of (5) for any given* p *and any given weight matrix* $\mathbf{V}$.

(1) The asymptotic bias of $\widehat{\theta}_p^*(x;h)$ is,

$$(i) if\ p=0, \qquad bias\{\widehat{\theta}_0^*(x;h)\} \approx h^2 \left\{ \theta^{(1)}(x) \frac{\sum_{j=1}^m v^{jj}(x) f_j^{(1)}(x)}{\sum_{j=1}^m v^{jj}(x) f_j(x)} + \frac{\theta^{(2)}(x)}{2} \right\};$$

$$(ii) if\ p=odd, \qquad bias\{\widehat{\theta}_p^*(x;h)\} \approx h^{p+1} \frac{\theta^{(p+1)}(x)}{(p+1)!} \boldsymbol{e}^T \mathbf{E}_p^{-1}(c) \mathbf{E}_c(p+1),$$

Note that $bias\{\widehat{\theta}_p^*(x;h)\}$ for odd p is free of $\mathbf{V}_i$ and $f_j(x)$;

(iii) if p=even and $p>0$,

$$bias\{\widehat{\theta}_p^*(x;h)\} \approx h^{p+2} \left\{ \frac{\theta^{(p+1)}(x)}{(p+1)!} \frac{\sum_{j=1}^m \partial\{T_j(x) f_j(x)\}/\partial x}{\sum_{j=1}^m T_j(x) f_j(x)} + \frac{\theta^{(p+2)}(x)}{(p+2)!} \right\} \boldsymbol{e}^T \mathbf{E}_p^{-1}(c) \mathbf{E}_c(p+2);$$

where $T_j(x) = \left[\mu^{(1)}\{\theta(x)\}\right]^2 v^{jj}(x)$.

(2) The asymptotic variance of $\widehat{\theta}_p^*(x;h)$ is

$$var\{\widehat{\theta}_p^*(x;h)\} \approx \frac{\gamma_K(0)}{nh} \frac{\sum_{j=1}^m \{v^{jj}(x)\}^2 \sigma_{jj}(x) f_j(x)}{\left[\mu^{(1)}\{\theta(x)\}\right]^2 \left[\sum_{j=1}^m v^{jj}(x) f_j(x)\right]^2} \boldsymbol{e}^T \mathbf{E}_p^{-1}(c) \mathbf{E}_p(\gamma) \mathbf{E}_p^{-1}(c) \boldsymbol{e}.$$

(3) The asymptotic variance of $\widehat{\theta}_p^*(x;h)$ is minimized when one assumes the working correlation matrix $\mathbf{R}=\mathbf{I}$ (independence), and is given in equation (7).

Part (3) of Theorem 4 gives the most important results. It suggests that under estimating equation (5), for any given local pth order polynomial, the most efficient kernel estimator $\widehat{\theta}_p^*(x;h)$ is obtained by simply assuming independence for kernel regression and methods (c) dominates methods (a) and (b). It is also interesting to notice that unlike estimating equation (4), methods (a) and (b) behave the same asymptotically under estimating equation (5). If $\mathbf{R}=\mathbf{I}$, the results in Theorem 4 reduce to those in Theorem 3.

It is of interest to compare the asymptotic performance of $\widehat{\theta}_p(x;h)$ and $\widehat{\theta}_p^*(x;h)$ when a general weight matrix $\mathbf{V}$ is specified. Such a comparison is difficult for any given local pth order polynomial. We hence restrict our attention to average kernel estimation ($p=0$). The results in Theorem 1 and Theorem 4 suggest that $\widehat{\theta}_0^*(x;h)$ replaces $v^{j\cdot}(x)$ in the bias and variance expressions of $\widehat{\theta}_0(x;h)$ by $v^{jj}(x)$. Consider the case when $f_j(\cdot)=f(\cdot)$ and $\sigma_{jj}(x)=\sigma(x)$. If $\sum_{j=1}^m \{v^{jj}(x)\}^2 / \left\{\sum_{j=1}^m v^{jj}(x)\right\}^2 < \sum_{j=1}^m \{v^{j\cdot}(x)\}^2 / \left\{\sum_{j=1}^m v^{j\cdot}(x)\right\}^2$, it is better to use $\widehat{\theta}_0^*(x;h)$, which has a smaller variance. If $v^{jj}(x)$ and $v^{j\cdot}(x)$ do not depend on j, e.g., under exchangeable working correlation or $AR(q)$ working correlation assumption, $\widehat{\theta}_0(x;h)$ and $\widehat{\theta}_0^*(x;h)$ have the same asymptotic variance. Furthermore, when $\mathbf{V}$ is a diagonal matrix, e.g., under methods (b) and (c), $\widehat{\theta}_p(x;h) = \widehat{\theta}_p^*(x;h)$ and they have the same asymptotic properties.

3.4 Selection of the Bandwidth Parameter

An important step in kernel smoothing is to choose the bandwidth parameter h. One approach is to use cross-validation by deleting one cluster datum at a time and choose h to minimize

$$\mathrm{CV}(h) = \sum_{i=1}^{n} \sum_{j=1}^{m_i} \frac{\left\{ Y_{ij} - \widehat{\mu}_{ij}^{(-i)}(X_{ij}) \right\}^2}{\widehat{\phi}_j w_{ij}^{-1} V \left\{ \widehat{\mu}_{ij}^{(-i)}(X_{ij}) \right\}},$$

where $\widehat{\mu}_{ij}^{(-i)}(.)$ is the estimate of $\mu_{ij}(.)$ calculated from the data leaving out the ith cluster. A difficulty in using cross-validation is that it is computationally intensive.

An alternative approach is to extend the Ruppert (1997) empirical bias bandwidth selection (EBBS) method to clustered data. Specifically, one calculates the empirical mean square errors $\mathrm{EMSE}(x, h)$ of $\widehat{\theta}(x; h)$ (either $\widehat{\theta}_p(x; h)$ or $\widehat{\theta}_p^*(x; h)$) at a series of values of x and h and chooses h to minimize $\mathrm{EMSE}(x, h)$ for each x. Calculations of $\mathrm{EMSE}(x_0, h_0)$ at any given value of x_0 and h_0 proceed by $\mathrm{EMSE}(x_0, h_0) = \widehat{\mathrm{bias}}^2 \left\{ \widehat{\theta}(x_0; h_0) \right\} + \widehat{\mathrm{var}} \left\{ \widehat{\theta}(x_0; h_0) \right\}$. Here $\widehat{\mathrm{bias}} \left\{ \widehat{\theta}(x_0; h_0) \right\}$ denotes the empirical bias of $\widehat{\theta}(x_0; h_0)$ at x_0 and h_0 and is estimated by fitting a polynomial regression

$$E \left\{ \widehat{\theta}(x_0; h) \right\} = \nu_0 + \nu_1 h^{p+1} + \ldots + \nu_t h^{p+t} \tag{8}$$

using the "data" $\{h, \widehat{\theta}(x_0; h)\}$ in a neighborhood of h_0 for a given integer t (e.g., $t = 1$ or 2). The empirical bias $\widehat{\mathrm{bias}} \left\{ \widehat{\theta}(x_0; h_0) \right\}$ is calculated as the estimated value of $\nu_1 h_0^{p+1} + \ldots + \nu_t h_0^{p+t}$. The variance $\widehat{\mathrm{var}} \left\{ \widehat{\theta}(x_0; h_0) \right\}$ can be easily calculated using the sandwich estimator in Section 3.1. We will use this method to choose h when analyzing the ACSUS data in Section 5.

3.5 Summary of Nonparametric Regression for Clustered Data

Our results in Sections 3.2-3.3 suggest that it is the best strategy to use (4) or (5) with $\mathbf{R} = \mathbf{I}$, i.e., entirely ignoring the within-cluster correlation. The proposal is extremely easy to compute: simply pool the data and compute a standard local polynomial kernel estimator in GLIMs, with weights depending on the cluster if the scale parameters ϕ_j are not constant.

In the "panel data" problem with $m_i \equiv m$, there is another estimator which can be considered, namely to compute $\widehat{\theta}_j(x; h)$ based only on the (Y_{ij}, X_{ij}) for fixed j, and then construct an optimal weighted average of the resulting estimators, where the optimal weights are the reciprocal of the $\mathrm{var}\{\widehat{\theta}_j(x; h)\}$. We call such an estimator "the weighted average estimator." A simple generalization of the results of Ruckstuhl, et al. (1999) shows that this estimator is asymptotically equivalent to method (c), "the weighted pooled estimator." The key step in proving this result is to show that $\mathrm{cov}\left\{ \widehat{\theta}_j(x; h), \widehat{\theta}_{j'}(x; h) \right\} = O(n^{-1})$ $(j \neq j')$ is of smaller order compared to

$\mathrm{var}\Big\{\widehat{\theta}_j(x;h)\Big\} = O\{(nh)^{-1}\}$. In other words, for asymptotic arguments, the individual estimators $\widehat{\theta}_j(x;h)$ are independent.

It seems that the technique of constructing separate estimators and then pooling them could be complex because asymptotically the optimal weights depend on the density functions of X_{ij} for $j = 1, ..., m$, which must then be estimated separately. In practice, this is not really that great an issue since standard kernel methods allow estimation of variances (and hence weights) via such techniques as the sandwich method. As we show later, the extra complication in the no measurement error case of having to estimate weights can be worthwhile when there is measurement error, since the weighted average estimator is asymptotically more efficient than the weighted pooled estimator.

4 SIMEX LOCAL POLYNOMIAL ESTIMATION WHEN THERE IS MEASUREMENT ERROR

We discuss in this section extending the kernel methods in Section 3 to the case when the covariate X is measured with error under the additive measurement error model (2). We use the SIMEX method (Cook and Stefanski, 1994) to correct measurement error. The results in Section 3 show that when X is accurately measured, it is the best strategy to entirely ignore the correlation and assume independence when calculating the kernel estimator of $\theta(x)$. In view of this result, we hence propose to calculate the naive kernel estimator by assuming independence in the simulation step of the SIMEX method.

