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

Xihong LIN and Raymond J. CARROLL

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 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 the 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 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.

1. INTRODUCTION

A vast literature has 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, as 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 recent interest in extending the existing parametric models to allow for nonparametric covariate effects (Severini and Staniswalis 1994; Wild and Yee 1996; Zeger and Diggle 1994). Such nonparametric regression allows for more flexible functional dependence of the outcome variable on the covariates and also can 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 (Tsatis, Degruutola, and Wolfsohn 1995) are often subject to measurement error. We consider here data from the AIDS Costs and Services Utilization Survey (ACSUS) (Berk, Maffeo, and Schur 1993). The ACSUS sampled 2,487 subjects in 10 randomly selected U.S. cities with the highest AIDS rates. A series of six interviews were conducted for each respondent every 3 months from 1991 to 1992. A main outcome of interest was whether an interviewee had had hospital admissions (yes/no) during the past 3 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. Analysis of this dataset entails two major complications. The first complication is that even though it is believed that a lower CD4 count is associated with a greater 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 changepoint, 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; for example, the coefficient of variation could be as large as 50% (Tsatis 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 rather the most recent CD4 count was abstracted from each respondent’s medical record using his or 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 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

© 2000 American Statistical Association
Journal of the American Statistical Association
June 2000, Vol. 95, No. 450, Theory and Methods

520

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
polynomial kernel methods and extend these methods to the measurement error case using the simulation–extrapolation (SIMEX) method (Cook and Stefanski 1994). We study the asymptotic biases and variances of the proposed estimators.

We develop two main striking results:

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; Verbly, Cullis, Kenward, and Welham 1999). We show that generally the best strategy is to ignore the correlation structure entirely, and pretend as if the data within a cluster were independent (i.e., the working independence model in generalized estimating equation terminology). Furthermore, correctly specifying the correlation structure in estimating the nonparametric function in fact has adverse effects; that is, 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). Although 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 follow-up. 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 of the data from different panels. This result is also dramatically different from parametric measurement error regression, where pooled analysis gives an asymptotically efficient estimator.

The article 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 article 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, Rice, Wu, and Yang (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. We apply the proposed methods to the analysis of the ACSUS data in Section 5, followed by discussion in Section 6.

2. THE MODEL

Suppose that the data consist of \( n \) clusters with the \( i \)th \((i = 1, \ldots, n)\) 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 \( j \)th \((j = 1, \ldots, m_i)\) observation in the \( i \)th cluster. The observations within the same cluster might be correlated. Given the true covariate \( X_{ij} \), the mean and variance of \( Y_{ij} \) are

\[
E(Y_{ij}|X_{ij}) = \mu_{ij} \quad \text{and} \quad \text{var}(Y_{ij}|X_{ij}) = \phi_j w_{ij}^{-1} V(\mu_{ij})
\]

where \( \phi_j \) is a scale parameter, \( w_{ij} \) is a weight, and \( V(\cdot) \) is a variance function. The marginal mean \( \mu_{ij} \) depends on \( X_{ij} \) through a known monotonic link function \( \mu(\cdot) \).

\[
\mu_{ij} = \mu(\{\theta(X_{ij})\})
\]

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 that the unobserved covariate \( X_{ij} \) is related to the observed covariate \( W_{ij} \) by an additive measurement error model

\[
W_{ij} = X_{ij} + U_{ij}
\]

where \( U_{ij} \) is a measurement error and \( U_i = (U_{1i}, \ldots, U_{mi})' \) follows normal(0, \( \Sigma_{u} \)). 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 \( j = 1, \ldots, m \) 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). 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 \( j \); for example, 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,
and so on be independent of $j$. In what follows we do our calculations as if special meaning was 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 $\theta(\cdot)$ is a parametric $p$th polynomial function $\theta(\cdot) = G_p(\cdot)^T \beta$, where $G_p(z) = (1, z, \ldots, z^p)^T$ and $\beta = (\beta_0, \ldots, \beta_p)^T$. Let $Y_i = (Y_{i1}, \ldots, Y_{im_i})^T$ and $G_{ip} = (G_p(X_{i1}), \ldots, G_p(X_{im_i}))^T$. The regression coefficients $\beta$ can be estimated using the conventional generalized estimating equations (GEEs) (Liang and Zeger 1986),

$$
\sum_{i=1}^{n} G_{ip}^T \Delta_i V_i^{-1} (Y_i - \mu_i) = 0,
$$

where $\mu_i = E(Y_i)$ with its $j$th component $\mu_{ij} = \mu(G_{ip}(X_{ij}))$, $\Delta_i = \text{diag}\{\mu^{(1)}(G_p(X_{ij}))\}$, $\mu^{(1)}(\cdot)$ is the first derivative of $\mu(\cdot)$, $V_i = S_i^{1/2} R_i(\delta) S_i^{1/2}$, $S_i = \text{diag}\{h^{-1}\delta_i^{-1} V\{\mu_{ij}\}\}$, and $R_i$ is an invertible working correlation matrix, possibly depending on a parameter vector $\delta$, which can be estimated using the method of moments. Liang and Zeger (1986) showed that the GEE estimator $\hat{\beta}$ is asymptotically consistent if the mean function $\mu(\cdot)$ is correctly specified even when the working correlation matrix $R_\delta$ is misspecified. The most efficient estimator of $\beta$ is obtained by correctly specifying $R_\delta$.

We now consider how to extend the parametric GEE (3) to (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 $K(h) = h^{-1} K(v/h)$. The idea is to approximate $\theta(x)$ at any given $x$ using a local polynomial satisfying $\theta(X) = \{G_p(X - x)\}^T \beta$, where $G_p(\cdot)$ and $\beta$ are as defined earlier. Having estimated $\beta$ at $x$, the estimated $\theta(x)$ satisfies $\hat{\theta}(x) = \hat{\beta}_0$.

Let $G_{ip}(x) = (G_p(X_{i1} - x), \ldots, G_p(X_{im_i} - x))^T$. 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 $\theta(x)$,

$$
\sum_{i=1}^{n} G_{ip}^T (x) \Delta_i (x) V_i^{-1} (x) K_{ih}(x) \{Y_i - \mu_i(x)\} = 0
$$

or

$$
\sum_{i=1}^{n} G_{ip}(x) \Delta_i (x) K_{ih}^{1/2}(x) V_i^{-1}(x) \times K_{ih}^{1/2}(x) \{Y_i - \mu_i(x)\} = 0,
$$

where $K_{ih}(x) = \text{diag}\{K_h(X_{ij} - x)\}$, and $\{\mu_i(x), \Delta_i(x), V_i(x), S_i(x)\}$ are the same as those in (3) except that they are evaluated at $\mu_i(x) = \mu(G_{ip}(X_{ij} - x) \beta)$. The working correlation matrix $R_i$ in $V_i(x)$ may depend on a parameter vector $\delta$, 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 $V_i(x)$ is a diagonal matrix (assuming independence). Equation (5) weights the residuals $\{Y_i - \mu_i(x)\}$ symmetrically, whereas (4) does not. They hence often give different estimators of $\theta(x)$. We let $\hat{\theta}_p(x; h)$ denote the local $p$th-order kernel estimator using (4) and let $\hat{\theta}_p(x; h)$ denote the local $p$th-order kernel estimator using (5). These two estimators are identical when $V_i(x)$ is a diagonal matrix. We show in Sections 3.2 and 3.3 that the symmetric property of (4) and (5) results in different asymptotic properties of $\hat{\theta}_p(x; h)$ and $\hat{\theta}_p(x; h)$.

We have allowed the scale parameters $\phi_j$ to depend on $j$. In many problems it is reasonable to suppose that they do not depend on $j$; then we can set $S_i(x) = \text{diag}\{h^{-1}\delta_i^{-1} V\{\mu_{ij}\}\}$. If the $\phi_j$ do depend on $j$, then they will have to be estimated, again by the method of moments.

