Statistica Sinica **10**(2000), 905-921

FLEXIBLE REGRESSION CALIBRATION FOR COVARIATE MEASUREMENT ERROR WITH LONGITUDINAL SURROGATE VARIABLES

C. Y. Wang

Fred Hutchinson Cancer Research Center

Abstract: In regression analysis, covariate measurement error occurs in many applications. If a covariate variable of interest for a subject is the long–term average of some measurements, then in practice repeated measurements are considered surrogates for the true covariate. Surrogate variables, which are longitudinal, may be modeled as the sum of the unobserved true covariate and longitudinal errors, where the errors are dependent with a continuous correlation function of time. In this paper, we consider a flexible modeling of the correlation of the surrogate variable. This proposed polynomial correlation modeling is not as sensitive as an exponential type autocorrelation. A refined regression calibration estimator is studied for logistic regression. Simulation studies were conducted to examine the finite sample performance of a cubic correlation–based regression calibration estimator for exponential and piecewise–linear correlation models. The asymptotic covariance of the proposed estimator is given. The proposed method is applied to a study of adult obesity in relation to childhood body mass index.

Key words and phrases: Maximum likelihood, measurement error, polynomial regression, regression calibration.

1. Introduction

The problem of covariate measurement error has been reviewed in Fuller (1987) for linear regression and Carroll, Ruppert and Stefanski (1995) for nonlinear regression. It occurs in many applications (Carroll, Spiegelman, Lan, Bailey and Abbott (1984), Rosner, Willett and Spiegelman (1989), Pierce, Stram, Vaeth and Schafer (1992), Prentice (1996)). Consider a regression model with a univariate outcome variable Y and covariate X. One problem of interest is that, instead of observing true covariate X_i $(i = 1, \ldots, n)$, longitudinal surrogate variables $W_{ij} \equiv W_i(t_{ij})$, $j = 1, \ldots, k_i$ are available. In an additive error model $W_i(t_{ij}) = X_i + U_i(t_{ij})$, X_i usually denotes a long–term average of $W_i(t)$, *i.e.*, $\sum_{i=1}^{k_i} W_i(t_{ij})/k_i$ for $k_i \to \infty$. If the $W_i(t_{ij})$ are independent, then it is well known that using observed averages of $W_i(t_{ij})$ may have an attenuation effect. Methods for independent $W_i(t_{ij})$ have been well-addressed; see the related papers cited in the two monographs described above.

A motivating example of this work is a study of childhood growth. Consider adult obesity as the outcome variable, and let the covariate variable be the long–term average body mass index (BMI) z–score between ages 1 and 4. Measurement error analysis is needed in connection with the long–term average. Furthermore, a dependent error structure is required since $W_i(t)$ is a continuous function of time.

Wang, Carroll and Liang (1996) described estimation with a special covariance matrix Σ_u, assuming equally–spaced times and an autocorrelation model. A more general setting for unequally–spaced times with an autoregression model was discussed in Wang and Pepe (2000). However, an autoregression error model will not hold in general. Furthermore, even if the autoregression model holds, a divergence problem may occur when the sample size is moderate (say $n = 100$).

In this paper, we consider estimation based on a more flexible regression model for the correlation of the error, which leads to a more flexible modeling of the correlation of $W_i(t_1)$ and $W_i(t_2)$ for times t_1 and t_2. Although this modeling of correlation may be applied to other methods, the focus of this paper is on the application to the regression calibration approach; see Carroll, Ruppert and Stefanski (1995, Chapter 3) for a general review. The idea in this paper is to model the correlation process as a moderate order polynomial function so the goodness of fit of the modeling can be examined with available data.

Section 2 describes the model and reviews some previously developed methods. In Section 3, a more flexible model for the correlation of observed surrogate variables is investigated. The simulation results are given in Section 4 for both linear and logistic regression. In Section 5, analyses of data from the study of childhood predictors of adult obesity are presented.

2. Model and Existing Methods

2.1. The model

Let Y_i be the outcome variable for the ith of n subjects, X_i be a corresponding covariate variable which is measured with error, and Z_i be a covariate vector which is measured without error. The regression model of interest is written as $P_\beta(Y|X,Z)$ for some unknown parameter β. Let the longitudinal surrogates be denoted by $\tilde{W}_i = (W_i(t_{i1}),\ldots,W_i(t_{ik_i}))$, assumed available. Assume that $P_\beta(Y|X,\tilde{W},Z) = P_\beta(Y|X,Z)$. This has been called the surrogacy condition and it means that $\tilde{W}$ offers no additional information regarding the outcome Y given data on the true covariate (X,Z). For notational simplicity, we suppress the notation for Z but note that conditioning on Z is assumed throughout. We consider the additive model $W_i(t) = X_i + U_i(t)$, where X_i has a normal distribution with mean μ_x and variance σ_x^2. The error process U_i is assumed to be Gaussian with $E\{U_i(t)\} = 0$ and $E[U_i(t_1)U_i(t_2)] = \sigma_u^2\rho_u(t_2 - t_1)$ for some correlation function ρ_u with unknown nuisance parameters. This implies that W_i is a Gaussian process with mean $\mu_x 1_{k_i}$ and variance–covariance matrix $\sigma_w^2\mathcal{G}_i$, where 1_{k_i} is an identity vector of length k_i, $\sigma_w^2 = \sigma_x^2 + \sigma_u^2$, and the (k, l)th element of matrix $\mathcal{G}_i$ is $\rho_w(t_{ik} - t_{il})$, where $\rho_w(t_{i2} - t_{i1}) = \{\sigma_x^2 + \sigma_u^2\rho_u(t_{i2} - t_{i1})\}/(\sigma_x^2 + \sigma_u^2)$.

The goal is to estimate β in the presence of some nuisance parameters, including σ_u^2, μ_x, σ_x^2 and the related parameters in modeling $\rho_u(t)$. The main point of the repeated measurements is to understand the variance of the measurement errors. Assume that $\rho_u(t) = \rho_u(-t)$, which implies that the correlation of $W_i(t_1)$ and $W_i(t_2)$ is a function of $|t_2 - t_1|$. In the classical additive model, $\rho_u(t)$ is assumed to be zero. This is usually not the case for longitudinal measurements because if $W_i(t)$ is continuous, then $W_i(t) \to W_i(t_0)$ if $t \to t_0$ and hence $\rho_w(t - t_0) \to 1$, or equivalently $\rho_u(t - t_0) \to 1$.

One approach to this problem is to replace X_i by the sample average $\overline{W}_{i\cdot}$. To demonstrate the bias problem of this naive estimator, we consider linear regression $E(Y_i|X_i) = \beta_0 + \beta_1 X_i$. It is easily seen that $E(Y_i|\overline{W}_{i\cdot}) = \beta_0 + \beta_1 E(X_i|\overline{W}_{i\cdot})$. Write $\text{cov}(\tilde{U}_i) = \sigma_u^2 M_i$, then $E(X_i|\overline{W}_i) = \lambda_i\overline{W}_{i\cdot} + (1 - \lambda_i)\mu_x$, where $\lambda_i = \sigma_x^2\{\sigma_x^2 + (\sigma_u^2/k_i^2)1_{k_i}^t M_i 1_{k_i}\}^{-1}$. As a result, $E(Y_i|\overline{W}_{i\cdot}) = \beta_{0*} + \beta_{1*}\overline{W}_{i\cdot}$, where $\beta_{1*} = \lambda_i\beta_1$, and $\beta_{0*} = \beta_0 + (1 - \lambda_i)\mu_x\beta_1$. Therefore, there is an attenuation effect if one ignores the measurement error problem. Wang, Carroll and Liang (1996) discuss the effect due to the ignorance of correlated $U_i(t_1)$ and $U_i(t_2)$.