This approach leads to two SIMEX estimators of $\theta(x)$: the SIMEX weighted pooled estimator and the SIMEX weighted average estimator. The former calculates the naive weighted pooled estimators in the simulation step, while the latter calculates the naive weighted average estimators in the simulation step and can only be applied to the "panel data" case. The most interesting result we have found is that unlike in the no measurement error case, where the two estimators have the same asymptotic properties, the SIMEX weighted average estimator has a smaller asymptotic variance than the SIMEX weighted pooled estimator in the presence of measurement error. We describe in Section 4.1 local polynomial kernel estimation using SIMEX, and propose the SIMEX weighted pooled estimator, whose asymptotic properties are studied in Section 4.2. We discuss the SIMEX weighted average estimator in Section 4.3.

4.1 The SIMEX Kernel Estimator

The Simulation-Extrapolation (SIMEX) estimator was developed by Cook & Stefanski (1994). The idea behind the SIMEX method is most clearly seen in simple linear regression when the independent variable is subject to measurement error. Suppose the regression model is $E(Y|X) = \alpha + \beta X$ and that $W = X + U$, rather than X, is observed where U has mean zero and variance σ_u^2, and the measurement error variance σ_u^2 is known. It is well known that the ordinary least squares estimate of the slope from regressing Y on W converges to $\beta\ \sigma_x^2(\sigma_x^2 + \sigma_u^2)^{-1}$ where $\sigma_x^2 = \mathrm{var}(X)$.

For any fixed $\lambda > 0$, suppose one repeatedly "adds on," via simulation, additional error with mean zero and variance $\sigma_u^2\lambda$ to W, computes the ordinary least squares slope each time and then takes the average. This simulation estimator consistently estimates $g(\lambda) = \beta\sigma_x^2/\{\sigma_x^2 + \sigma_u^2(1+\lambda)\}$. Since, formally at least, $g(-1) = \beta$, the idea is to plot $g(\lambda)$ against $\lambda \geq 0$, fit a model to this plot, and then extrapolate back to $\lambda = -1$. Cook and Stefanski (1994) show that this procedure will yield a consistent estimate of β if one fits the model $g(\lambda) = \gamma_0 + \gamma_1(\gamma_2 + \lambda)^{-1}$.

The SIMEX estimator for nonparametric regression is constructed as follows. We discuss only the case that measurement error covariance matrices $\mathbf{\Sigma}_{i,uu}$ are known, and we will keep track of these variances by means of the shorthand "$\mathbf{\Sigma}_{uu}$". In practice, the $\mathbf{\Sigma}_{i,uu}$ will have to be estimated, but estimating such parameters occurs at a parametric rate faster than the rate of convergence of any nonparametric estimator. Thus the theory is unchanged by estimating $\mathbf{\Sigma}_{uu}$.

Fix $D > 0$ to be a large but finite integer (50–200 in practice), and consider estimation of $\theta(x)$ in (1). For $d = 1, \ldots, D$ and any $\lambda > 0$, let $(\epsilon_{ijd})_1^n$ be a set of independent standard normal random variables which are then transformed to have sample mean zero, variance one and to be uncorrelated with the Y's and the W's. Let $\mathbf{\Sigma}_{i,uu}^{1/2}$ be the matrix square root of $\mathbf{\Sigma}_{i,uu}$. Define $\{W_{i1d}(\lambda), ..., W_{im_id}(\lambda)\}^T = \{W_{i1}, ..., W_{i,m_i}\}^T + \lambda^{1/2}\mathbf{\Sigma}_{i,uu}^{1/2}\{\epsilon_{i1d}, ..., \epsilon_{im_id}\}^T$. We calculate the GEE kernel estimator, which solves either (4) or (5), from these simulated data and denote it by $\widehat{\theta}_d\{x, (1+\lambda)\mathbf{\Sigma}_{uu}\}$. The average of these estimates over $d = 1, \ldots, D$ is denoted by $\widehat{\theta}\{x, (1+\lambda)\mathbf{\Sigma}_{uu}\}$, We run the SIMEX algorithm with D simulation replications at each value λ in a finite set $\mathbf{\Lambda}$. We extrapolate $\widehat{\theta}\{x, (1+\lambda)\mathbf{\Sigma}_{uu}\}$ using a polynomial of order q_s back to $\lambda = -1$. This gives the SIMEX local polynomial estimator $\widehat{\theta}_{sx}(x)$.

In view of the results in Section 3, we propose to calculate the naive estimators $\widehat{\theta}_d\{x, (1+\lambda)\mathbf{\Sigma}_{uu}\}$ using the weighted pooled estimator by assuming independence of observations within a cluster. The resulting estimator is called the SIMEX weighted pooled estimator and is denoted by $\widehat{\theta}_{sx,wpe}(x)$.

4.2 Asymptotic Theory for the SIMEX Weighted Pooled Estimator

The SIMEX estimator has a more complex theory for the weighted pooled estimator than in the independent data case that $m_i \equiv 1$, because the marginal distributions of W_{ij} and the conditional distributions of X_{ij} given W_{ij} may depend on j, for example because the distributions of X_{ij} or the measurement error may depend on j. This means that the "naive" regression for Y_{ij} on W_{ij} ignoring measurement error may have a mean $\zeta_j(w, \mathbf{\Sigma}_{uu}) = E(Y_{ij}|W_{ij} = w)$ depending on j.

In the case that $m_i \equiv m$, the following is easily shown. Let $f_{jW}(\cdot, \mathbf{\Sigma}_{uu})$ be the marginal density of W_{ij}. Let $\phi_j(\mathbf{\Sigma}_{uu})$ be the limiting value of estimates of ϕ_j ignoring measurement error. Then the naive estimate of $\theta(w)$ converges to $\theta_N(w, \mathbf{\Sigma}_{uu})$ given by

$$\mu\{\theta_N(w, \mathbf{\Sigma}_{uu})\} = \{\sum_{j=1}^{m} \zeta_j(w, \mathbf{\Sigma}_{uu}) f_{jW}(w, \mathbf{\Sigma}_{uu})/\phi_j(\mathbf{\Sigma}_{uu})\}\{\sum_{j=1}^{m} f_{jW}(w, \mathbf{\Sigma}_{uu})/\phi_j(\mathbf{\Sigma}_{uu})\}^{-1}. \tag{9}$$

Let $\boldsymbol{s}(\lambda)$ be the (q_s+1)–vector with jth element λ^{j-1}, $\mathbf{E}_s$ be the $(q_s+1)\times(q_s+1)$ matrix whose elements are zero except that the first element equals one, and $\boldsymbol{c}^T(\Lambda) = \boldsymbol{s}(-1)^T\{\sum_{\lambda\in\mathbf{\Lambda}} \boldsymbol{s}(\lambda)\boldsymbol{s}^T(\lambda)\}^{-1}$. The results are unchanged, and the theory simplifies tremendously, if we assume that for each λ, the same bandwidth h_λ for all SIMEX replicates. In our theory, we also require that the polynomial extrapolation is exact, i.e., $\boldsymbol{c}^T(\mathbf{\Lambda})\sum_{\lambda\in\mathbf{\Lambda}} \theta_N\{x; (1+\lambda)\mathbf{\Sigma}_{uu}\}\boldsymbol{s}(\lambda) = \theta(x)$. Hence the extrapolation results in a consistent estimate of $\theta(x)$. This is only exactly true, of course, in special cases. The bias that results from the extrapolation changes only the bias expression in the results given below, but not the variance expression.

Let the SIMEX weighted pooled estimator at x be denoted by $\widehat{\theta}_{sx,wpe}(x)$. The naive weighted pooled estimator which ignores measurement error is given by $\widehat{\theta}_{N,wpe}(x)$. Finally, define

$$Q(w, \mathbf{\Sigma}_{uu}) = \frac{\sum_{j=1}^{m}\{\mathcal{U}_j(w, \mathbf{\Sigma}_{uu}) + \Gamma_j(w, \mathbf{\Sigma}_{uu})\} f_{jW}(w, \mathbf{\Sigma}_{uu})/\phi_j^2(\mathbf{\Sigma}_{uu})}{[\mu^{(1)}\{\theta_N(w, \mathbf{\Sigma}_{uu})\}\sum_{j=1}^{m} f_{jW}(w, \mathbf{\Sigma}_{uu})/\phi_j(\mathbf{\Sigma}_{uu})]^2},$$

where $\mathcal{U}_j(w, \mathbf{\Sigma}_{uu}) = [\zeta_j(w, \mathbf{\Sigma}_{uu}) - \mu\{\theta_N(w, \mathbf{\Sigma}_{uu})\}]^2$ and $\Gamma_j(w, \mathbf{\Sigma}_{uu}) = \mathrm{var}(Y_{ij}|W_{ij} = w)$. In Appendix A.5, we sketch an argument which gives the following approximate bias and variance expansions, assuming that the number of SIMEX replicates D is large. For simplicity, the bias expressions given below assume p is odd.

$$\begin{aligned}
\mathrm{bias}\{\widehat{\theta}_{N,wpe}(x)\} &\approx \theta_N(x, \mathbf{\Sigma}_{uu}) - \theta(x) + h_0^{p+1}\theta_N^{(p+1)}\{x, \mathbf{\Sigma}_{uu}\}\{\boldsymbol{e}^T\mathbf{E}_p^{-1}(c)\mathbf{E}_c(p+1)\}; \\
\mathrm{bias}\{\widehat{\theta}_{sx,wpe}(x)\} &\approx \frac{\boldsymbol{c}^T(\mathbf{\Lambda})}{(p+1)!}\sum_{\lambda\in\mathbf{\Lambda}} h_\lambda^{p+1}\theta_N^{(p+1)}\{x, (1+\lambda)\mathbf{\Sigma}_{uu}\}\boldsymbol{s}(\lambda)\{\boldsymbol{e}^T\mathbf{E}_p^{-1}(c)\mathbf{E}_c(p+1)\}; \quad (10) \\
\mathrm{var}\{\widehat{\theta}_{N,wpe}(x)\} &\approx (nh_0)^{-1}Q(x, \mathbf{\Sigma}_{uu})\left\{\boldsymbol{e}^T\mathbf{E}_p^{-1}(c)\mathbf{E}_p(\gamma)\mathbf{E}_p^{-1}(c)\boldsymbol{e}\right\} \quad (11) \\
\mathrm{var}\{\widehat{\theta}_{sx,wpe}(x)\} &\approx (nh_0)^{-1}Q(x, \mathbf{\Sigma}_{uu})\left\{\boldsymbol{e}^T\mathbf{E}_p^{-1}(c)\mathbf{E}_p(\gamma)\mathbf{E}_p^{-1}(c)\boldsymbol{e}\boldsymbol{c}^T(\mathbf{\Lambda})\mathbf{E}_s c(\mathbf{\Lambda})\right\}. \quad (12)
\end{aligned}$$

Equations (11)–(12) are the most surprising, because they say that the variance of the SIMEX estimate is asymptotically the same as if measurement error were ignored, but multiplied by the factor $\boldsymbol{c}^T(\mathbf{\Lambda})\mathbf{E}_s\boldsymbol{c}(\mathbf{\Lambda})$, a factor which is independent of the problem. Thus, we can easily compare the various extrapolants on the basis of variance. For instance, suppose that the set of possible values of $\Lambda = (0, .5, 1.0, 1.5, 2.0)$. Then direct calculation shows that use of the quadratic extrapolant leads to an estimator which is 9 times more variable than that based on the linear extrapolant, while the cubic extrapolant is 52 times more variable than the linear extrapolant. Of course, such increases in variance have to be balanced by decreases in bias, and it is our experience in other problems (Carroll, Maca and Ruppert, 1999) that the excess bias of the linear extrapolant is sufficiently large that many times the quadratic extrapolant is preferred in terms of mean squared error.