Application of the Fisher scoring algorithm to (4) shows that the estimator $\hat{\beta}$ can be updated using iteratively reweighted least squares,

$$
\left[ \sum_{i=1}^{n} G_{ip}(x)^T C_i(x) G_{ip}(x) \right] \beta = \sum_{i=1}^{n} G_{ip}(x)^T C_i(x) y_i,
$$

where $C_i(x) = \Delta_i(x) V_i^{-1}(x) K_{ih}(x) \Delta_i(x)$ is a working weight matrix and $y_i = G_{ip}(x)^T \beta + \Delta_i^{-1}(x) \{Y_i - \mu_i(x)\}$ is a working vector. The variance of $\hat{\theta}_p(x; h)$ is equal to $\text{var}\{\hat{\beta}_p(x; h)\}$ and can be estimated using a sandwich estimator, which takes the form $\text{cov}\{\hat{\beta}_p(x; h)\} = e^T \Omega_1^{-1} \Omega_2 \Omega_1^{-1} e$, where $e = (1, 0, \ldots, 0)^T$ and

$$
\Omega_1 = \sum_{i=1}^{n} G_{ip}(x)^T \Delta_i(x) V_i^{-1}(x) K_{ih}(x) \Delta_i(x) G_{ip}(x)
$$

and

$$
\Omega_2 = \sum_{i=1}^{n} G_{ip}(x)^T \Delta_i(x) V_i^{-1}(x) K_{ih}(x) \{Y_i - \mu_i(x)\} \times \{Y_i - \mu_i(x)\}^T K_{ih}(x) V_i^{-1}(x) \Delta_i(x) G_{ip}(x).
$$

A similar Fisher scoring algorithm can be constructed to solve (5) for $\hat{\theta}_p(x; h)$ and to calculate its variance. Specifically, one simply replaces $V_i(x) K_{ih}(x)$ by $K_{ih}^{1/2}(x) V_i(x) K_{ih}^{1/2}(x)$ in $(C_i, \Omega_1, \Omega_2)$.

Some versions of (4) have been proposed earlier. There are three obvious choices: (I) let $R_i$ be an estimator of the actual correlation matrix; (II) let $V_i^{-1}$ be the diagonal values of the inverse of the covariance matrix of $Y_i$; and (III) let $R_i$ be the identity matrix, thus effectively ignoring the correlation structure within clusters. We call method (III) the weighted pooled estimator. Method (I) was proposed by Severini and Staniswalis (1994) in their equation (18) for average kernel $p = 0$. Method (II) is a generalization,
from the Gaussian case to generalized linear models, of the modified quasi-likelihood proposal of Ruckstuhl et al. (1999). Method (III) 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 \( \phi \) depending on the value of \( j \). In Section 3.5 we consider another estimator called the "weighted average estimator."

Ruckstuhl et al. (1999) considered (4) for Gaussian data; that is, \( w_{ij} = 1 \) and \( \mathbf{V}(\mu_{ij}) = 1 \). They showed that under the simple variance component model \( \mathbf{V}_i = \phi \mathbf{I} + \delta_i \mathbf{I} \), where \( \mathbf{I} \) is an identity matrix and \( \mathbf{J} \) is a matrix of 1's, when \( m_i = m, p = 1 \) so that local linear regression is used and the \( X_{ij} \)'s are iid, methods (II) and (III) are asymptotically equivalent and have uniformly smaller asymptotic mean squared errors than method (I). Methods (II) and (III) 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 \( \hat{\theta}_p(x; h) \) and \( \hat{\theta}_p^*(x; h) \) under the kernel GEEs (4) and (5). This investigation will allow us to compare the asymptotic performance of methods (I)-(III) and to identify an optimal working correlation matrix \( \mathbf{R}_i \). Our main conclusions from the asymptotic analyses are as follows:

1. The two kernel estimators \( \hat{\theta}_p(x; h) \) and \( \hat{\theta}_p^*(x; h) \) often have different asymptotic properties, and the asymptotic properties of \( \hat{\theta}_p(x; h) \) 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 \( \hat{\theta}_p(x; h) \) and \( \hat{\theta}_p^*(x; h) \) are obtained when ignoring the within-cluster correlation entirely; that is, assuming working independence \( \mathbf{R}_i = \mathbf{I} \). Correctly specifying the correlation matrix in fact results in an asymptotically less efficient estimator of \( \theta(x) \).

3.2 Asymptotic Theory for the Kernel Estimator \( \hat{\theta}_p(x; h) \)

From (4)

Asymptotic bias and variance analysis of \( \hat{\theta}_p(x; h) \) under (4) is often difficult for general local \( p \)th 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 (\( p = 1 \)) for both Gaussian and non-Gaussian data (Theorem 1), and then study local kernel estimation (\( p = 1 \)) for Gaussian data (Theorem 2). If working independence \( \mathbf{R}_i = \mathbf{I} \) is assumed, then asymptotic bias and variance analysis of general local \( p \)th 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 \( \mathbf{X}_i = (X_{i1}, \ldots, X_{im})^T \) to be correlated unless stated otherwise, and let \( j_i^*(\cdot) \) denote the marginal density of \( X_i \). We further assume that the \( (Y_{ij}, X_i) \) \((i = 1, \ldots, n)\) are iid pairs with a continuous density function, and \( \mathbf{V}_i(\mu_i, \delta) = \mathbf{V}(\mu_i, \delta) \). Let \( \hat{g}^{(r)}(\cdot) \) denote the \( r \)th derivative of \( g(\cdot) \), and let \( \nu^k \) denote the \( (j, k) \)th element of \( \mathbf{V}^{-1} \). Let \( c_K(r) = \int z^r K(z) \, dz \), with \( c_K(0) = c_K(2) = 1 \). \( \gamma_K(r) = \int z^r K^2(z) \, dz \). \( \mathbf{E}_c(L) = \{c_K(L), c_K(L + 1), \ldots, c_K(L + p)\}^T \), and \( \mathbf{E}_p(c) \) and \( \mathbf{E}_p(\gamma) \) the \((p + 1) \times (p + 1)\) matrices with \((j, k)\) element \( c_K(j + k - 2) \) and \( \gamma_K(j + k - 2) \). We further assume that \( nh \to \infty \) as \( n \to \infty \) and \( h \to 0 \).

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

a. The asymptotic bias and variance of \( \hat{\theta}_0(x; h) \) are given by

\[
\text{bias}\{\hat{\theta}_0(x; h)\} \approx h^2 \left\{ \phi^{(1)}(x) \sum_{j=1}^m \nu^j(x) f_j^{(1)}(x) + \frac{\phi^{(2)}(x)}{2} \right\}
\]

and

\[
\text{var}\{\hat{\theta}_0(x; h)\} \approx \frac{\gamma_{K(0)}}{nh} \sum_{j=1}^m \nu^j(x) \sigma_{jj}(x) f_j(x)
\]

where \( \sigma_{jj}(x) = \text{var}(Y_{ij}, X_i) = \nu_j^{-1} \phi_j \mathbf{V}[\mu(\theta(x))] \), and \( \nu^j(x) = \sum_{i=1}^m \nu^i(x) \). If \( f_j(\cdot) = f(\cdot) \), the bias of \( \hat{\theta}_0(x; h) \) is free of \( \mathbf{V} \).