2.2. Expected estimating equation approach

The likelihood–based approach has been proposed by Schafer and Purdy (1996) for a general measurement error problem, but they did not specifically consider the correlated error process. In this special problem if the estimating score is obtained directly from the likelihood, then it is equivalent to the expected estimating equation (EEE) approach proposed by Wang and Pepe (2000). These authors also considered the covariate measurement error problem in marginal or partly conditional regression of longitudinal data. It is noted that the likelihood–based score is $\partial/(\partial\beta)\text{log}P(Y|\tilde{W}) = E\{\partial/(\partial\beta)\text{log}P(Y|X)|Y, \tilde{W}\}$. Thus, the maximum likelihood (ML) estimator solves

$$\sum_{i=1}^{n} E\{S_\beta(Y_i, X_i)|Y_i, \tilde{W}\} = 0. \tag{1}$$

Note that $P(Y_i|X_i, \tilde{W}_i)P(\tilde{W}_i|X_i)P(X_i) = P(Y_i|X_i)P(\tilde{W}_i|X_i)P(X_i)$. Hence, (1) may be written as

$$\sum_{i=1}^{n} \frac{\int S_\beta(Y_i, x)P_\beta(Y_i|x)P_x(\tilde{W}_i)P(x)dx}{\int P_\beta(Y_i|x)P_x(\tilde{W}_i)P(x)dx} = 0.$$

One finds that, (1) requires numerical integration in nonlinear regression settings, but it is not needed in linear regression with a normally distributed error. In an illustration, Wang and Pepe (2000) considered $\rho_u(t) = \rho^{|t|}$. The estimation of the nuisance parameters may be based on the maximum likelihood estimator or the method of mements. They applied the method to marginal or partly conditional regression of longitudinal outcome and covariate variables.

2.3. Regression calibration

An alternative approach to this problem is regression calibration, where one replaces the unobservable X_i by $E(X_i|\tilde{W}_i)$. Under the Gaussian process model described in Section 2.1, we have

$$\begin{pmatrix} X_i \\ \tilde{W}_i \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_x \\ \mu_x 1_{k_i} \end{pmatrix}, \sigma_w^2 \begin{pmatrix} 1 & \sigma_x^2\sigma_w^{-2} 1_{k_i} \\ \sigma_x^2\sigma_w^{-2} 1_{k_i}^t & \mathcal{G}_i \end{pmatrix}\right).$$

Note that if Y given X is normal and all the nuisance parameters were known, then the RC estimator is the ML estimator. Wang and Pepe (2000) assumed an autoregression model with $\rho_u(t) = \rho^t$ for some positive ρ. Then we have

$$E(X_i|\tilde{W}_i) = \mu_x + (\sigma_x^2 1_{k_i})^t \{\sigma_x^2 I_{k_i} + \sigma_u^2 M_i\}^{-1} (\tilde{W}_i - \mu_x 1_{k_i}),$$

where the (l, m)th element of M_i is $\rho^{|t_{il} - t_{im}|}$ and I_{k_i} is the $k_i \times k_i$ identity matrix. Let $W_{ij}^* = W_{i(j+1)} - W_{ij}$. The RC estimator may be obtained by solving the estimating equations for $\Theta = (\beta_0, \beta_1, \sigma_u, \rho, \mu_x, \sigma_x)^t$:

$$\begin{cases} \sum_{i=1}^{n} S_\beta\{Y_i, E(X_i|\tilde{W}_i)\}; \\ \sum_{i=1}^{n} k_i(\overline{W}_{i\cdot} - \mu_x); \\ \sum_{i=1}^{n} k_i \left\{(\overline{W}_{i\cdot} - \mu_x)^2 - \sigma_x^2 - (1_{k_i}^t \sigma_u^2 M_i 1_{k_i})/k_i^2\right\}; \\ \sum_{i=1}^{n} \Big[\sum_{j=1}^{k_{i-1}} \left\{W_{ij}^{*2} - 2\sigma_u^2(1 - \rho^{|t_{i(j+1)} - t_{ij}|})\right\}\Big]; \\ \sum_{i=1}^{n} \Big[\sum_{j=1}^{k_i - 2} \left\{W_{ij}^* W_{i(j+1)}^* - \sigma_u^2(\rho^{|t_{i(j+2)} - t_{i(j+1)}|} - 1 - \rho^{|t_{i(j+2)} - t_{ij}|} + \rho^{|t_{i(j+1)} - t_{ij}|})\right\}\Big]. \end{cases} \tag{2}$$

In (2), the last four equations are from moment calculations for $(\sigma_u, \rho, \mu_x, \rho_x)$. The advantage of the RC estimator is that it is efficient in linear regression and it has good performance with small mean square error in logistic regression if the relative risk is not too large. However, there are two associated problems: (i) a bias problem especially when the relative risk is large; (ii) it is very sensitive to the assumed model $\rho_u(t) = \rho^{|t|}$. The first problem will be addressed next, the second later.

2.4. Refined RC estimator for binary outcome

For binary outcome regression, the RC estimator may have a bias problem when either β_1 or $\text{var}(X|\tilde{W})$ is large, and especially in the former case. The bias problem of the RC estimator for nonlinear regression can be seen from the following Taylor series expansion. Assume $E(Y_i|X_i) = \Psi(\beta_0 + \beta_1 X_i)$ for some function Ψ. Let $(\partial^2/\partial^2 x)\Psi(x) = \Psi''(x)$. Then $E(Y_i|\tilde{W}_i) = E\{\Psi(\beta_0 + \beta_1 X_i)|\tilde{W}_i\}$ $\approx \Psi\{\beta_0 + \beta_1 E(X_i|\tilde{W}_i)\} + (\beta_1^2/2)\Psi''\{\beta_0 + \beta_1 E(X_i|\tilde{W}_i)\}E[\{X_i - E(X_i|\tilde{W}_i)\}^2|\tilde{W}_i]$. To reduce this bias problem, a refined analysis seeks a more precise approximation of $E(Y_i|\tilde{W}_i)$, as a function of $E(X_i|\tilde{W}_i)$ and $\text{var}(X_i|\tilde{W}_i)$. First, we consider probit regression: $\text{pr}(Y = 1|X) = \Phi(\beta_0 + \beta_1 X)$ where Φ is a standard normal distribution function. As in Carroll, Spiegelman, Lan, Bailey and Abbott (1984), Zeger, Liang and Albert (1988) and Liang and Liu (1991),

$$\text{pr}\{Y_i = 1|\tilde{W}_i\} = \int \Phi(\beta_0 + \beta_1 x) f(x|\tilde{W}_i)dx = \int \Phi(\beta_0 + \beta_1 x) d\Phi\left(\frac{x - \mu_{x|\tilde{W}_i}}{\sigma_{x|\tilde{W}_i}}\right),$$

where $\mu_{x|\tilde{W}_i} = E(X|\tilde{W}_i)$ and $\sigma^2_{x|\tilde{W}_i} = (\sigma_x^2) - (\sigma_x^2 1_{k_i})^t\{\sigma_x^2 I_{k_i} + \sigma_u^2 M_i\}^{-1}(\sigma_x^2 1_{k_i})$. If $\Phi(\dfrac{x - \mu_{x|\tilde{W}_i}}{\sigma_{x|\tilde{W}}}) = s$, it can be shown that

$$\begin{aligned}\text{pr}(Y_i = 1|\tilde{W}_i) &= \int_0^1 \Phi(\beta_0 + \beta_1 \Phi^{-1}(s)\sigma_{x|\tilde{W}_i} + \beta_1 \mu_{x|\tilde{W}_i})ds \\ &= \Phi\left((\beta_0 + \beta_1 \mu_{x|\tilde{W}_i})/(1 + \beta_1^2 \sigma^2_{x|\tilde{W}_i})^{1/2}\right).\end{aligned}$$