Variance estimation of the SIMEX regression function can be performed in two ways. First, one can use the sandwich formula described previously to estimate the variance for the naive estimator which ignored measurement error, and then multiply it by the factor $\boldsymbol{c}^T(\boldsymbol{\Lambda})\mathbf{E}_s c(\boldsymbol{\Lambda})$ in (12) to account for the extrapolation. An alternative method uses the sandwich formula and the SIMEX replicates, see Section 5.4 of Stefanski & Cook (1995).

4.3 The SIMEX Weighted Average Estimator

The weighted pooled estimator in Section 4.2 is applicable in great generality. In particular, different cluster sizes are easily accommodated, and it is not required that there be a natural ordering, so that the jth observation in one cluster is somehow linked with the j observation in any other cluster. However, when such a natural ordering exists, the fact is that the variance of the SIMEX weighted pooled estimator is inflated by the terms $\mathcal{U}_j(\cdot)$. These terms are an artifact, arising only because that while the regression of Y_{ij} on X_{ij} does not depend on j in the presence of measurement error, the regression of Y_{ij} on W_{ij} may exhibit such a dependence. It seems sensible therefore to explore circumstances under which less variable methods can be constructed.

One such circumstance occurs in the "panel data" problem with $m_i \equiv m$, e.g., in a panel study where subjects are observed at the same time points. In such a situation, one could instead estimate the regression function $\theta(x)$ separately using SIMEX for each of $j = 1, ..., m$, and then average the estimates using some weights. Since each SIMEX estimate is an approximately consistent estimate, this device should in principle help us avoid an artificial variance inflation. We term the resulting estimator the SIMEX weighted average estimator and denote it by $\widehat{\theta}_{sx,wae}(x)$.

To see how this might work, suppose that the bandwidths in the jth observation are h_λ, the same as for the weighted pooled estimator. Then applying (12) but for a single observation, the asymptotic variance in the jth observation of the SIMEX estimate $\widehat{\theta}_{sx,j}(x)$, is proportional to $(nh_0)^{-1}\Gamma_j(x, \boldsymbol{\Sigma}_{uu})\left\{[\mu^{(1)}\{\theta_j(x, \boldsymbol{\Sigma}_{uu})\}]^2 f_{jW}(x, \boldsymbol{\Sigma}_{uu})\right\}^{-1}$, where $\theta_j(x, \boldsymbol{\Sigma}_{uu}) = \mu^{-1}\{\xi_j(x, \boldsymbol{\Sigma}_{uu})\}$. The constant of proportionality being enclosed in brackets in (12). We construct the SIMEX weighted average estimator $\widehat{\theta}_{sx,wae}(x)$ as the optimal linear combination of the individual estimators as

$$\widehat{\theta}_{sx,wae}(x) = \sum_{j=1}^{m} \alpha_j \widehat{\theta}_{sx,j}(x), \tag{13}$$

where $\alpha_j \propto \left\{[\mu^{(1)}\{\theta_j(x, \mathbf{\Sigma}_{uu})\}]^2 f_{jW}(x, \mathbf{\Sigma}_{uu})\right\} \{\Gamma_j(x, \mathbf{\Sigma}_{uu})\}^{-1}$ and $\sum_{j=1}^{m} \alpha_j = 1$. Assuming the polynomial extrapolation is exact for each j, i.e., $\boldsymbol{c}^T(\mathbf{\Lambda}) \sum_{\lambda \in \mathbf{\Lambda}} \theta_j\{x; (1+\lambda)\mathbf{\Sigma}_{uu}\}\boldsymbol{s}(\lambda) = \theta(x)$, the asymptotic bias of $\widehat{\theta}_{sx,wae}(x)$ is

$$\text{bias}\{\widehat{\theta}_{sx,wae}(x)\} \approx \frac{\boldsymbol{c}^T(\mathbf{\Lambda})}{(p+1)!} \sum_{j=1}^{m} \sum_{\lambda \in \mathbf{\Lambda}} \alpha_j h_\lambda^{p+1} \theta_j^{(p+1)}\{x, (1+\lambda)\mathbf{\Sigma}_{uu}\}\boldsymbol{s}(\lambda)\{\boldsymbol{e}^T \mathbf{E}_p^{-1}(c)\mathbf{E}_c(p+1)\}. \tag{14}$$

It is difficult to compare its bias with the bias of the SIMEX weighted pooled estimator $\widehat{\theta}_{sx,wpe}(x)$. However, if $h_\lambda = h$, assuming the q_sth order polynomial extrapolation is exact for both $\widehat{\theta}_{sx,wpe}(x)$ and $\widehat{\theta}_{sx,wae}(x)$, equations (10) and (14) are identical and are simplified as

$$\text{bias}\{\widehat{\theta}_{sx,wpe}(x)\} = \text{bias}\{\widehat{\theta}_{sx,wae}(x)\} \approx \frac{h^{p+1}}{(p+1)!} \theta^{(p+1)}(x)\{\boldsymbol{e}^T \mathbf{E}_p^{-1}(c)\mathbf{E}_c(p+1)\}.$$

This means that the asymptotic bias of the SIMEX estimators $\widehat{\theta}_{sx,wpe}(x)$ and $\widehat{\theta}_{sx,wae}(x)$ is the same as that when X is observed.

The variance of the weighted average estimator $\widehat{\theta}_{sx,wae}(x)$ is proportional to

$$\text{var}\{\widehat{\theta}_{sx,wae}(x)\} \propto (nh_0)^{-1} \left\{ \sum_{j=1}^{m} \left[\mu^{(1)}\{\theta_j(x, \mathbf{\Sigma}_{uu})\}\right]^2 f_{jW}(x, \mathbf{\Sigma}_{uu}) \left[\Gamma_j(x, \mathbf{\Sigma}_{uu})\right]^{-1} \right\}^{-1}, \tag{15}$$

where again the constant of proportionality is enclosed in brackets in (12). The proof of equation (15) again has used the fact that the covariance $\text{cov}\{\widehat{\theta}_{sx,j}(x), \widehat{\theta}_{sx,j'}(x)\} = O(n^{-1})$ for $(j \neq j')$, which is of smaller order compared to $\text{var}\{\widehat{\theta}_{sx,j}(x)\} = O\{(nh_0)^{-1}\}$. In other words, the individual SIMEX estimates $\widehat{\theta}_{sx,j}(x)$ are independent asymptotically. In Appendix A.6, we show that the variance of the SIMEX weighted pooled estimator $\widehat{\theta}_{sx,wpe}(x)$ is greater than or equal to the variance of the SIMEX weighted average estimator $\widehat{\theta}_{sx,wae}(x)$. Of course, in the case that the distribution of (Y, W, X) is independent of j, the two expressions are equal.

Because of the complex nature of the bias expressions for SIMEX estimators, it is generally not possible to compare the SIMEX weighted pooled estimator and the SIMEX weighted average estimator in terms of mean squared error. However, when $h_\lambda = h$, such a comparison is possible and our calculations suggest that the latter should be used if there are major observed differences as a function of j in the regression functions.

Since the weights used to calculate the SIMEX weighted average estimate $\widehat{\theta}_{sx,wae}(x)$ depend on the unknown density functions $f_{iW}(x, \mathbf{\Sigma})$ and the unknown conditional variances $\Gamma_j(x, \mathbf{\Sigma}_{uu})$, it is difficult to calculate $\widehat{\theta}_{sx,wae}(x)$ using (13) in practice. We hence propose the following procedure, which yields an asymptotically equivalent estimate. For the dth simulated SIMEX data set, we first calculate the naive weighted average estimate $\widehat{\theta}_{N,wae,d}\{x,(1+\lambda)\mathbf{\Sigma}_{uu}\} = \sum_{j=1}^{m} \alpha_{jd}\widehat{\theta}_{N,jd}(x)\{x,(1+\lambda)\mathbf{\Sigma}_{uu}\}$, where $\widehat{\theta}_{N,jd}\{x,(1+\lambda)\mathbf{\Sigma}_{uu}\}$ is the naive kernel estimate using the simulated jth observation data $W_{ijd}(\lambda)$, and α_{jd} is the reciprocal of the variance of $\widehat{\theta}_{N,jd}\{x,(1+\lambda)\mathbf{\Sigma}_{uu}\}$ obtained from standard kernel regression, e.g., the sandwich estimate. We then calculate the average of these estimates over $d = 1,\cdots,D$ and extrapolate it back to $\lambda = -1$. To compute the variance of the resulting estimate, we only need to calculate the variance of the weighted average estimate $\widehat{\theta}_{N,wae,d}\{x,(1+\lambda)\mathbf{\Sigma}_{uu}\}$ using the sandwich method (see Section 3.1) and then apply the SIMEX standard error method of Stefanski and Cook (1995).