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

\[
\min_V \left\{ \text{var}\{\hat{\theta}_0(x; h)\} \right\} \approx \frac{\gamma_{K(0)}}{nh} \left[ \left[ \phi^{(1)}(x) \right]^2 \sum_{j=1}^m \left( f_j(x)/\sigma_{jj}(x) \right) \right]^{-1}
\]

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 \( \mu(\cdot) \) and \( \mathbf{V}(\cdot) \). 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 \( \hat{\theta}_{1,C}(x; h) \) be the solution of (4) for Gaussian data with \( \mathbf{V}(\cdot) = 1, w_{ij} = 1, \) and \( p = 1 \) and any given weight matrix \( \mathbf{V} \).

a. The asymptotic bias and variance of \( \hat{\theta}_{1,C}(x; h) \) are

\[
h^2 \phi^{(3)}(x)/2 \text{ and } c(\mathbf{V}(nh)) \text{, where the expression of } c \text{ is complicated and is given in Appendix A.2 and Ruckstuhl et al. (1999). Note that the asymptotic bias of } \hat{\theta}_{1,C}(x; h) \text{ is free of the distribution of the } X_{ij} \text{ and } \mathbf{V} \text{.}
\]

b. If the \( X_{ij} \) are iid with common density \( f(\cdot) \), the asymptotic variance of \( \hat{\theta}_{1,C}(x; h) \) is minimized when one assumes the working correlation matrix \( \mathbf{R} = \mathbf{I} \).
(independence) and is equal to

$$\min V \{ \text{var} \{ \hat{\theta}_{1,C}(x; h) \} \} \approx \{ \gamma_K(0)/nh \} \left[ f(x) \sum_{j=1}^{m} \{ 1/\sigma_{jj} \} \right]^{-1}.$$  

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

Part (b) of 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 (III) dominates methods (I) and (II). In other words, the asymptotically most efficient estimators $\hat{\theta}_0(x; h)$ and $\hat{\theta}_{1,C}(x; h)$ 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 $p$th polynomial estimation under a general working correlation matrix $R$ in (4) is difficult, even for Gaussian data. However, such calculations are possible when assuming independence $R = I$—that is, for the weighted pooled estimator [method (III)]. These results are stated in Theorem 3, whose proof is given in Appendix A.3.

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

a. The asymptotic bias of $\hat{\theta}_{p,\text{wpe}}(x; h)$ is

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

if $p = 0$,

$$\text{bias}\{ \hat{\theta}_{p,\text{wpe}}(x; h) \} = h^{p+1} \frac{\theta^{(p+1)}(x)}{(p+1)!} e^T E_p^{-1}(c) E_c(p+1);$$

if $p = \text{even}$ and $p > 0$,

$$\text{bias}\{ \hat{\theta}_{p,\text{wpe}}(x; h) \} \approx h^{p+2} \left\{ \frac{\theta^{(p+2)}(x)}{(p+2)!} + \frac{\theta^{(p+1)}(x)}{(p+1)!} \right\} e^T E_p^{-1}(c) E_c(p+2),$$

where $L_j(x) = \{ \mu^{(1)}(\theta(x))/\sigma_{jj}(x) \}$ and $\sigma_{jj}(x) = \text{var}(Y_{ij} | X_{ij} = x) = w_j^{-1}\phi_jV[\mu(\theta(x))]$.

b. The asymptotic variance of $\hat{\theta}_{p,\text{wpe}}(x; h)$ is

$$\text{var}\{ \hat{\theta}_{p,\text{wpe}}(x; h) \} \approx \frac{\gamma_K(0)}{nh} \left( \frac{\mu^{(1)}(\theta(x))^2}{\sum_{j=1}^{m} \{ f_j(x)/\sigma_{jj}(x) \}} \right)^{-1} \times e^T E_p^{-1}(c) E_p(e)^{-1}(c);$$

(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 $\hat{\theta}_{1,\text{wpe}}(x; h)$ are

$$\text{bias}\{ \hat{\theta}_{1,\text{wpe}}(x; h) \} \approx \frac{\gamma_K(0)}{nh}$$

and

$$\text{var}\{ \hat{\theta}_{1,\text{wpe}}(x; h) \} \approx \left( \frac{\mu^{(1)}(\theta(x))^2}{\sum_{j=1}^{m} \{ f_j(x)/\sigma_{jj}(x) \}} \right)^{-1}.$$  

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

We study in Theorem 4 the asymptotic bias and variance of $\hat{\theta}_p^*(x; h)$, the solution of the estimating equation (5), for a general local $p$th-order polynomial and a general weight matrix $V_i$ for both Gaussian and non-Gaussian cases. Unlike $\hat{\theta}_0(x; h)$, whose bias and variance analysis under this general condition is difficult, such a general analysis is feasible for $\hat{\theta}_p^*(x; h)$, and the results are much simpler and are different from those of $\hat{\theta}_0(x; h)$. 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 $R_i$.

The key result in Theorem 4 is given in part c; that is, the asymptotically most efficient estimator $\hat{\theta}_p^*(x; h)$ is obtained by entirely ignoring the within-cluster correlation and assuming that the data were independent. Note that under working independence, the two kernel estimators $\hat{\theta}_0(x; h)$ 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 that $\int s^r K^{1/2}(s) ds < \infty$ for integers $r \leq p$. Let $\hat{\theta}_p^*(x; h)$ be the solution of (5) for any given $p$ and any given weight matrix $V$.

a. The asymptotic bias of $\hat{\theta}_p^*(x; h)$ is

$$\text{bias}\{ \hat{\theta}_0(x; h) \} \approx h^2 \left\{ \frac{\theta^{(1)}(x)}{\sum_{j=1}^{m} \{ v_j^{(1)}(x)/\sigma_{jj}(x) \}} + \frac{\theta^{(2)}(x)}{2} \right\};$$

if $p = 0$,

$$\text{bias}\{ \hat{\theta}_p^*(x; h) \} \approx h^{p+1} \frac{\theta^{(p+1)}(x)}{(p+1)!} e^T E_p^{-1}(c) E_c(p+1);$$

if $p = \text{even}$ and $p > 0$,

$$\text{bias}\{ \hat{\theta}_p^*(x; h) \} \approx h^{p+2} \left\{ \frac{\theta^{(p+2)}(x)}{(p+2)!} + \frac{\theta^{(p+1)}(x)}{(p+1)!} \right\} e^T E_p^{-1}(c) E_c(p+2),$$

where $L_j(x) = \{ \mu^{(1)}(\theta(x))/\sigma_{jj}(x) \}$ and $\sigma_{jj}(x) = \text{var}(Y_{ij} | X_{ij} = x) = w_j^{-1}\phi_jV[\mu(\theta(x))].$
choose \( h \) to minimize

\[
CV(h) = \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{\{Y_{ij} - \hat{\mu}^{(-t)}_{ij}(X_{ij})\}^2}{\hat{\phi}_j w_{ij}^{-1} V_{\hat{\mu}^{(-t)}_{ij}(X_{ij})}}
\]

where \( \hat{\mu}^{(-t)}_{ij}(\cdot) \) is the estimate of \( \mu_{ij}(\cdot) \) calculated from the data leaving out the \( i \)th 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 squared errors \( \text{EMSE}(x,h) \) of \( \hat{\theta}(x,h) \) (either \( \hat{\theta}_p(x,h) \) or \( \hat{\theta}_p^*(x,h) \)) at a series of values of \( x \) and \( h \) and chooses \( h \) to minimize \( \text{EMSE}(x,h) \) for each \( x \). Calculations of \( \text{EMSE}(x_0,h_0) \) at any given value of \( x_0 \) and \( h_0 \) proceed by \( \text{EMSE}(x_0,h_0) = \text{bias}^2(\hat{\theta}(x_0,h_0)) + \text{var}^2(\hat{\theta}(x_0,h_0)) \). Here \( \text{bias}(\hat{\theta}(x_0,h_0)) \) denotes the empirical bias of \( \hat{\theta}(x_0,h_0) \) at \( x_0 \) and \( h_0 \) and is estimated by fitting a polynomial regression,

\[
E\{\hat{\theta}(x_0,h)\} = v_0 + v_1 h^{p+1} + \cdots + v_t h^{p+t}, \tag{8}
\]

using the “data” \( \{\hat{\theta}(x_i,h)\} \) in a neighborhood of \( h_0 \) for a given integer \( t \) (e.g., \( t = 1 \) or 2). The empirical bias, \( \text{bias}(\hat{\theta}(x_0,h_0)) \), is calculated as the estimated value of \( v_0 h_0^{p+1} + \cdots + v_t h_0^{p+t} \). The variance, \( \text{var}(\hat{\theta}(x_0,h_0)) \), can be easily calculated using the sandwich estimator in Section 3.1. We use this method in Section 5 to choose \( h \) when analyzing the ACSUS data.

### 3.5 Summary of Nonparametric Regression for Clustered Data

Our results in Sections 3.2 and 3.3 suggest that it is the best strategy to use (4) or (5) with \( R = I \); that is, 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 generalized linear models (GLMs), with weights depending on the cluster if the scale parameters \( \phi_j \) are not constant.