Therefore, for probit regression, a refined estimator may be obtained by solving

$$n^{-1/2}\sum_{i=1}^n \begin{pmatrix} a_i \\ b_i \end{pmatrix}\{Y_i - \Phi(a_i\beta_0 + b_i\widehat{\mu}_{x|\tilde{W}_i}\beta_1)\} = 0,$$

where $a_i = (1 + \beta_1^2\widehat{\sigma}^2_{X|\tilde{W}_i})^{-1/2}$, $b_i = (a_i - \beta_1^2 a_i^3 \widehat{\sigma}^2_{X|\tilde{W}_i})\widehat{\mu}_{X|\tilde{W}_i} - \beta_0\beta_1 a_i^3 \widehat{\sigma}^2_{X|\tilde{W}_i}$.

For logistic regression, $\text{pr}(Y_i = 1|X_i) = H(\beta_0 + \beta_1 X_i)$, where $H(u) = \{1 + \exp(-u)\}^{-1}$. As in Liang and Liu (1991) and Carroll, Ruppert and Stefanski (1995, p.65), $\Phi(x) \approx H(1.7x)$, so

$$\text{pr}(Y_i = 1|\tilde{W}_i) \approx H\Big(\frac{\beta_0 + \beta_1 \mu_{X_i|\tilde{W}_i}}{(1 + \beta_1^2 \sigma^2_{X|\tilde{W}_i}/2.89)^{1/2}}\Big).$$

Let $\widehat{\mu}_{x|\tilde{W}_i}$ and $\widehat{\sigma}^2_{x|\tilde{W}_i}$ be consistent estimators of $\mu_{x|\tilde{W}_i}$ and $\sigma_{x|\tilde{W}_i}$ respectively. Let $c_i = \{1 + \beta_1^2\widehat{\sigma}^2_{X|\tilde{W}_i}/2.89\}^{-1/2}$, $d_i = (\partial/\partial\beta_1)(c_i\beta_0 + c_i\beta_1\widehat{\mu}_{X|\tilde{W}_i}) = (c_i - \beta_1^2c_i^3\widehat{\sigma}^2_{X|\tilde{W}_i}/2.89)\widehat{\mu}_{X|\tilde{W}_i} - \beta_0\beta_1c_i^3\widehat{\sigma}^2_{X|\tilde{W}_i}/2.89$. Then the refined regression calibration (RRC) solves

$$n^{-1/2}\sum_{i=1}^{n}\begin{pmatrix} c_i \\ d_i \end{pmatrix}\{Y_i - H(c_i\beta_0 + c_i\widehat{\mu}_{x|\tilde{W}_i}\beta_1)\} = 0.$$

This analysis can reduce the bias significantly for large relative risk. However, the sensitivity due to the assumption that $\rho_u(t) = \rho^{|t|}$ needs to be considered and we turn to that.

3. Polynomial Correlation Regression

Modeling the correlation $\rho_u(t)$ of the error process, or $\rho_w(t)$ of $W(t)$, is essential in our approach. In this section we consider a more flexible model for the correlation by modeling $\rho_w(|t_2 - t_1|)$ as a polynomial function of $|t_2 - t_1|$, for example, a cubic or quartic correlation regression. Two advantages are as follows:

(i) A cubic function or a quartic function can reasonably model an exponentially correlated model, *i.e.*, $\rho_w(|t_2-t_1|) = (\sigma_x^2 + \sigma_u^2\rho^{|t_2-t_1|})/(\sigma_x^2+\sigma_u^2)$, but not *vice versa.*

(ii) Although a nonparametric smoother may relax the model assumption on $\rho_w(t)$, it is technically difficult to solve finite sample difficulties when estimating β. A cubic or quartic function may be enough to approximate the correlation curve, and is simple to implement.

We now consider estimation of the correlation function $\rho_w(t)$. It can be shown that $E\{W_i(t) - \mu_x\}^2 = \sigma_u^2 + \sigma_x^2 = \sigma_w^2$ and $E[\{W_i(t_1) - \mu_x\}\{W_i(t_2) - \mu_x\}] = \sigma_w^2\rho_w(t_2 - t_1)$. Let $\overline{W}_{\cdot\cdot} = \sum_{i=1}^n k_i\overline{W}_{i\cdot}/N$, where $N = \sum_{i=1}^n k_i$. Note that $\widehat{\mu}_x = \overline{W}_{\cdot\cdot}$ is a consistant estimator of μ_x. By some algebra, as $n \to \infty$,

$$\begin{aligned} &E\{W_i(t) - \overline{W}_{\cdot\cdot}\}^2 = \sigma_w^2 + O(n^{-1}); \\ &E[\{W_i(t_1) - \overline{W}_{\cdot\cdot}\}\{W_i(t_2) - \overline{W}_{\cdot\cdot}\}] = \sigma_w^2\rho_w(|t_1 - t_2|) + O(n^{-1}). \end{aligned} \tag{3}$$

A consistent estimator of σ_w^2 is $\widehat{\sigma}_w^2 = N^{-1}\sum_{i=1}^n\sum_{j=1}^{k_i}\{W_i(t_{ij}) - \overline{W}_{\cdot\cdot}\}^2$. Let $t_{ij,m} = |t_{ij} - t_{im}|$, $V_{ij,m} = \{W_i(t_{ij}) - \overline{W}_{\cdot\cdot}\}\{W_i(t_{im}) - \overline{W}_{\cdot\cdot}\}/\widehat{\sigma}_w^2$. By (3), it is easily seen that $E(V_{ij,m}|t_{ij,m}) = \rho_w(t_{ij,m}) + O(n^{-1})$. The proposal here is to model ρ_w by a polynomial function of order q based on the correlation regression:

$$\rho_w(t) = 1 + \sum_{s=1}^{q}\gamma_s t^s. \tag{4}$$

The choice of q depends on the problem of interest, but in practice it should not be too large, say larger than the maximum of the number of replicates. On the other hand, if there are lots of replicates, a naive estimator will work since the sample average $\overline{W}_{i\cdot}$ would estimate $E(X|\tilde{W})$ well in this case. Here, the intercept in (4) is restricted to 1 because $\rho_w(0) = 1$. Applying the least squares estimate of γ in (4) using data $\{V_{ij,m}, t_{ij,m}\}$, $\widehat{\rho}_w(t) = 1 + \sum_{s=1}^{q} \widehat{\gamma}_s t^s$. We further note that $\sigma_x^2 = \sigma_w^2 \text{lim}_{t\to\infty}\rho_w(t)$. Hence σ_x^2 can be consistently estimated without the moment calculations similar to the last three equations of (2), as long as $\max\{t_{ij,m} : i = 1,\ldots,n; j,m = 1,\ldots,k_i\} \equiv T^*$ is large enough such that the correlation of $\rho_u(T^*)$ is ignorable. This is probably true in lots of applications. For example, in an autoregression model if the errors are moderately correlated with $\rho = 0.2$ for equally–spaced data and $k_i = 4$, then the correlation between U_{i1} and U_{i4} would be only .008. In general, one could draw the curve $\widehat{\rho}_w(t)$ to examine this assumption. If this is not the case then it is necessary to apply the method of moments similar to (2), replacing ρ^t with a polynomial function of t.