5 APPLICATION TO THE ACSUS DATA

We applied the proposed SIMEX local polynomial kernel method to analyzing the ACSUS data described in Section 1. Since the risk of hospitalization depends on various covariates, such as HIV status, treatments, race and gender, and we only allow a single covariate in model (1), we limited our analysis to a subset of homogeneous subjects. Specifically, we restricted our attention to 273 white male patients who were HIV positive at entry into the study and were treated with antiretroviral drugs. The study participants were interviewed about every three months for about 18 months and were asked whether they had had hospital admissions (yes/no) during the interviews. The question of main interest was how the CD4 counts affected the risk of hospitalization. The total number of observations was 1059 with each patient contributing from 1 to 6 observations over time. The major covariate of interest, CD4 count, ranged from 1 to 2131 and 90% of these patients had CD4 count less than 500. As discussed in Section 1, the CD4 counts were measured with error since the most recent CD4 counts prior to each interview were used and they were subject to substantial lab errors. Since the investigator does not know in what fashion the risk of hospitalization decreases with the CD4 counts and is interested in identifying the form of the functional dependence (see Section 1 for discussion), we would like to make such dependence as flexible as possible by assuming a nonparametric function in order to properly identifying the functional form. Note that the other covariates included interview time and age. Examination of the data suggests the dependence of the risk of hospitalization on time and age is slight. We hence did not include them in the model.

We fit model (1) using the logit link with a single covariate W defined as $W = \log(CD4/100)$, a transformation which reduces the marked skewness of CD4 counts. We assumed the measurement errors U_{ij} were independent and normally distributed with mean 0 and variance σ_u^2. However, W

itself is left–skew and so an assumption that X is normally distributed would be inappropriate. The power of the SIMEX idea is that no assumptions need be made about the distribution of X. To estimate the measurement error variance σ_u^2, one needs to have either a validation study or replicates of CD4 count measures. However, these were not available in the ACSUS and hence we were not able to estimate σ_u^2 using the the ACSUS. We hence conducted a sensitivity analysis by assuming σ_u^2 equal to 1/4 and 1/2 of the variance of W, i.e, assuming $\sigma_u^2 = 0.34$ and $\sigma_u^2 = 0.68$. Wulfsohn and Tsiatis (1995) estimated the measurement error variance of log(CD4) as $\sigma_u^2 = 0.39$ using data from a clinical trial conducted by Burroughs-Wellcome. Our assumption of $\sigma_u^2 = 0.34$ is similar to their finding. As in Wulfsohn and Tsiatis (1995), we assumed that the measurement errors were independent and the measurement error variance σ_u^2 was a constant. If we had validation or replication data we could of course assess the possibility of correlated measurement errors, additivity, constant measurement variance and whether a different transformation of CD4 counts is required by using the techniques of Nusser, et al. (1996) and Eckert, Carroll and Wang (1997).

Since different subjects had different numbers of observations, calculations of the weighted average estimate of $\theta(x)$ was difficult. We calculated the SIMEX weighted pooled estimate of $\theta(x)$, letting $\lambda = (0.0, 1.0, 1.5, 2.0)$. We used the empirical bias bandwidth selection (EBBS) method discussed in Section 2.4 to select the bandwidth parameter h for each simulated data set and assumed $t = 2$ in (8). We further treated σ_u^2 as fixed and known. We used a quadratic extrapolation function in the SIMEX procedure and calculated the standard errors of the SIMEX estimates $\widehat{\theta}_{sx,wpe}(x)$ using the standard error estimation method of Stefanski and Cook (1995). The SIMEX method was applied with $D = 100$. Analysis of each simulated data set including estimating the bandwidth parameter h took only 16 seconds on a SPARC Ultra.

Figures 1-3 plot the estimated $\theta(x)$ against $x = \log(CD4/100)$ with it 95% confidence intervals assuming $\sigma_u^2 = 0.0$ (naive estimate ignoring measurement error), $\sigma_u^2 = 0.34$ and 0.68 respectively. The results suggest that the risk of hospitalization decreases as the CD4 count increases, but not in a linear fashion. It decreases more quickly when CD4 count is relatively low ($CD4 < 14$, $\log(CD4/100) < -2$) or high ($CD4 > 100$, $\log(CD4/100) > 0$) and is fairly stable when CD4 count takes middle values, e.g., between 14 and 100. Ignoring measurement error clearly affects the estimated risk of hospitalization. The naive curve is attenuated towards 0 compared to the SIMEX curves, especially for small and large values of CD4 counts. As expected, an increase in the measurement error variance leads to more change in the SIMEX estimate.

As a further check on the results, instead of kernel regression we fit the model by smoothing splines with the GAM procedure in Splus by assuming independence for each simulated SIMEX data and calculated the SIMEX estimate of $\theta(x)$. The fitted model ignoring measurement error, as well as the two SIMEX fits, were well within accord with Figures 1–3.

To examine whether a simple parametric model can fit the data equally well as the nonparametric model, we fit a simple linear model and a quadratic model using the GEE method assuming working independence (Liang and Zeger, 1986) and calculated the SIMEX estimates to account for measurement error. For illustration, we compare in Figure 4 the SIMEX kernel estimate with the SIMEX linear and quadratic estimates when $\sigma_u^2 = 0.34$. Figure 4 shows that the SIMEX local polynomial kernel estimator seems to have nonlinearity detected by neither the linear model nor the quadratic model. To test whether this extra nonlinearity is simply a figment of noise, we fit a cubic model to the data. Table 1 shows the naive and the SIMEX regression coefficient estimates of the cubic model assuming $\sigma_u^2 = (0, 0.34, 0.68)$, along with 95% bootstrap confidence intervals based on 2000 bootstrap samples. The coefficient of the cubic term is marginally statistically significant in naive regression when measurement error is ignored, and they are statistically significant for both SIMEX analyses after accounting for measurement error.

6 DISCUSSION

We have discussed in this paper local polynomial kernel regression methods for clustered data in the absence/presence of measurement error. We have emphasized that our work is specific to the case of random regressors with a bounded number of observations per cluster, while the number of clusters becomes large. We developed two main results. First, in the absence of measurement error, methods based on ignoring within–cluster correlations generally improve upon methods which attempt to use these correlations. Furthermore, correctly specifying correlation in estimation results in an asymptotically less efficient estimator. This is mainly due to the fact that kernel methods, being local, then essentially act as if the data were independent. A referee suggested that one might gain additional insight into the explanation of this result by considering a sequence of models wherein the within-cluster correlation approach one as $n \to \infty$. It should be noted that our results in this paper assume the working covariance matrix $\mathbf{V}$ is invertible and they might not be applied directly to this situation. Our second main result is in the "panel data" context with measurement error, where it can be preferable to fit separate functions to each time period and then combine the methods via weighted averaging, rather than try to perform a single pooled measurement error analysis. For simplicity, we assume in this paper a single nonparametric function. We conjecture that our results are applicable to models involving several continuous nonparametric functions, e.g., in the generalized additive model context.

Our results may have implications outside the realm of kernel smoothing, for example to spline smoothing, because of the well–known "equivalent kernel" results of Silverman (1984). These results say that linear and cubic smoothing splines behave away from the boundary like a Nadaraya–Watson kernel regression estimator with a locally chosen bandwidth and a higher–order kernel. Using this equivalent kernel, our results on kernel smoothing suggest that even for splines, it may be more efficient statistically, and is certainly easier computationally, to ignore the correlation structure within clusters and simply compute a weighted smoothing spline for GLIM's with weights inversely proportional to the ϕ_j.

Our results thus may have a direct impact on recent very active developments in modeling longitudinal curve data using smoothing splines via a linear mixed effects model formulation (Brumback and Rice, 1998; Wang, 1998; Verbyla, et al, 1999). These authors account for the within-cluster correlation using random effects while estimating the nonparametric function using a smoothing spline. An advantage of this approach is that the smoothing spline estimators can be written as a linear combination of fixed effects and random effects and hence an enlarged linear mixed model can be used to fit a linear random effects smoothing spline model. Our results however show that the smoothing spline estimator obtained in this way could possibly be asymptotically less efficient compared to that obtained by ignoring correlation. These suggestions are of course all conjectures, based on an equivalence in the non–clustered framework between local polynomial estimation and smoothing spline estimation. However, it would appear important for smoothing spline methodologists to show explicitly that accounting for correlation within clusters is a worthwhile endeavor. We would not expect our results to apply to nonlinear random effects smoothing spline models, e.g., generalized additive mixed models (Lin and Zhang, 1999).

Our results do not of course apply to the time series context, where the predictors are the fixed observation times, with the number of such times converging to infinity. It is well–known that one can construct estimators which take advantage of the autocorrelation structure in this case (Hart, 1991) and the asymptotic variance of the estimator of the nonparametric function depends on the correlation function.

In view of the no measurement error results, we have considered in the measurement error case estimation of the nonparametric function using the SIMEX approach by ignoring the within-cluster correlation in calculating the naive kernel estimators in the simulation step. It is unclear whether this strategy is the best strategy, i.e., whether ignoring correlation yields the most efficient SIMEX estimator. More research is needed, although we expect the theory would be extremely difficult.

An advantage of the SIMEX method is that it makes no distributional assumption on the unobserved covariate X. It is clearly of substantial interest for future work to develop methods which allow for an assumed parametric distribution for X. It is known (in models without correlated responses) that correct specification of a distribution for X can allow substantial gains in efficiency (Carroll, et al., 1999), albeit at the price of a loss of robustness to misspecification of the distributions of X.

REFERENCES

Berk, M. L., Maffeo, C., and Schur, C. L. (1993), *Research Design and Analysis Objectives. AIDS Cost and Services Utilization Survey Report, No. 1.* Rockville, MD: Agency for Health Care Policy and Research.

Brumback, B. A. and Rice, J. A. (1998). "Smoothing Spline Models for the Analysis of Nested and Crossed Samples of Curves (with discussion)." *Journal of the American Statistical Association*, **93**, 961-994.

Carroll, R. J., Ruppert, D. and Stefanski, L. A. (1995), *Measurement Error in Nonlinear Models.* London: Chapman and Hall.

Carroll, R. J., Ruppert, D. and Welsh, A. (1998). "Local Estimating Equations," *Journal of the American Statistical Association*, **93**, 214–227.

Carroll, R. J., Maca, J. D., and Ruppert, D. (1999), "Nonparametric Estimation in the Presence of Measurement Error," *Biometrika*, to appear.

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.

Eckert, R. S., Carroll, R. J. and Wang, N. (1997). Transformations to additivity in measurement error models. *Biometrics*, **53**, 262-272.