In the “panel data” problem with \( m_i \equiv m \), another estimator can be considered—to compute \( \hat{\theta}_j(x,h) \) 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 var\{\hat{\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 (III), the weighted pooled estimator. The key step in proving this result is to show that \( \text{cov}\{\hat{\theta}_j(x,h),\hat{\theta}_j(x,h)\} = O(n^{-1}) \) if \( j \neq j' \) is of smaller order compared to \( \text{var}\{\hat{\theta}_j(x,h)\} = O((nh)^{-1}) \). In other words, for asymptotic arguments, the individual estimators \( \hat{\theta}_j(x,h) \) 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 \( j = 1, \ldots, m \), which must then be estimated separately. In practice, this is not really that important an issue, because 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, as the weighted average estimator is asymptotically more efficient than the weighted pooled estimator.

4. SIMEX LOCAL POLYNOMIAL ESTIMATION WHEN THERE IS MEASUREMENT ERROR

In this section we discuss 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 propose calculating 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, whereas the latter calculates the naive weighted average estimators in the simulation step and can be applied only 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 local polynomial kernel estimation using SIMEX and propose the SIMEX weighted pooled estimator in Section 4.1, and study the asymptotic properties of this estimator in Section 4.2. We discuss the SIMEX weighted average estimator in Section 4.3.

4.1 The SIMEX Kernel Estimator

The SIMEX estimator was developed by Cook and Stefanski (1994). The idea behind the SIMEX method is seen most clearly in simple linear regression when the independent variable is subject to measurement error. Suppose that the regression model is \( E(Y\mid X) = \alpha + \beta X \) and that \( W = X + U \), rather than \( X \), is observed where \( U \) has mean 0 and variance \( \sigma_u^2 \) and the measurement error variance \( \sigma_u^2 \) is known. It is well known that the ordinary least squares estimate of the slope from regressing \( Y \) on \( W \) converges to \( \beta \) with variance \( \sigma_u^2 (\sigma_u^2 + \sigma^2) \), where \( \sigma^2 = \text{var}(X) \).

For any fixed \( \lambda > 0 \), suppose that one repeatedly “adds on,” via simulation, additional error with mean 0 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

\[
\gamma(\lambda) = \beta \sigma_u^2 \sigma_u^2 (1 + \lambda),
\]

Because, formally at least, \( g(\lambda) \) against \( \lambda \geq 0 \), fit a model to this plot, and then extrapolate back to \( \lambda = -1 \). Cook and Stefanski (1994) showed that this procedure will yield a consistent estimate of \( \beta \) 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 where measurement error covariance matrices \( \Sigma_{u,u} \) are known, and we keep track of these variances by means of the shorthand “\( \Sigma_{u,u} \).” In practice, the \( \Sigma_{u,u} \) 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 \( \Sigma_{u,u} \).

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 \( (\varepsilon_{id}) \) be a set of independent standard normal random variables that are then transformed to have sample mean 0 and variance 1 and to be uncorrelated with the \( Y \)'s and the \( W \)'s. Let \( \Sigma_{u,u}^{1/2} \) be the matrix square root of \( \Sigma_{u,u} \). Define

\[
\{W_{11}(\lambda), \ldots, W_{1D}(\lambda)\}^T = \{W_{11}, \ldots, W_{1D}\}^T + \lambda^{1/2} \Sigma_{u,u}^{1/2} \varepsilon_{1d}, \ldots, \varepsilon_{md}\}^T.
\]

We calculate the GEE kernel estimator, which solves either (4) or (5), from these simulated data and denote it by \( \theta_{g}(x, (1 + \lambda) \Sigma_{u,u}) \). The average of these estimates over \( d = 1, \ldots, D \) is denoted by \( \theta_{g}(x, (1 + \lambda) \Sigma_{u,u}) \). We run the SIMEX algorithm with \( D \) simulation replications at each value \( \lambda \) in a finite set \( A \). We extrapolate \( \theta_{g}(x, (1 + \lambda) \Sigma_{u,u}) \) using a polynomial of order \( q_\lambda \) back to \( \lambda = -1 \). This gives the SIMEX local polynomial estimator \( \theta_{g}(x) \).

In view of the results in Section 3, we propose calculating the naive estimators \( \theta_{g}(x, (1 + \lambda) \Sigma_{u,u}) \) using the weighted pooled estimator by assuming independence of observations within a cluster. The resulting estimator, called the SIMEX weighted pooled estimator, is denoted by \( \theta_{s}(x, \text{wpe}) \).

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 where \( m_i \equiv 1 \), 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 \( \zeta_j(w, \Sigma_{u,u}) = E(Y_{ij}|W_{ij} = w) \) depending on \( j \).

In the case where \( m_i \equiv m \), the following is easily shown. Let \( f_{W}(y, \Sigma_{u,u}) \) be the marginal density of \( W_{ij} \). Let \( \phi_j(\Sigma_{u,u}) \) be the limiting value of estimates of \( \phi_j \) ignoring measurement error. Then the naive estimate of \( \theta(w) \) converges to \( \theta_N(w, \Sigma_{u,u}) \) given by

\[
\mu(\theta_N(w, \Sigma_{u,u})) = \left\{ \sum_{j=1}^{m} \zeta_j(w, \Sigma_{u,u}) f_{W}(w, \Sigma_{u,u})/\phi_j(\Sigma_{u,u}) \right\}^{-1} \times \left\{ \sum_{j=1}^{m} f_{W}(w, \Sigma_{u,u})/\phi_j(\Sigma_{u,u}) \right\}.
\]

Let \( s(\lambda) \) be the \((q_s + 1)\)-vector with \( j \)th element \( \lambda^{-j} \). Let \( E_s \) be the \((q_s + 1) \times (q_s + 1) \) matrix whose elements are 0.
except that the first element equals 1, and let \( q^T(A) = s(-1)^T \{ \sum_{\lambda \in A} s(\lambda)s^T(\lambda) \}^{-1} \). The results are unchanged, and the theory simplifies tremendously, if we assume that for each \( \lambda \), the same bandwidth \( h_\lambda \) for all SIMEX replicates.

In our theory we also require that the polynomial extrapolation is exact; that is, \( q^T(A) \sum_{\lambda \in A} \theta_N(x; (1 + \lambda)\Sigma_uu) s(\lambda) = \theta(x) \). Hence the extrapolation results in a consistent estimate of \( \theta(x) \). This is exactly true, of course, only in special cases. The bias that results from the extrapolation changes only the bias expression in the results given later, but not the variance expression.

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

\[
Q(w, \Sigma_uu) = \sum_{j=1}^{m} \{ U_j(w, \Sigma_uu) + \Gamma_j(w, \Sigma_uu) \}
= \mu^{(1)}(\theta_N(w, \Sigma_uu)) \sum_{j=1}^{m} f_j(w, \Sigma_uu) / \phi_j^2(\Sigma_uu),
\]

where \( U_j(w, \Sigma_uu) = [\Gamma_j(w, \Sigma_uu) - \mu(\theta_N(w, \Sigma_uu))]^2 \) and \( \Gamma_j(w, \Sigma_uu) = \text{var}(Y_{ij} | W_{ij} = w) \). In Appendix A.5 we sketch an argument that gives the following approximate bias and variance expansions, assuming that the number of SIMEX replicates \( D \) is large. For simplicity, the bias expressions given here assume that \( p \) is odd:

\[
\text{bias}\{\hat{\theta}_{N,wpe}(x)\} \approx \theta_N(x, \Sigma_uu) - \theta(x) + \mu^{(p+1)}(\theta_N)^{(p+1)}(x, \Sigma_uu) \times \{ e^T \mathbf{E}_p^{-1}(c) \mathbf{E}_c(p+1) \},
\]

\[
\text{bias}\{\hat{\theta}_{sx,wpe}(x)\}
\approx \frac{q^T(A)}{(p+1)!} \sum_{\lambda \in A} h_\lambda^{p+1}(\theta_N)^{(p+1)}(x, (1 + \lambda)\Sigma_uu) s(\lambda) \times \{ e^T \mathbf{E}_p^{-1}(c) \mathbf{E}_c(p+1) \},
\]

and

\[
\text{var}\{\hat{\theta}_{N,wpe}(x)\} \approx (nh_0)^{-1} Q(x, \Sigma_uu) \{ e^T \mathbf{E}_p^{-1}(c) \mathbf{E}_p(\gamma) \mathbf{E}_p^{-1}(c) e \},
\]

and

\[
\text{var}\{\hat{\theta}_{sx,wpe}(x)\} \approx (nh_0)^{-1} Q(x, \Sigma_uu) \times \{ e^T \mathbf{E}_p^{-1}(c) \mathbf{E}_p(\gamma) \mathbf{E}_p^{-1}(c) e \} q^T(A) Q(A).
\]

Equations (11) and (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 \( q^T(A) E_s q(A) \), a factor that 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 \( A = \{0, 1, 2, 3, 4\} \). Then direct calculation shows that use of the quadratic extrapolant leads to an estimator that is 9 times more variable than that based on the linear extrapolant, whereas 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 so 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 ignores measurement error, and then multiply it by the factor \( q^T(A) E_s q(A) \) in (12) to account for the extrapolation. An alternative method uses the sandwich formula and the SIMEX replicates (see Stefanski and Cook 1995, sec. 5.4).