The RC estimator can be implemented by noting

$$\widehat{E}(X_i|\tilde{W}_i) = \widehat{\mu}_x + (\widehat{\sigma}_x^2 1_{k_i})^t \{\widehat{\sigma}_w^2 \widehat{\mathcal{G}}_i\}^{-1} (\tilde{W}_i - \widehat{\mu}_x 1_{k_i}) = \widehat{\mu}_x + \{\widehat{\rho}_w(T^*) 1_{k_i}\}^t \widehat{\mathcal{G}}_i^{-1} (\tilde{W}_i - \widehat{\mu}_x 1_{k_i}),$$

where $\widehat{\mu}_x$ and $T^* = \max\{t_{ij,m}; i = 1,\ldots,n; j,m = 1,\ldots,k_i\}$ were defined above, and $\widehat{\mathcal{G}}_i$ is a $k_i \times k_i$ matrix with the (j,m)th element being $\widehat{\rho}_w(t_{ij,m})$. Similarly, for logistic regression a refined RC estimator can be obtained by the procedure described in Section 2.4.

The limiting distribution of the RC estimator is given in the Appendix. When Y given X is linear, the asymptotic covariance of $n^{1/2}(\widehat{\beta} - \beta)$ is given in (7). The corresponding formula for logistic regression is given in (8). The asymptotic covariance of the RRC estimator is omitted because it is similar to that of the RC estimator.

4. Simulation Studies

The small sample behavior of the proposed estimator is examined in some Monte-Carlo studies. The regression models for Y given X were linear and logistic, respectively. There were four replicates of W_{ij}, $W_{ij} = X_i + U_i(t_{ij})$, where X_i is normal with mean μ_x and variance σ_x^2; t_{ij} were randomly distributed in [0,4]. The error process U_i will be described later. In Tables 1–5, the true parameters and sample sizes n used in each simulation study are shown. In the tables, "bias" was calculated by taking the average of $\widehat{\beta} - \beta$ from 500 replicates, "s.d." denotes the sample standard deviation of the estimates, "mean(s.e.)" denotes the average of the estimated standard errors of the estimates. The 95% confidence interval coverage probabilities are also included.

4.1. Linear regression

Tables 1 and 2 consider linear regression. The methods considered are (i) a naive estimator which replaces X_i by $\overline{W}_{i\cdot}$; (ii) the RC estimator assuming $\rho_u(t) = \rho^{|t|}$, denoted by RC$_{ar}$; (iii) the RC estimator using a cubic correlation regression model, denoted by RC$_{cu}$. The RC$_{ar}$ estimates were obtained using estimating equations in (2) and the RC$_{cu}$ estimates were obtained from the cubic correlation function described in Section 3. To avoid a finite sample performance problem in estimating $\sigma_x^2 \equiv \sigma_w^2\rho_w(T^*)$, we used the 90th percentile point of $\{t_{ij,m}; i = 1, \ldots, n; j, m = 1, \ldots, k_i\}$ as T^*, since the polynomial correlation may have poor estimation in the upper 10% of the $t_{ij,m}$ points. In Table 1, the error process $U_i(t_{ij})$ is an autoregression (AR) model such that $\text{corr}\{U_i(t_1), U_i(t_2)\} = \rho^{|t_2-t_1|}$. Observe that the RC$_{ar}$ is a pseudo–ML estimator since we consider linear regression of Y given X, and the RC estimator is obtained from the likelihood of Y_i on $\tilde{W}_i$. The parameters are $\beta = (-1, 1)$, $(\sigma_u, \mu_x, \sigma_x) = (0.5, 0, 1)$ and $\rho = 0.2$. The naive estimator underestimates β_1 but not β_0 because $\mu_x = 0$ in our setting. The RC estimator assuming a cubic correlation coefficient is almost as good as the RC estimator assuming a correctly specified AR model. Note that although RC$_{ar}$ was obtained under a correct error model, the Newton Raphson algorithm solving (2) did not converge in about 5% of the cases with $n = 100$. The results for RC$_{ar}$ with $n = 100$ were calculated from 500 data sets in which RC$_{ar}$ converged. The divergence problem occurred mainly because the $\rho^{|t|}$ value involved in the computing algorithm came with a negative ρ and a non–integer t.

Table 1. Linear regression with exponential correlation.

		$n = 300$			$n = 100$		
		Naive	RC$_{ar}$	RC$_{cu}$	Naive	RC$_{ar}$	RC$_{cu}$
β_0	bias	0.003	0.003	0.003	-0.005	-0.004	-0.003
	s.d.	0.058	0.058	0.059	0.103	0.104	0.104
	mean(s.e.)	0.060	0.060	0.060	0.104	0.104	0.105
	95% cov.	0.948	0.954	0.946	0.944	0.948	0.942
β_1	bias	-0.084	0.000	0.002	-0.088	0.005	-0.003
	s.d.	0.058	0.065	0.067	0.098	0.113	0.119
	mean(s.e.)	0.058	0.065	0.069	0.100	0.116	0.124
	95% cov.	0.710	0.934	0.952	0.866	0.968	0.970

Note: Parameters are $(\beta_0, \beta_1) = (-1, 1)$; $(\sigma_u, \mu_x, \sigma_x) = (0.5, 0, 1)$ and $\rho_u(t) = \rho^{|t|}$ with $\rho = .2$. The RC$_{ar}$ estimator here is the ML estimator under the correct specification of the distributions of $X_i|\tilde{W}_i$ and $\tilde{W}_i$. The RC$_{cu}$ models $\rho_w(t)$ as a cubic function of t. The result was from 500 replicates.

Table 2. Linear regression with piecewise linear correlation.

		$n = 300$			$n = 100$		
		Naive	RC$_{ar}$	RC$_{cu}$	Naive	RC$_{ar}$	RC$_{cu}$
β_0	bias	0.001	—	0.002	0.008	—	0.008
	s.d.	0.065	—	0.065	0.108	—	0.110
	mean(s.e.)	0.060	—	0.061	0.104	—	0.105
	95% cov.	0.938	—	0.934	0.940	—	0.932
β_1	bias	-0.103	—	-0.001	-0.098	—	-0.027
	s.d.	0.062	—	0.077	0.096	—	0.125
	mean(s.e.)	0.057	—	0.072	0.100	—	0.119
	95% cov.	0.538	—	0.938	0.834	—	0.926

Note: Parameters are $(\beta_0, \beta_1) = (-1, 1)$; $(\sigma_u, \mu_x, \sigma_x) = (0.5, 0, 1)$ and $\rho_u(t) = 1 - .5|t|$ for $|t| \in [0, 2]$, $= 0$, for $|t| > 2$. The RC$_{ar}$ estimator here misspecifies ρ_u as an AR type process with $\rho_u(t) = \rho^{|t|}$ does not converge in most of our data generated. The RC$_{cu}$ estimator models $\rho_w(t)$ as a cubic function of t. The result was from 500 replicates.