Hart, J. D. (1991),"Kernel Regression Estimation with Time Series Errors," *Journal of the Royal Statistical Society, Series B*, **53**, 173-187.

Hoover, D. R., Rice, J. A., Wu, C. O. & Yang, Y. (1998). "Nonparametric Smoothing Estimates of Time–Varying Coefficient Models with Longitudinal Data," *Biometrika*, **85**, 809–822.

Liang, K. Y. and Zeger, S. L. (1986), "Longitudinal Data Analysis Using Generalized Linear Models," *Biometrika*, **73**, 13-22.

Lin, X. and Zhang, D. (1999). "Inference in Generalized Additive Mixed Models Using Smoothing Splines." *Journal of the Royal Statistical Society, Series B*, **61**, 381-400.

Nusser, S. M., Carriquiry, A. L., Dodd, K. W.and Fuller, W. A. (1996). "A Semiparametric Transformation Approach to Estimating Usual Daily Intake Distributions." *Journal of the American Statistical Association*, **91**, 1440-1449.

Ruckstuhl, A. F., Welsh, A. H. and Carroll, R. J. (1999). "Nonparametric Function Estimation of the Relationship Between Two Repeatedly Measured Variables," *Statistica Sinica*, to appear.

Ruppert, D. (1997). "Empirical-Bias Bandwidths for Local Polynomial Nonparametric Regression and Density Estimation," *Journal of the American Statistical Association*, **92**, 1049-1062.

Severini, T. A. and Staniswalis, J. G. (1994). 'Quasilikelihood Estimation in Semiparametric Models," *Journal of the American Statistical Association*, **89**, 501–511.

Silverman, B. (1984). "Spline Smoothing: the Equivalent Variable Kernel Method," *Annals of Statistics*, **12**, 898-916.

Stefanski, L. A. and Cook, J. R. (1995). "Simulation–Extrapolation: The Measurement Error Jackknife," *Journal of the American Statistical Association*, **90**, 1247–1256.

Tsiatis, A. A., Degruttola, V. and Wulfsohn, M. S. (1995). "Modeling the Relationship of Survival to Longitudinal Data Measured with Error. Applications to Survival and CD4 Counts in Patients with AIDS," *Journal of American Statistics Association* **90**, 27-37.

Verbyla, A. P., Cullis, B. R., Kenward, M. G. and Welham, S. J. (1999). "The Analysis of Designed Experiments and Longitudinal Data Using Smoothing Splines (with discussion)." *Applied Statistics*, **48**, 269-311.

Wang Y. (1998). "Mixed-Effects Smoothing Spline ANOVA. *Journal of the Royal Statistical Society B*, **60**, 159-174.

Wild, C. J. and Yee, T. W. (1996), "Additive Extensions to Generalized Estimating Equation Methods," *Journal of the Royal Statistical Society, Series B*, **58**, 711-725.

Wu, C. O., Chiang, C. T. & Hoover, D. R. (1998). "Asymptotic Confidence Regions for Kernel Smoothing of a Varying Coefficient Model with Longitudinal Data," *Journal of the American*

Statistical Association, **93**, 1388–1402.

Wulfsohn, M. S. and Tsiatis, A. A. (1997), “A Joint Model for Survival and Longitudinal Data Measured with Error,” *Biometrics*, **53**, 330-339.

Zeger, S. L. and Diggle, P. J. (1994), “Semi-parametric Models for Longitudinal Data With Application to CD4 Cell Numbers in HIV Seroconverters,” *Biometrics*, 50, 689-699. **93**, 710-719.

Appendix A THEORY FOR KERNEL METHODS

A.1 Proof of Theorem 1

For $p = 0$, a simple Taylor expansion of (4) shows that the its solution $\widehat{\beta}_0 = \widehat{\theta}_0(x;h)$ satisfies $\widehat{\theta}_0(x;h) - \theta(x) \approx B_n^{-1}A_n$, where

$$\begin{aligned} B_n &= n^{-1}\sum_{i=1}^{n} \mathbf{1}^T \mathbf{\Delta}_i(x)\mathbf{V}_i^{-1}(x)\mathbf{K}_{ih}(x)\mathbf{\Delta}_i(x)\mathbf{1} \\ A_n &= n^{-1}\sum_{i=1}^{n} \mathbf{1}^T \mathbf{\Delta}_i(x)\mathbf{V}_i^{-1}(x)\mathbf{K}_{ih}(x)\left[\mathbf{Y}_i - \mu\{\theta(x)\}\mathbf{1}\right], \end{aligned}$$

and $\mathbf{1}$ is an $m \times 1$ vector of ones. Let $B = \lim_{n\to\infty} B_n$. The asymptotic bias of $\widehat{\theta}_0(x;h)$ is $B^{-1}E(A_n)$ and the asymptotic variance of $\widehat{\theta}_0(x;h)$ is $\mathrm{var}(A_n)/B^2$.

Specifically, some calculations give

$$\begin{aligned} B &= E\left\{\sum_{j=1}^{m}\left[\mu^{(1)}\{\theta(x)\}\right]^2 v^{j\cdot}K_h(X_j - x)\right\} = \left[\mu^{(1)}\{\theta(x)\}\right]^2 \sum_{j=1}^{m} v^{j\cdot} f_j(x) + O(h) \\ E(A_n) &= E\left\{\sum_{j=1}^{m}\mu^{(1)}\{\theta(x)\}v^{j\cdot}K_h(X_j - x)\left[\mu\{\theta(X_j)\} - \mu\{\theta(x)\}\right]\right\} \\ &= h^2\left[\mu^{(1)}\{\theta(x)\}\right]^2 \sum_{j=1}^{m} v^{j\cdot}\left\{f_j^{(1)}(x)\theta^{(1)}(x) + f_j(x)\theta^{(2)}(x)/2\right\} + o(h^2) \\ \mathrm{var}(A_n) &\approx n^{-1}E\left\{\sum_{j=1}^{m}\sum_{l=1}^{m}\left[\mu^{(1)}\{\theta(x)\}\right]^2 v^{j\cdot}v^{l\cdot}\sigma_{jl}K_h(X_j - x)K_h(X_l - x)\right\} \\ &= \frac{\gamma_K(0)\left[\mu^{(1)}\{\theta(x)\}\right]^2}{nh}\sum_{j=1}^{m}\{v^{j\cdot}\}^2\sigma_{jj}f_j(x) + o\{(nh)^{-1}\}, \end{aligned}$$

where $\mathbf{\Sigma} = \mathrm{cov}(\mathbf{Y}_i)$ and σ_{jl} is the (j,l)th element of $\mathbf{\Sigma}$. Part (1) follows immediately. A direct application of the Cauchy-Schwartz inequality leads to part (2).

A.2 Proof of Theorem 2

For part (1), see Theorem 2 of Ruckstuhl et al. (1999). We here prove part (2). The results of Appendix A.3 of Ruckstuhl et al. (1999) show that when the X_j are independent with density $f_j(\cdot)$, the asymptotic variance of $\widehat{\theta}_{1,G}(x)$ is the first diagonal element of $\boldsymbol{B}^{-1}\text{cov}(\mathbf{A}_n)\left(\boldsymbol{B}^{-1}\right)^T$, where

$$\boldsymbol{B} = \begin{bmatrix} B_{00} & hB_{01} \\ h^{-1}B_{01} & B_{11} \end{bmatrix} \qquad \text{cov}(\mathbf{A}_n) \approx \frac{\gamma_0}{nh}\begin{bmatrix} A_{00} & h^{-1}A_{01} \\ h^{-1}A_{01} & h^{-2}A_{11} \end{bmatrix}.$$

Here

$$\begin{aligned} B_{00} &= \sum_{j=1}^m v^{j\cdot} f_j(x) & B_{01} &= \sum_{j=1}^m v^{j\cdot} f_j^{(1)}(x) \\ B_{10} &= \sum_{j=1}^m \sum_{l\neq j} v^{jl} E(X_l - x) f_j(x) & B_{11} &= \sum_{j=1}^m \sum_{l\neq j} v^{jl} E(X_l - x) f_j^{(1)}(x) + \sum_{j=1}^m v^{jj} f_j(x), \\ A_{00} &= \sum_{j=1}^m \{v^{j\cdot}\}^2 \sigma_{jj} f_j(x) & A_{01} &= \sum_{j=1}^m v^{j\cdot} \sigma_{jj} f_j(x) \sum_{l\neq j} v^{jl} E(X_l - x) \\ A_{11} &= \sum_{j=1}^m \sigma_{jj} f_j(x) \left[\sum_{l\neq j} \{v^{jl}\}^2 E(X_l - x)^2 + \left\{ \sum_{l\neq j} v^{jl} E(X_l - x) \right\}^2 \right]. \end{aligned}$$

Some calculations show that $\text{var}\{\widehat{\theta}_{1,G}(x;h)\} \approx M\gamma_K(0)/nhH^2$, where $M = A_{00}B_{11}^2 - 2A_{01}B_{11}B_{01} + A_{11}B_{01}^2$ and $H = B_{00}B_{11} - B_{10}B_{01}$, which is

$$H \;=\; \left\{\sum_{j=1}^m v^{j\cdot} f_j\right\}\left\{\sum_{j=1}^m v^{jj} f_j\right\} + \left\{\sum_{j=1}^m v^{j\cdot} f_j^{(1)}\right\}\left\{\sum_{j=1}^m v^{jj} f_j\right\} - \left\{\sum_{j=1}^m v^{j\cdot} f_j\right\}\left\{\sum_{j=1}^m v^{jj} f_j^{(1)}\right\}.$$

If the X_j are independent and identically distributed with common density $f(\cdot)$, some calculations show that H and M can be simplied as

$$\begin{aligned} H &= \left(\sum_{j=1}^m v^{j\cdot}\right)\left(\sum_{j=1}^m v^{jj}\right) f^2(x) \\ M &= f(x)\left[\left(\sum_{j=1}^m \{v^{j\cdot}\}^2\sigma_{jj}\right)\left(\sum_{j=1}^m v^{jj}\right)^2 \{f(x) - af^{(1)}(x)\}^2 + \left(\sum_{j=1}^m \{v^{jj}\}^2\sigma_{jj}\right)\left(\sum_{j=1}^m v^{j\cdot}\right)^2 \{af^{(1)}(x)\}^2 \right. \\ &\quad -2\left(\sum_{j=1}^m v^{j\cdot}\right)\left(\sum_{j=1}^m v^{jj}\right)\left(\sum_{j=1}^m v^{j\cdot}v^{jj}\sigma_{jj}\right)\{f(x) - af^{(1)}(x)\}af^{(1)}(x) \\ &\quad \left. + \left(\sum_{j=1}^m \sigma_{jj}\sum_{l\neq j}\{v^{jl}\}^2\right)\left(\sum_{j=1}^m v^{j\cdot}\right)^2 \{f^{(1)}(x)\}^2 E(X-x)^2 \right] \end{aligned}$$

where $a = E(X-x)$. Noting the last term of M is nonnegative and using the Cauchy-Schwartz inequality for the first three terms, some calculations show that $M \geq f^3(x)\left|\sum_{j=1}^m v^{jj}\right|\left|\sum_{j=1}^m v^{j\cdot}\right|\{\sum_{j=1}^m 1/\sigma_{jj}\}^{-1}$. It follows that $\text{var}(\widehat{\theta}_{1,G}(x;h)) \geq \gamma_K(0)\{nhf(x)\sum_{j=1}^m 1/\sigma_{jj}\}^{-1}$. This completes the proof of part (2).