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 a natural ordering is not required, so that the jth observation in one cluster is somehow linked with the jth observation in any other cluster. However, when such a natural ordering exists, the fact is the variance of the SIMEX weighted pooled estimator is inflated by the terms \( U_j(.) \). These terms are an artifact, arising only because that although 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 \( m_t \equiv m \); for example, 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 \( j = 1, \ldots, m \), and then average the estimates using some weights. Because 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 \( \hat{\theta}_{sx,wave}(x) \).

To see how this might work, suppose that the bandwidths in the jth observation are \( h_\lambda \), 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 \( \hat{\theta}_{sx,j}(x) \), is proportional to \( (nh_0)^{-1} \Gamma_j(x, \Sigma_uu) \{ \mu^{(1)}(\theta_j(x, \Sigma_uu)) \}^2 f_j(x, \Sigma_uu))^{-1} \), where \( \theta_j(x, \Sigma_uu) = \mu^{-1}(\gamma_j(x, \Sigma_uu)) \). The constant of proportionality is enclosed in brackets in (12). We construct the SIMEX weighted average estimator \( \hat{\theta}_{sx,wave}(x) \) as the optimal linear combination of the individual estimators as

\[
\hat{\theta}_{sx,wave}(x) = \sum_{j=1}^{m} \alpha_j \hat{\theta}_{sx,j}(x),
\]

where \( \alpha_j \propto \{ \mu^{(1)}(\theta_j(x, \Sigma_uu)) \}^2 f_j(x, \Sigma_uu)) \). Assuming that the poly-
nominal extrapolation is exact for each \( j \)—that is, \( q^T(\mathbf{A}) \sum_{\lambda \in \Lambda} \theta_j(x, (1 + \lambda) \Sigma_{uu}) s(\lambda) = \theta(x) \)—the asymptotic bias of \( \hat{\theta}_{x,w}(x) \) is

\[
\text{bias}\{\hat{\theta}_{x,w}(x)\} \\
\approx \frac{q^T(\mathbf{A})}{(p + 1)!} \sum_{j=1}^{m} \sum_{\lambda \in \Lambda} \alpha_{j,\lambda} h_{\lambda}^{p+1} \theta_j^{(p+1)}(x) \times \{x, (1 + \lambda) \Sigma_{uu}\} s(\lambda) \left(e^T E_p^{-1}(c) E_c(p + 1)\right).
\]

(14)

It is difficult to compare its bias with the bias of the SIMEX weighted pooled estimator \( \hat{\theta}_{x,w}(x) \). However, if \( h_x = h \), assuming that the \( q \)th order polynomial extrapolation is exact for both \( \hat{\theta}_{x,w}(x) \) and \( \hat{\theta}_{x,w}(x) \), then (10) and (14) are identical and are simplified as

\[
\text{bias}\{\hat{\theta}_{x,w}(x)\} = \text{bias}\{\hat{\theta}_{x,w}(x)\} \\
\approx \frac{h_{\lambda}^{p+1}}{(p + 1)!} \theta_j^{(p+1)}(x) \left(e^T E_p^{-1}(c) E_c(p + 1)\right).
\]

This means that the asymptotic bias of the SIMEX estimators \( \hat{\theta}_{x,w}(x) \) and \( \hat{\theta}_{x,w}(x) \) is the same as that when \( X \) is observed.

The variance of the weighted average estimator \( \hat{\theta}_{x,w}(x) \) is proportional to

\[
\text{var}\{\hat{\theta}_{x,w}(x)\} \\
\propto (nh_x)^{-1} \left\{ \sum_{j=1}^{m} \left[ \mu^{(1)} \{\theta_j(x, \Sigma_{uu})\}\right]^2 \times f_{jW}(x, \Sigma_{uu}) \Gamma_j(x, \Sigma_{uu}) \right\}^{-1},
\]

(15)

where again the constant of proportionality is enclosed in brackets in (12). The proof of (15) again has used the fact that the covariance \( \text{cov}\{\hat{\theta}_{x,j}(x), \hat{\theta}_{x,j'}(x)\} = O(n^{-1}) \) for \( j \neq j' \), which is of smaller order than \( \text{var}\{\hat{\theta}_{x,j}(x)\} = O\{(nh_x)^{-1}\} \). In other words, the individual SIMEX estimates \( \hat{\theta}_{x,j}(x) \) are independent asymptotically. In Appendix A.6 we show that the variance of the SIMEX weighted pooled estimator \( \hat{\theta}_{x,w}(x) \) is greater than or equal to the variance of the SIMEX weighted average estimator \( \hat{\theta}_{x,w}(x) \). Of course, the result 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 \( h_x = h \), 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.

Because the weights used to calculate the SIMEX weighted average estimator \( \hat{\theta}_{x,w}(x) \) depend on the unknown density functions \( f_{jW}(x, \Sigma) \) and the unknown conditional variances \( \Gamma_j(x, \Sigma_{uu}) \), it is difficult to calculate \( \hat{\theta}_{x,w}(x) \) using (13) in practice. We hence propose the following procedure, which yields an asymptotically equivalent estimate. For the \( d \)th simulated SIMEX dataset, we first calculate the naive weighted average estimate \( \hat{\theta}_{N,w,d}(x, (1 + \lambda) \Sigma_{uu}) = \sum_{j=1}^{m} \alpha_{jd} \theta_{N,j,d}(x, (1 + \lambda) \Sigma_{uu}) \), where \( \theta_{N,j,d}(x, (1 + \lambda) \Sigma_{uu}) \) is the naive kernel estimate using the simulated \( j \)th observation data \( W_{ij,d}(\lambda) \), and \( \alpha_{jd} \) is the reciprocal of the variance of \( \theta_{N,j,d}(x, (1 + \lambda) \Sigma_{uu}) \) obtained from standard kernel regression (e.g., the sandwich estimate). We then calculate the average of these estimates over \( d = 1, \ldots, D \) 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 \( \hat{\theta}_{N,w,d}(x, (1 + \lambda) \Sigma_{uu}) \) using the sandwich method (see Sec. 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. Because the risk of hospitalization depends on various covariates, such as HIV status, treatments, race, and gender, and we allow only 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 3 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 1,059, with each patient contributing from 1 to 6 observations over time. The major covariate of interest, CD4 count, ranged from 1 to 2,131, and 90% of these patients had CD4 count below 500. As discussed in Section 1, the CD4 counts were measured with error, because the most recent CD4 counts prior to each interview were subject to substantial lab errors. Because 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 Sec. 1 for discussion), we would like to make such dependence as flexible as possible by assuming a nonparametric function to properly identify the functional form. Note that the other covariates included interview time and age. Examination of the data suggests only slight dependence of the risk of hospitalization on time and age, and we did not include them in the model.