Table 2 demonstrates the sensitivity to the assumption on the error process. The error process $U_i(t)$ has $\rho_u(t) = 1 - .5|t|$ for $|t| \in [0, 2]$ and 0 elsewhere. The RC$_{ar}$ estimator here has a misspecification problem, and is so sensitive that a Newton–Raphson algorithm solving estimating equation (2) did not converge in most of our cases. The RC$_{cu}$ still has good performance even when a cubic function is not a perfect fit for this specific correlation function.

4.2. Logistic regression

Tables 3–5 consider logistic regression. The intention is to understand the performance of the RC analysis and the refined RC analysis; both applied the cubic correlation function. In Table 3, the error process is an AR type model as in Table 1. The parameters are $\beta = (-\ln(2), \ln(2))$, $(\sigma_u, \mu_x, \sigma_x) = (1, 0, 1)$ and $\rho = 0.2$. In this case with moderate relative risk, the RC$_{cu}$ estimator is very good, has small biases, and is more efficient than the refined estimator RRC$_{cu}$.

The data generated for Table 4 are the same as those in Table 3 except that $(\beta_0, \beta_1) = (-\ln(5), \ln(5))$. The RC estimator has a bias problem in this case. The RRC estimator has smaller biases especially when $n = 300$. Observe also that the bias of RRC decreases when n increases, while it does not for RC. With larger sample size, the RRC estimator is better in terms of bias and coverage probability.

Table 5 looks at misspecification of the Gaussian process $W_i(t)$, with $\rho = 0.2$ and 0.5 considered. The data were generated as in Table 3 except that X was uniform on $[-\sqrt{3}, \sqrt{3}]$. Comparing with Table 3 when $\rho = 0.2$, there is almost no difference although X is no longer normal. This is perhaps due to the fact that the approximation for $E(X_i|\tilde{W}_i)$ based on a multivariate model is a best linear approximation; see Carroll, Ruppert and Stefanski (1995, Chapter 3) for the case with an independent error process. However, when the errors have the higher correlation $\rho = 0.5$, the biases increase. This is due to the fact that $\rho_u(T^*) = 0.125 \neq 0$ which violates Assumption (A3) of the Appendix; see the discussion below (4) in Section 3.

Table 3. Logistic regression with moderate relative risk.

		$n = 300$			$n = 100$		
		Naive	RC_{cu}	RRC_{cu}	Naive	RC_{cu}	RRC_{cu}
β_0	bias	0.014	0.014	-0.002	0.023	0.022	0.000
	s.d.	0.124	0.125	0.129	0.232	0.239	0.254
	mean(s.e.)	0.128	0.129	0.133	0.224	0.227	0.238
	95% cov.	0.944	0.946	0.958	0.944	0.946	0.950
β_1	bias	-0.191	-0.011	0.006	-0.178	0.034	0.065
	s.d.	0.115	0.163	0.175	0.209	0.306	0.345
	mean(s.e.)	0.117	0.158	0.170	0.206	0.289	0.320
	95% cov.	0.622	0.942	0.946	0.800	0.952	0.958

Note: Parameters are $(\beta_0, \beta_1) = (-\ln(2), \ln(2))$; $(\sigma_u, \mu_x, \sigma_x) = (1, 0, 1)$ and $\rho_u(t) = \rho^{|t|}$ with $\rho = 0.2$. The RC_{cu} estimator models $\rho_w(t)$ as a cubic function of $|t|$ and RRC_{cu} is the refined estimator. The result was from 500 replicates.

Table 4. Logistic regression with larger relative risk.

		$n = 300$			$n = 100$		
		Naive	RC_{cu}	RRC_{cu}	Naive	RC_{cu}	RRC_{cu}
β_0	bias	0.151	0.150	-0.026	0.152	0.152	-0.049
	s.d.	0.175	0.178	0.248	0.294	0.298	0.413
	mean(s.e.)	0.174	0.175	0.235	0.303	0.305	0.431
	95% cov.	0.830	0.832	0.962	0.894	0.884	0.964
β_1	bias	-0.529	-0.141	0.044	-0.541	-0.126	0.099
	s.d.	0.166	0.242	0.365	0.259	0.374	0.573
	mean(s.e.)	0.164	0.220	0.313	0.285	0.389	0.592
	95% cov.	0.128	0.842	0.946	0.474	0.910	0.952

Note: Parameters are $(\beta_0, \beta_1) = (-\ln(5), \ln(5))$; $(\sigma_u, \mu_x, \sigma_x) = (1, 0, 1)$ and $\rho_u(t) = \rho^{|t|}$ with $\rho = 0.2$. The RC_{cu} estimator models $\rho_w(t)$ as a cubic function of $|t|$ and RRC_{cu} is the refined estimator. The result was from 500 replicates.

Table 5. Logistic regression with moderate relative risk and uniformly $[-\sqrt{3}, \sqrt{3}]$ distributed X.

		$n = 300$			$n = 100$		
		Naive	RC$_{cu}$	RRC$_{cu}$	Naive	RC$_{cu}$	RRC$_{cu}$
$\rho = 0.2$							
β_0	bias	0.016	0.015	-0.001	0.006	0.007	-0.014
	s.d.	0.127	0.127	0.132	0.234	0.234	0.245
	mean(s.e.)	0.128	0.129	0.133	0.226	0.227	0.239
	95% cov.	0.956	0.956	0.958	0.950	0.948	0.948
β_1	bias	-0.180	-0.002	0.016	-0.169	0.033	0.062
	s.d.	0.117	0.160	0.172	0.216	0.306	0.341
	mean(s.e.)	0.115	0.156	0.168	0.204	0.283	0.312
	95% cov.	0.640	0.942	0.948	0.822	0.940	0.952
$\rho = 0.5$							
β_0	bias	0.016	0.016	0.003	0.015	0.012	-0.004
	s.d.	0.130	0.130	0.134	0.228	0.229	0.238
	mean(s.e.)	0.128	0.128	0.132	0.224	0.225	0.233
	95% cov.	0.934	0.926	0.932	0.950	0.954	0.958
β_1	bias	-0.237	-0.087	-0.074	-0.239	-0.081	-0.062
	s.d.	0.112	0.150	0.160	0.196	0.276	0.299
	mean(s.e.)	0.108	0.144	0.152	0.189	0.254	0.293
	95% cov.	0.414	0.882	0.892	0.710	0.904	0.916

Note: Parameters are $(\beta_0, \beta_1) = (-\ln(2), \ln(2))$; $(\sigma_u, \mu_x, \sigma_x) = (1, 0, 1)$ and $\rho_u(t) = \rho^{|t|}$. The RC$_{cu}$ estimator models $\rho_w(t)$ as a cubic function of $|t|$ and RRC$_{cu}$ is the refined estimator.