A.3 Proof of Theorem 3

For simplicity, we provide the proof by assuming that p is odd. When p is even, bias calculations are similar but are more complex (see Appendix A.4 and Carroll, et al., 1998). Let $\Psi(Y, s) = \{Y - \mu(s)\}\mu^{(1)}(s)/V(s)$. Then, using the techniques of Carroll, et al. (1998), it can be shown that $\widehat{\theta}_{p,wpe}(x;h)$ has the following expansion:

$$\begin{aligned}\widehat{\theta}_{p,wpe}(x,h) - \theta(x) \approx \quad & h^{p+1}\frac{\theta^{(p+1)}(x)}{(p+1)!}\boldsymbol{e}^T\mathbf{E}_p^{-1}(c)\mathbf{E}_c(p+1) + \left\{\left[\mu^{(1)}\{\theta(x)\}\right]^2\sum_{j=1}^m f_j(x)/\sigma_{jj}\right\}^{-1} \quad \text{(A. 1)}\\ & \times n^{-1}\sum_{i=1}^n\sum_{j=1}^m \boldsymbol{e}^T\mathbf{E}_p^{-1}(c)\phi_j^{-1}K_h(X_{ij}-x)\boldsymbol{G}_p\{(X_{ij}-x)/h\}\Psi\{Y_{ij},\theta(X_{ij})\},\end{aligned}$$

where $\sigma_{jj} = \phi_j w_j^{-1}V[\mu\{\theta(x)\}]$. The bias of $\widehat{\theta}_{p,wpe}(x,h)$ is the first term in (A. 1) and the variance is

$$\text{var}\{\widehat{\theta}_{p,wpe}(x,h)\} \approx \gamma_K(0)(nh)^{-1}\left\{[\mu\{\theta(x)\}]^2\sum_{j=1}^m f_j(x)/\sigma_{jj}\right\}^{-1}\boldsymbol{e}^T\mathbf{E}_p^{-1}(c)\mathbf{E}_p(\gamma)\mathbf{E}_p^{-1}(c)\boldsymbol{e}.$$

A.4 Proof of Theorem 4

Reparametrize $\boldsymbol{G}_p(X_{ij}-x)$ in equation (5) as $\boldsymbol{G}_p\{(X_{ij}-x)/h\}$ and $\boldsymbol{\beta}$ as $\boldsymbol{\alpha}$ whose jth component $\alpha_j = h^j\theta^{(j)}(x)/j!$. Then $\widehat{\theta}_p^*(x;h) = \widehat{\alpha}_0$. A Taylor expansion of (5) gives $\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha} = \boldsymbol{B}_n^{-1}\mathbf{A}_n$, where

$$\begin{aligned}\boldsymbol{B}_n &= n^{-1}\sum_{i=1}^n \boldsymbol{G}_{ip}^T(x)\boldsymbol{\Delta}_i(x)\mathbf{K}_{ih}^{1/2}(x)\mathbf{V}_i^{-1}(x)\mathbf{K}_{ih}^{1/2}(x)\boldsymbol{\Delta}_i(x)\boldsymbol{G}_{ip}(x)\\ \mathbf{A}_n &= \sum_{i=1}^n \boldsymbol{G}_{ip}^T(x)\boldsymbol{\Delta}_i(x)\mathbf{K}_{ih}^{1/2}(x)\mathbf{V}_i^{-1}(x)\mathbf{K}_{ih}^{1/2}(x)(\mathbf{Y}_i - \boldsymbol{\mu}_i).\end{aligned}$$

Since $(\mathbf{Y}_i, \mathbf{X}_i)$ are independent and identically distributed, we suppress below the subscript i. Let $\boldsymbol{B} = \lim_{n\to\infty}\boldsymbol{B}_n = E\left\{\boldsymbol{G}_p^T(x)\boldsymbol{\Delta}(x)\mathbf{K}_h^{1/2}(x)\mathbf{V}^{-1}(x)\mathbf{K}_h^{1/2}(x)\boldsymbol{\Delta}(x)\boldsymbol{G}_p(x)\right\}$. The (r_1, r_2)th component of $\boldsymbol{B}$ is

$$B_{r_1,r_2} = E\left\{\sum_{j=1}^m\sum_{l=1}^m \mu_j^{(1)}\mu_l^{(1)}v^{jl}K_h^{1/2}(X_j-x)K_h^{1/2}(X_l-x)\left(\frac{X_j-x}{h}\right)^{r_1-1}\left(\frac{X_l-x}{h}\right)^{r_2-1}\right\},$$

where $\mu_j^{(1)} = \mu_j^{(1)}\left[\boldsymbol{G}_p^T\{(X_j-x)/h\}\boldsymbol{\alpha}\right]$. Some calculations give

$$\begin{aligned}B_{r_1,r_2} &= \sum_{j=1}^m\int\left\{\mu_j^{(1)}\right\}^2 v^{jj}K(s_j)f_j(x+s_jh)s_j^{r_1+r_2-2}ds_j + o(h)\\ &= \left[\mu^{(1)}\{\theta(x)\}\right]^2\sum_{j=1}^m v^{jj}f_j(x)c_K(r_1+r_2-2) + o(h).\end{aligned}$$

It follows that $\boldsymbol{B} = \left[\mu^{(1)}\{\theta(x)\}\right]^2 \sum_{j=1}^m v^{jj} f_j(x)\mathbf{E}_p(c) + o(h)$.

The rth component of $E(A_n)$ is

$$\begin{aligned} & E\left[\sum_{j=1}^m \sum_{l=1}^m \left\{\mu_j^{(1)}\right\}^2 v^{jl} K_h^{1/2}(X_j - x) K_h^{1/2}(X_l - x) \left(\frac{X_j - x}{h}\right)^{r-1}\right. \\ & \qquad \left.\left\{\frac{h^{p+1}\theta^{(p+1)}(x)}{(p+1)!}\left(\frac{X_l - x}{h}\right)^{p+1} + \frac{h^{p+2}\theta^{(p+2)}(x)}{(p+2)!}\left(\frac{X_l - x}{h}\right)^{p+2}\right\}\right] + o\left(h^{p+2}\right) \\ & = \sum_{j=1}^m \int \{\mu_j^{(1)}\}^2 v^{jj} K(s_j) f_j(x + s_j h)\left\{\frac{h^{p+1}\theta^{(p+1)}(x)}{(p+1)!} s_j^{r+p} + \frac{h^{p+2}\theta^{(p+2)}(x)}{(p+2)!} s_j^{r+p+1}\right\} + o(h^{p+2}). \end{aligned} \tag{A. 2}$$

If $p = 0$, noting $c_K(2) = 1$, some calculations show equation (A. 2) becomes

$$h^2\left[\mu^{(1)}\{\theta(x)\}\right]^2 \left\{\theta^{(1)}(x)\sum_{j=1}^m v^{jj} f_j^{(1)}(x) + \frac{\theta^{(2)}(x)}{2}\sum_{j=1}^m v^{jj} f_j(x)\right\} + o(h^2).$$

If $p > 0$, equation (A. 2) becomes

$$\begin{aligned} & h^{p+1}\frac{\theta^{(p+1)}(x)}{(p+1)!}\left[\mu^{(1)}\{\theta(x)\}\right]^2 \sum_{j=1}^m v^{jj} f_j(x) c_K(r+p) \\ & + h^{p+2}\left\{\frac{\theta^{(p+1)}(x)}{(p+1)!}\sum_{j=1}^m \frac{\partial[T_j(x) f_j(x)]}{\partial x} + \frac{\theta^{(p+2)}(x)}{(p+2)!}\sum_{j=1}^m T_j(x) f_j(x)\right\} c_K(r+p+1), \end{aligned}$$

where $T_j(x) = \left[\mu^{(1)}\{\theta(x)\}\right]^2 v^{jj}(x)$. Noting that $c_K(s) = 0$ and the $(1, s+1)$ elements of $\boldsymbol{B}$ and $\boldsymbol{B}^{-1}$ are zero if s is odd, using bias$\left\{\widehat{\theta}_p^*(x;h)\right\} = \boldsymbol{e}^T\boldsymbol{B}^{-1}E(\mathbf{A}_n)$, some calculations give the bias expressions of $\widehat{\theta}_p^*(x;h)$ stated in Theorem 4.