We fit model (1) using the logit link with a single covariate \( W \) defined as \( W = \log(\text{CD4}/100) \), a transformation that reduces the marked skewness of CD4 counts. We assumed the measurement errors \( U_{ij} \) were independent and normally distributed with mean 0 and variance \( \sigma^2_U \). However, \( W \) itself is left skewed, 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 \( \sigma^2_U \), one needs either a validation study or
replicates of CD4 count measures. But these were not available in the ACSUS, and hence we were not able to estimate \( \sigma_u^2 \) using the ACSUS. We thus conducted a sensitivity analysis by assuming \( \sigma_u^2 \) equal to 1/4 and 1/2 of the variance of \( W \); that is, assuming \( \sigma_u^2 = .34 \) and \( \sigma_u^2 = .68 \). Wulfsohn and Tsiatis (1995) estimated the measurement error variance of \( \log(\text{CD4}) \) as \( \sigma_u^2 = .39 \) using data from a clinical trial conducted by Burroughs-Wellcome. Our assumption of \( \sigma_u^2 = .34 \) is similar to their finding. Following Wulfsohn and Tsiatis (1995), we assumed that the measurement errors were independent and the measurement error variance \( \sigma_u^2 \) was a constant. If we had validation or replication data, then 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, Carriquiry, Dodd, and Fuller (1996) and Eckert, Carroll, and Wang (1997).

Because different subjects had different numbers of observations, calculation of the weighted average estimate of \( \hat{\theta}(x) \) was difficult. We calculated the SIMEX weighted pooled estimate of \( \hat{\theta}(x) \), letting \( \lambda = (0, 1.0, 1.5, 2.0) \). We used the EBBS method discussed in Section 2.4 to select the bandwidth parameter \( h \) for each simulated dataset and assumed \( t = 2 \) in (8). We further treated \( \sigma_u^2 \) as fixed and known. We used a quadratic extrapolation function in the SIMEX procedure and calculated the standard errors of the SIMEX estimates \( \hat{\theta}_{x,x,\text{wpe}}(x) \) using the standard error estimation method of Stefanski and Cook (1995). The SIMEX method was applied with \( D = 100 \). Analysis of each simulated dataset including estimating the bandwidth parameter \( h \) took only 16 seconds on a SPARC Ultra.

Figures 1–3 plot the estimated \( \hat{\theta}(x) \) against \( x = \log(\text{CD4}/100) \) with 95% confidence intervals assuming \( \sigma_u^2 = 0 \) (naive estimate ignoring measurement error), \( \sigma_u^2 = .34 \) and \( \sigma_u^2 = .68 \). 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 \((\text{CD4} < 14, \log(\text{CD4}/100) < -2)\) or high \((\text{CD4} > 100, \log(\text{CD4}/100) > 0)\) and is fairly stable when CD4 count takes middle values, for example, between 14 and 100. Ignoring measurement error clearly affects the estimated risk of hospitalization. The naive curve is attenuated toward 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 S-PLUS by assuming independence for each simulated SIMEX data and calculated the SIMEX estimate of \( \hat{\theta}(x) \). 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 as 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, Figure 4 compares the SIMEX kernel estimate with the SIMEX linear and quadratic estimates when $\sigma^2_n = .34$. Figure 4 shows that the SIMEX local polynomial kernel estimator seems to have nonlinearity detected neither by the linear model nor by 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 $\sigma^2_n = (0, .34, .68)$, along with 95% bootstrap confidence intervals based on 2,000 bootstrap samples. The coefficient of the cubic term is marginally statistically significant in naive regression when measurement error is ignored, and it is statistically significant for both SIMEX analyses after accounting for measurement error.

### Table 1. Naive and SIMEX Estimates of the Regression Coefficients of the Cubic Models for the ACSUS Data

<table>
<thead>
<tr>
<th></th>
<th>Naive $\hat{\beta}$</th>
<th>SIMEX($\sigma^2_n = .34$)</th>
<th>SIMEX($\sigma^2_n = .68$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-2.19</td>
<td>-1.84</td>
<td>-1.66</td>
</tr>
<tr>
<td>Linear</td>
<td>-5.54</td>
<td>-6.65</td>
<td>-7.5</td>
</tr>
<tr>
<td>Quadratic</td>
<td>-2.9</td>
<td>-6.65</td>
<td>-7.8</td>
</tr>
<tr>
<td>Cubic</td>
<td>-0.06</td>
<td>-1.14</td>
<td>-1.7</td>
</tr>
<tr>
<td>Cubic, 2.5% bootstrap quantile</td>
<td>-1.17</td>
<td>-3.38</td>
<td>-4.5</td>
</tr>
<tr>
<td>Cubic, 97.5% bootstrap quantile</td>
<td>0.007</td>
<td>-0.02</td>
<td>-0.02</td>
</tr>
</tbody>
</table>

**NOTE:** The 4% and 96% bootstrap quantiles of the naive cubic term are $-1.53$ and $-0.01$.

6. DISCUSSION

We have discussed 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 on methods that attempt to use these correlations. Furthermore, correctly specifying correlation in estimation results in an asymptotically less efficient estimator. This is due mainly 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 approaches 1 as $n \to \infty$. It should be noted that our results in this article assume that the working covariance matrix $V$ is invertible and distributions of $Y_i$ and $X_i$ are continuous, 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 a single nonparametric function. We conjecture that our results are applicable to models involving several continuous nonparametric functions; for example, 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 GLIMs with weights inversely proportional to the $\phi_j$.

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; Verbyla et al. 1999; Wang 1998). 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. But our results show that the smoothing spline estimator obtained in this way possibly could be asymptotically less efficient than that obtained by ignoring correlation. These suggestions are, of course, all conjectures, based on an equivalence in the nonclustered framework between local polynomial estimation and smoothing spline estimation. But 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, such as 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 that 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; that is, whether ignoring correlation yields the most efficient SIMEX estimator. More research is needed, although we expect the theory to 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 importance for future work to develop methods that 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 mis specification of the distributions of X.

APPENDIX: THEORY FOR KERNEL METHODS

A.1 Proof of Theorem 1

For $p = 0$, a simple Taylor expansion of (4) shows that its solution $\hat{\beta}_0 = \beta_0(x; h)$ satisfies $\hat{\beta}_0(x; h) - \beta(x) \approx B_n^{-1} A_n$, where

$$B_n = \sum_{j=1}^{m} 1^T \Delta_j(x) V_i^{-1}(x) K_n h(x) \Delta_j(x) 1$$

and

$$A_n = \sum_{j=1}^{m} 1^T \Delta_j(x) V_i^{-1}(x) K_n h(x) [Y_i - \mu(\theta(x))] 1,$$

and $1$ is an $m \times 1$ vector of 1's. Let $B = \lim_{n \to \infty} B_n$. The asymptotic bias of $\hat{\beta}_0(x; h)$ is $B^{-1} E(A_n)$ and the asymptotic variance of $\hat{\beta}_0(x; h)$ is $\text{var}(A_n) / B^2$.

Specifically, some calculations give

$$B = E \left\{ \sum_{j=1}^{m} [\mu^{(1)}(\theta(x))]^2 v^j K_n (X_j - x) \right\} = [\mu^{(1)}(\theta(x))]^2 \sum_{j=1}^{m} v^j f_j(x) + O(h),$$

$$E(A_n) = E \left\{ \sum_{j=1}^{m} \mu^{(1)}(\theta(x)) v^j K_n (X_j - x) [\mu(\theta(X_j)) - \mu(\theta(x))] \right\}.$$

A.2 Proof of Theorem 2

For part (a), see theorem 2 of Ruckstuhl et al. (1999). Here we prove part (b). The results of appendix A.3 of Ruckstuhl et al. (1999) show that when the $X_i$ are independent with density $f_j(\cdot)$, the asymptotic variance of $\hat{\beta}_{1,c}(x)$ is the diagonal element of $B^{-1} \text{cov}(A_n)(B^{-1})^T$, where

$$B = \begin{bmatrix} B_{00} & hB_{01} \\ h^{-1}B_{01} & B_{11} \end{bmatrix}$$

and

$$\text{cov}(A_n) \approx \frac{n}{nh} \begin{bmatrix} A_{00} & h^{-1}A_{01} \\ h^{-1}A_{10} & h^{-2}A_{11} \end{bmatrix}.$$
where $\sigma_{jj} = \phi_j w_j^{-1} V\{\mu(\theta(x))\}$. The bias of $\hat{\theta}_{p,\text{wpe}}(x; h)$ is the first term in (A.1), and the variance is

$$\text{var}\{\hat{\theta}_{p,\text{wpe}}(x; h}\}$$

$$\approx \gamma_k(0)(nh)^{-1} \left\{ \frac{\|\mu(\theta(x))\|^2}{\sigma_{jj}} \sum_{i=1}^m f_j(x)/\sigma_{jj} \right\}^{-1} \times E_p^{-1}(c)E_p(\gamma)E_p^{-1}(c).$$