These simulation results suggest that if the true correlation function is either exponential or piecewise linear (Table 2), then modeling it by a cubic function leads to ignorable biases in estimating β. However, one should be cautious in checking Assumption (A3) if the errors are highly correlated. Also, it is conceivable that in some situations we may need a higher order polynomial for the correlation function of $W(t)$. The choice of the order of the polynomial is considered in the next data analyses. For binary outcome regression, a refined analysis is preferred if the relative risk parameter is large and the sample size is large.

5. Data Analysis

This section brings in data from 563 subjects included in a retrospective longitudinal study of childhood growth, where each subject had at least one measurement of body mass index (BMI) z–score between ages 1 and 4 years.

The outcome variable of interest is adult obesity. See Whitaker, Wright, Pepe, Seidel and Dietz (1997) for details of the study. In brief, the study included all subjects born at Group Health Cooperative, a health maintenance organization in Seattle, between 1/1/65 – 1/1/71, and who had at least one outpatient visit on or after their 21st birthday. Subjects were categorized as obese or non-obese as adults using their average BMI between 21 and 29 years of age. The objective here is to determine the extent to which average BMI z–score between ages 1 and 4 years is predictive of adult obesity. The measurement error analysis is considered since one only approximates the true long–term average BMI z–score between ages 1 and 4 years. Thirteen percent of the adults in this study were classified as obese.

For subject i let BMI measurements be denoted by W_{ij} available at times $\{t_{ij}, j = 1, \ldots, k_i\}$ in the interval 1 to 4 years of age. Assume that the data follow a logistic regression model which is linear in X. According to the model assumption in Section 2, we assume that $\tilde{W}_i$ given X_i is multivariate normal with mean $X_i 1_{k_i}$ and variance Σ_{U_i}, where the (j, m)th element of Σ_{U_i} is $\sigma_u^2 \rho_u(t_{ij} - t_{im})$ for $j, m \in \{1, \ldots, k_i\}$.

Table 6. Logistic regression analysis of child growth data.

	Naive	RC$_{ar}$	RC$_{quar}$	RRC$_{quar}$
β_0	-1.843 (0.125)	-1.853 (0.123)	-1.842 (0.123)	-1.849 (0.124)
β_1	0.464 (0.159)	0.653 (0.186)	0.553 (0.159)	0.556 (0.161)

Note: There were 563 subjects used in the analysis, β_0, β_1 are the logistic regression parameters of adulthood obesity on childhood BMI. The RC$_{quar}$ and RC$_{quar}$ estimates were obtained from modeling $\rho_w(t)$ as a quartic function.

Results are in Table 6. We note that the RC$_{ar}$ did not have a divergence problem in solving (2). Under an AR error model, $\rho_u(t) = \rho^{|t|}$, we obtained $\widehat{\rho} = 0.19$, $\widehat{\sigma}_u^2 = 0.33$ and $\widehat{\sigma}_x^2 = 0.42$. These led to $\widehat{\sigma}_w^2 = 0.75$. The time points $|t_{ij} - t_{im}|$ used to estimate the cubic correlation model run from 0 to 3. As described earlier, for the purpose of reducing finite sample performance, we excluded the 10% points with $|t_{ij} - t_{im}| > 2.04$ when we applied our method to estimate σ_x^2, which led to $\widehat{\sigma}_x^2 = 0.48$. The correlation process estimated by an AR type exponential function, and polynomials of order 3, 4, 5, respectively, for $\rho_w(t)$, $t \in [0, 2.04]$, are in Figure 1. It seems that a cubic function does not model the correlation well, while a quartic correlation is reasonable–it is similar to an order–5 polynomial but it oscillates less. Table 6 shows results from the naive estimator, RC$_{ar}$, and the RC$_{quar}$ and RRC$_{quar}$ estimators (the subindex $_{quar}$ denotes the β estimates from a quartic correlation estimation). The RC$_{quar}$ and RRC$_{quar}$ estimates are in close agreement because the relative risk parameter indicates that the odds ratio e^{β_1} is less than 2. This is moderate and still in the range where the RC estimator performs well; see the summary from the simulation. The RC$_{ar}$ estimate of β_1 is larger than that from the RC$_{quar}$ estimate because the estimate of σ_x is smaller, which leads to a larger attenuation effect.

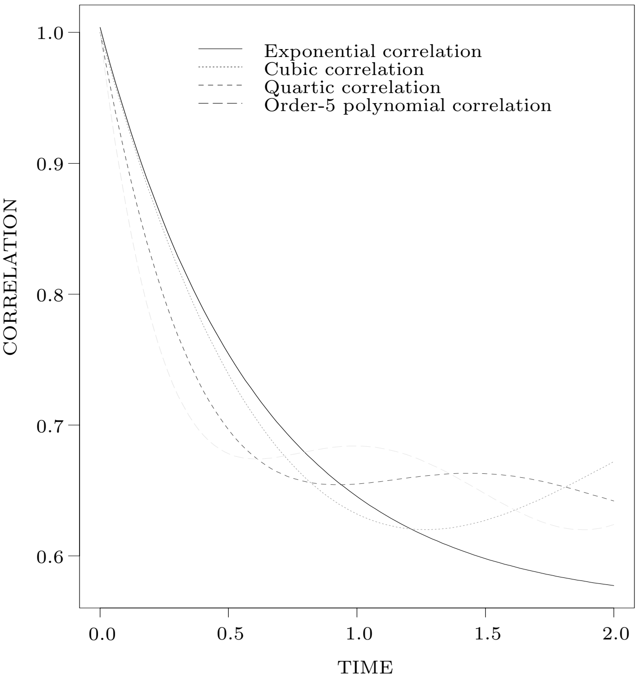

Figure 1. Correlation analysis of child growth data.

Acknowledgements

This research was supported by US National Institutes of Health grants AG 15026, CA 53996 and HL 57288. The author is thankful to Margaret Pepe and a referee for helpful comments, and to Bob Whitaker for the data on child growth.

Appendix. Technical Proofs

A. Asymptotic distribution for the RC estimator in linear regression

Assumptions.

(A1) The surrogate variable $W_i(t_1)$ is independent of $W_j(t_2)$ for $i \neq j$ and any t_1, t_2.

(A2) The correlation process $\text{corr}\{W_i(t_1), W_i(t_2)\} = \rho_w(t_2 - t_1)$ is a function of $|t_2 - t_1|$ with $\rho_w(t) = 1 + \sum_{s=1}^{q} \gamma_s |t|^s$.

(A3) The correlation process satisfies $\rho_u(T^*) = 0$, where $T^* = \max\{t_{ij,m}; i = 1,\ldots,n; j,m = 1,\ldots,k_i\}$.