To calculate the asymptotic variance of $\widehat{\theta}_p^*(x;h)$, we first calculate cov$(\mathbf{A}_n)$ as

$$\text{cov}(\mathbf{A}_n) = \frac{1}{n}E\left(\boldsymbol{G}_p^T \boldsymbol{\Delta}\mathbf{K}_h^{1/2}\mathbf{V}^{-1}\mathbf{K}_h^{1/2}\boldsymbol{\Sigma}\mathbf{K}_h^{1/2}\mathbf{V}^{-1}\mathbf{K}_h^{1/2}\boldsymbol{\Delta}\boldsymbol{G}_p\right) + o\left\{(nh)^{-1}\right\},$$

where $\boldsymbol{\Sigma} = \text{cov}(\mathbf{Y}_i|\mathbf{X}_i = x\mathbf{1})$ and σ_{jk} is the (j,k)th element of $\boldsymbol{\Sigma}$. The (r_1, r_2)th component of the first term is

$$\begin{aligned} & n^{-1}E\left\{\sum_{j=1}^m\sum_{k=1}^m\sum_{l=1}^m\sum_{q=1}^m \mu_j^{(1)}\mu_l^{(1)} v_{jk}\sigma_{kl}v_{lq} K_h^{1/2}(X_j - x)K_h^{1/2}(X_k - x)K_h^{1/2}(X_l - x)K_h^{1/2}(X_q - x)\right. \\ & \qquad \left.\left(\frac{X_j - x}{h}\right)^{r_1 - 1}\left(\frac{X_q - x}{h}\right)^{r_2 - 1}\right\} \\ = \;& (nh)^{-1}\sum_{j=1}^m \int \left\{\mu_j^{(1)}\right\}^2 \{v^{jj}\}^2 \sigma_{jj} K^2(s_j) f_j(x + s_j h) s_j^{r_1 + r_2 - 2} ds_j + o\left\{(nh)^{-1}\right\} \\ = \;& (nh)^{-1}\left[\mu^{(1)}\{\theta(x)\}\right]^2 \sum_{j=1}^m \{v^{jj}\}^2 \sigma_{jj} f_j(x)\gamma_K(r_1 + r_2 - 2) + o\left\{(nh)^{-1}\right\}. \end{aligned}$$

Using cov$\{\widehat{\theta}_p^*(x;h)\} = \boldsymbol{e}^T\boldsymbol{B}^{-1}\text{cov}(\mathbf{A}_n)\boldsymbol{B}^{-1}\boldsymbol{e}$, we have the expression of cov$\{\widehat{\theta}_p^*(x;h)\}$ as that given in part (2). A direct application of the Cauchy-Schwartz inequality gives part (3).

A.5 Distribution of the Weighted Pooled Estimator Under Measurement Error

To develop the SIMEX theory, we need an asymptotic expansion for the naive estimator. In expressions which follow, the argument $(\cdot)$ refers to $G_p^T\{(W_{ij}-w)/h\}\boldsymbol{\beta}$, the argument $(\bullet)$ refers to $\theta_N(W_{ij},\boldsymbol{\Sigma}_{uu})$, and the argument $(\circ)$ refers to $\theta_N(w,\boldsymbol{\Sigma}_{uu})$. The first $p+1$ terms of the Taylor series expansion of $\theta_N(W_{ij},\boldsymbol{\Sigma}_{uu})$ about $\theta_N(w,\boldsymbol{\Sigma}_{uu})$ are given by $G_p^T\{(W_{ij}-w)/h\}\boldsymbol{\beta}$. We solve $\boldsymbol{\beta}$ by

$$n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{Y_{ij}-\mu(\cdot)}{\phi_j(\boldsymbol{\Sigma}_{uu})V(\cdot)}\mu^{(1)}(\cdot)K_h(W_{ij}-w)G_p\{(W_{ij}-w)/h\}=0.$$

It is easily seen by a first order Taylor expansion and using (9) that $\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}=B_n^{-1}A_n$, where

$$\begin{aligned}
B_n &= n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\{\mu^{(1)}(\cdot)\}^2}{\phi_j(\boldsymbol{\Sigma}_{uu})V(\cdot)}K_h(W_{ij}-w)G_p\{(W_{ij}-w)/h\}G_p^T\{(W_{ij}-w)/h\}\\
&\approx \mathbf{E}_p(c)[\{\mu^{(1)}(\circ)\}^2/V(\circ)]\sum_{j=1}^{m}f_{jW}(w,\boldsymbol{\Sigma}_{uu})/\phi_j(\boldsymbol{\Sigma}_{uu})+o_p(1)=B+o_p(1);\\
A_n &= A_{n1}+A_{n2};\\
A_{n1} &= n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{Y_{ij}-\mu(\bullet)}{\phi_j(\boldsymbol{\Sigma}_{uu})V(\cdot)}\mu^{(1)}(\cdot)K_h(W_{ij}-w)G_p\{(W_{ij}-w)/h\};\\
A_{n2} &= n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\mu(\bullet)-\mu(\cdot)}{\phi_j(\boldsymbol{\Sigma}_{uu})V(\cdot)}\mu^{(1)}(\cdot)K_h(W_{ij}-w)G_p\{(W_{ij}-w)/h\}.
\end{aligned}$$

It is easily seen that

$$A_{n2}\approx[\{\mu^{(1)}(\circ)\}^2/V\{\mu(\circ)\}]\{h^{p+1}\theta_N^{(p+1)}(w)/(p+1)!\}\mathbf{E}_c(p+1)\sum_{j=1}^{m}f_{jW}(w,\boldsymbol{\Sigma}_{uu})/\phi_j(\boldsymbol{\Sigma}_{uu}),$$

and hence that

$$\widehat{\theta}_N(w)-\theta_N(w)\approx h^{p+1}\theta_N^{(p+1)}(w)\boldsymbol{e}^T\mathbf{E}_p^{-1}(c)\mathbf{E}_c(p+1)/(p+1)!+\boldsymbol{e}^T\boldsymbol{B}^{-1}\mathbf{A}_{n1}. \tag{A. 3}$$

Remembering that $E(Y_{ij}|W_{ij}=w)=\zeta_j(w)$ and using (9), a tedious but straightforward calculation shows that $E(\mathbf{A}_{n1})=0$. Hence, the first term in (A. 3) is the bias expansion for the naive estimate.

It is also easily seen that we can replace the argument $(\circ)$ by $(\bullet)$ in the definition of A_{n1} leading to the expression $A_{n1}=A_{n11}+A_{n12}$, where

$$\begin{aligned}
A_{n11} &= n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{Y_{ij}-\zeta_j(W_{ij},\boldsymbol{\Sigma}_{uu})}{\phi_j(\boldsymbol{\Sigma}_{uu})V(\bullet)}\mu^{(1)}(\bullet)K_h(W_{ij}-w)G_p\{(W_{ij}-w)/h\};\\
A_{n12} &= n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\zeta_j(W_{ij},\boldsymbol{\Sigma}_{uu})-\mu(\bullet)}{\phi_j(\boldsymbol{\Sigma}_{uu})V(\bullet)}\mu^{(1)}(\bullet)K_h(W_{ij}-w)G_p\{(W_{ij}-w)/h\}.
\end{aligned}$$

Since A_{n12} is a function only of the W's, these two terms are uncorrelated.

A direct calculation shows that A_{n1} has variance asymptotically equivalent to

$$\frac{\{\mu^{(1)}(\circ)\}^2\mathbf{E}_p(\gamma)}{nhV^2(\circ)}\sum_{j=1}^{m}\{\mathcal{U}_j(w,\mathbf{\Sigma}_{uu})+\Gamma_j(w,\mathbf{\Sigma}_{uu})\}f_{jW}(w,\mathbf{\Sigma}_{uu})/\phi_j^2(\mathbf{\Sigma}_{uu}).$$

We have thus shown (11), namely that the variance of $\widehat{\theta}_N(w,\mathbf{\Sigma}_{uu})$ is asymptotically

$$\mathrm{var}\{\widehat{\theta}_N(w,\mathbf{\Sigma}_{uu})\}\approx(nh)^{-1}Q(w,\mathbf{\Sigma}_{uu})\boldsymbol{e}^T\mathbf{E}_p^{-1}(c)\mathbf{E}_p(\gamma)\mathbf{E}_p^{-1}(c)\boldsymbol{e}.$$

In the case that the (Y,X,W)'s are marginally identically distributed, although not necessarily independent, simplification occurs because $\mathcal{U}_j(w,\mathbf{\Sigma}_{uu})=0$, $\zeta_j=\mu(\theta_N)$ and none of the terms Γ_j, ϕ_j or f_{jW} depend on j.

We are now in a position to verify (12). The expansion (A. 3), with A_{n1} replaced by $A_{n11}+A_{n12}$ can be analyzed using the same techniques as in Carroll, et al. (1999). Since the calculations are similar, although tedious, in the interest of space we have chosen not to provide them. The key step in the proof is is to show $\mathrm{var}\{\widehat{\theta}_N(x;(1+\lambda)\mathbf{\Sigma}_{uu})\}=O\{(nhD)^{-1}\}+O(n^{-1})$ for $\lambda>0$, which is of smaller order than $\mathrm{var}\{\widehat{\theta}_N(x;\mathbf{\Sigma}_{uu})\}=O\{(nh)^{-1}\}$.

A.6 Comparison of the Variances of $\widehat{\theta}_{sx,wpe}(x)$ and $\widehat{\theta}_{sx,wae}(x)$

Using the Cauchy-Schwartz inequality, we have

$$\left\{\sum_{j=1}^{m}\frac{\Gamma_j(\cdot)f_{jW}(\cdot)}{\phi_j^2}\right\}\left\{\sum_{j=1}^{m}\frac{\left[\mu^{(1)}\{\theta_j(\cdot)\}\right]^2f_{jW}(\cdot)}{\Gamma_j(\cdot)}\right\}\geq\left\{\sum_{j=1}^{m}\frac{\left|\mu^{(1)}\{\theta_j(\cdot)\}\right|f_{jW}(\cdot)}{\phi_j}\right\}^2\geq\left\{\sum_{j=1}^{m}\frac{\mu^{(1)}\{\theta_j(\cdot)\}f_{jW}(\cdot)}{\phi_j}\right\}^2.$$

Using equation (9) and noting $\xi_j(\cdot)=\mu\{\theta_j(\cdot)\}$, the last term is $\left[\mu^{(1)}\{\theta_N(\cdot)\}\sum_{j=1}^m f_{jW}(\cdot)/\phi_j\right]^2$. We have

$$\frac{\sum_{j=1}^m\Gamma_j(\cdot)f_{jW}(\cdot)/\phi_j^2}{\left\{\mu^{(1)}\{\theta_N(\cdot)\}\sum_{j=1}^m f_{jW}(\cdot)/\phi_j\right\}^2}\geq\frac{1}{\sum_{j=1}^m\left[\mu^{(1)}\{\theta_j(\cdot)\}\right]^2f_{jW}(\cdot)/\Gamma_j(\cdot)}.$$

Further noting that $U_j(\cdot)\geq 0$, we have $\mathrm{var}\{\widehat{\theta}_{sx,wpe}(x)\}\geq\mathrm{var}\{\widehat{\theta}_{sx,wae}(x)\}$.