A.4 Proof of Theorem 4

Reparameterize $G_p(X_i\in x)$ as $G_p\{((X_i-x)/h)\}$ and $\alpha$ as $\alpha_h = h^{\frac{1}{2}}D(\mu(x))/h$. Then $\hat{\theta}_{p,h}(x) = \alpha_h$. A Taylor expansion of (5) gives $\alpha_h = \alpha - \alpha = B_n^{-1}A_n$, where

$$B_n = n^{-1} \sum_{i=1}^n G_i^T(x)\Delta_i(x)K_{th}^{-1}(x)V^{-1}(x)$$

$$\times K_{th}^{-1}(x)\Delta_i(x)G_i(x)$$

and

$$A_n = \sum_{i=1}^n G_i^T(x)\Delta_i(x)K_{th}^{1/2}(x)V^{-1}(x)K_{th}^{1/2}(x)(Y_i - \mu),$$

Because $(Y_i, X_i)$ are iid, we suppress the subscript $i$. Let $B = \lim_{n\to\infty} B_n = \mu(\theta(x))\Delta(x)K_{th}^{-1}(x)V^{-1}(x)K_{th}^{1/2}(x)\Delta(x)G_p(x)$. The $(r_1, r_2)$th component of $B$ is

$$B_{r_1, r_2} = E\left[ \sum_{i=1}^m \sum_{j=1}^m \mu_j^{(1)}(x)\rho_j^{1/2}(X_i - x)K_{th}^{1/2}(X_i - x) \times \left( X_j - x \right)^{-1} \left( X_i - x \right)^{-1} \right],$$

where $\mu_j^{(1)} = \mu(\theta(x))G_i^T((X_j - x)/h)\alpha_h$. Some calculations give

$$B_{r_1, r_2} = \sum_{i=1}^m \int \left[ \mu_j^{(1)}(x)\rho_j^{1/2}(X_i - x)K_{th}^{1/2}(X_i - x) \right] \times \left( X_j - x \right)^{-1} \left( X_i - x \right)^{-1} ds_j + o(h)$$

$$= \mu(\theta(x))\rho_j^{1/2}(x)\sum_{i=1}^m \int \rho_j^{1/2}(x)\epsilon_k(r_1, r_2 - 2) + o(h).$$

It follows that $B = \mu(\theta(x))\sum_{i=1}^m \rho_j^{1/2}(x)E_p(c) + o(h)$. The nth component of $E(A_n)$ is

$$E\left[ \sum_{i=1}^m \sum_{j=1}^m \mu_j^{(1)}(x)\rho_j^{1/2}(X_i - x)K_{th}^{1/2}(X_i - x) \times \left( X_j - x \right)^{-1} \left( X_i - x \right)^{-1} \right]$$

$$\times \left[ h^{p+1}\rho^{(p+1)}(x) \left( \frac{X_i - x}{h} \right)^{p+1} \right] + o(h^{p+2})$$

$$= \sum_{i=1}^m \int \left( \mu_j^{(1)}(x)\rho_j^{1/2}(x) \times \left( X_j - x \right)^{-1} \left( X_i - x \right)^{-1} \right) ds_j \times \left[ h^{p+1}\rho^{(p+1)}(x) \left( \frac{X_i - x}{h} \right)^{p+1} \right] + o(h^{p+2}).$$

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 more complex (see App. A.4 and Carroll, Ruppert, and Welsh 1998). Let $\Psi(Y_i, s) = (Y - \mu(s))\mu^{(1)}(s)/V(s)$. Then, using the techniques of Carroll et al. (1998), it can be shown that $\hat{\theta}_{p,\text{wpe}}(x; h)$ has the expansion

$$\hat{\theta}_{p,\text{wpe}}(x, h) - \theta(x)$$

$$\approx h^{p+1}\frac{\rho^{(p+1)}(x)}{\rho^{(p+1)}(h)} E_p^{-1}(c)E_p(c) + o(h^{p+1})$$

$$+ \left\{ \frac{\rho^{(p+1)}(x)}{\rho^{(p+1)}(h)} \sum_{j=1}^m f_j(x)/\sigma_{jj} \right\}^{-1}$$

$$\times \frac{1}{n^{-1} \sum_{i=1}^m \sum_{j=1}^m e^{T}E_p^{-1}(c)\phi_j^{-1} K_h(X_i - x)$$

$$\times G_p((X_i - x)/h)\Psi(Y_i, \theta(X_i))). \quad \text{(A.1)}$$

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
If \( p = 0 \), then, noting that \( c_K(2) = 1 \), some calculations show that (A.2) becomes
\[
 h^2 \mu^{(1)} \{ \theta(x) \}^2 
 \times \left\{ \theta^{(1)}(x) \sum_{j=1}^{m} v_j^{(+)f_j(x)} + \frac{\theta^{(2)}(x)}{2} \sum_{j=1}^{m} v_j^{(+)f_j(x)} \right\} + o(h^2).
\]
If \( p > 0 \), then (A.2) becomes
\[
 h^{p+1} \frac{\theta^{(p+1)}(x)}{(p+1)!} \left[ \mu^{(1)} \{ \theta(x) \}^2 \sum_{j=1}^{m} v_j^{(+)f_j(x)} c_K(r + p) \right. 
+ h^{p+2} \left. \left\{ \frac{\theta^{(p+1)}(x)}{(p+1)!} \sum_{j=1}^{m} \frac{\theta^{(1)}(x)}{\partial x} f_j(x) \right\} c_K(r + p + 1) \right.
\]
where \( T_j(x) = \mu^{(1)} \{ \theta(x) \}^2 v_j^{(+)f_j(x)} \). Noting that \( c_K(s) = 0 \) and that the \( (1,s+1) \) elements of \( \mathbf{B} \) and \( \mathbf{B}^{-1} \) are 0 if \( s \) is odd, using \( \theta^{(s)}[x; h] = e^{T} \mathbf{B}^{-1} E(\mathbf{A}_n) \), some calculations give the bias expressions of \( \hat{\theta}^{(s)}[x; h] \) stated in Theorem 4.

To calculate the asymptotic variance of \( \hat{\theta}^{(s)}[x; h] \), we first calculate \( \text{cov}(\mathbf{A}_n) \) as
\[
\text{cov}(\mathbf{A}_n) = \frac{1}{n} E (G_p^T \Delta K_h^{-1/2} \Delta V^{-1} K_h^{-1/2} \Delta V^{-1} K_h^{-1/2} \Delta G_p) + o \left( \{nh \}^{-1} \right),
\]
where \( \Sigma = \text{cov}(Y_i, X_i = x_1) \) and \( \sigma_{jk} \) is the \((j,k)\)th element of \( \Sigma \). The \((r_1,r_2)\)th component of the first term is
\[
n^{-1} E \left\{ \sum_{j=1}^{m} \sum_{k=1}^{m} \sum_{l=1}^{m} \sum_{j=1}^{m} \mu_j^{(1)} \mu_k^{(1)} v_{jk} \sigma_{vl} v_l K_h^{-1/2} (X_j - x) \right. 
\times K_h^{-1/2} (X_k - x) K_h^{-1/2} (x_2 - x_1) K_h^{-1/2} (x_1 - x) 
\left. \times \left( \frac{x_2 - x_1}{x_2 - x_1} \right)^{r_1-1} \left( \frac{x_1 - x_1}{x_2 - x_1} \right)^{r_2-1} \right\} 
\]
\[
= (nh)^{-1} \sum_{j=1}^{m} \int \mu_j^{(1)} \{ \theta(x) \}^2 v_j^{(+)2} \sigma_{jj} K^{(2)}(s_j f_j(x + s_j h)) 
\times s_j^{r_1+2} d s_j + o ( \{nh \}^{-1} ) 
\]
\[
= (nh)^{-1} [ \mu^{(1)} \{ \theta(x) \}^2 ] \sum_{j=1}^{m} v_j^{(+)2} \sigma_j K(r_1 + r_2 - 2) 
+ o ( \{nh \}^{-1} ).
\]
Using \( \text{cov}(\hat{\theta}^{(s)}[x; h]) = e^{T} \mathbf{B}^{-1} \text{cov}(\mathbf{A}_n) \mathbf{B}^{-1} e \), we have the expression of \( \text{cov}(\hat{\theta}^{(s)}[x; h]) \) as that given in part (b). A direct application of the Cauchy–Schwartz inequality gives part (c).