Let

$$U_n(\beta,\mu_x,\rho_w) = n^{-1/2}\sum_{i=1}^n \begin{pmatrix} 1 \\ X_i^* \end{pmatrix} (Y_i - \beta_0 - \beta_1 X_i^*),$$

where $X_i^* = E(X_i|\tilde{W}_i) = \mu_x + (\sigma_x^2 1_{k_i})^t\{\sigma_x^2 I_{k_i} + \sigma_u^2 M_i\}^{-1}(\tilde{W}_i - \mu_x 1_{k_i})$. Under Assumptions (A2) and (A3), $\rho_w(T^*) = \sigma_x^2/\sigma_w^2$ and hence $X_i^* = \mu_x + (\rho_w(T^*)1_{k_i})^t\mathcal{G}_i^{-1}$ $(\tilde{W}_i - \mu_x 1_{k_i})$, where the (j,m)th element of matrix $\mathcal{G}_i$ is $\rho_w(t_{ij,m})$. Therefore, X_i^* is a function of parameters μ_x, γ, but not σ_w^2. Let $N_* = \sum_{i=1}^n k_i^2$ and $T_{ij,m} = (t_{ij,m},\ldots,t_{ij,m}^q)^t$. Let A be the probability limit of A_n as $n \to \infty$, where $A_n = (N_*)^{-1}\sum_{i=1}^n\sum_{j,m=1}^{k_i} T_{ij,m}T_{ij,m}^t$. Recall that $\widehat{\gamma}$ is a least square estimator of γ using data $\{V_{ij,m}, t_{ij,m}\}$, $i = 1,\ldots,n$, $j,m = 1,\ldots,k_i$. Define $N^{*-1/2}\sum_{i=1}^n\sum_{j,m=1}^{k_i} T_{ij,m}\{V_{ij,m} - \rho_w(t_{ij,m})\} \equiv \Psi(\gamma)$, zero when evaluated at $\widehat{\gamma}$. Taking a Taylor series expansion of $\Psi(\widehat{\gamma})$ at γ and using some simple algebra, it can be shown (see Carroll, Ruppert and Stefanski (1995, Appendix A.3.1)) that

$$N_*^{1/2}(\widehat{\gamma} - \gamma) = A^{-1}N^{*-1/2}\sum_{i=1}^n\sum_{j,m=1}^{k_i} T_{ij,m}\{V_{ij,m} - \rho_w(t_{ij,m})\} + o_p(1),$$

where $V_{ij,m} = \{W_i(t_{ij}) - \overline{W}_{\cdot\cdot}\}\{W_i(t_{im}) - \overline{W}_{\cdot\cdot}\}/\widehat{\sigma}_w^2$; $\widehat{\sigma}_w^2 = N^{-1}\sum_{i=1}^n\sum_{j=1}^{k_i}\{W_i(t_{ij}) - \overline{W}_{\cdot\cdot}\}^2$. Define $\widehat{\rho}_w(t) = 1 + \sum_{s=1}^q\widehat{\gamma}_s|t|^s$. Note that $N^{1/2}(\widehat{\mu}_x - \mu_x) = N^{-1/2}\sum_{i=1}^n\sum_{j=1}^{k_i}$ $(W_{ij} - \mu_x)$. The RC estimator $\widehat{\beta}$ solves $U_n(\beta,\widehat{\mu}_x,\widehat{\rho}_w) = 0$, where

$$U_n(\beta,\widehat{\mu}_x,\widehat{\rho}_w) = n^{-1/2}\sum_{i=1}^n \begin{pmatrix} 1 \\ \widehat{X}_i^* \end{pmatrix} (Y_i - \beta_0 - \beta_1 \widehat{X}_i^*);$$
$$\widehat{X}_i^* = \widehat{\mu}_x + (\widehat{\rho}_w(T^*)1_{k_i})^t\widehat{\mathcal{G}}_i^{-1}(\tilde{W}_i - \widehat{\mu}_x),$$

and the (j,m)th element of $\widehat{\mathcal{G}}_i$ is $\widehat{\rho}_w(t_{ij,m})$. Let

$$G_n(\beta,\mu_x,\rho_w) = n^{-1}\sum_{i=1}^n \begin{pmatrix} 1 \\ X_i^* \end{pmatrix} (1, X_i^*);$$
$$R_n(\beta,\mu_x,\rho_w) = n^{-1}\sum_{i=1}^n \begin{pmatrix} \beta_1 X_{i\mu_x}^* \\ 2\beta_1 X_i^* X_{i\mu_x}^* - (Y_i - \beta_0)X_{i\mu_x}^* \end{pmatrix};$$
$$B_n(\beta,\mu_x,\rho_w) = n^{-1}\sum_{i=1}^n \begin{pmatrix} \beta_1 X_{i\gamma}^* \\ 2\beta_1 X_i^* X_{i\gamma}^* - (Y_i - \beta_0)X_{i\gamma}^* \end{pmatrix}; \qquad (5)$$

where $X_{i\mu_x}^* = 1 - (\rho_w(T^*)1_{k_i})^t\mathcal{G}_i^{-1}1_{k_i}$, and

$$X_{i\gamma}^* = \Big[\{(\partial/\partial\gamma)\rho_w(T^*)1_{k_i}\}^t\mathcal{G}_i^{-1} - (\rho_w(T^*)1_{k_i})^t\mathcal{G}_i^{-1}\{(\partial/\partial\gamma)\mathcal{G}_i\}\mathcal{G}_i^{-1}\Big](\tilde{W}_i - \mu_x).$$

By a Taylor expansion,

$$\begin{aligned}0 &= U_n(\widehat{\beta}, \widehat{\mu}_x, \widehat{\rho}_w)\\ &= U_n(\beta, \mu_x, \rho_w) - G_n(\beta, \mu_x, \rho_w) n^{1/2}(\widehat{\beta} - \beta) - R_n(\beta, \mu_x, \rho_w) n^{1/2}(\widehat{\mu}_x - \mu_x)\\ &\quad -B_n(\beta, \mu_x, \rho_w) n^{1/2}(\widehat{\gamma} - \gamma) + o_p(1).\end{aligned}$$

Observe that G_n, R_n and B_n are all sums of independent random variables and hence each converges by the Law of Large Numbers. Let $G(\beta, \mu_x, \rho_w)$, $R(\beta, \mu_x, \rho_w)$ and $B(\beta, \mu_x, \rho_w)$ be the probability limit of $G_n(\beta, \mu_x, \rho_w)$, $R_n(\beta, \mu_x, \rho_w)$ and $B_n(\beta, \mu_x, \rho_w)$, respectively. Hence, if $n/N \to \lambda_1$, $n/N_* \to \lambda_2$,

$$\begin{aligned}&n^{1/2}(\widehat{\beta} - \beta)\\ &= G^{-1}(\beta, \mu_x, \rho_w)\Big[U_n(\beta, \mu_x, \rho_w) - R(\beta, \mu_x, \rho_w)\lambda_1 n^{-1/2}\sum_{i=1}^{n}\sum_{j=1}^{k_i}(W_{ij} - \mu_x)\\ &\quad -B(\beta, \mu_x, \rho_w)\lambda_2 A^{-1} n^{-1/2}\sum_{i=1}^{n}\sum_{j=1}^{k_i} T_{ij,m}\{V_{ij,m} - \rho_w(t_{ij,m})\}\Big] + o_p(1)\\ &= G^{-1}(\beta, \mu_x, \rho_w) n^{-1/2}\sum_{i=1}^{n}\Big[\begin{pmatrix}1\\ X_i^*\end{pmatrix}(Y_i - \beta_0 - \beta_1 X_i^*) - \lambda_1 R(\beta, \mu_x, \rho_w)\sum_{j=1}^{k_i}(W_{ij} - \mu_x)\\ &\quad -\lambda_2 B(\beta, \mu_x, \rho_w) A^{-1}\sum_{j,m=1}^{k_i} T_{ij,m}\{V_{ij,m} - \rho_w(t_{ij,m})\}\Big] + o_p(1)\\ &\equiv G^{-1}(\beta, \mu_x, \rho) n^{-1/2}\sum_{i=1}^{n} U_{i*}(\beta, \mu_x, \rho_w) + o_p(1). \qquad (6)\end{aligned}$$