Table 1 Naive and SIMEX estimates of the Regression Coefficients of the Cubic Models for the ACSUS Data

	Naive	SIMEX($\sigma_u^2 = 0.34$)	SIMEX($\sigma_u^2 = 0.68$)
Intercept	-2.19	-1.84	-1.66
Linear	-0.54	-0.65	-0.75
Quadratic	-0.29	-0.65	-0.78
Cubic	-0.06	-0.14	-0.17
Cubic, 2.5% bootstrap quantile	-0.17	-0.38	-0.45
Cubic, 97.5% bootstrap quantile	0.007	-0.02	-0.02

Note: The 4% and 96% bootstrap quantiles of the naive cubic term are -0.153 and -0.001.

LIST OF ILLUSTRATIONS

Figure 1

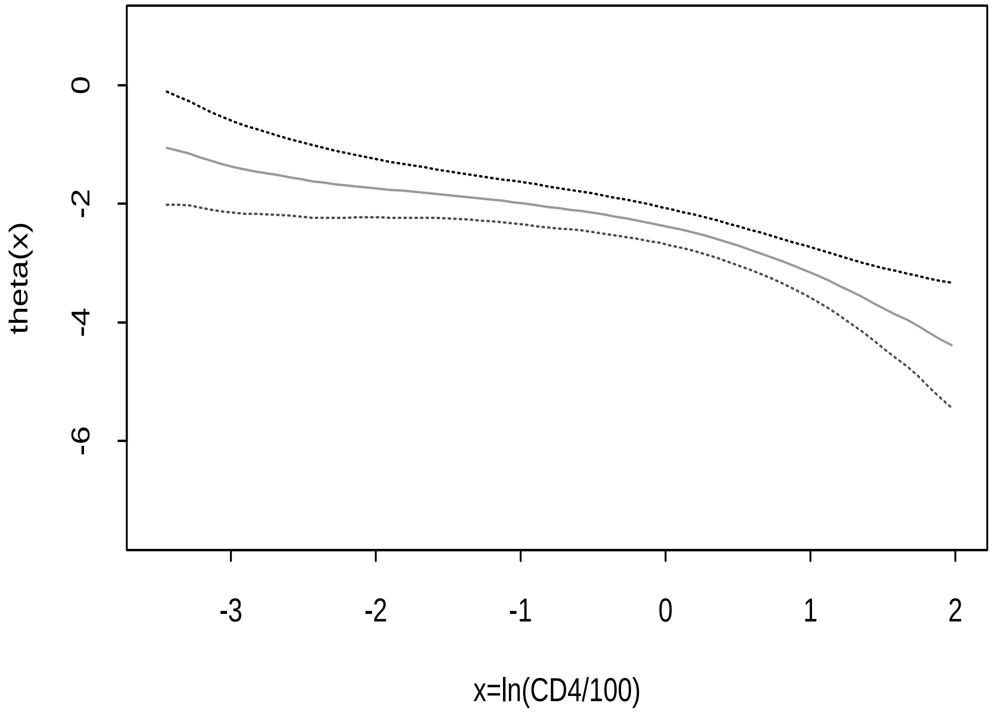


Figure 2

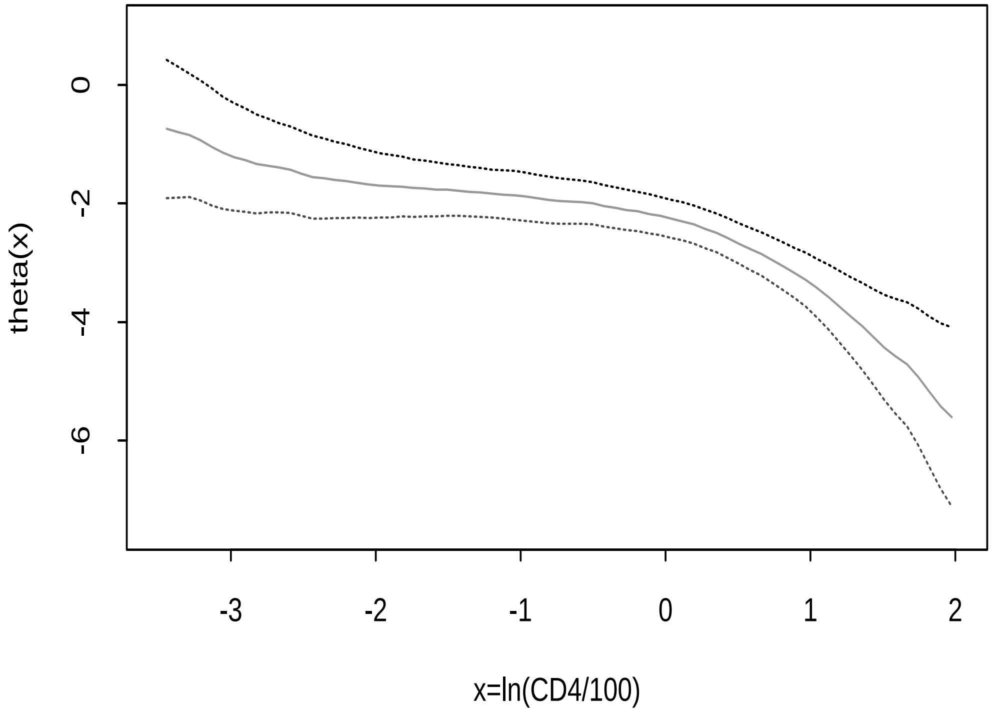


Figure 3

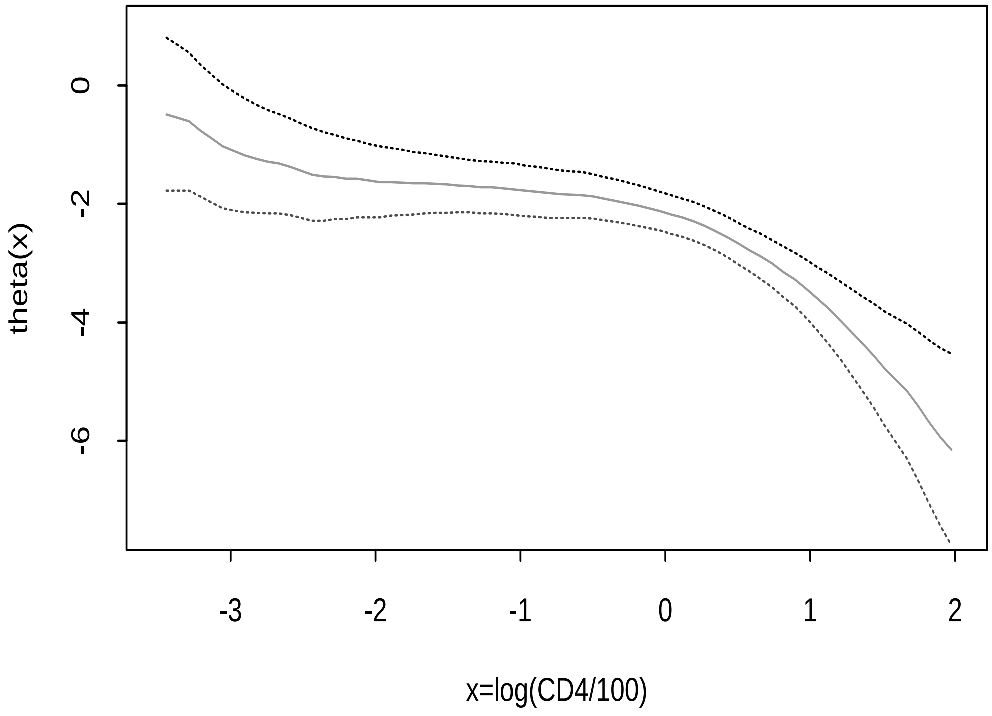
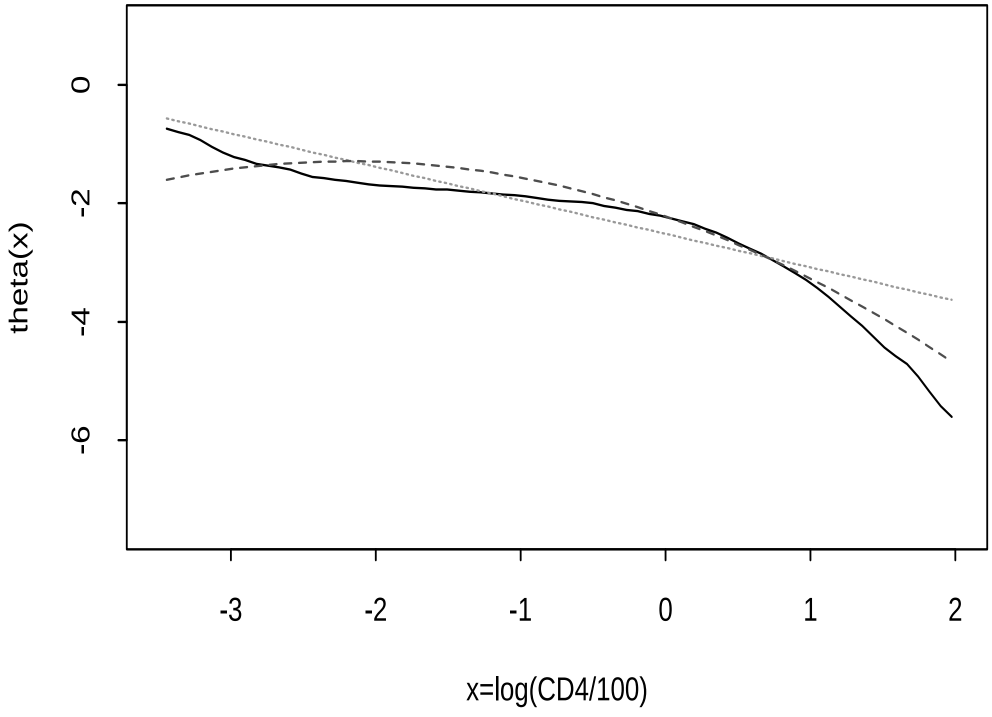


Figure 4


theta(x)
x=log(CD4/100)