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 the followings, the function that follows, the argument (•) refers to \( G_p^T \{ (W_{ij} - w)/h \} \beta \), the argument (•) refers to \( \theta_N(W_{ij}, \Sigma_{u}) \), and the argument (•) refers to \( \theta_N(w, \Sigma_{u}). \) The first \( p + 1 \) terms of the Taylor series expansion of \( \theta_N(W_{ij}, \Sigma_{u}) \) about \( \theta_N(w, \Sigma_{u}) \) are given by \( G_p^T \{ (W_{ij} - w)/h \} \beta \). We solve \( \beta \) by
\[
n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{Y_{ij} - \mu_1(\cdot)}{\sigma_j(\Sigma_{u})} V(\cdot) \mu_1(\cdot) K_h(W_{ij} - w) 
\times G_p \{ (W_{ij} - w)/h \} = 0.
\]
It is easily seen by a first-order Taylor expansion and using (9) that \( \beta - \hat{\beta} = B_n^{-1} A_n \), where
\[
B_n = n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{\{Y_{ij}(\cdot)\}^2}{\sigma_j(\Sigma_{u})} V(\cdot) K_h(W_{ij} - w) 
\times G_p \{ (W_{ij} - w)/h \} G_p^T \{ (W_{ij} - w)/h \} 
\approx E_p(\sigma) \{ \mu^{(1)}(\cdot)^2 / V(\cdot) \} \sum_{i=1}^{m} f_{ij}(W, \Sigma_{u}) / \phi_j(\Sigma_{u}) 
\times o_p(1) = B + o_p(1),
\]
\[
A_n = A_n(1) + A_n(2),
\]
\[
A_{n(1)} = n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{Y_{ij} - \mu(\cdot)}{\phi_j(\Sigma_{u})} V(\cdot) \mu(\cdot) 
\times K_h(W_{ij} - w) G_p \{ (W_{ij} - w)/h \},
\]
\[
A_{n(2)} = n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{\mu(\cdot) - \mu(\cdot)}{\phi_j(\Sigma_{u})} V(\cdot) \mu^1(\cdot) 
\times K_h(W_{ij} - w) G_p \{ (W_{ij} - w)/h \},
\]
It is easily seen that
\[
A_{n(2)} \approx \{ \mu^{(1)}(\cdot)^2 / V(\cdot) \} \{ h^{p+1} \theta^{(p+1)}(\cdot)(w)/(p+1) \} 
\times E_p(\sigma) \{ (w - \mu(\cdot))/V(\cdot) \} \sum_{j=1}^{m} f_{ij}(w, \Sigma_{u}) / \phi_j(\Sigma_{u}),
\]
and hence that
\[
\hat{\theta}_N(w) - \theta_N(w) \approx h^{p+1} \theta^{(p+1)}(\cdot)(w) e^T E_p(\sigma) \{ e(\cdot) + o(1) \}.
\]
Remembering that \( E(Y_{ij}|W_{ij} = w) = \zeta_i(w) \) and using (9), a tedious but straightforward calculation shows that \( E(\mathbf{A}_n) = 0 \). 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 (•) by (•) in the definition of \( A_{n(1)} \) leading to the expression \( A_{n(1)} = A_{n(1)} + A_{n(2)} \), where
\[
A_{n(1)} = n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{Y_{ij} - \zeta_i(W_{ij}, \Sigma_{u}) - \mu(\cdot)}{\phi_j(\Sigma_{u})} V(\cdot) \mu(\cdot) 
\times K_h(W_{ij} - w) 
\times G_p \{ (W_{ij} - w)/h \}.
\]
and
\[
A_{n(2)} = n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{\zeta_i(W_{ij}, \Sigma_{u}) - \mu(\cdot)}{\phi_j(\Sigma_{u})} V(\cdot) \mu^1(\cdot) 
\times K_h(W_{ij} - w) 
\times G_p \{ (W_{ij} - w)/h \}.
\]
Because \( A_{n(2)} \) 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
\[
\frac{(\mu^{(1)}(\cdot))^2 E_p(\gamma)}{n_h V^2(\cdot)} \times \sum_{j=1}^{m} (U_j(w, \Sigma_{uw}) + \Sigma_j(w, \Sigma_{uw})) f_j w(w, \Sigma_{uw}) / \phi_j^2(\Sigma_{uw}).
\]
We have thus shown (11), namely that the variance of $\hat{\theta}_N(w, \Sigma_{uw})$ is asymptotically
\[
\text{var}(\hat{\theta}_N(w, \Sigma_{uw})) \approx (nh)^{-1} Q(w, \Sigma_{uw}) e^T E_p^{-1}(c) E_p(\gamma) E_p^{-1}(c) e.
\]
In the case where the $(Y, X, W)$'s are marginally identically distributed, although not necessarily independent, simplification occurs because $U_j(w, \Sigma_{uw}) = 0$, $\zeta_j = \mu(\theta_N)$ and none of the terms $\Gamma_j, \phi_j$, or $f_j w$ depends 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 used by Carroll et al. (1999). Because the calculations are similar, although tedious, in the interest of space we have chosen not to provide them here. The key step in the proof is to show that \[\text{var}(\hat{\theta}_N(x; (1 + \lambda) \Sigma_{uw})) = O((nhD)^{-1}) + O(n^{-1})\]
for $\lambda > 0$, which is of smaller order than \[\text{var}(\hat{\theta}_N(x; \Sigma_{uw})) = O((nh)^{-1}).\]

A.6 Comparison of the Variances of $\hat{\theta}_{x_{W}}, \hat{\theta}_{x_{W_{0}}} (x)$ and $\hat{\theta}_{x_{W_{0}}} (x)$

Using the Cauchy–Schwarz inequality, we have
\[
\left\{ \sum_{j=1}^{m} \frac{\Gamma_j(w)}{\phi_j^2} \right\} \left\{ \sum_{j=1}^{m} \frac{\mu^{(1)}(\theta_j)}{\phi_j} f_j w(w) \right\} \geq \left\{ \sum_{j=1}^{m} \mu^{(1)}(\theta_j) f_j w(w) \right\}^2 \geq \left\{ \sum_{j=1}^{m} \mu^{(1)}(\theta_j) f_j w(w) \right\}^2 \geq \sum_{j=1}^{m} \frac{\mu^{(1)}(\theta_j)}{\phi_j^2} f_j w(w).\]
Using equation (9) and noting $\zeta_j(\cdot) = \mu(\theta_j(\cdot))$, the last term is $[\mu^{(1)}(\theta_N(\cdot)) \sum_{j=1}^{m} f_j w(w) / \phi_j^2]$. We have
\[
\sum_{j=1}^{m} \frac{\mu^{(1)}(\theta_N(\cdot)) \sum_{j=1}^{m} f_j w(w) / \phi_j^2}{\phi_j^2} \geq \sum_{j=1}^{m} \frac{1}{\phi_j^2} f_j w(w).\]
Further noting that $U_j(\cdot) \geq 0$, we have \[\text{var}(\hat{\theta}_{x_{W_{0}}}(x)) \geq \text{var}(\hat{\theta}_{x_{W_{0}}}(x)).\]

[Received November 1998. Revised October 1999.]