Therefore, $n^{1/2}(\widehat{\beta} - \beta)$ is asymptotically normally distributed with mean 0 and covariance

$$G^{-1}(\beta, \mu_x, \rho_w)\Big\{n^{-1}\sum_{i=1}^{n} U_{i*}(\beta, \mu_x, \rho_w) U_{i*}^t(\beta, \mu_x, \rho_w)\Big\}\{G^{-1}(\beta, \mu_x, \rho_w)\}^t. \qquad (7)$$

The asymptotic covariance of $n^{1/2}(\widehat{\beta} - \beta)$ can be consistently estimated by noting that $G_n(\widehat{\beta}, \widehat{\mu}_x, \widehat{\rho}_w)$, $R_n(\widehat{\beta}, \widehat{\mu}_x, \widehat{\rho}_w)$, $B_n(\widehat{\beta}, \widehat{\mu}_x, \widehat{\rho}_w)$ are consistent estimates of $G(\beta, \mu_x, \rho_w)$, $R(\beta, \mu_x, \rho_w)$ and $B(\beta, \mu_x, \rho_w)$, respectively.

B. Asymptotic Covariance for the RC Estimator in Logistic Regression

Let $H(u) = \{1 + \exp(-u)\}^{-1}$ and $H^{(1)}(u) = H(u)\{1 - H(u)\}$. Similar to (5), define

$$G_n(\beta, \mu_x, \rho_w) = n^{-1}\sum_{i=1}^{n}\begin{pmatrix}1\\ X_i^*\end{pmatrix}(1, X_i^*) H^{(1)}(\beta_0 + \beta_1 X_i^*);$$

$$R_n(\beta,\mu_x,\rho_w)=n^{-1}\sum_{i=1}^n\begin{pmatrix}\beta_1X^*_{i\mu_x}H^{(1)}(\beta_0+\beta_1X_i^*)\\ \beta_1X_i^*X^*_{i\mu_x}H^{(1)}(\beta_0+\beta_1X_i^*)-X^*_{i\mu_x}\{Y_i-H(\beta_0+\beta_1X_i^*)\}\end{pmatrix};$$

$$B_n(\beta,\mu_x,\rho_w)=n^{-1}\sum_{i=1}^n\begin{pmatrix}\beta_1X^*_{i\mu_\gamma}H^{(1)}(\beta_0+\beta_1X_i^*)\\ \beta_1X_i^*X^*_{i\mu_\gamma}H^{(1)}(\beta_0+\beta_1X_i^*)-X^*_{i\mu_\gamma}\{Y_i-H(\beta_0+\beta_1X_i^*)\}\end{pmatrix}.$$

Let $G(\beta,\mu_x,\rho_w)$, $R(\beta,\mu_x,\rho_w)$, $B(\beta,\mu_x,\rho_w)$ be the probability limits of $G_n(\beta,\mu_x,\rho_w)$, $R_n(\beta,\mu_x,\rho_w)$ and $B_n(\beta,\mu_x,\rho_w)$, respectively. Let the asymptotic limit of $\widehat{\beta}$ be β_*, which solves

$$E\Big[\begin{pmatrix}1\\ X^*\end{pmatrix}\{Y_i-H(\beta_0+\beta_1X^*)\}\Big]=0.$$

Similar to (6), it can be shown that

$$n^{1/2}(\widehat{\beta}-\beta_*)=G^{-1}(\beta_*,\mu_x,\rho_w)n^{-1/2}\sum_{i=1}^n U_{i*}(\beta_*,\mu_x,\rho_w)+o_p(1),$$

where

$$U_{i*}(\beta,\mu_x,\rho_w)=\Big[\begin{pmatrix}1\\ X_i^*\end{pmatrix}\{Y_i-H(\beta_0+\beta_1X_i^*)\}-\lambda_1R(\beta,\mu_x,\rho_w)\sum_{j=1}^{k_i}(W_{ij}-\mu_x)$$
$$-\lambda_2B(\beta,\mu_x,\rho_w)A^{-1}\sum_{j,m=1}^{k_i}T_{ij,m}\{V_{ij,m}-\rho_w(t_{ij,m})\}\Big].$$

The asymptotic covariance of $n^{1/2}(\widehat{\beta}-\beta_*)$ is thus

$$G^{-1}(\beta_*,\mu_x,\rho_w)\Big\{n^{-1}\sum_{i=1}^n U_{i*}(\beta_*,\mu_x,\rho_w)U^t_{i*}(\beta_*,\mu_x,\rho_w)\Big\}\{G^{-1}(\beta_*,\mu_x,\rho_w)\}^t. \quad (8)$$

References

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

Carroll, R. J., Spiegelman, C. H., Lan, K. K., Bailey, K. T. and Abbott, R. D. (1984). On errors–in–variables for binary regression models. *Biometrika* **71**, 19-26.

Fuller, W. A. (1987). *Measurement Error Models*. John Wiley, New York.

Liang, K. Y. and Liu, X. H. (1991). Estimating equations in generalized linear models with measurement error. In *Estimating Functions* (Edited by V. P. Godambe), 47-63. Clarendon Press, Oxford.

Pierce, D. A., Stram, D. O., Vaeth, M. and Schafer, D. (1992). The errors in variables problem: considerations provided by radiation dose-response analyses of the A-bomb survivor data. *J. Amer. Statist. Assoc.* **87**, 351-359.

Prentice, R. L. (1996). Dietary fat and breast cancer: measurement error and results from analytic epidemiology. *J. Nat. Cancer Inst.* **88**, 1738-1747.

Rosner, B., Willett, W. C. and Spiegelman, D. (1989). Correction of logistic regression relative risk estimates and confidence intervals for systematic within-person measurement error. *Statist. Medicine* **8**, 1051-1070.

Schafer, D. W. and Purdy, K. (1996). Likelihood analysis for error–in–variables regression with replicate measurements. *Biometrika* **83**, 813-824.

Wang, C. Y. and Pepe, M. S. (2000). Expected estimating equations to accommodate covariate measurement error. *J. Roy. Statist. Soc. Ser. B* in press.

Wang, N., Carroll, R. J. and Liang, K. Y. (1996). Quasi-likelihood estimation in measurement error models with correlated replicates. *Biometrics* **52**, 401-411.

Whitaker, R. C., Wright, J. A., Pepe, M. S., Seidel, K. D. and Dietz, W. H. (1997). Predicting adult obesity from childhood and parent obesity. *New England J. Medicine* **13**, 869-873.

Zeger, S. L., Liang, K. Y. and Albert, P. S. (1988). Models for longitudinal data: a generalized estimating equation. *Biometrics* **44**, 1049-1060.

Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, P.O. Box 19024,
Seattle, WA 98109-1024, U.S.A.

E-mail: cywang@fhcrc.org

(Received July 1998; accepted September 1999)